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

Volume 16 Issue 7
July 2025



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

SAI
www.ijacsa.thesai.org

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 16 Issue 7 July 2025
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

Arun Kulkarni
University of Texas at Tyler
Domain of Research: Machine Vision, Artificial Intelligence, Computer Vision, Data Mining, Image Processing, Machine Learning, Neural Networks, Neuro-Fuzzy Systems

Dr Ronak AL-Haddad
Anglia Ruskin University / Cambridge
Domain of Research : Technology Trends,Communication,Security,Software Engineering and Quality,Computer Networks,Cyber Security,Green Computing,Multimedia Communication,Network Security,Quality of Service

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

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: Enhancing Trust in Human-AI Collaboration: A Conceptual Review of Operator-AI Teamwork
Authors: Abduljaleel Hosawi, Richard Stone
PAGE 1 – 15

Paper 2: Designing an Empathetic Conversational Agent for Student Mental Health: A Pilot Study
Authors: Kaichi Minami, Choi Dongeun, Panote Siriaraya, Noriaki Kuwahara
PAGE 16 – 23

Paper 3: Cross-Context Evaluation of an Indoor–Outdoor AR Navigation System in a University Campus Environment
Authors: Toma Marian-Vladut, Pascu Paul, Turcu Corneliu Octavian
PAGE 24 – 33

Paper 4: Brightness-Aware Generative Adversarial Network for Low-Light Image Enhancement
Authors: Huafei Zhao, Mideth Abisado
PAGE 34 – 43

Paper 5: Power-Aware Video Transmission in 5G Telemedicine: Challenges, Solutions, and Future Directions
Authors: Qiuhong SHI, Mingjing CAO, Yun BAI
PAGE 44 – 54

Paper 6: Evaluating Intangible Software Quality Metrics for Effective Project Management Information Systems
Authors: Gu Xin, Rozi Nor Haizan Nor, Nur Ilyana Ismarau Tajuddin, Khairi Azhar Aziz
PAGE 55 – 63

Paper 7: CNN-LinATFormer: Enhancing PM2.5 Prediction Through Feature Assessment and Linear Attention Mechanism
Authors: Yuchen Zhang, Rajermani Thinakaran
PAGE 64 – 74

Paper 8: Transformer Model Optimization Method for Multi-Modal Data Fusion
Authors: Shanshan Yang, Jie peng
PAGE 75 – 86

Paper 9: Emotional Analysis and Interpretation of Music Conducting Works Based on Artificial Intelligence
Authors: He Huang, Chengcheng Zhang, Yun Liu, Liyuan Liu
PAGE 87 – 96

Paper 10: Object Recognition in Pond Environments Using Deep Learning
Authors: Suhaila Sari, Ng Wei Jie, Nik Shahidah Afifi Md Taujuddin, Hazli Roslan, Nabilah Ibrahim, Mohd Helmy Abd Wahab
PAGE 97 – 105

Paper 11: FoodSharePro: An Integrated Mobile Platform for Sustainable Food Donation and Decentralized Composting
Authors: Jamil Abedalrahim Jamil Alsayaydeh, Rex Bacarra, Shamsul Fakhar Bin Abd Gani, Serhij Mamchenko, Safarudin Gazali Herawan
PAGE 106 – 121

Paper 12: Benefits and Challenges of Cloud Computing System in Malaysian Public Healthcare Organizations
Authors: Nurul Izzatty Ismail, Juhari Noor Faezah, Muhammad Syukri Abdullah, Masrina Nadia Mohd Salleh
PAGE 122 – 128

Paper 13: niCNN: A Novel Neuromorphic Approach to Energy-Efficient and Lightweight Human Activity Recognition on Edge Devices
Authors: Preeti Agarwal
PAGE 129 – 138

Paper 14: Improved YOLOv8 Model for Enhanced Small-Sized Breast Mass Detection on Magnetic Resonance Imaging
Authors: Feiyan Wu, Chia Yean Lim, Sau Loong Ang, Jiaxin Zheng
PAGE 139 – 145

Paper 15: SE-Pruned ResNet-18: Balancing Accuracy and Efficiency for Object Classification on Resource-Constrained Devices
Authors: Zeyad Farisi
PAGE 146 – 151

Paper 16: Advanced AI for Liver Cancer Detection: Vision Transformers, XAI and Contrastive Learning
Authors: B C Anil, Jayasimha S R, Samitha Khaiyum, T L Divya, Rakshitha Kiran P, Vishal C
PAGE 152 – 159

Paper 17: A Qualitative Constructivist Framework for Assessing Knowledge Transfer in Enterprise System Projects: Insights from Expert Interviews
Authors: Jamal M. Hussien, Riza bin Sulaiman, Ali H Hassan, Mansoor Abdulhak, Hasan Kahtan, Basit Shahzad
PAGE 160 – 171

Paper 18: Investigating Space-Time Dynamics in Live Memory Forensics Using Hybrid Transformer Approaches
Authors: Sarishma Dangi, Kamal Ghanshala, Sachin Sharma
PAGE 172 – 184

Paper 19: Detecting Fake News Images Using a Hybrid CNN-LSTM Architecture
Authors: Dina R. Salem, Abdullah A. Abdullah, AbdAllah A. AlHabshy, Kamal A. ElDahshan
PAGE 185 – 193

Paper 20: Heuristics-Based Clustering of Internet of Vehicle Based on Effective Approximate QoS Guideline for Message Dissemination
Authors: Tanuja Kayarga, S Ananda Kumar, Lakshmi B S, Anil Kumar B H
PAGE 194 – 201

Paper 21: SOM-Based Leader Selection Strategies for Cooperative Spectrum Sensing in Multi-Band Multi-User 6G CR IoT
Authors: Mayank Kothari, Suresh Kurumbanshi
PAGE 202 – 209

Paper 22: High-Speed Fiber-Optic Communication Performance Utilizing Fiber Bragg Grating-Based Dispersion Compensation Schemes
Authors: Kripa Kalkala Balakrishna, Karthik Palani
PAGE 210 – 220

Paper 23: Bio-Inspired Metaheuristic Framework for DNA Motif Discovery Using Hybrid Cluster Based Walrus Optimization
Authors: M. Shilpa, C. Nandini
PAGE 221 – 228

Paper 24: Detection of Autism Spectrum Disorder (ASD) Using Lightweight Ensemble CNN Based on Facial Images for Improved Diagnostic Accuracy
Authors: Andi Kurniawan Nugroho, Jajang Edi Priyanto, D. S. P. Vinski
PAGE 229 – 236

Paper 25: An Adaptive SVR-Based Framework for Multimodal Corpus Classification
Authors: Yuhui Wang
PAGE 237 – 246

Paper 26: Speech Emotion Recognition from Audio Data Using LSTM Model
Authors: Md. Mahbub-Or-Rashid, Akash Kumar Nondi, Abdullah Al Sadnun, Md. Anwar Hussen Wadud, T M Amir Ul Haque Bhuiyan, Md.Saddam Hossain
PAGE 247 – 254

Paper 27: AI-Driven Textual Feedback Analysis in E-Training Using Enhanced RoBERTa
Authors: Rakan Saad Alotaibi, Fahad Mazyed Alotaibi, Sameer Abdullah Nooh, Abdulaziz A. Alsulami
PAGE 255 – 266

Paper 28: A Deep Learning-Based Dual-Model Framework for Real-Time Malware and Network Anomaly Detection with MITRE ATT&CK Integration
Authors: Migara H. M. S, Sandakelum M. D. B, Maduranga D. B. W. N, Kumara D. D. K. C, Harinda Fernando, Kavinga Abeywardena
PAGE 267 – 272

Paper 29: Method for Maternal Health Risk Assessment with Smartwatch-Based Vital Sign Measurements
Authors: Kohei Arai, Diva Kurnianingtyas
PAGE 273 – 279

Paper 30: Multi-Step Cross-Domain Aspect-Based Sentiment Generation with Error Correction Mechanism
Authors: Ningning Mao, Xuanliang Zhu, Yadi Xu
PAGE 280 – 290

Paper 31: Perceptual Hash Techniques for Audio Copyright Protection in Decentralized Systems
Authors: N. Kavitha, Rashmika S J, Reshika A S
PAGE 291 – 301

Paper 32: Advancing Precision Livestock Farming: Integrating Hybrid AI, IoT, Cloud and Edge Computing for Enhanced Welfare and Efficiency
Authors: Hakim Jebari, Siham Rekiek, Kamal Reklaoui
PAGE 302 – 311

Paper 33: AI-Powered Skin Disease Detection Using Adaptive Particle Swarm Intelligent Optimization and Hyper-Convolutional Neural Networks
Authors: N Annalakshmi, S Umarani
PAGE 312 – 323

Paper 34: Deep Learning Optimization Conception: Less Data, Less Time, More Performance
Authors: Mohamed Amine MEDDAOUI, Moulay AMZIL, Imane KARKABA, Mohammed ERRITALI
PAGE 324 – 332

Paper 35: A Novel Hybrid HO-CAL Framework for Enhanced Stock Index Prediction
Authors: Zeren Shi, Othman Ibrahim, Hanini Ilyana Che Hashim
PAGE 333 – 342

Paper 36: The Text Mining Model for Lecturer Performance Evaluation: A Comparative Study
Authors: Anita Ratnasari, Vina Ayumi, Mariana Purba, Wachyu Hari Haji, Handrie Noprisson, Marissa Utami
PAGE 343 – 348

Paper 37: Design and Evaluation of a Biometric IoT-Based Smart Lock System with Real-Time Monitoring and Alert Mechanisms
Authors: Jamil Abedalrahim Jamil Alsayaydeh, Mohd Faizal Yusof, Serhij Mamchenko, Rostam Affendi Hamzah, Safarudin Gazali Herawan
PAGE 349 – 362

Paper 38: Modelling Cloud Computing Adoption in the Malaysian Healthcare
Authors: Normilah Mohd Noh, Nurhizam Safie Mohd Satar, Hasimi Sallehudin, Ibrahim Hassan Mallam, Surya Sumarni Hussein, Nur Azaliah Abu Bakar
PAGE 363 – 376

Paper 39: Potential Variables in Pharmaceutical Drug Prediction Research with Machine Learning Approach: A Literature Review
Authors: Gunadi Emmanuel, Yulyani Arifin, Ilvico Sonata, Muhammad Zarlis
PAGE 377 – 389

Paper 40: Scalable Graph Learning with Graph Convolutional Networks and Graph Attention Networks: Addressing Class Imbalance Through Augmentation and Optimized Hyperparameter Tuning
Authors: Chaima Ahle Touate, Rachid El Ayachi, Mohamed Biniz
PAGE 390 – 400

Paper 41: Enhancing Banking Data Classification Through Hybrid L2 Regularisation and Early Stopping in Artificial Neural Networks
Authors: Khairul Nizam Abd Halim, Abdul Syukor Mohamad Jaya, Fauziah Kasmin, Azlan Abdul Aziz
PAGE 401 – 408

Paper 42: Empirical Validation and Enhancement of ADiBA: A Framework for Big Data Analytics Implementation
Authors: Norhayati Daut, Naomie Salim, Sharin Hazlin Huspi, Anazida Zainal, Chan Weng Howe, Muhammad Aliif Ahmad, Siti Zaiton Mohd Hashim, Masitah Ghazali, Mohd Adham Isa, Rashidah Kadir, Nuremira Ibrahim, Norazlina Khamis
PAGE 409 – 425

Paper 43: An In-Depth Analysis of Security Flaws in Advanced Authentication Protocols for the Internet of Medical Things
Authors: Haewon Byeon
PAGE 426 – 431

Paper 44: Enhancing Patient Health Through Smart IoT Technologies in Healthcare
Authors: Monica Bhutani, Osman Elwasila, Rajermani Thinakaran, Yonis Gulzar
PAGE 432 – 441

Paper 45: Modeling an Adaptive and Collaborative E-Learning System with Artificial Intelligence Tools
Authors: Kawtar Zargane, Hassane Kemouss, Mohamed Khaldi
PAGE 442 – 453

Paper 46: Determinant Factors of Success for the Village Information System in Providing Sustainable Services and Governance
Authors: Sutia Dwi Santika, Tuga Mauritsius
PAGE 454 – 466

Paper 47: Machine Learning Methods for Detecting Fake News: A Systematic Literature Review of Machine Learning Applications in Key Domains
Authors: Nur Ida Aniza Rusli, Nur Atiqah Sia Abdullah, Fatin Nabila Abd Razak, Nor Haniza Ramli
PAGE 467 – 481

Paper 48: SpatialSolar-Net: A Multi-Site Collaborative Framework for Solar Power Forecasting with Adaptive Spatial Correlation Assessment
Authors: Yiming Liu, Mugambigai Darajah, Christopher Gan
PAGE 482 – 495

Paper 49: AFL-BERT : Enhancing Minority Class Detection in Multi-Label Text Classification with Adaptive Focal Loss and BERT
Authors: Zakia Labd, Said Bahassine, Khalid Housni
PAGE 496 – 501

Paper 50: Attention Aware Dual-Path Autoencoder with Asymmetric Loss for Recognition in Complex Scenes
Authors: Hashim Rosli, Rozniza Ali, Muhamad Suzuri Hitam, Ashanira Mat Deris, Noor Hafhizah Abd Rahim
PAGE 502 – 511

Paper 51: Secure Data Sharing Using Blockchain Technology: A Systematic Literature Review
Authors: Azman Azmi, Farashazillah Yahya, Nur Afrina Azman, Hazlina Jalil
PAGE 512 – 525

Paper 52: Enhancing Firefighter PPE Compliance Through Deep Learning and Computer Vision
Authors: Asmaa Alayed, Razan Talal Alqurashi, Samah Hamoud Alhelali, Asrar Yousef Khadawurdi, Bashayer Fayez Khan
PAGE 526 – 534

Paper 53: Multi-Agent Deep Reinforcement Learning Algorithms for Distributed Charging Station Management
Authors: Li Junda, Wang Tianan, Zhang Dingyi, Wu Quancai, Liu Jian
PAGE 535 – 543

Paper 54: A Review of Federated Learning Attacks: Threat Models and Defence Strategies
Authors: Fizlin Zakaria, Shamsul KamalAhmad Khalid
PAGE 544 – 554

Paper 55: Intelligent Logistics Vehicle Scheduling Based on MPHIGA
Authors: Xinxin Gao, Qing Wang
PAGE 555 – 563

Paper 56: Fuzzy Delphi Method: A Step-by-Step Guide to Obtaining Expert Consensus on Mobile Tourism Acceptance Culture
Authors: Syaifullah, Shamsul Arrieya Ariffin, Norhisham Mohamad Nordin
PAGE 564 – 577

Paper 57: Optimized Automatic Temperature and Humidity Control for Tobacco Storage Using TwinCAT and Deep Reinforcement Learning
Authors: Zhen Liu, Jili Wang, Shihao Song, Qiang Hua
PAGE 578 – 588

Paper 58: Adaptive AI-Driven Enterprise Resource Planning for Scalable and Real-Time Strategic Decision Making
Authors: Ghayth AlMahadin
PAGE 589 – 599

Paper 59: Enhancing Intent Recognition for Mixed Script Queries Using Roman Transliteration
Authors: Anu Chaudhary, Rahul Pradhan, Shashi Shekhar
PAGE 600 – 611

Paper 60: A Hybrid Deep Learning and Optimization Approach for Accurate Channel Estimation in 5G MIMO-OFDM Systems
Authors: Mohammed Fakhreldin
PAGE 612 – 628

Paper 61: Explainable Approach Using Semantic-Guided Alignment for Radiology Imaging Diagnosis
Authors: Fatima Cheddi, Ahmed Habbani, Hammadi Nait-Charif
PAGE 629 – 639

Paper 62: 3D Reconstruction from JPG Images
Authors: Youssif Mohamed Mostafa, Maryam N. Al-Berry, Howida A. Shedeed
PAGE 640 – 646

Paper 63: A Focused Survey of ECG Datasets for Artificial Intelligence-Based Atrial Fibrillation Detection
Authors: ASSALHI Imane, Bybi Abdelmajid, Oulad Hamdaoui Hanaa, Ebobisse Djene Yves Frederic, Drissi Lahssini Hilal
PAGE 647 – 658

Paper 64: DLCA-CapsNet: Dual-Lane CDH Atrous CapsNet for the Detection of Plant Diseases
Authors: Steve Okyere-Gyamfi, Michael Asante, Yaw Marfo Missah, Kwame Ofosuhene Peasah, Vivian Akoto-Adjepong
PAGE 659 – 673

Paper 65: A Comparative Study of Machine Learning Techniques for AE-Based Corrosion Detection with Emphasis on Transformer Models
Authors: Osama Shahid Ali, Lukman B A Rahim
PAGE 674 – 684

Paper 66: Smartphone-Integrated Sensor-Based DFU Risk Assessment Using CatBoost and Deep Neuro-Fuzzy Intelligence
Authors: Jayashree J, Vijayashree J, Perepi Rajarajeswari, Saravanan S
PAGE 685 – 691

Paper 67: Decoding Sales Order Anomalies: Advanced Predictive Modeling and Discrepancy Resolution Utilizing Machine Learning Algorithms
Authors: Amit Kumar Soni, Pooja Jain
PAGE 692 – 698

Paper 68: Home Network Attached Storage (HOMENAS) Using Raspberry Pi with Telegram Bot Notification
Authors: Nurul Najwa Abdul Rahid @ Abdul Rashid, Syafnidar Abdul Halim, Siti Maisarah Md Zain, Nik Aiman Shafiq Nik Shukri
PAGE 699 – 705

Paper 69: Enhancing Data Management for Decision Support Systems in Indonesian Government Internal Audit: A DMBOK Approach
Authors: Febrian Imanda Effendy, Nilo Legowo
PAGE 706 – 715

Paper 70: Game Theory Meets Explainable AI: An Enhanced Approach to Understanding Black Box Models Through Shapley Values
Authors: Mouad Louhichi, Redwane Nesmaoui, Mohamed Lazaar
PAGE 716 – 725

Paper 71: Anomaly Detection and Fault Diagnosis of Power Distribution Line Point Cloud Data Based on Deep Learning
Authors: Jiangshun Yu, Poyu You, Jian Zhao, Xianzhe Long, Yuran Chen
PAGE 726 – 739

Paper 72: HGWWO: A Hybrid Grey Wolf–Whale Optimizer for Load Balancing in Cloud Computing Environments
Authors: Yameng BAI, Junxia MENG, Shuai ZHAO, Ruoyu REN
PAGE 740 – 749

Paper 73: Automated Bubble Detection in Contact Lenses Using a Hybrid Deep Learning Framework
Authors: Chee Chin Lim, Yen Fook Chong, Vikneswaran Vijean, Gei Ki Tang
PAGE 750 – 758

Paper 74: Integration of Grey Wolf Optimizer Algorithm with Combinatorial Testing for Test Suite Generation
Authors: Muhamad Asyraf Anuar, Rosziati Ibrahim, Mazidah Mat Rejab, Nurezayana Zainal
PAGE 759 – 771

Paper 75: A Multi-Level Stacking Ensemble Model Optimized by Soft Set Theory for Customer Churn Prediction
Authors: Nurul Nadzirah Adnan, Mohd Khalid Awang
PAGE 772 – 775

Paper 76: A Dual-Path Gated Attention-Based Deep Learning Model for Automated Essay Scoring Using Linguistic Features
Authors: Qin Jie, Congling Huang
PAGE 776 – 788

Paper 77: Comparative Analysis of Cybersecurity Frameworks in Educational Institutions: Towards a Tailored Security Model
Authors: Syarif Hidayatulloh, Aedah Binti Abd. Rahman
PAGE 789 – 797

Paper 78: The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data
Authors: Jinzhu Lin, Tianwei Ni
PAGE 798 – 807

Paper 79: AI-Driven Firewall Log Analysis: Enhancing Threat Detection with Deep Learning Techniques
Authors: Yasmine ABOUDRAR, Khalid BOURAGBA, Mohamed OUZZIF
PAGE 808 – 815

Paper 80: Automated Dried Fish Classification Using MobileNetV2 and Transfer Learning
Authors: Rajmohan Pardeshi, Rajermani Thinakaran, Sanjay Kharat
PAGE 816 – 825

Paper 81: Enhancing Portfolio Optimization with Weighted Scoring for Return Prediction Through Machine Learning and Neural Networks
Authors: Ruili Sun, Qiongchao Xia, Shiguo Huang
PAGE 826 – 840

Paper 82: CeC-SMOTE: A Clustering and Centroid-Based Adaptive Oversampling Method for Imbalanced Data
Authors: Xiaoling Gao, Marshima Mohd Rosli, Muhammad Izzad Ramli, Nursuriati Jamil
PAGE 841 – 850

Paper 83: Towards Robust IoT Security: The Impact of Data Quality and Imbalanced Data on AI-Based IDS
Authors: Hiba El Balbali, Anas Abou El Kalam
PAGE 851 – 865

Paper 84: Automated Anatomical Analysis of Wood Cross Sections Using Macroscopic Images
Authors: Khanh Nguyen-Trong, Thanh Nhan Nguyen-Thi
PAGE 866 – 876

Paper 85: Air Quality Prediction Based on VMD-CNN-BiLSTM-Attention
Authors: Huang Xinxin, Mohd Suffian Sulaiman, Marshima Mohd Rosli
PAGE 877 – 887

Paper 86: Cardio-Edge: Hardware-Software Co-design Implementation of LSTM Based ECG Classification for Continuous Cardiac Monitoring on Wearable Devices
Authors: Nousheen Akhtar, Abdul Rehman Buzdar, Jiancun Fan, Muhammad Umair Khan
PAGE 888 – 899

Paper 87: Robust Particle Filter for Accurate WiFi-Based Indoor Positioning in the Presence of Outlier-Corrupted Sensor Data
Authors: Mohamed Aizad Bin Mohamed Ghazali, Aroland Kiring, Lyudmila Mihaylova, Hoe Tung Yew, Seng Kheau Chung, Farrah Wong
PAGE 900 – 909

Paper 88: Design and Analysis of Smart Lighting System for Room Environments Using Simulation Supporting Diverse Light Bulbs
Authors: Husnul Ajra, Mazlina Abdul Majid, Md. Shohidul Islam
PAGE 910 – 917

Paper 89: Deep Learning-Driven DNA Image Encryption with Optimal Chaotic Map Selection
Authors: Sara Bentouila, Kamel Mohamed Faraoun
PAGE 918 – 929

Paper 90: Older Adults and Technology Design from the HCI Perspective
Authors: Hasan Ali Sagga, Richard Stone
PAGE 930 – 938

Paper 91: Forecast COVID-19 Epidemics by Strengthening Deep Learning Models with Time Series Analysis
Authors: Warapree Tangseefa, Tepanata Pumpaibool, Paisit Khanarsa, Krung Sinapiromsaran
PAGE 939 – 954

Paper 92: Cyber Deception Across Domains: A Comprehensive Survey of Techniques, Challenges, and Perspectives
Authors: Amal Sayari, Slim Rekhis
PAGE 955 – 986

Paper 93: Gender and Age Estimation from Facial Images Based on Multi-Task and Curriculum Learning
Authors: Toma Brezovan, Claudiu Ionut, Popîrlan
PAGE 987 – 999

Paper 94: Improving Cross-Patient Epilepsy Detection via EEG Decomposition into Canonical Brain Rhythms with Deep Learning
Authors: Jose Yauri, Elinar Carrillo-Riveros, Edith Guevara-Morote, Juan Carlos Carreño-Gamarra, Karel Peralta-Sotomayor, Pelayo Quispe-Bautista
PAGE 1000 – 1007

Paper 95: Evaluating and Interpreting Pooling Techniques in Spectrogram-Based Audio Analysis Using Diverse Metrics
Authors: Supun Bandara, Uthayasanker Thayasivam
PAGE 1008 – 1018

Paper 96: Recommendation Engine for Amazon Magazine Subscriptions
Authors: Sushil Khairnar, Deep Bodra
PAGE 1019 – 1026

Paper 97: Empowering Accessibility: IoT-Driven Smart Buildings for Elderly and Disabled Individuals
Authors: Zeyad Alshboul, Burhan Mahmoud Hamadneh, Turki Mahdi Alqarni, Bajes Zeyad Aljunaeidia, Methaq Khadum
PAGE 1027 – 1034

Paper 98: Enhancing Dendritic Cell Algorithm by Integration with Multi-Layer Perceptron for Anomaly Detection
Authors: Yousra Abudaqqa, Zulaiha Ali Othman, Azuraliza Abu Bakar
PAGE 1035 – 1044

Enhancing Trust in Human-AI Collaboration: A Conceptual Review of Operator-AI Teamwork

Abduljaleel Hosawi[1], Richard Stone[2]
Dept. of HCI, Iowa State University Ames, USA[1]
Dept. of Industrial Engineering, Iowa State University Ames, USA[2]

***Abstract*—Trust is vital to collaborative work between operators and AI. Yet, important elements of its nature remain to be investigated, including the dynamic process of trust formation, growth, decline, and even death between an operator and an AI. This review analyzes how the dynamic development of trust is determined by Team performance and its complex interaction with factors related to AI system characteristics, Operator competencies, and Contextual factors. This review summarizes current concepts, theories, and models to propose a genuine framework for enhancing trust. It analyzes the current understanding of trust in human-AI collaborations, highlighting key gaps and limitations, such as a lack of robustness, poor explainability, and effective collaboration design. The findings emphasize the importance of key components in this collaborative environment, including Operator capabilities and AI technology characteristics, underscoring their impact on trust. This study advances understanding of the nature of Operator-AI collaboration and the Dynamics of trust in trust calibration. Through a multidisciplinary approach, it also emphasizes the impact of Explainability, Transparency, and trust repair mechanisms. It highlights how Operator-AI systems can be improved through Design principles and developing Human competencies to enhance collaboration.**

Keywords*—*Operator-AI collaboration; trust calibration; trust dynamics; explainability; transparency; trust repair mechanisms; cross-cultural trust; Clinical Decision Support Systems (CDSS); AI autonomy and influence; ethical considerations in AI; team performance; AI system characteristics; operator competencies; contextual factors; framework; limitations; robustness; Human-AI teaming; design principles; human competencies; predictability; reliability; understandability; over-reliance; automation bias; under-trust; trust measurement; trust erosion

I. Introduction

The catastrophic failure of HAL 9000 in the movie "2001: A Space Odyssey" highlights the loss of trust between humans and artificial intelligence [1]. As humans cede more control to Al in vital areas, fostering and nurturing trust becomes critical. The line between beneficial cooperation and disastrous dependence is fine and requires careful consideration. The ongoing effort to achieve greater integration of Artificial Intelligence (AI) into complex operational environments is rapidly transforming industries and reshaping the nature of work. AI, can be defined as machines or systems capable of performing tasks that typically require human intelligence. For instance, the ability to learn, solve problems, perceive, and take decisions offers unprecedented potential for human, such as enhancing our efficiency, augmenting our capabilities, as well as tackling previously intractable problems [2]–[5]. However, to deploy and utilize these powerful technologies successfully, particularly in collaborative settings with human operators, must be on the establishment, maintenance, and calibration of trust [6]. As AI systems are increasingly guaranteeing autonomy and responsibility in high-stakes areas, such as healthcare, finance, transportation, manufacturing, and defense [4], [7], [8]. Gaining a deep understanding the dynamics of trust within operator-AI teams today is paramount not only for ensuring safety, but also effectiveness, and the realization of AI's collaborative potential [9]–[11].

The concept of trust is multifaceted and dynamic in the context of human-AI interaction and has its own characteristics. Even though It's drawing parallels with interpersonal trust, it still presenting unique challenges stemming from the AI's non-human nature [6]. Generally, it is understood as an attitude or psychological state that involve a willingness to accept vulnerability (under conditions of uncertainty and risk) based on positive expectations regarding the AI's future behavior and capabilities. That includes its predictability, competence, reliability, as well as the alignment with the operator's goals [12]. There are three main aspects that are likely to affect this evolving attitude including the AI's perceived performance. For example, observed reliability, accuracy, and consistency. The second aspect is process in particular understandability, transparency, predictability of its model. Third is the purpose and how far it aligns with operato's goals, perceived benevolence, and integrity of its design [13], [14]. These three facets dynamically interact; for instance, if an AI's reasoning is explainable (process transparency), may lead to better assessment of its competence by the operator and may also be easier in predictability (performance). That potentially fosters faster trust emergence or aids calibration after errors [15].

How the level of operator's trust is calibrated in the capabilities of AI in a given context is critical, as it significantly impacts adoption behavior and overall team performance [16]. Distrust or (insufficient trust) could lead to the disuse (underutilization) or even abandonment of valuable AI assistance. Likewise, if an operator over-reliance or his/ her automation bias (excessive trust) results in issues such as complacency, reduced monitoring, and override AI errors or failure to detect them. This misuse, potentially, may have significant negative consequences [17].

According to Fig. 1, operators-AI collaboration could be best when the operator's level of trust is appropriately aligned with the AI' true capabilities. Failure to do so could lead to misuse by overtrusting the AI or under trust. For this reason, understanding trust dynamics (the entire lifecycle) is essential. In particular, how and when trust emerges, strengthened, and maintained before it decays or erodes. Trust can grow as a result of good behavior, and it can also be weakened as a

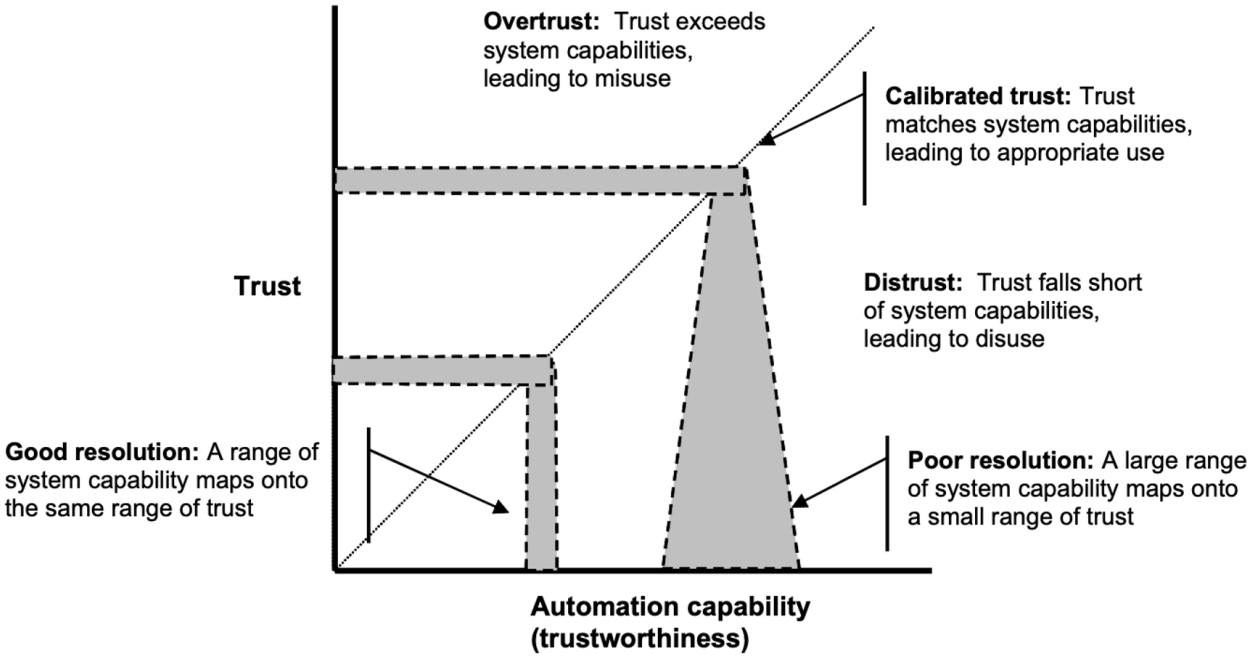


Fig. 1. Calibrating operator trust according to the true AI capabilities. Note: This figure is taken from Lee and See [12].

result of errors and biases. In addition, knowing trust repair strategies when it is damaged is just as important as knowing its life cycle, and both are essential for designing systems that promote safe, effective, and appropriately calibrated human-AI collaboration [18].

The dynamic process of trust formation between the operator and AI throughout team performance is influenced by a complex interplay of factors related not only to the AI system, but also to operator, interaction and contextual factors. Factors related to the AI system include objective performance metrics, communication strategies employed by the AI, how the design of transparency and explainability are featured , and even AI's perceived cooperativity or embodiment [19]. Human operator factors encompass many significant elements such as individual differences including propensity to trust, domain expertise. In addition to cognitive biases, required competencies for interacting with AI, age, cultural background, and emotional responses [20], [21]. For the factors related to interaction and contextuality , literature suggests involvement in the nature of the task, the development of shared mental models, and team composition. Finally, the overall organizational environment suporting human-AI teaming [22].

Despite rapid growth and advancements in research interest to explore the trust dynamics in the Human-AI settings, significant gaps remain. There is still a persistent need for more longitudinal studies tracking trust dynamics over time [23], [24] , deeper exploration of cross-cultural variations in trust also still demanded [25], [26], validated approaches not only to measure trust but also to capture its dynamics and reveal its calibrated nature [27], [28], better understanding of trust erosion and repair mechanisms following AI errors, failures or bias [29], [30],continued examination of ethical considerations, (for example, regarding autonomy and influence) [21], [31] , and developing new generation of AI systems with true teaming competencies [8], [9], [32].

The present literature review aims to synthesize current knowledge on the dynamics of trus, its emergence, growth, and decay within operator-AI teams. Specifically, this review focuses primarily on new literature that reflects the latest advancements in the field. It examines the various types of AI systems relevant to this context of collaboration with human operators, explores the theoretical underpinnings of trust, analyzes the key factors influencing trust dynamics in addition to discussing strategies for building and maintaining trust. The review also investigates mechanisms of trust erosion and repair, and considers the implications for designing effective and trustworthy human-AI partnerships. Through consolidating research from diverse fields and highlighting persistent gaps, this review seeks to provide a comprehensive overview of the current state of understanding and inform future research directions. The aim is to contribute in enhancing trust and collaboration in the increasingly prevalent operator-AI teams.

A. Structure of the Review

The review is organized into several sections, as shown in Fig. 2: Section I: structured exploration of trust dynamics in operator-AI teams. Section II: classifying AI systems pertinent to operator-AI collaboration, and defining the scope of AI. Section III: primary theoretical frameworks informing trust, from different fields. Section IV: analysis of factors influencing trust dynamics. Factors categorized by AI, human operator, and context. Section V: temporal processes of trust (formation, calibration, decay, repair). Section VI: case study on Clinical Decision Support Systems (CDSS). Section VII: critical

discussion and synthesis, challenges, limitations, unresolved questions. Section VIII: conclusion, core arguments, gaps, methodological approach, and future directions.

Thus, having explored the foundational dynamics of trust in operator-AI teams, it is now crucial to understand the diverse types of AI systems involved. Therefore, Section II will discuss classifying AI systems to identify the types relevant to operator-AI collaboration and define the AI scope in our review.

II. Defining the Scope: AI Systems in Operator-AI Teams

The term "Artificial Intelligence" encompasses a wide spectrum of technologies with different capabilities and functionalities [2]. In order to effectively discuss trust dynamics in operator-AI teams, it is crucial to specify the types of AI systems most relevant to collaborative contexts. Researchers need to ensure that interconnectedness and common goals exist or are possible [6], [9]. AI systems can be classified along several dimensions, including their capabilities relative to human intelligence, their core functionality based on how they process information, and finally, based on their intended role and scope within human interaction [3], [33]. Table I synthesizes common classifications relevant to operator-AI teaming. It's drawn from the original draft and sources including Jutel [2], Makarius [33], Bansal [34]–[36](Refer to Table I).

Table I [2], [33] illustrates AI systems classification in the context of operator-AI teaming. The table classifies the AI systems into three general groups: capability, functionality, and scope or role. Each group is divided further into categories, including Narrow AI (ANI), General AI (AGI), and Super AI (ASI), according to their capabilities. The second group contains AI systems classified based on functionality and encompasses Reactive Machines, Limited Memory (Alteration/Symbiotic under the Scope/Role classification), Theory of Mind, and Self-Aware. The last group is based on the scope or role of the AI model and contains six different types of AI systems. The review will focus on the Limited Memory AI systems as this category represents AI systems currently interacting with operators and possess sophisticated capabilities within specific domains. This category also can collaborate with humans to adapt and share goals with medium novelty ,content changing scope. Furthermore, these systems also can learn from experience, enabling advanced collaborations from the perspective of organizational integration and human roles [33]. Operator-AI teaming often involves tasks such as Augmentation, Alteration, and potentially Autonomous systems, where the relationship is complementary, symbiotic. The tasks also could involve careful monitoring and intervention by the human operator. The human role shifts from simply controlling a tool, coaching, or overseeing to collaborating with a true and more capable AI partner. In the future, Artificial General Intelligence (AGI) or Superintelligence (ASI) will need more examinations as they remain largely theoretical. Based on the above, this review focuses primarily on interactions with Narrow AI systems exhibiting Limited Memory functionality. These systems are more suitable for our study because they operate in roles that involve complementarity, symbiosis, or monitored independence (Augmentation, Alteration, Autonomous). In such systems, dynamic trust, reliance calibration, and effective teaming processes are most critical and currently researched. While future AGI might introduce radically different trust dynamics, investigating trust with current and near-future interactive Narrow AI could provide the necessary foundation. The principles discussed, particularly regarding transparency, reliability, communication, and human factors, are likely to remain relevant, even if the are modified as AI capabilities advance.

Having established the scope and types of AI systems in this context, it is time to understand the theoretical foundations for studying trust. Therefore, Section III delves into the primary theoretical frameworks drawn from various fields that inform our understanding of trust in the context of human-AI collaboration.

III. Theoretical Foundations of Trust in Human-AI Collaboration

Several theoretical frameworks represent a starting point to understand trust dynamics in operator-AI teams.

1) First, Social Exchange Theory (SET): Originating in sociology [22], according to SET, social relationships are built on the exchange of resources. Individuals seeking to maximize benefits and minimize costs. Human operators are more likely to cooperate if the perceived benefits of teaming with AI (e.g., improved decision making, reduced time, increased efficiency) outweigh the costs (e.g., risk of errors, loss of control, reduced cognitive effort). Trust develops when the perceived benefits consistently outweigh the costs, and enhances willingness to continue cooperation (reliance) with the AI [37], [38]. Several factors such as perceived fairness, reliability, and competence of the AI can influence this cost-benefit analysis and subsequent trust [39]. However, applying SET directly faces various challenges due to the AI's lack of social factors and the opacity of its "intentions" [40].

2) Cognitive Load Theory (CLT): Focuses on the limitations of human working memory, and distinguishes between three types of memories: intrinsic load (task complexity), extraneous load (interface/information presentation), and germane load (schema construction/learning). CLT developed in educational psychology Sweller. In the context of human-AI interaction, CLT states that a poorly designed AI systems can impose high extraneous cognitive load (e.g., confusing interfaces, overly complex explanations). This can hinder understanding and trust formation [41], [42]. Conversely, well-designed AI can reduce cognitive load by automating routine tasks or presenting information clearly and concisely. A well-designed AI model more likely help in freeing up cognitive resources for operators to better evaluate the AI's output and build trust [43]. Managing cognitive load is thus crucial for effective collaboration and trust [44].

3) Human-Computer Interaction (HCI) and Technology Acceptance Models: HCI research provides numerous models for understanding user interactions with various technologies.The Technology Acceptance Model (TAM) and its more recent extensions such as "AI Adoption TAM", "User Acceptance of AI Systems", or "Trust and Technology Acceptance Models for AI" are some of those basic models. These models emphasize perceived usefulness and perceived ease of use as key elements that determinants of technology acceptance. Both elements

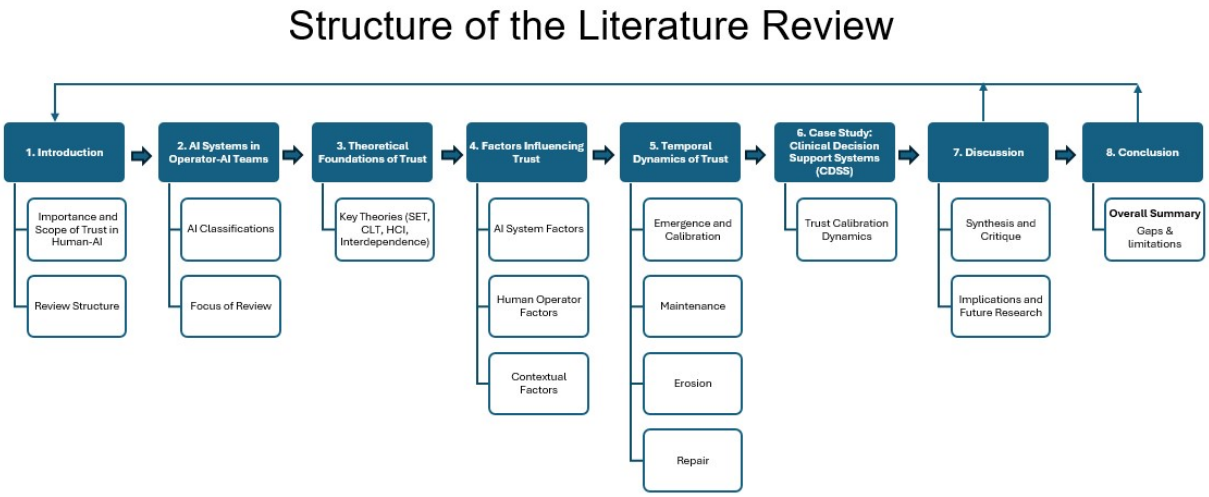


Fig. 2. Organization of the literature review.

are essential and may be considered related to trust more recent models, including the Unified Theory of Acceptance and Use of Technology (UTAUT), incorporate additional factors. Some important factors are social influence and facilitating conditions [45], [46]. In the field of HCI, multi-dimensional models of trust in automation and AI have emerged (e.g., competence, benevolence, encompassing, integrity, predictability, and dynamic). These models highlight the importance of system characteristics such as performance, and design, user characteristics such as disposition to trust, and finally contextual factors in shaping trust over time [42], [47], [48].

4) Theory of Mutual Dependence or Interdependence Theory: Based on these theories, interdependence among team members is crucial for achieving common goals. This applies whether this interdependence is within traditional teamwork between humans or between humans and artificial intelligence [49]. In the team, trust is cornerstone and can act as an auxiliary worker in interdependent relationships, facilitating coordination, communication, and willingness to rely on others [50]. In human-AI teams, this interdependence should go both ways, requiring operators to trust the AI's capabilities within its designated role, while the AI (through its design) must reliably fulfill that role [41]. Correctly calibrating trust dynamics in light of AI capabilities is crucial and a significant challenge, especially given the impact of influential factors. Also, it's important to establish shared awareness and mental models between human and AI agents [51]–[53]. These frameworks collectively highlight that trust in operator-AI teams is a complex phenomenon influenced by several factors such as perceived utility, cognitive feasibility, system design, user psychology, and the dynamics of interdependent collaboration.

These diverse theoretical perspectives provide a robust foundation for understanding trust in human-AI contexts. We can now turn to a detailed analysis with these theoretical frameworks established. Section IV systematically analyzes the multifaceted factors identified in the literature that influence trust dynamics.

IV. Factors Influencing Trust in Operator-AI Teams

Trust is not static but emerges and evolves based on interactions between the operator, the AI system, and the context. Key influencing factors can be categorized as follows:

A. AI System Factors

1) Performance and Reliability: Performance is arguably the most fundamental factor in building trust. It must be consistent, accurate, and reliable [12], [39], [42]. Predictability, or the extent to which the AI behaves as expected, is also crucial [40], [54]. In addition, errors or inconsistent performance significantly erode trust between operators and AI systems [24], [29].

2) Automation and Trust: Fig. 3 illustrates a model of trust in automation [12], [55], indicating the factors that directly impact how a human operator builds and maintains trust in automated systems and how that trust influences the operator's reliance on automation. At the top of the figure, the large box indicates the general context associated with the formation

TABLE I. CLASSIFICATIONS OF AI SYSTEMS RELEVANT TO OPERATOR-AI TEAMING

Classification Axis	Category	Description	Examples Relevant to Teaming	Typical Human Role
Based on Capability	**Narrow AI (ANI)**	Designed for specific tasks; cannot perform outside training domain.	Facial recognition, Driving assistance, Diagnostic aids, NLP chatbots, Go programs	Controller, Collaborator
	General AI (AGI)	Possesses human-like cognitive abilities; can learn and perform various intellectual tasks (mostly hypothetical).	Hypothetical versatile assistants or partners.	Co-creator, Comprehend
	Super AI (ASI)	Surpasses human intelligence across all domains (purely hypothetical).	Speculative superintelligences.	Comprehend
Based on Functionality	**Reactive Machines**	Reacts to stimuli; no memory of past experiences.	Early game AI (e.g., Deep Blue).	Supervisor (limited teaming)
	Limited Memory	Learns from past data/experience to inform future decisions; most current AI applications.	Chatbots, CDSS, Recommendation systems, Autonomous vehicles.	Controller, Collaborator, Coach
	Theory of Mind	Understands emotions, beliefs, intentions (conceptual/developing).	Advanced virtual assistants, Social robots.	Collaborator, Co-creator
	Self-Aware	Possesses consciousness, sentience (theoretical).	Hypothetical conscious AI.	not specified
Based on Scope/Role	**Automation / Substitution**	AI replaces human tasks; low novelty, content-changing scope.	Assembly line robots, Basic DSS.	Controller
	Amplification / Supplementary	AI enhances human analysis; low novelty, context-changing scope.	Predictive analytics tools.	Conductor
	Augmentation / Complementary	AI assists humans in complex tasks; medium novelty, content-changing scope.	Surgical robots, Advanced DSS, Go programs, Chip design AI.	Collaborator
	Alteration / Symbiotic	AI and humans co-create/adapt; medium novelty, context-changing scope.	Deep learning systems needing expert interaction.	Co-creator
	Autonomous / Independence	AI operates largely independently; high novelty, content-changing scope.	Self-driving vehicles (higher levels).	Keep-in-Check (Monitor)
	Authentic / Singularity	AI potentially surpasses human roles; high novelty, content-changing scope (hypothetical).	Superintelligence (ASI).	Comprehend

Note: Capability and Functionality classifications adapted from Jutel [2], common AI taxonomies, scope and role classification adapted from Makarius [33].

and reliance of trust, including the individual's characteristics (such as experience and personality), the organization's organizational structure (such as policies and procedures), culture (differences in shared norms and values regarding trust in technology), and the work environment context (the circumstances and situations in which automation is used).

In the same box, the figure shows the factors that influence the level of automation implementation, which in turn contributes to the operator's perception and level of trust. For example, these factors include what others say, rumors, and interface design quality. Other factors include workload, effort required to participate, self-confidence and personal capabilities, perceived risks, and system setup errors related to automation. In the middle of the model, four boxes appear from the left to the right, representing the mechanism through which the trust audit process progresses, from its creation to its adoption. This process begins with gathering information, perceptions, and beliefs about automation, followed by an evaluation by the operator in preparation for transforming them into practical actions by determining how to deal with the automation and, finally, whether to adopt it. Feedback between some stages may alter perception, evaluation, and the operator's readiness to embrace or avoid the automation. This development occurs through dynamic interaction with the application context and influencing factors.

In the lower right of the model, we find the automated system and its related capabilities and usage procedures. On the opposite side (bottom right), the model depicts the user interface and its impact, including ease of use, information reception and clarity, and ease of feedback. In the middle part, the model demonstrates that the operator receives information about the automation that may contribute to their understanding and confidence, such as the purpose of the automation, how it is implemented, and the quality of performance (refer to Fig. 2).

3) Transparency and Explainability: Explainability, the ability of human operators to understand how an AI arrives at a decision or recommendation, and having insight into its processes, true capabilities, and limitations (transparency) are critical for building calibrated trust. Transparency is vital, especially for complex or "black box" models [56], [57], [56] ExplainableAI. Explainable AI (XAI) is a technique that aims to provide transparency insight, which is particularly important in high-stakes domains [58], [23], [30]. Even the availability of explanations, whether accessed or not, can increase perceived trustworthiness [24], [59].

4) Interface and Interaction Design: Several essential elements, including User-friendly interfaces, precise feedback mechanisms, and intuitive controls, contribute significantly to perceived ease of use and reduce extraneous cognitive load, fostering positive experiences and trust [6], [40], [47]. Human-centered design (HCD) principles are critical and emphasize

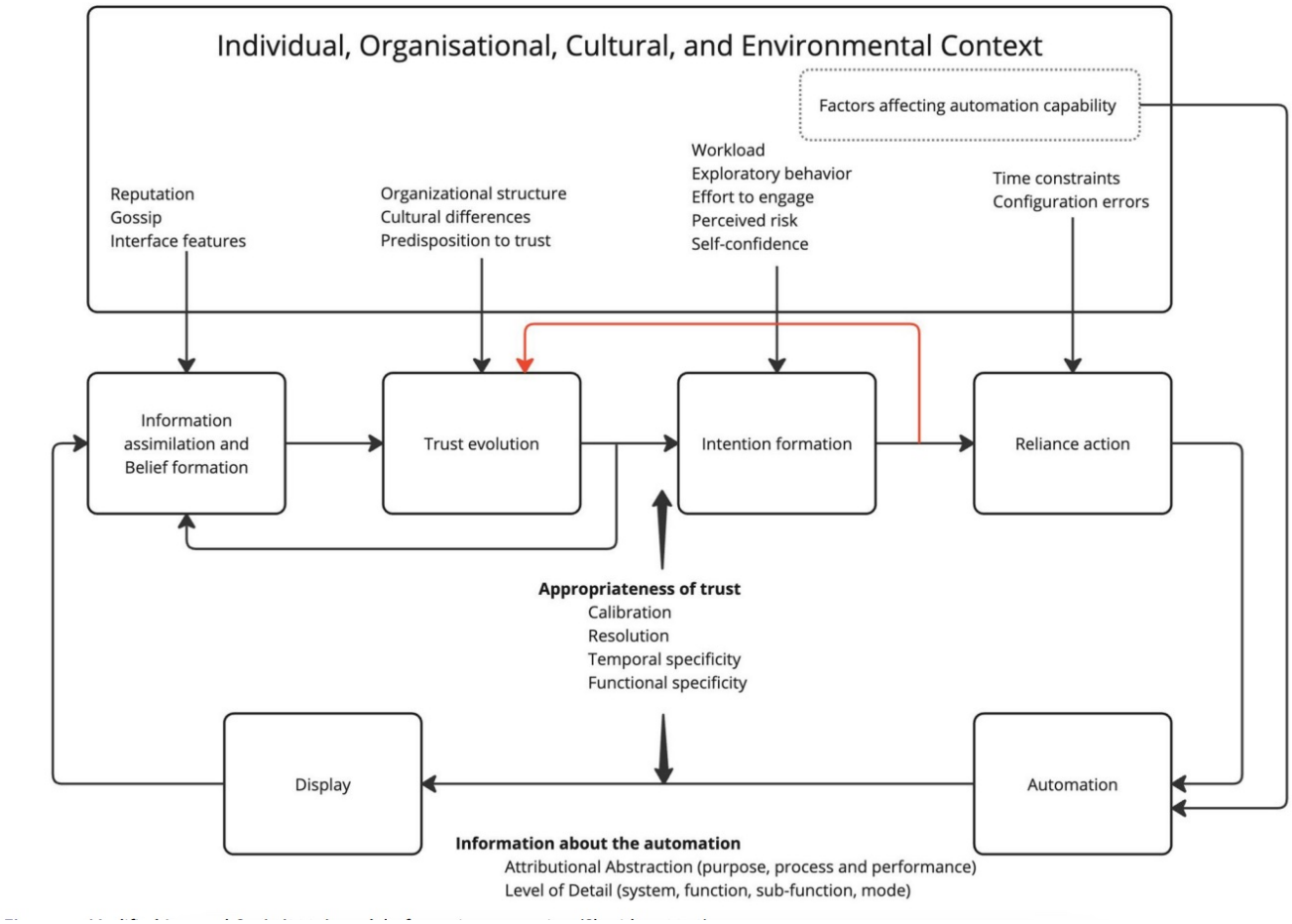


Fig. 3. The model of trust in automation, as originally proposed by [12] and modified by Sheridan [55]

tailoring the AI system to operator needs, capabilities, and context [6], [40]. Anthropomorphism refers to the design of AI with human-like characteristics (e.g., appearance, communication style) can sometimes foster emotional trust and rapport. However, it can also lead to unrealistic expectations or uncanny feelings if not implemented carefully [37], [40], [60].

B. Human Operator Factors

1) Disposition to Trust: Individuals vary in their inherent propensity to trust others or even to reassure to use technology, influenced by personality traits and past experiences [37], [42]. Experience and Expertise: Prior experience with AI or similar technologies, as well as domain expertise, are significant factors and can shape mental models and expectations of AI systems, influencing trust levels [61], [62]. Training can help develop accurate mental models but must be carefully designed [43], [63].

2) Cognitive Biases: Operators are susceptible to biases like automation bias (over-reliance on automation) or confirmation bias, leading to miscalibrated trust [6], [43]. Cultural Background: Cultural norms and values can significantly influence expectations, communication styles, and perceptions of technology, leading to cross-cultural differences in trust towards AI (GSD Venture Studios, n.d.), [26], [40]. AI design and interaction must consider these cultural nuances [26].

C. Contextual and Task Factors

1) Task Complexity and Criticality: Trust requirements in the operator-AI settings are high and often higher for complex tasks or in high-stakes environments, especially where the consequences of failure are severe (e.g., healthcare diagnostics, autonomous flying, military operations) [42], [64], [65].

2) Risk and Uncertainty: The capabilities of AI systems vary depending on the model and its characteristics. Higher perceived risk or uncertainty in the operational environment often necessitates greater trust in the AI's ability to perform reliably [62], [66].

3) Organizational Culture: The organizational context, including policies regarding AI use, transparency norms, and support structures, can influence operator trust [67].

Understanding these categories of influencing factors is essential, but trust is also a dynamic process that changes over time. Therefore, Section V focuses on the temporal processes of trust, examining how it initially forms and is calibrated and maintained over time.

V. Temporal Dynamics of Trust: Emergence, Calibration, Maintenance, Erosion, and Repair

Trust in an operator-AI team is dynamic rather than static. It is created and evolves over time, up and down, due to ongoing interactions with surrounding factors [68]. Operators assess trust and interpret AI behavior in specific contexts. Understanding the lifecycle of trust is crucial: how it is created, grows, or erodes due to adverse events, the strategies for repairing it when it is damaged, and how it is calibrated and maintained. The above elements are essential for developing strong and resilient partnerships in human-AI teams.

A. Trust Emergence and Initial Calibration

Initial trust between the operator and the AI is formed before or during early interactions [12]. Usually, it is shaped by factors such as relying on heuristics, pre-existing attitudes, and surface-level characteristics rather than deep experience. Other factors that may influence initial trust include the operator's disposition to trust technology in general [69], preconceptions based on reputation or prior experiences with AI or similar systems, the perceived purpose and design pedigree of the AI (e.g., developed by a reputable source), in addition to interface characteristics that signal professionalism or usability [12]. Initial interactions are critical; early positive experiences demonstrating reliability and competence can establish a positive trajectory. Similarly, early failures can disproportionately damage nascent trust [12], [69].

Operators begin trust calibration as the interactions with the AI progresses. They adjust their level of trust to align with the AI's perceived capabilities and limitations in the context [12]. AI processes and boundaries require transparent and accurate feedback on their performance to succeed in the calibration effectively. Providing explanations, particularly those relevant to the task and understandable to the operator (e.g., example-based or counterfactual explanations according to the contexts), can aid this process by supporting operators in building more accurate mental models of the AI's reasoning [6], [16]. However, ineffective or overly complex explanations can do the opposite. Poor explanations can hinder calibration. Explicitly revealing AI uncertainty or confidence levels can also be profitable in aiding calibration, even though it may sometimes decrease overall trust levels [6], [70]. Several models are being explored and incorporating Bayesian learning, which refers to applying probability theory to learning from data, especially regarding trust and interpretability [71]. The way operators update their appraisal of AI capabilities could be captured through the exploration based on observed outcomes [72] . To avoid the pitfalls of over-trust or under-trust, successful calibration is essential.

B. Trust Maintenance

The process of maintaining trust begins once it is established and calibrated between the operator and the AI [42]. Sustained reliability and predictable performance are crucial [12], [42]. Operators can anticipate AI actions and integrate them smoothly into their workflow through predictability. [54]. The importance of transparency goes beyond just the initial calibration but also effective communication, and ongoing verification, understanding. This importance is vital in a changeable environment or when AI's behavior is modified. Helpful strategies, such as proactive updates or confirmations, can reinforce perceptions of reliability and teamwork, contributing to trust maintenance [70]. Furthermore, AI systems designed with intuitive human-centered principles that align with operator workflows and cognitive limits are crucial and more likely to sustain trust over time by ensuring good user experiences [6], [11]. Sharing learning, an environment where both humans and AI can exchange knowledge and experience through collaboration, strengthen relationships, and preserve trust [11].

C. Trust Erosion

Trust can be described as "hard-won but easily lost". It has several actors that lead to trust erosion: First, AI errors and failures are the most direct reasons trust decreases. Although trust can be easily affected be errors, the type of error, its frequency, and its severity may have the greatest impact [12], [42], [54]). Sometimes, occasional errors, especially critical or unexpected ones, can significantly affect or damage trust.

The next cause can be the lack of transparency and explainability [57], as operators need to know everything while collaborating with their intelligent partners. Issues can arise when AI fails or behaves unexpectedly. Black-Box issues can be a real problem by hindering users' understanding of the cause of system errors. As a result, users may feel frustrated, uncertain, confused, and lack control [12], [16], [73]. Also, poor interpretations can lead to similar harms and results, leading to dissatisfaction or overtrust of the system.

Unreliable performance can make predictability difficult [42], negatively impacting the operator's trust in the system. For instance, surprises in AI system behavior can cause system performance to fluctuate, leading to a gradual erosion of trust over time [12], [63]. Other harmful reasons are perceived bias and unfairness. AI biased or unfair actions would cause damage to trust, particularly if they negatively impact the operator or specific groups. Trust can be severely damaged [33], [63]. Similarly, a discrepancy between operator expectations and the actions of an AI system can negatively impact trust between the two. Sometimes, AI capabilities are exaggerated, leading to feelings of disappointment or distrust, especially in early interactions with the system [12]. Trust can also be decreased as a result of poor usability or communication and can lead to trust decreased [12]. Communication is vital for operators while teaming with an AI partner. Awkward interfaces, ambiguous feedback, or ineffective communication protocols may create friction, confusion, and frustration. Leading to negative impacts on the user experience and indirectly eroding trust [12], [16], [70]. Finally, the context also matters because errors in high-stakes situations typically cause more significant trust erosion than errors in low-stakes contexts [42], [73].

D. Trust Repair

Trust may be easily damaged and must quickly be repaired [74]. Rebuilding trust after damage is not an easy challenge in human-AI interaction. Arguably, it can be even more so than in human relationships due to the AI's lack of genuine social recourse (e.g., repentance, sincere apology, remorse). Research indicates that simple apologies or explanations for

AI failure have shown only marginal effectiveness in restoring trust [29]. The study found apologies and explanations to be the most effective relative strategies compared to not taking action or no repair; however, their overall impact was not very influential.

First, acknowledging and explaining failures is a must action. Transparency about the source of an error is crucial, moving beyond simple apologies [6], [16]. Tailored explanations that address the specific failure context may be needed. Also, demonstrating corrective action could be a good strategy [74]. We need to confirm to the operators that the underlying cause of the failure has been handled (e.g., communication, policy, algorithm update, retraining) and that affected trust can be rebuilt faster. Nevertheless, It is essential to work towards consistent future performance. Reestablishing a track history of reliable and predictable behavior after a failure and overwriting the negative experience can be a strategy for repairing trust [6], [12]. Additionally, an innovative solution could be providing a means of control to the operator or the appropriate agency as support; this may help restore damaged trust [75]. This support could provide transparency and regain trust. Allowing operators to understand, verify, and potentially override AI actions, such transparent procedure can aid in restoring a sense of control and mitigate the vulnerability associated with trust decrease [31]. Longitudinal studies demonstrate that explanations accelerate trust restoration after a malfunction. This acceleration can be beneficial even if not always accessed [24]. However, We must acknowledge a need to develop accurate and effective solutions that guarantee systematic trust repair strategies for human-AI teams. This demand remains an open research area.

E. Trust Measurements

Trust is multifaceted and changes over time, so no single metric may be sufficient to assess it or measure the performance of the operator and AI team. Table II and Table III summarize the most important metrics used to assess trust and evaluate the effectiveness of the operator and AI team.

Table II shows several metrics for measuring trust drawn from various fields (human factors, HCI, and psychology research). Questionnaires and surveys are the most commonly used metrics due to their ease of use and are a branch of self-report metrics. Standardized trust scales are also used to assess predictive ability, perceived reliability, and trustworthiness, particularly in automation and artificial intelligence. There are two other questionnaire and survey scales: Likert-type and semantic differential scales, which ask the user to rate their level of agreement with a specific statement and use bipolar adjectives such as trustworthy-untrustworthy, respectively. The final scale used in questionnaires and surveys is Interview/Focus Groups, which focus on identifying causes and explaining phenomena, such as which factors specifically led to a high or low level of trust. The second type of scale that can be used to measure trust is Behavioral Measures (objective), which focus on observing how the operator interacts with the artificial intelligence system. Three measures under this category are Reliance/Adherence, Task Performance, and Interaction Patterns. Each has its purpose and nature; some are used for measuring the frequency of a particular phenomenon, others for error rate or the extent of the operator's need to search for information such as clarification or assistance. The final category of trust metrics that can be employed in an operator-AI team are Physiological Measures (Objective but more complex). These metrics involve observing a human operator's vital signs, such as Heart Rate Variability (HRV), Galvanic Skin Response (GSR)/ Electrodermal Activity (EDA), and Eye Tracking.

Table III shows which metrics are used for measuring operator-AI teamwork. Operator-AI Collaboration is a crucial measure, and its metrics overlap with trust metrics [76], but they focus on the effectiveness and nature of joint work. Operator-AI Collaboration is measured using the Team Performance Metrics, which measure efficiency, accuracy, and workload—interaction Quality Metrics for evaluating communication effectiveness, mutual understanding, role clarity, allocation, and adaptability. For Subjective Collaboration Measures, usually asks perators to reate some aspects. For instance, "I felt the AI was a bad manager" or "I comprehended how the AI was finishing the task". In addition, Subjective Collaboration Measures could also be used to find out how well the operator feels integrated with the AI as a team (refer to Table III).

The important question is what is the appropriate measure for measuring trust between operators and artificial intelligence (AI)? In addition, what is the best way to assess their collaborative relationship? Given the fluid and multifaceted nature of trust, more than one type of measure may be needed depending on the scope of the application and its extent, which underscores the importance of context. Trust must be measured at various levels for the trust enhancement model proposed following this review, which will primarily focus on enhancing trust in the operator-AI environment. There will likely be a need for subjective (e.g. questionnaires) measures with objectives (e.g., behavioral) related to the human element, the AI system, and context. These measures are also essential in most fields, such as ethical principles, measuring AI's true capabilities, and the human operator's readiness. By integrating these measures, weaknesses that have arisen in previous studies due to the absence of necessary measures can be overcome. Trust formation, decay, and repair mechanisms are fundamental, especially in high-stakes environments. To illustrate these concepts in practice, Section VI presents a case study focused on Clinical Decision Support Systems (CDSS). The CDSS provides a window to examine the evolution of trust through its various stages.

VI. Case Study: Trust Calibration Dynamics in Clinical Decision Support Systems

Many challenges of trust dynamics are uncovered in the CDSS case study [77], particularly in the calibration. These challenges and the impact of system design are vividly described in the context of Clinical Decision Support Systems (CDSS). In a high-stakes environment where decisions require high accuracy and reliability (examination and prescribing chemotherapy to patients), the study examined the impact of XAI's explanations on trust levels among 41 medical practitioners. Four categories of explanations were compared (example-based, global information-based, local context-based, and unrealistic explanations). Participants interacted with scenarios featuring both correct and incorrect AI recommenda-

TABLE II. METRICS FOR TRUST IN OPERATOR-AI COLLABORATION

Metric Category	Measurement Methods	Focus of Measurement
Self-Report	Questionnaires/Surveys, Interviews/Focus Groups	Subjective perceptions and beliefs about AI trustworthiness.
Behavioral	Reliance/Adherence, Task Performance, Interaction	Observable actions and patterns of operator behavior when interacting with AI.
Physiological	HRV, GSR/EDA, Eye Tracking	Physiological responses that may correlate with trust, stress, or cognitive load related to the AI.

TABLE III. METRICS FOR MEASURING OPERATOR-AI COLLABORATION

Metric Category	Measurement Methods	Focus of Measurement
Team Performance	Efficiency, Accuracy, Workload (Subjective/Objective), Situation Awareness	Effectiveness of the combined human-AI system in achieving task goals.
Interaction Quality	Communication Effectiveness, Mutual Understanding, Role Clarity/Allocation, Adaptability	Nature and quality of the interaction between the operator and the AI system.
Subjective Collaboration	Collaboration Experience Questionnaires, Perceived Team Cohesion	Operator's perceptions and feelings about working collaboratively with the AI.

Note: Tables II and III are adapted from [76]. Table II focuses on trust measures, and Table III shows the measures used to evaluate team performance.

tions, then trust was assessed using self-reported cognitive dimensions (understandability, reliability, competence) and behavioral indicators (agreement with AI, switching decisions, overall human-AI performance).

1) Trust Emergence and Understandability: The case study results indicated the importance of the type of explanation. The study found that the initial impressions were significantly influenced by the type of explanation provided. Example-based and unrealistic explanations were easier to understand and distinguish than other explanations (public and local). According to the results, explanations that included concrete examples or "what if" scenarios were more straightforward to comprehend than abstract scores. This finding aligns with psychological principles suggesting humans favor familiar, causal explanations. In addition, some participants have to separate local and global interpretations or misinterpret them due to difficulty in understanding them, ultimately hindering the possibility of these interpretations strengthening initial trust. This finding is consistent with psychological principles that humans favor familiar, known causal explanations. Furthermore, some practitioners have been forced to misuse local and global explanations because they are difficult to understand, which ultimately leads to these explanations not strengthening initial trust. This finding suggests that humans prefer only well-known logical explanations.

2) Trust Calibration Challenges: While providing explanations (regardless of class) improved overall human-AI task performance compared to no explanation, a critical finding emerged regarding calibration: explanations did not significantly help practitioners recognize incorrect AI recommendations. Explanations did not help practitioners substantially recognize incorrect AI recommendations. Participants found defective AI suggestions, including the no-explanation baseline, not easy to disagree with across all explanation conditions. Furthermore, providing explanations significantly increased participants' agreement with AI recommendations compared to not providing any explanations. Therefore, suggesting explanations may inadvertently lead to overreliance or confirmation bias. Participants may have used explanations heuristically (System 1 thinking) as a signal of AI competence rather than engaging in deeper critical analysis (System 2 thinking), especially given workflow pressures.

3) Trust Decay and Contextual Needs: Qualitative feedback revealed critical factors that could lead to trust erosion or prevent trust from developing appropriately, even with explanations. Participants provided feedback expressing their concern that in their feedback, participants voiced concern when explanations seemed disconnected from clinical context or task constraints (For example, explanations that are unrelated to the patient's condition or seem unrealistic). In addition, participants highlighted the need for assurances regarding the validity and capability of the explanation method (e.g., data integrity used, consistency) and allowing for desired tailoring and customization to fit professionals' specific workflow and information needs. The lack of multi-step explainability or the ability to ask follow-up questions about an explanation was also a barrier to deeper understanding and trust validation.

This critical case study powerfully demonstrates that simply providing explanations is insufficient for ensuring appropriate trust calibration in operator-AI teams. The study shows that explanations' type, design, usability, and contextual relevance critically influence how operators perceive, understand, and ultimately rely on AI systems. It highlights the risk of explanations fostering over-reliance if not carefully designed and integrated. In addition, the findings underscore the need for interfaces that support critical thinking, address usability constraints, provide necessary assurances, and allow for user adaptation and deeper inquiry. This case study illustrates the complexities of confidence calibration in a credit decision support system. Based on these specific examples, we now turn to Section VII, which provides a critical discussion and summarizes the key findings from all the previous sections.

VII. DISCUSSION

AI teams' evolving and expanding landscape requires operators to understand the dynamics of trust, a critical factor in successful collaboration. This literature review explores this vital topic, highlighting the shift toward viewing AI as a team partner rather than simply a tool, revealing the complex

nature of trust in these contexts, synthesizing various theoretical perspectives and research findings, and revealing key themes, significant challenges, and critical tips for future work. Building upon this overview, the following section delves into synthesizing the key findings regarding the multi-determined nature of trust dynamics.

A. Synthesis: The Multi-Determined Nature of Trust Dynamics

Researchers have consistently demonstrated that trust has a lifecycle through which it emerges, grows, and begins to decay. Trust is not driven by any single factor. Instead, it results from a complex interplay between the AI system's characteristics, the human operator's attributes, and the interaction context [12], [42]. AI performance and reliability are the foundation of trust [12], [37]. But these important factors are just as important as other key factors, such as transparency and explainability. This importance is particularly true in high-risk environments or sensitive systems [6], [16], [73]. Clear explanations facilitate operator roles and help operators calibrate trust; in other words, while performance and reliability form the foundation of solid trust, it is important to remember that "reliability" is not a fixed measure. Its perception can be highly subjective. An AI system may be objectively reliable, but trust can quickly erode if the operator fails to make decisions or if occasional errors are not explained and addressed effectively. This finding suggests that subjective perceptions of reliability, influenced by transparency and interpretability, are just as important, if not more so, than objective technical performance. On the other hand, complex or poorly designed explanations should be avoided. They can hinder usability, cause serious problems, and increase cognitive load. Poor explanations can also lead to over-reliance on AI or avoiding it over time.

While AI capabilities form a foundation for trust, the role of the human operator is equally vital. The human element is equally critical. Individual differences, willingness to trust, expertise, cognitive styles, and competencies, all these factors affect how operators perceive AI [12], [20], [78]. Cultural background emerges as vital in trust dynamics, nevertheless often Ignored. An operator's artistic background is an essential factor that can significantly influence trust perceptions and expectations of the team in trust [26]. The frequent neglect of cultural and even artistic backgrounds in current research is significant. This gap may be due to the predominant focus on the technical aspects of AI. However, the fundamental truth is that this technology ultimately approaches humans from an inherently cultural and individual perspective. For example, an operator from a particular culture may prioritize AI's ability to integrate with group systems seamlessly. In contrast, an operator from an individualistic culture may emphasize security and privacy more. Similarly, an artist may be more attuned to AI's creative potential and fine-grained judgment, while an engineer may prioritize precision and logical reasoning. These differences can lead to significant variations in trust formation and team dynamics. Developing shared mental models and how cooperative they seem can greatly affect the teamwork and trust levels [27], [22], [70]. The Hanabi experiment highlights a crucial point. The game experiment highlights how subjective perceptions of teamwork and agent interpretability can diverge from objective performance. Sometimes, rule-based agents are preferred over objectively comparable, but less predictable, learning-based agents [27], [79]. In other words, people such as operators might choose to work with an AI that they "understand" and "feel comfortable" with, even if that AI isn't the one that is better or will not necessarily achieve the highest score. Trust and perceived teamwork matter. The way in which an AI agent is predictable and interpretable can influence them.

Although the above synthesis presents a clear picture of trust dynamics and the factors influencing them, it is important to recognize that there are still many ongoing challenges and limitations in the operator-AI environment that merit further discussion.

B. Critique: Persistent Challenges and Limitations

Despite the advances, several critical challenges continue in operator-AI settings. Firstly, the problem of trust calibration stays central as indicated by Naiseah and Vössing in [16] and [6]. According to Naiseah and Vössing, even well-intentioned transparency mechanisms, such as explanations, don't automatically guarantee appropriate reliance; explanations can sometimes lead to over-trust or fail to help users detect AI errors. A significant limitation is designing adequate interfaces and precise interactions that promote critical evaluation (System 2 thinking) rather than heuristic acceptance (System 1 thinking).

Another significant issue is that we are still unable to understand the deep problem of trust repair, especially after significant AI failures, and it's still challenging to achieve it successfully. Current strategies have not gone beyond Apology, and basic explanations are still insufficient [29]. One of the challenges that complicates repair efforts is the asymmetry of the operator-AI relationship. AI lacks genuine remorse and social accountability, which is taken from interpersonal trust literature. Additional research is needed on robust repair mechanisms. The new studies potentially involve verifiable demonstrations of learning and correction by AI.

Researchers conduct most studies in AI from a Western perspective, often ignoring other perspectives. Furthermore, much of the researches are conducted in laboratories and controlled environments. Often, non-experts participate, or participants perform simple roles. For instance, while useful, the Hanabi game [27] may not capture high-stakes complexities. Longitudinal studies can help us better understand trust dynamics, especially when done in real-world settings. These studies allow us to track changes and influencing factors over time. These studies are rare, but they are essential [24], [59]. Furthermore, a significant portion of research on AI technologies originates from Western cultural contexts. This central view can limit the generalizability of findings and potentially lead to culturally inappropriate AI designs [25], [26]. Beyond these practical challenges, broader cultural issues also complicate the field.

Fourth, the very definition and measurement of "teaming" and "trust" in the human-AI context are still evolving, which confirms the need for more extensive and accurate studies. Metrics that primarily focus on measuring task performance fail to measure the subjective quality of the interaction, potentially impacting long-term adoption and effectiveness [27], [79]. There is still a great need for vital multidimensional measures that can capture tangible cooperation, shared importance, and measured dependence. Given these challenges, the findings

compiled in this review have direct implications for AI design, training, and policy.

C. Implications for Design, Training, and Policy

The synthesized results of our review confirm the direct implications of trust on AI design, particularly concerning human-centered approaches [6], [58]. The repercussions mean prioritizing not just AI System performance but also designing transparency and explainability features that are genuinely useful and understandable to the specific operator in their context [6], [16], [30]. To achieve this, it is required to involve end-users in the design process and considering their needs and preferences. As well as cognitive load, usability, and design for effective communication [70]. Incorporating mechanisms for behavioral synchrony [19] could also enhance teaming. Trust dynamics also significantly impact how tasks are allocated between humans and AI. When operators do not trust AI, they often rely on manual methods, decreasing efficiency. Conversely, over-reliance on AI can lead to unjustified trust and potentially fatal errors. Therefore, understanding and managing trust goes beyond just a psychological issue; it is a fundamental aspect of effective work design in operator-AI teams. In addition, considering the division of labor and interaction patterns from an organizational design perspective can provide structure [11], [41]. Crucially, designs must be sensitive to potential cultural differences in expectations and interaction styles.

For operator training and competency development, the focus should be on the key factors that enhance operator trust. In particular, there is a demand to focus on the key appropriate mental models of AI capabilities. Critical evaluation skills are essential for operators; it is important for training to go beyond technical skills. Operators need "trust literacy", meaning they can critically evaluate AI, calibrate their trust appropriately, and understand when and how to intervene. In addition, operators must develop the competencies necessary to interact with AI, including digital literacy, adaptability, collaboration, and problem-solving skills [78]. Finally, training should explicitly address foundational issues such as automation bias and calibration strategies. Such capabilities have implications for curriculum development and assessment across multiple disciplines.

Form organizational policy, promoting and fostering a new culture that supports transparency, ethical oversight, and learning from successes and failures in human-AI teaming is vital [6], [33]. As a necessary direction for policies for the manufacture and employment of human-based artificial intelligence, clear guidelines on roles, responsibilities, and accountability within human-AI teams are essential, especially in high-stakes domains [6], [73]. Who is responsible when something goes wrong in a human-AI team? The dynamics of trust can blur the lines of accountability. Who bears responsibility if an operator gives AI too much trust and follows incorrect advice? The operator or the AI designer? These legal and ethical considerations need to be studied and regulated. Policies should also consider the potential for differential impacts of AI systems adoption across demographic groups, e.g., age [20]. These findings highlight several gaps in the existing literature, warranting further future investigations in several areas.

D. Future Research Directions

The existing body of literature highlights significant gaps that necessitate future investigation in several areas. First, there is a need for cross-cultural trust dynamics studies, especially for conducting empirical studies across diverse cultural contexts. These studies are essential to comprehending how cultural values, norms, and communication styles interact with trust dynamics and how these cultural elements shape trust emergence, calibration, and repair in human-AI teams. Developing culturally adaptive AI interaction designs is another study area. Future research also needs to look at longitudinal studies. Literature highlights the importance of implementing long-term studies, especially in realistic operational settings, to observe how trust evolves. Also, to learn how to adapt to changing AI capabilities or task demands and recover (or fail to recover) from significant failures over time. In addition, current research underscores the importance of conducting a study to explore effective trust repair mechanisms. There is a need to move beyond simple apologies and explanations. It is time to investigate and develop robust trust repair strategies, potentially involving demonstrable AI learning and correction, adaptive transparency levels, or mechanisms for shared responsibility.

The next area that the thorough review reveals a compelling need to examine is calibration interventions. Literature highlights the need for evaluating and designing interfaces and training methods specifically aimed at improving trust calibration, helping users critically assess AI output, and avoiding over-reliance and under-reliance. These topics are significant in areas that require a cognitive load or time pressure. The development of diverse and effective AI error detection tools is also required. Their role should extend beyond identifying correct outputs to include error checking.

For the coming AI systems, researchers must further define and investigate the critical issues beyond task performance, including "teaming competencies". AI agents need to be perceived as effective collaborators by humans. Specifically, proactive communication, adaptivity, predictability, and expressing intent. Ethical frameworks for trust are more necessary than ever. Researchers must continue to explore and connect such frameworks to aspects of our lives in general and to the trustworthy cooperation of the operator AI in particular. The focus must address potential issues of manipulation, undue influence, and the balance between fostering trust through ethics and privacy, and security and between maintaining appropriate operator's vigilance and autonomy. The findings also underscore the importance of developing future AI models capable of tracking trust and addressing its dynamics, particularly within operator-AI team settings. These models must feature advanced capabilities that address bias, emotions, and social influence and have the ability to learn over time [72].

E. Solutions from Existing Literature

This review highlights significant gaps in the current understanding of trust dynamics in operator-AI collaboration, particularly regarding enhancing trust by increasing AI transparency and explainability (XAI), the need for adaptive trust calibration mechanisms, designing effective human-AI collaborations and

teams, ensuring AI robustness, communicating uncertainty, and incorporating ethical AI principles and governance into design. The current literature offers emerging and promising directions for addressing these challenges. For example, the explorations of the need to develop human-centered XAI designs, presented in [80], can be adapted to provide insights into the problem of black-box systems, which are one of the reasons why decisions are difficult for operators to understand, undermining trust and complicating problem diagnosis. New designs provide understandable explanations (a posteriori and a priori), focusing on how these explanations are presented, such as through visual interfaces or natural language. The goal is to enhance operator understanding and trust. Similarly, implementing the mechanisms proposed by [81] can support adaptive trust calibration solutions for dynamic systems where operators rely on AI. These systems use real-time feedback about AI trust and uncertainty and targeted training interventions to help operators maintain appropriate trust levels. Similarly, to meet the requirements for effective human-AI teamwork and collaboration models, a potential solution may be [82]. It can provide a valuable perspective for reconsidering the weaknesses of human-AI collaboration caused by unclear roles, communication breakdowns, and a lack of shared understanding of tasks and constraints that can be addressed. According to this work, human-AI teamwork can be enhanced by focusing on joint task design where roles are optimally distributed based on strengths. This includes enabling shared mental models through AI design and implementing flexible autonomy for dynamic control transitions. Trust erosion is another gap caused by miscommunication or errors in AI. The work presented in [83], [84] can help address this trust erosion. AI systems are known to fail unexpectedly or behave unpredictably, significantly eroding user trust and potentially leading to risks with dire consequences. Integrating ethical AI principles and governance into all phases of the AI development lifecycle is the missing solution to this problem. Implementing bias detection techniques, fairness metrics, and privacy preservation will be helpful. Success in achieving this could lead to building systematic trust through responsible design and auditing, as proposed in [84], [85]. We still need to expand the scope of ethical principles. This expansion will address broader ethical concerns such as bias, fairness, and accountability. Addressing these concerns increases operator trust and facilitates AI adoption. These challenges and limitations pave the way for future research. Finally, Section VIII concludes with a review and summarizes the main arguments.

To summrize, scholarly articles in the review consistently highlighted essential and inspiring subjects. Handling these dreaming topics through continued interdisciplinary collaboration is the only way to bring them into reality. It is necessary to continue combining insights from computer science, human factors, psychology, sociology, and organizational sciences. Only in this way can we realize this bright future vision, one in which more reliable AI systems collaborate with a greater understanding of the dynamics of trust. Even more importantly, we can address imminent risks, such as those witnessed in HAL 9000 from the movie "2001: A Space Odyssey" story, which reminds us of the importance of calibrating and enhancing trust between humans and their artificial companions at all stages. While these gaps exist, the current literature offers emerging and promising directions for addressing these challenges.

VIII. Conclusions

Trust is dynamic, not static. It has a starting point and grows and wanes according to the influencing factors that intertwine to define its parameters. These significant factors include human perception, AI capabilities, and changing context. Given the importance of trust, we must know the best ways to maintain it. Indeed, we need to dig deeper to know more about the factors that lead to its rise and decline and the best strategies for repairing it if it is damaged. Despite significant progress, our current methods fall short of capturing the intricacies of trust, emphasizing the need for more work to explore the mysteries of trust in the new and rapidly expanding work environments. AI performance is often referred to as the foundation upon which trust in human-AI collaboration rests. Other equally essential elements include transparency, explainability, and accuracy. However, a closer look reveals that these foundations suffer from cognitive holes and require continued efforts to bridge existing gaps, several of which were revealed by this review. AI system transparency is an important and influential factor in collaborative work in human-AI teams. However, it was found that it is not always sufficient when we looked closely at the results of a case study on an XAI used as a Clinical Decision Support System to assess the importance and impact of transparency. In this study, operators (healthcare practitioners) were presented with explanations varying in clarity levels, ranging from start forward to semi-clear to inconsistent. The results demonstrated that simply providing explanations for transparency does not transform a black box problem into a white box and, most importantly, does not ensure proper trust calibration. In the study, some explanations reinforced overreliance on the system, while participants ignored others, especially unclear or inaccurate ones. These conclusions indeed raise our concern in such a high-stakes environment, and it is a wake-up call—people focused on providing information, not necessarily ensuring its compatibility with human understanding and critical evaluation. Digging deeper, it was found several factors influencing the nature of trust, some related to AI systems, others to operators and context. These multiple factors interact in a complex dynamic that ultimately enhances or degrades trust. Operators have inherent skill capabilities and influential cognitive limitations, including individual propensities, cognitive biases, and cultural backgrounds, whose role in shaping trust should not be overlooked. In addition, there are the required competencies for interacting with AI. While understanding the human factors involved in communication and teamwork is vital, this is insufficient www.ijacsa.thesai.org 12 — P a g e (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 07, 2025 in a human-machine environment devoid of human emotions and social stimuli. On the contrary, this relationship with unstable and renewable factors presents new challenges, many of which still need to be explored. Why do the cross-cultural variations remain a glaring gray spot when partnering with AI? Similarly, are Western-centric models universal, or are studies overlooking crucial cultural nuances that drastically alter human-AI interactions? And how can people expect operators to develop calibrated trust when training often neglects to equip them with the critical skills for evaluating AI? Also, what would happen

if we left operators without enough knowledge or when leaving them vulnerable to automation bias and misinterpretations? Many fundamental questions related to the human aspect need scientific answers. As highlighted in the literature, we need to transform operators from passive users into active, discerning partners who can interpret and measure AI trust and not just accept the system as a black box. To successfully measure trust across its various stages, different metrics will need to be used depending on the context and complexity, in addition to each factor that must be measured in both the human and machine elements. Our review highlights several important gaps that require further research. First, we urgently need new requirements for the responsible design and deployment of AI systems, especially given the rapid developments in this field. Furthermore, a practical methodology is needed to ensure that the design adheres to the new requirements and the principles of ethics, transparency, and accountability to achieve responsible and trustworthy AI. During the collaboration lifecycle, the dynamic nature of trust (its fragile emergence, difficult maintenance, rapid erosion, and arduous repair) further complicates matters. Current attempts at trust repair are still limited to apologies or basic explanations, which appear woefully inadequate. Can an apology or clear explanation mend trust damaged by a system perceived as unreliable or biased? Or do we need to explore more human-like elements of trust repair? Likewise, despite growing recognition that AI is moving beyond a simple tool to a teammate, existing theoretical models struggle to keep pace. Existing frameworks are often helpful but fail to capture the unique challenges of human-AI interdependence fully. Similarly, there is a need to bridge the widening gap toward developing future systems capable of comprehensively replacing humans or transforming their role to become more like a tool used in a specific context when needed. So far, the hope to accurately define and measure "trust" and "operator-AI collaboration". The current metrics still fail to capture the essence of this relationship. The limitations we face in this regard are not merely academic controversies; they are realities with dire consequences unless more is done to bridge the gaps. Miscalibrating the future of human-AI teams will lead to disastrous consequences. This threat is most acute in high-risk areas, including healthcare, transportation, and defense. Society is entering an era of risks cloaked in complacency and misunderstanding. Innovative solutions need to be prioritized over the promise of AI efficiency. The review adopted a systematic approach to analyzing the existing literature by collecting and categorizing research papers in the selected field (the dynamics of trust in the collaborative environment between operators and artificial intelligence) to reveal areas of research focus and, conversely, absence. This task was followed by reviewing and examining relevant theoretical foundations that addressed the relationship between humans and modern technologies, such as AI, within the framework of collaborative work. The methodology identified areas where findings conflicted or where different studies used incompatible definitions. Synthesizing the findings, it was possible to determine whether certain aspects of "trust" or "operator-AI collaboration" were consistently overlooked or only superficially addressed. The synthesized outcome allowed for a comprehensive mapping of the current landscape of human-AI trust research, revealing underexplored areas and methodological inconsistencies. The identified gaps provide a rationale for the need for a framework, and these findings directly shape the aspects that the framework this review aims to address. The review confirms that the current approach is insufficient and that we need a fundamental shift beyond the stereotypical view of trust as merely a given emotion. Instead, trust must become a valuable ranking acquired through careful design of future operators' behaviors, perceptions, and ethics. This recommendation is consistent with the review gaps and emphasizes the need for a new approach to address the human aspects of operator-AI partnerships effectively. Therefore, exploring an innovative framework to address these needs, informed by cognitive and affective factors, will go beyond an academic endeavor to a vital solution. Through it, the hope is to contribute to safeguarding the future of this rapidly transformative technology. The new framework promises to transcend current limitations and enhance trust in human-AI collaborations. The development of this framework will be the focus of my following methodological study.

REFERENCES

[1] A. C. Clarke, *2001: a space odyssey*. Penguin, 2016.

[2] M. Jutel, M. Zemelka-Wiacek, M. Ordak, O. Pfaar, T. Eiwegger, M. Rechenmacher, and C. A. Akdis, "The artificial intelligence (ai) revolution: How important for scientific work and its reliable sharing." *Allergy*, vol. 78, no. 8, 2023.

[3] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2021.

[4] P. Grover, A. K. Kar, and Y. K. Dwivedi, "Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions," *Annals of Operations Research*, vol. 308, no. 1, pp. 177–213, 2022.

[5] F. Kitsios and M. Kamariotou, "Artificial intelligence and business strategy towards digital transformation: A research agenda," *Sustainability*, vol. 13, no. 4, p. 2025, 2021.

[6] S. Schmager, I. Pappas, and P. Vassilakopoulou, "Understanding human-centred ai: a review of its defining elements and a research agenda," *Behaviour & Information Technology*, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0144929X.2024.2448719

[7] C. Dragomirescu-Gaina, D. Philippas, and M. G. Tsionas, "Trading off accuracy for speed: Hedge funds' decision-making under uncertainty," *International Review of Financial Analysis*, vol. 75, p. 101728, 2021.

[8] N. J. Mcneese, B. G. Schelble, L. Canonico, and M. Demir, "Who/what is my teammate? team composition considerations in human–ai teaming," *IEEE Transactions on Human-Machine Systems*, vol. 51, pp. 288–299, 2021.

[9] V. Hagemann, M. Rieth, A. Suresh, and F. Kirchner, "Human-ai teams—challenges for a team-centered ai at work," *Frontiers in Artificial Intelligence*, vol. 6, p. 1252897, 2023.

[10] H.-A. Teaming, "State-of-the-art and research needs," *National Academies of Sciences, Engineering and Medicine, Washington DC*, vol. 10, p. 26355, 2022.

[11] M. J. Sáenz, E. Revilla, and C. Simón, "Designing ai systems with human-machine teams," *MIT Sloan Management Review*, vol. 61, no. 3, pp. 1–5, spring 2020. [Online]. Available: https://www.proquest.com/scholarly-journals/designing-ai-systems-with-human-machine-teams/docview/2392464050/se-2

[12] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of Human Factors and Ergonomics Society, Human Factors, 46(1), 50–80. URL: https://doi.org/10.1518/hfes.46.1.50_30392*, 2004.

[13] The Interaction Design Foundation. (2025) What Is Human-Centered AI (HCAI)? Accessed: May 23, 2025. [Online]. Available: https://www.interaction-design.org/literature/topics/human-centered-ai

[14] A. Korinek and A. Balwit, "Aligned with Whom? Direct and Social Goals for AI Systems," National Bureau of Economic Research, Working Paper 30017, 2022. [Online]. Available: https://www.nber.org/system/files/working_papers/w30017/w30017.pdf

[15] O. Vereschak, G. Bailly, and B. Caramiaux, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–39, 2021.

[16] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "How the different explanation classes impact trust calibration: The case of clinical decision support systems," *Int. J. Hum. Comput. Stud.*, vol. 169, p. 102941, 2023.

[17] J. Reyes, A. U. Batmaz, and M. Kersten-Oertel, "Trusting ai: does uncertainty visualization affect decision-making?" *Frontiers in Computer Science*, vol. 7, p. 1464348, 2025.

[18] M. B. Junaid and M. B. Hassan, "Ai and human-robot interaction: A review of recent advances and challenges," *GSC Advanced Research and Reviews*, vol. 18, pp. 321–330, 2024. [Online]. Available: https://gsconlinepress.com/journals/gscarr/sites/default/files/GSCARR-2024-0070.pdf

[19] M. Y. M. Naser and S. Bhattacharya, "Empowering human-ai teams via intentional behavioral synchrony," *Frontiers in Neuroergonomics*, vol. 4, 2023.

[20] M. C. Dorton and S. A. Harper, "The trust calibration maturity model for characterizing and communicating trustworthiness of ai systems," *arXiv preprint arXiv:2503.15511*, 2025. [Online]. Available: https://arxiv.org/pdf/2503.15511

[21] D. D. Scholz, J. Kraus, and L. Miller, "Measuring the propensity to trust in automated technology: Examining similarities to dispositional trust in other humans and validation of the ptt-a scale," *International Journal of Human–Computer Interaction*, 2025. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10447318.2024.2307691

[22] A. Bostrom, J. L. Demuth, C. D. Wirz, M. G. Cains, A. Schumacher, D. Madlambayan, A. S. Bansal, A. Bearth, R. Chase, K. M. Crosman, D. J. Gagne II, S. L. Henderson, A. McGovern, R. J. Redmon, Y. Rao, and I. Ebert-Uphoff, "Trust and trustworthy artificial intelligence: A research agenda for ai in the environmental sciences," *Risk Analysis*, vol. 44, pp. 1364–1378, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/risa.14245

[23] S. E. Kase, R. S. G. III, W. A. III, and C. S. Morris, "Trust and trustworthiness in human-ai interaction: A review and proposed research agenda," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1, 2022, pp. 1473–1477.

[24] K. Okamura and S. Yamada, "Adaptive trust calibration for human-ai collaboration," *Plos one*, vol. 15, no. 2, p. e0229132, 2020.

[25] S. Benk, J. M. Becker, and A. Kluge, "Defining human-ai teaming the human-centered way: A scoping review and network analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1250725, 2024. [Online]. Available: https://doi.org/10.3389/frai.2023.1250725

[26] S. Kang, A.-E. Potinteu, and N. Said, "Explainitai: When do we trust artificial intelligence? the influence of content and explainability in a cross-cultural comparison," *arXiv preprint arXiv:2503.17158*, 2025.

[27] C. Attig, P. Wollstadt, T. Schrills, T. Franke, and C. B. Wiebel-Herboth, "More than task performance: Developing new criteria for successful human-ai teaming using the cooperative card game hanabi," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–11.

[28] N. K. Dang, M. T. Spaan, and D. Lo, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. Article 426, 1–32, 2021. [Online]. Available: https://doi.org/10.1145/3479570

[29] C. Esterwood and L. P. Robert, "Repairing trust in robots?: A meta-analysis of hri trust repair studies with a no-repair condition," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 410–419.

[30] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1 – 45, 2021.

[31] A. Laitinen and O. Sahlgren, "Ai systems and respect for human autonomy," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[32] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems," *arXiv preprint arXiv:2105.03354*, 2021.

[33] E. E. Makarius, D. Mukherjee, J. D. Fox, and A. K. Fox, "Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization," *Organizations & Markets: Policies & Processes eJournal*, 2020.

[34] R. Bansal and A. Sangwan, "A comprehensive review of artificial intelligence," *International Journal For Multidisciplinary Research*, vol. 6, pp. 149–158, 2024. [Online]. Available: https://www.ijfmr.com/research-paper.php?id=28932

[35] M. H. Haag, S. R. Schmidthuber, D. F. Dietl, and M. R. R. Bitsch, "Exploring generative artificial intelligence: A taxonomy and types," in *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*. University of Hawaii at Manoa, 2024. [Online]. Available: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/fa9a6175-9ff2-4ad4-868e-fec5127cd430/content

[36] M. M. Al Rashed, M. S. Uddin, A. K. M. L. Khan, P. K. Roy, M. M. Islam, and M. S. Hossain, "Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems," *Frontiers in Artificial Intelligence*, vol. 5, p. 803098, 2022.

[37] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of management annals*, vol. 14, no. 2, pp. 627–660, 2020.

[38] E. Georganta and A.-S. Ulfert, "My colleague is an ai! trust differences between ai and human teammates," *Team Performance Management: An International Journal*, vol. 30, no. 1/2, pp. 23–37, 2024.

[39] J. Liu, X. Liu, D. Wang, and Y. Zhang, "Grading by AI makes me feel fairer? how different evaluators affect college students' perception of fairness," *Frontiers in Psychology*, vol. 15, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1221177/full

[40] A. G. Glikson and A. Kaplan, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 15, pp. 627–665, 2021. [Online]. Available: https://journals.aom.org/doi/full/10.5465/annals.2018.0057

[41] P. Puranam, "Human–ai collaborative decision-making as an organization design problem," *Journal of Organization Design*, vol. 10, pp. 75 – 80, 2020.

[42] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 30 3, pp. 286–97, 2000.

[43] R. Parasuraman and D. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 52, pp. 381 – 410, 2010.

[44] F. M. Cau and L. D. Spano, "Exploring the impact of explainable ai and cognitive capabilities on users' decisions," *arXiv preprint arXiv:2505.01192*, 2025.

[45] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management science*, vol. 46, no. 2, pp. 186–204, 2000.

[46] J. Dingel, A.-K. Kleine, J. Cecil, A. L. Sigl, E. Lermer, and S. Gaube, "Predictors of health care practitioners' intention to use ai-enabled clinical decision support systems: Meta-analysis based on the unified theory of acceptance and use of technology," *Journal of medical internet research*, vol. 26, p. e57224, 2024.

[47] M. Madsen and S. Gregor, "Measuring human-computer trust," vol. 53, pp. 6–8, 2000.

[48] E. Glikson and A. W. Woolley, "A systematic literature review of user trust in AI-enabled systems: An HCI perspective," *Human-Computer Interaction*, vol. 35, no. 10, pp. 933–1001, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/07370024.2020.1750577

[49] S. Tannenbaum and E. Salas, *Teams that work: The seven drivers of team effectiveness*. Oxford University Press, 2020.

[50] A. C. Costa, R. A. Roe, and T. Taillieu, "Trust within teams: The relation between individual trustworthiness, team trust, and team performance," *European Journal of Work and Organizational Psychology*, vol. 10, no. 3, pp. 225–244, 2001. [Online]. Available: https://doi.org/10.1080/13594320143000654

[51] M. R. Endsley, "Situation awareness misconceptions and misunderstandings," *Journal of Cognitive Engineering and Decision*

Making, vol. 9, no. 1, pp. 4–32, 2015. [Online]. Available: https://doi.org/10.1177/1555343415572631

[52] P. Esmaeilzadeh, "Use of ai-based tools for healthcare purposes: A survey study from consumers' perspectives," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–21, 2022. [Online]. Available: https://doi.org/10.1186/s12911-022-01891-8

[53] S. Kim, S. Choo, D. Park, H. Park, C. S. Nam, J.-Y. Jung, and S. Lee, "Designing an xai interface for bci experts: A contextual design for pragmatic explanation interface based on domain knowledge in a specific context," *International Journal of Human-Computer Studies*, vol. 174, p. 103009, 2023.

[54] J. G. G. De Visser, J. C. D. Van Der Meij, and M. W. B. Van Der Meij, "Inferring trust from users' behaviours: Agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration," *Frontiers in Robotics and AI*, vol. 8, p. 642201, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2021.642201/full

[55] T. B. Sheridan, *Humans and automation: System design, responsibility, and allocation of function*, 2nd ed. John Wiley & Sons, 2019.

[56] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, and S. Gil-López and Daniel Molina and Randy Benjamins and Raja Chatila and Francisco Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[57] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2017.

[58] S. Berretta, A. Tausch, G. Ontrup, B. Gilles, C. Peifer, and A. Kluge, "Defining human-ai teaming the human-centered way: a scoping review and network analysis," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[59] P. Malinowska and E. Alberdi, "Trust development and explainability: A longitudinal study with a personalized assistive system," *Algorithms*, vol. 8, no. 3, p. 20, 2024.

[60] D. Kishnani, "The uncanny valley: An empirical study on human perceptions of ai-generated text and images," Master's thesis, Massachusetts Institute of Technology, 2025. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/159096/kishnani-deepalik-sm-sdm-2025-thesis.pdf?sequence=-1&isAllowed=y

[61] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, P. Szolovits, and M. Ghassemi, "Do as ai say: susceptibility in deployment of clinical decision-aids," *npj Digital Medicine*, vol. 4, no. 1, p. 31, 2021. [Online]. Available: https://doi.org/10.1038/s41746-021-00385-9

[62] J. B. Lyons, K. Stokes, and M. D. McNeese, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 15, no. 2, pp. 627–652, 2021.

[63] S. Dhuliawala, V. Zouhar, M. El-Assady, and M. Sachan, "A diachronic perspective on user trust in ai under uncertainty," *arXiv preprint arXiv:2310.13544*, 2023.

[64] Z. Chen, Y. Luo, and M. Sra, "Higher stakes, healthier trust? an application-grounded approach to assessing healthy trust in high-stakes human-ai collaboration," *arXiv preprint arXiv:2503.03529*, 2025. [Online]. Available: https://arxiv.org/abs/2503.03529

[65] I. Gkikopouli and S. Kontogiannis, "A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger," *Electronics*, vol. 13, no. 18, p. 3660, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/18/3660

[66] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[67] U.S. Department of Health & Human Services, "Trustworthy AI (TAI) Playbook: Executive Summary," 2021, accessed: [Insert Date Accessed, e.g., May 23, 2025]. [Online]. Available: https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf

[68] M. J. McGrath, A. Duenser, J. Lacey, and C. Paris, "Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration," *arXiv preprint arXiv:2404.01615*, 2024. [Online]. Available: https://arxiv.org/abs/2404.01615

[69] J. Rehm, M. Söllner, and A. Kluge, "Defining human-ai teaming the human-centered way: a scoping review and network analysis," *Computers in Human Behavior Reports*, vol. 11, p. 100318, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10570436/

[70] L. Roeder, P. Hoyte, J. van der Meer, L. Fell, P. Johnston, G. Kerr, and P. Bruza, "A quantum model of trust calibration in human–ai interactions," *Entropy*, vol. 25, no. 9, p. 1362, 2023.

[71] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015. [Online]. Available: https://www.nature.com/articles/nature14541

[72] D. Gopinath, T. Kulesza, L. Ma, J. S. Siegel, P. Stone, and J. Zinman, "Bayesian modeling of human–ai complementarity," *Proceedings of the National Academy of Sciences*, vol. 119, no. 10, p. e2111547119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2111547119

[73] M. Cascella, J. Montomoli, V. Bellini, A. Vittori, H. Biancuzzi, F. Dal Mas, and E. G. Bignami, "Crossing the ai chasm in neurocritical care," *Computers*, vol. 12, no. 4, p. 83, 2023.

[74] S. Gao and Y. Yan, "Verbal or written? the impact of apology on the repair of trust: Based on competence- vs. integrity-based trust violation," *Frontiers in Psychology*, vol. 13, p. 884867, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.884867/full

[75] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[76] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, pp. 53–71, 2000. [Online]. Available: https://www.researchgate.net/publication/247502831_Foundations_for_an_Empirically_Determined_Scale_of_Trust_n_Automated_Systems

[77] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "How the different explanation classes impact trust calibration: The case of clinical decision support systems," *International Journal of Human-Computer Studies*, vol. 169, p. 102941, 2023.

[78] I. Popa, M.-M. Cioc, A. Breazu, and C. Popa, "Identifying sufficient and necessary competencies in the effective use of artificial intelligence technologies," *Amfiteatru Economic*, 2024.

[79] K. Chen, C. Gu, and J. Huang, "Calibrating trust in ai-assisted decision making," UC Berkeley School of Information, Capstone Project Report, 2020. [Online]. Available: https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/humanai_capstonereport-final.pdf

[80] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–15.

[81] N. Srinivasan and E. Thomason, "Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance," *arXiv preprint arXiv:2502.13321*, 2025.

[82] J.-H. Ju and S. Aral, "Collaborating with ai agents: Field experiments on teamwork, productivity, and performance," *Management Science (Forthcoming)*, 2025.

[83] H. Sun, Z. Li, J. Zhang, J. Fu, and J. Li, "Robustness in ai-generated detection: Enhancing resistance to adversarial attacks," *arXiv preprint arXiv:2505.03435*, 2025.

[84] M. Teufel, R. Pinsler, J. E. S. W. Reinders, A. Maier, and C. L. Buckley, "Improving counterfactual truthfulness for molecular property prediction through uncertainty quantification," *arXiv preprint arXiv:2504.02606*, 2025.

[85] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," *arXiv preprint arXiv:1906.01409*, 2021.

Designing an Empathetic Conversational Agent for Student Mental Health: A Pilot Study

Kaichi Minami[1], Choi Dongeun[2], Panote Siriaraya[3], Noriaki Kuwahara[4]*
Graduate School of Science and Technology, Kyoto Institute of Technology, Kyoto, Japan[1, 2, 3, 4]
Center for Social and Biomedical Engineering, Kyoto Institute of Technology, Kyoto, Japan[3, 4]

***Abstract*—This study presents the design and evaluation of a conversational agent aimed at supporting university students' mental health. We implemented two variants of a chatbot, referred to as A1 and A2, using large language models (LLMs). A1 employed a baseline prompt reflecting a structured yet neutral counseling style, while A2 was an enhanced version incorporating feedback from psychiatrists and findings from a preliminary study. Emotionally rich expressions, conversational variation, and mild self-disclosure are introduced in A2. A mixed-method user study with 18 participants was conducted to compare A1, A2, and human interactions. Results indicated that A2 significantly improved users' perception of empathy and engagement compared to A1, though human-level rapport was not fully achieved. These findings highlight the role of prompt design in creating emotionally responsive AI companions for mental health support.**

Keywords*—Conversational agent; mental health; large language models; prompt design; empathy; chatbot evaluation

I. Introduction

University students are increasingly vulnerable to psychological distress caused by academic pressures, social isolation, and the lack of immediate access to mental health resources. According to the national report published in Japan in 2022, nearly 20% of private university students have no one they can talk to about their problems [1], contributing to lowered academic performance and well-being [2].

AI-powered conversational agents have emerged as potential tools for addressing these issues. Such systems can provide scalable, anonymous, and readily available support [3], making them particularly attractive for non-clinical, and university-level applications [4]. Previous work focused on emotionally expressive responses and explored the design of empathetic conversational agents using GPT models [5]. However, recent studies have identified limitations in the emotional responsiveness of existing chatbots [6], particularly regarding inconsistent tone, inappropriate responses, and lack of sensitivity to crisis language [7, 8]. This study addresses the emotional limitations of traditional chatbots by proposing prompt-based interventions to improve empathy and user engagement—key challenges in mental health support.

Recent advances in large language models (LLMs) like GPT-4 have made it possible to develop more contextually aware chatbots. Even so, how prompt design affects the quality of mental health support in these systems remains an open question. To explore this, we developed and evaluated two chatbot variants: A1, a baseline chatbot with a structured but emotionally neutral prompt; and A2, an enhanced chatbot that incorporates empathetic phrasing, conversational variability, and light self-disclosure.

This study investigates how these design differences influence user perceptions of empathy, trust, and effectiveness in emotionally supportive dialogues. We aim to contribute new insights into prompt-based control of LLMs for mental health support and provide practical design guidelines for emotionally responsive AI companions.

This paper is organized as follows: Section II presents related work, Section III details the system design, Section IV describes the methodology, and Section V discusses the results.

By demonstrating the effectiveness of prompt-based strategies in enhancing user empathy, this study highlights the practical utility of design interventions in real-world mental health support systems.

II. Related Works

The integration of conversational agents into mental health support systems has garnered significant attention, particularly in the context of an increasing number of university students experiencing psychological distress. Several studies have explored the development, implementation, and evaluation of AI-driven chatbots aimed at providing accessible and empathetic support.

Liu et al. proposed ChatCounselor—an LLM trained on real-world counseling transcripts—which demonstrated the feasibility of generating psychologically informed responses [9]. Similarly, Lai et al. introduced the Psy-LLM framework, combining pre-trained LLMs with structured psychological Q&A data to support large-scale mental health consultations [10]. In another domain-specific study, Yao et al. evaluated three chatbot systems targeting postpartum mood and anxiety disorders, highlighting the importance of context-aware responses tailored to sensitive health needs [11].

The clinical efficacy of AI chatbots in mental health care is also gaining empirical support. Jacobson's team reported that their chatbot Therabot led to significant improvements in symptoms of depression and anxiety in a randomized controlled trial involving college students and other vulnerable groups [12]. Similarly, Ahmed et al. conducted a pilot clinical trial that demonstrated improvements in students' overall well-being through structured AI-based interactions [13].

Beyond algorithmic design, prompt engineering has been identified as a critical factor in improving the performance and compliance of AI chatbots in mental health domains. Waaler et al. employed a multi-agent architecture to ensure consistent adherence to prompt instructions, especially in educational contexts [14]. In Sahoo et al.'s review [15], prompt-engineering techniques such as Few-shot and Chain-of-Thought are reported to be effective for generating empathetic responses. Building on these findings, this study proposes an original five-stage dialogue flow—Introduction → Rapport Building → Problem Exploration → Empathetic Response → Closing—to structure both the prompts and the system's replies.

Nonetheless, several studies warn against overestimating the empathetic capabilities of AI systems. Raczka argues that while AI therapists provide accessibility, they lack the nuanced interpersonal understanding necessary for deep emotional support [7]. Roshanaei M. et al. further demonstrate that chatbots may exhibit gender-based biases in performing empathy, pointing to ethical challenges in human–AI emotional interaction [8].

These ethical concerns extend to data privacy and surveillance. As Vincent notes, AI therapy tools may operate as covert surveillance systems in the absence of robust data protection frameworks [16]. Additionally, Ferguson reports that exposing LLMs to traumatic narratives—such as those involving war or violence—can induce "anxiety-like" outputs, potentially compromising the stability and interpretability of their responses [17].

Lastly among these studies, several comprehensive reviews provide theoretical grounding for the emotional dimensions of human–AI interaction. A. S. Raamkumar and Y. Yang conducted a systematic review on emotionally aware conversational agents, identifying key factors contributing to perceived empathy and engagement [18]. X. Zheng et al. evaluated ChatGPT's current capabilities and limitations in mental health support, arguing that prompt design plays a major role in user satisfaction [19]. In an empirical CHI study, B. Liu and S. S. Sundar investigated which conversational behaviors in chatbots most closely approximate human-like support in therapeutic contexts [20].

Taken together, these studies underscore the challenges and opportunities of deploying AI chatbots for mental health support in higher education. Our study builds upon this foundation by experimentally evaluating how variations in prompt design affect users' perceptions of empathy, trust, and emotional support in an LLM-powered chatbot designed for university students' mental health assistance.

To improve citation clarity, the references included in this section have been chosen and grouped based on their relevance to specific thematic aspects, such as LLM-based design, prompt engineering, and ethical concerns. Although some references appear at the end of paragraphs for structural flow, each cited work has been carefully selected to support the preceding discussion. Recent literature (2023–2025) has been prioritized, and older works are included only when they provide essential historical or conceptual context.

While prior work explored empathetic dialogue generation, few studies have directly compared structured prompt designs within a controlled experimental setting.

III. System Design

A. System Overview

Our system is designed as a web-based chatbot interface built using Streamlit [21] and integrated with the GPT-4 API [22] via OpenAI's endpoint. Users engage in one-on-one conversations with the chatbot, which responds in real time using prior dialogue context. The primary goal is to simulate supportive, human-like conversations that address student concerns related to emotional stress, anxiety, and academic challenges.

As shown in Fig. 1, the system comprises a front-end chat interface, a conversation history manager, and an LLM-based response engine. Depending on the assigned variant, either prompt A1 or A2 is used to condition the output behavior of the chatbot.

B. Prompt Design Strategy

To explore the effects of prompt structure and tone, we developed two chatbot variants:

1) A1 (Baseline): This variant follows a formal, neutral tone and uses a five-step structure derived from general mental health counseling principles [15]. It asks users to describe their concerns, confirms their understanding, and provides structured suggestions. Emotional expressions are minimal, and responses are often informational.

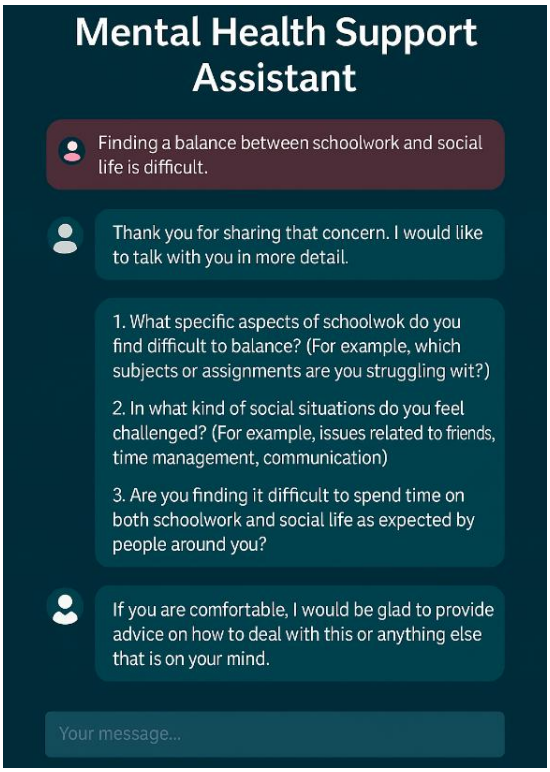


Fig. 1. User Interface example of the mental health support chatbot used in the preliminary study (A1 Condition).

2) A2 (Enhanced): This variant integrates feedback from psychiatrists and results from a preliminary user study. Key improvements include:

- Use of emotionally expressive language. (e.g., "That must have been very difficult for you.")
- Friendly, conversational tone with casual phrasing.
- Occasional light self-disclosure to build trust. (e.g., "I've heard similar stories from other students.")
- Dynamic phrasing to avoid repetition and pattern rigidity.
- An added step to check if the user wants more or fewer suggestions.

To inform our design of the enhanced prompt (A2), we first conducted a pilot study with 10 university students interacting with the baseline version (A1). Their feedback—low empathy ratings, overly factual replies, and rigid dialogue patterns—was complemented by a psychiatrist's expert recommendations on emotionally expressive language, conversational variation, and supportive follow-ups. Drawing on these insights, we refined the prompt to encourage deep support and light self-disclosure while avoiding mechanical phrasing. Table I presents a comparison of A1 and A2, highlighting enhancements in emotional integration, supportive questioning, balanced confirmations, empathic self-disclosure, advice-need checks, and more natural prose style.

TABLE I. IMPROVEMENTS FROM A1 TO A2

Improvement No.	Description
1. Fact & Emotion Mix	Add emotional elements to factual replies to boost understanding and empathy.
2. Supportive Questions	Pair questions with supportive phrasing to lighten user burden.
3. Optimized Questioning & Confirmation	Balance question and confirmation frequency for more natural dialogue.
4. Empathic Remarks & Shared Experience	Use brief self-disclosures to convey empathy and make users feel heard.
5. Checking Need for Detailed Advice	Prompt users before offering detailed advice to avoid overwhelm.
6. Reduced Bulleted Lists	Replace bullet points with natural prose to enhance warmth.

C. Implementation Details

The system uses the GPT-4 model with a temperature setting of 0.7 to balance creativity and reliability. Prompt text is stored as a system message and carried across turns via the messages array. The conversation interface maintains memory via st.session_state, enabling multi-turn interaction. Each response is generated in real time using current and prior messages, and user input is constrained to ensure ethical and safe interactions.

Fig. 2 shows the response flow used during conversation, where user input is appended to the message history, passed to the model along with the prompt, and a context-aware reply is generated.

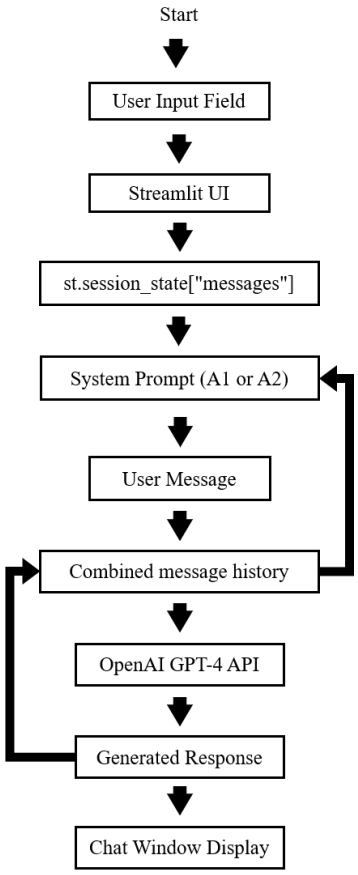


Fig. 2. Message flow of the conversational agent.

D. Ethical and Safety Considerations

The A2 prompt includes structured fallbacks in case of high-risk language (e.g., suicidal ideation), redirecting users with phrases such as, "Your safety is the most important. Please consider talking to a professional." In addition, suggestions are limited in number, and overly directive language is avoided to reduce psychological burden. These safeguards are informed by concerns raised in [7, 16, 17] about the ethical use of AI in mental health contexts.

E. Design Limitations and Future Enhancements

The system does not currently adapt its tone based on individual user preferences. Furthermore, emotional state tracking is not explicitly modeled, and the chatbot's personality is static. In future versions, we plan to introduce user-selectable tone styles and affect-sensitive dialogue modeling [8, 19].

IV. USER STUDY

A. Objectives

The purpose of this user study was to evaluate how variations in prompt design (A1 vs. A2) affect users' perceptions of empathy, effectiveness, and conversational naturalness in a mental health chatbot. Specifically, we sought to determine whether the enhanced prompt in A2, with its emotionally rich language and conversational variability, would improve the user experience in comparison to the baseline A1.

B. Participants and Conditions

A total of 18 university students (9 female, 9 male, and aged 19–25) were recruited for the study. Participants were randomly assigned to interact with three different agents: a human (control), A1 (baseline chatbot), and A2 (enhanced chatbot), under a within-subjects design. Each participant engaged in a single conversation with each condition, and the order of conditions was counterbalanced using a Latin square to control for order effects.

C. Experimental Procedure

Participants were instructed to imagine they were seeking advice for a personal concern related to university life (e.g., stress, relationships, and academic pressure). They engaged in a text-based chat lasting approximately 5–10 minutes per condition.

- For the human condition, a trained confederate acted as the conversation partner.
- For A1 and A2, participants chatted through LINE messenger, with the experimenter forwarding messages to the chatbot system and relaying the responses. Participants were blinded to whether the conversation partner was human or AI. LINE is a popular messaging app in Japan, selected for its familiarity and real-time chat functionality.

The overall experimental procedure and interaction flow are illustrated in Fig. 3. This includes the order of interactions, the chat method, and post-conversation survey collection.

Following each chat session, participants completed a short questionnaire to assess their impressions.

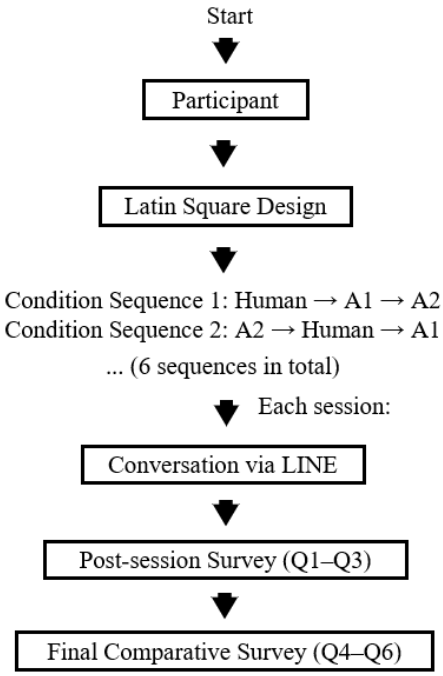


Fig. 3. Overview of the experimental procedure. Each participant engaged in three conditions (Human, A1, A2) using a within-subjects latin square design.

D. Measures

1) Quantitative data and analysis: The following subjective measures were used, all on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree):

- Q1. Perceived understanding: "I felt that the conversation partner understood my concern."
- Q2. Perceived helpfulness: "The conversation helped me to think about or cope with my issue."
- Q3. Perceived empathy: "I felt emotionally supported during the conversation."

Participants also completed a separate comparative questionnaire asking about:

- Q4. Perceived trustworthiness: "Which conversation partner felt most trustworthy?"
- Q5. Perceived emotional attunement: "Which one felt most emotionally attuned?"
- Q6. Perceived empathy: "Which one felt most empathetic toward your concern?"

While both Q3 and Q6 assess perceived empathy, Q3 measures the subjective feeling of emotional support experienced during each individual conversation, using a 5-point Likert scale. In contrast, Q6 asks participants to comparatively judge which agent felt the most empathetic overall across all conditions. Thus, Q3 captures within-condition depth, whereas Q6 reveals between-condition preference.

2) Qualitative data and analysis: We analyzed participants' open-ended feedback and complete chat transcripts by coding each utterance into six predefined categories (Table II) and achieved high inter-rater reliability (Cohen's $\kappa = 0.92$) before reconciling any discrepancies. Category frequencies were compared across conditions using Friedman tests, and a grounded-theory thematic analysis of free-text comments revealed user perceptions such as "mechanical" versus "friendly" tone and varying levels of perceived empathy and support. This approach ensured both statistical rigor and rich insight into how A2's modifications enhanced the chatbot's human-like qualities.

TABLE II. DEFINITION OF THE SIX CONVERSATIONAL CATEGORIES USED FOR QUALITATIVE CODING

Category	Description
Factual Information	General knowledge or advice not tailored to the user.
Light Support	Empathy or encouragement without personalized advice.
Deep Support	Empathy or encouragement with situation-specific advice.
Self-Disclosure	Statements revealing personal experiences or thoughts.
Question	Prompts asking the user to elaborate or clarify their feelings or issues.
Confirmation	Paraphrasing or repeating user input for understanding; no advice given.

E. Data Collection and Ethical Considerations

All participants gave informed consent prior to participation. The study was reviewed and approved by the Ethics Committee of Kyoto Institute of Technology.

No sensitive or personally identifiable information was stored. All dialogues were anonymized and stored securely for analysis.

To minimize potential psychological stress, participants were informed that the conversations were simulated and not therapeutic in nature. Participants were also debriefed after the study and given resources for professional mental health support if needed.

V. Results

A. Quantitative Results

We first analyzed participants' subjective ratings for each of the three conditions (Human, A1, A2) on three evaluation criteria: perceived understanding, perceived helpfulness, and perceived empathy. Fig. 4 shows the mean scores and standard deviations across conditions.

As shown in Fig. 4, Human received the highest scores across all measures, followed by A2, and then A1. Notably, the empathy scores for A2 showed a marked improvement over A1, indicating that the enhanced prompt design had a positive impact.

To assess statistical significance, we conducted a repeated measures ANOVA for each criterion. The results showed significant effects of condition on:

- perceived understanding: $F(2,34) = 5.12$, $p < 0.01$
- perceived helpfulness: $F(2,34) = 4.65$, $p < 0.05$
- perceived empathy: $F(2,34) = 9.27$, $p < 0.001$

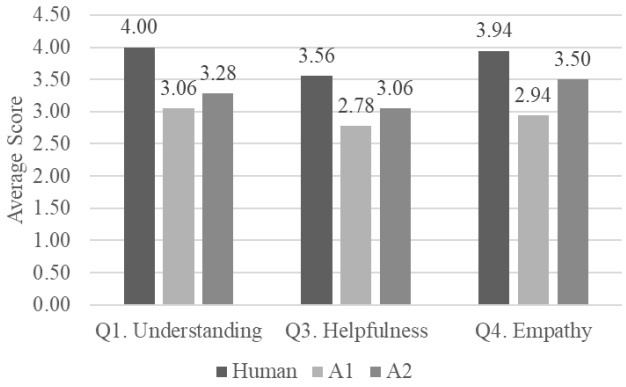


Fig. 4. Average participant ratings for perceived understanding, helpfulness, and empathy across the three conditions (Human, A1, A2).

Post hoc comparisons using the Bonferroni method indicated that A2 scored significantly higher than A1 in perceived empathy ($p < 0.05$), while both were significantly lower than the human condition ($p < 0.01$).

B. Comparative Preference Results

Participants were also asked to select which condition felt most trustworthy, emotionally attuned, and human-like. Fig. 5 summarizes the distribution of preference responses across the three agents.

- Trustworthiness: 11 participants selected human, 5 selected A2, and 2 selected A1
- Emotional attunement: 14 selected human, 4 selected A2, 0 selected A1
- Empathy: 13 selected human, 5 selected A2, 0 selected A1

These results suggest that while A2 was preferred over A1 in all aspects, the human agent was still perceived as superior in trust and empathy.

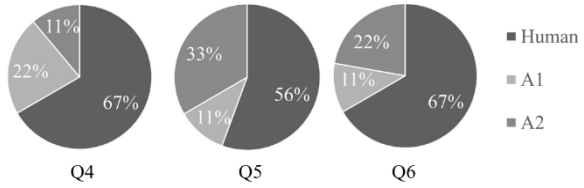


Fig. 5. Participants' comparative preferences across three conditions for trustworthiness, emotional attunement, and perceived empathy.

Pairwise binomial tests were conducted to compare participant preferences among the human condition (Human), the baseline AI (A1), and the improved AI (A2) for trustworthiness (Q4), emotional attunement (Q5), and empathy (Q6). Human was consistently rated highest across all dimensions. A1 showed statistically significant inferiority to Human on all three items ($p < 0.05$). In contrast, A2 demonstrated improvement upon the baseline; while significant differences remained between A2 and Human in trustworthiness and empathy, no significant difference was observed in emotional attunement (Q5), indicating that prompt enhancement in A2 led to partial gains in perceived emotional responsiveness.

C. Qualitative Results

To explore interactional differences, we categorized chatbot and human utterances using a predefined six-category coding scheme: Factual Information, Light Support, Deep Support, Self-Disclosure, Question, and Confirmation.

Fig. 6 shows the proportions of utterance categories for each condition. A2 produced more deep support and self-disclosure utterances than A1, aligning with its design. A1, in contrast, tended to overuse factual responses and confirmations, which may have contributed to its lower empathy scores.

In addition, open-ended feedback revealed that participants often described A1 as "robotic" or "too structured," while A2 was described as "friendlier but still a bit unnatural." Several participants noted that A2 sometimes sounded "playful" or "too casual," which may have impacted their trust in some cases.

The Friedman test revealed significant overall differences in the use of factual information, deep support, and self-disclosure across conditions. Specifically, in the Factual Information category, Human responses were significantly lower than both A1 ($p < 0.01$) and A2 ($p < 0.01$), with no difference between A1 and A2. In the Deep Support category, Human exceeded A1 ($p < 0.01$) and A2 ($p < 0.01$), and A2 also exceeded A1 ($p < 0.05$),

indicating that the improved AI increased deep support compared to its baseline. Self-Disclosure also differed overall ($p < 0.01$), with Human exceeding both A1 ($p < 0.05$) and A2 ($p < 0.05$), while A1 and A2 did not differ significantly. Although the Friedman test for question usage was not significant ($p > 0.05$), a marginal pairwise difference showed Human surpassing A2 ($p < 0.05$). Finally, there were no significant differences among Human, A1, and A2 in Light Support or Confirmation categories.

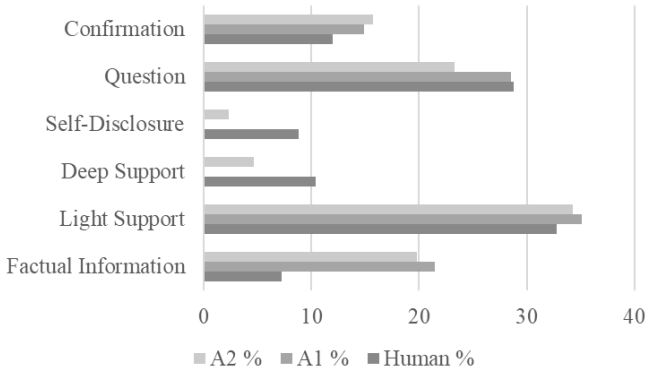


Fig. 6. Frequency of utterance types (factual information, support, self-disclosure, etc.) across the three conditions.

D. Correlation Analysis

To explore the relationships between dialogue behaviors and subjective user ratings, we calculated Spearman's rank correlations (ρ) between each behavior's frequency and three Likert-scale measures (perceived understanding, helpfulness, and empathy), adjusting all p-values via Bonferroni correction. In the Human condition, factual responses were negatively correlated with both perceived understanding ($\rho = -0.555$, $p = 0.017$) and empathy ($\rho = -0.683$, $p = 0.002$), while self-disclosure showed a positive correlation with helpfulness ($\rho = 0.471$, $p = 0.049$) and question-asking was positively associated with understanding ($\rho = 0.637$, $p = 0.004$). In the A1 condition, higher question frequency corresponded to lower understanding ($\rho = -0.486$, $p = 0.041$). In A2, factual responses remained negatively correlated with empathy ($\rho = -0.649$, $p = 0.004$). These results confirm that reducing purely factual replies and encouraging self-disclosure and targeted questioning can bolster users' comprehension, perceived support, and empathy.

VI. Discussion

A. Summary of Findings

The results of the user study demonstrated that the enhanced chatbot variant (A2) outperformed the baseline (A1) in terms of perceived empathy, helpfulness, and understanding. While neither AI-based agent surpassed the human control, A2 significantly narrowed the gap, particularly in empathy-related measures.

These findings indicate that relatively simple modifications to prompt design—such as adding emotionally expressive phrases, conversational variability, and light self-disclosure—can meaningfully improve the user experience in AI-mediated mental health support.

B. Design Implications

The positive impact of A2 provides several implications for the design of empathetic conversational agents:

- Emotionally expressive language matters: The increased empathy ratings suggest that incorporating warm, validating phrases makes users feel heard and supported. This effect is commonly documented in human counseling interactions [18, 20].
- Conversational variability enhances naturalness: Avoiding rigid or templated responses helped reduce the "robotic" impression users associated with A1, which supports prior findings in LLM dialogue systems [19].
- Deep support and self-disclosure are key cues: The correlation analysis confirmed that deeper, context-aware support strategies are associated with higher emotional satisfaction in users, while factual overload (as seen with A1) detracts from perceived empathy.

These results align with those in earlier research highlighting the importance of emotional attunement and personalization in AI-based supportive dialogue [7, 8, 15].

C. Trust vs. Friendliness Trade-off

While A2 was rated higher than A1 in empathy, some users perceived its tone as overly casual or playful. This introduces a design trade-off between friendliness and trustworthiness—a chatbot that sounds too familiar may be emotionally engaging but risk being taken less seriously.

Designers must consider tone adaptation as a dynamic component that possibly allows users to select or tune the emotional intensity of the chatbot according to their comfort level.

D. Remaining Challenges

Despite improvements in A2, A1 and A2 still falls short in following areas compared to human interactions. These include:

- Lack of deep contextual understanding: While A2 could simulate empathy, it still lacked true comprehension of nuanced emotional cues.
- Static personality: Both A1 and A2 followed scripted prompt logic without learning from prior interactions.
- No adaptation to individual user preferences: All participants received identical tone and style, absent of reflection on personal communication preferences.

These limitations are consistent with known boundaries of current LLM-based systems and suggest that hybrid models incorporating affective state tracking or memory may be needed [19].

E. Ethical Considerations

This study also underscores the importance of ethical safeguards. Although A2 included fallback responses for high-risk expressions (e.g., suicidal ideation), the system cannot

replace human professionals in crisis situations. Meanwhile, The Verge reports that in environments lacking robust data protection frameworks, users may place excessive trust in AI chatbots, potentially leading to privacy breaches or a false sense of security [16], and Live Science has shown that exposing LLMs to traumatic narratives can induce "anxiety-like" outputs, underscoring the risks of emotional overreliance [17].

Designing transparent boundaries—such as clear disclaimers and escalation pathways—remains essential in real-world deployment.

VII. Conclusion and Future Work

A. Conclusion

In this study, we designed and evaluated two versions of a conversational agent—A1 and A2—aimed at providing mental health support for university students. While A1 followed a structured but emotionally neutral prompt, A2 was enhanced with emotionally expressive phrasing, light self-disclosure, and conversational variability.

Through a mixed-method user study involving 18 participants, we found that A2 significantly had improved user perceptions of empathy and engagement compared to A1. Quantitative results showed meaningful gains across all subjective measures, particularly in empathy. Qualitative analysis and correlation data further confirmed that the presence of deep support and reduced factuality in A2's responses contributed to its improved reception.

These findings highlight the importance of prompt-level control in shaping large language model (LLM) behavior, especially in emotionally sensitive domains such as mental health. Even small design changes at the prompt level can have a significant impact on how AI is perceived by users.

However, persistent gaps between AI and human interaction remain. Participants still preferred human agents in terms of trust and emotional attunement, and the current system lacks personalization, emotional awareness, and adaptability.

B. Future Work

Based on these findings, we propose several directions for future research:

1) Personalized tone control: Enabling users to select or dynamically adjust the chatbot's tone according to preference or emotional state.

2) Emotion-aware dialogue modeling: Incorporating affective computing techniques to detect and respond to user emotions in real time.

3) Longitudinal evaluation: Studying the long-term effects of AI-mediated support across repeated interactions.

4) Safety and escalation mechanisms: Enhancing detection of high-risk language and implementing automatic escalation protocols.

5) Cross-cultural validation: Investigating how empathy is perceived and evaluated across cultural or linguistic groups.

By addressing these challenges, future conversational agents can become not only more effective but also more trustworthy and emotionally responsive companions.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP 24K02977.

References

[1] Japan Private University Federation, *Private University Student Life White Paper 2022*, pp.68–70, 2022.

[2] D. Eisenberg, E. Golberstein, and J. B. Hunt, "Mental health and academic success in college," *The B.E. Journal of Economic Analysis & Policy*, vol. 9, no. 1, pp. 1–35, 2009.

[3] J. M. Drazen, "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," *N. Engl. J. Med.*, vol. 388, pp. 1233–1239, 2023.

[4] K. Otsu, "Dialogue-based interventions for personalized healthcare using solution-focused approaches," *J. Human Interface Soc.*, vol. 24, no. 4, pp. 285–299, 2022. https://doi.org/10.11184/his.24.4_285

[5] N. Kuwahara, Y. Tanaka, and P. Siriaraya, "Exploring Photo-Based Dialogue Between Elderly Individuals and Conversational Agents Using Image Generation," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, 2024.

[6] A. Naik, J. Thomas, T. Sree, and H. Reddy, "Artificial Empathy: AI based Mental Health," *arXiv preprint* arXiv:2506.00081, 2025. [Online]. Available: https://arxiv.org/abs/2506.00081

[7] R. Raczka, "AI Therapists Can't Replace the Human Touch," *The Guardian*, May 2025. [Online]. Available: https://www.theguardian.com/society/2025/may/11/ai-therapists-cant-replace-the-human-touch

[8] Roshanaei M., Rezapour R., and Seif El-Nasr M., "Talk, Listen, Connect: Navigating Empathy in Human-AI Interactions," arXiv preprint arXiv:2409.15550, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.15550/

[9] J. M. Liu et al., "ChatCounselor: A Large Language Model for Mental Health Support," *arXiv preprint arXiv:2309.15461*, 2023.

[10] T. Lai et al., "Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models," *arXiv preprint arXiv:2307.11991*, 2023.

[11] X. Yao et al., "Development and Evaluation of Three Chatbots for Postpartum Mood and Anxiety Disorders," *arXiv preprint arXiv:2308.07407*, 2023.

[12] N. Jacobson, "Study Finds AI Chatbot Can Improve Mental Health," *Psychology Today*, Mar. 2025. [Online]. Available: https://www.psychologytoday.com/us/blog/the-future-brain/202503/study-finds-ai-chatbot-can-improve-mental-health

[13] S. Ahmed et al., "Effectiveness of AI Chatbots in Improving Students' General Wellbeing: A Clinical Pilot Trial," *MedPath*, May 2025.

[14] P. N. Waaler et al., "Prompt Engineering an Informational Chatbot for Education on Mental Health Using a Multiagent Approach: Algorithm Development and Validation," *JMIR AI*, vol. 4, 2025. [Online]. Available: https://ai.jmir.org/2025/1/e69820

[15] P. Sahoo, J. Green, and K. Price, "Prompt Engineering for Digital Mental Health: A Short Review," *Front. Digit. Health*, vol. 3, pp. 1–10, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1410947/full

[16] J. Vincent, "AI Therapy is a Surveillance Machine in a Police State," *The Verge*, May 2025. [Online]. Available: https://www.theverge.com/policy/665685/ai-therapy-meta-chatbot-surveillance-risks-trump

[17] N. Ferguson, "Study: Talking to AI about War and Violence Makes It Anxious," *Live Science*, Mar. 2025. [Online]. Available: https://www.livescience.com/technology/artificial-intelligence/traumatizing-ai-models-by-talking-about-war-or-violence-makes-them-more-anxious

[18] A. S. Raamkumar and Y. Yang, "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities," IEEE Trans. Affect. Comput., vol. 14, no. 4, pp. 2722–2739, 2023.

[19] X. Zheng, Z. Li, X. Gui, and Y. Luo, "Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots," *Proc. 2025 CHI Conf. on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, Apr. 2025,
doi: 10.1145/3706598.3713453.

[20] B. Liu and S. S. Sundar, "Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot," *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 10, pp. 625–636, Oct. 2018.
https://doi.org/10.1089/cyber.2018.0110

[21] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," [Online]. Available: https://streamlit.io/

[22] OpenAI, "GPT-4 Technical Report," Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4

Cross-Context Evaluation of an Indoor–Outdoor AR Navigation System in a University Campus Environment

Toma Marian-Vladut[1], Pascu Paul[2], Turcu Corneliu Octavian[3]
Computers, Electronics and Automation Department, Stefan cel Mare University of Suceava, Suceava, Romania[1, 3]
Department of Economics-Economic Informatics and Business Administration,
Stefan cel Mare University of Suceava, Suceava, Romania[2]

***Abstract*—This paper presents a comparative evaluation of a mobile augmented reality (AR) navigation application designed for both indoor and outdoor university environments. Building on a previously validated system for indoor guidance, the current study extends the deployment to outdoor campus spaces without relying on GPS or additional infrastructure. Using visual positioning and spatial anchors, the same application provides real-time AR cues and audio instructions to support wayfinding across different spatial contexts. The principal aim of this study is to determine whether a unified AR navigation system can deliver a consistent, infrastructure-free user experience in both indoor and outdoor university environments. A within-subjects study was conducted with 256 university students who completed both indoor and outdoor navigation tasks. Usability and user acceptance were assessed using the System Usability Scale (SUS) and constructs from the Technology Acceptance Model (TAM). Results revealed consistent user experience across both contexts, with no statistically significant differences in perceived intuitiveness, usefulness, engagement, behavioral intention to reuse, or localization accuracy. A significant difference was found only in perceived AR content loading speed, which was rated slightly higher indoors. These findings demonstrate the feasibility of a unified AR navigation system for academic campuses and provide practical insights into its scalability and user-centered design.**

***Keywords*—Augmented reality; hybrid navigation; spatial computing; usability testing; indoor-outdoor contexts; ARCore; higher education**

I. Introduction

Indoor navigation remains a significant challenge in large, unfamiliar environments such as hospitals, airports, and university campuses. Traditional wayfinding methods—including signage, printed maps, and kiosks—often fall short in addressing the spatial complexity and dynamic layouts of such settings [1]. In academic institutions, students and visitors frequently struggle to locate classrooms, offices, or facilities within multi-story buildings or across expansive outdoor areas. Augmented Reality (AR) has emerged as a promising solution by superimposing navigational cues—such as arrows, labels, and landmarks—onto the real-world view through mobile devices, enhancing spatial cognition and real-time orientation [2]. Navigating unfamiliar academic campuses, particularly during orientation periods, open days, or student admissions, is a recurring challenge that affects thousands of users each academic year. AR-based navigation tools offer an accessible, real-time solution that does not require technical training, printed maps, or assistance from staff—making them ideal for practical deployment in higher education settings.

While AR navigation systems for outdoor use have benefited from reliable GPS signals and integration with digital mapping services [3], indoor environments present unique technical constraints. GPS signals tend to degrade or become unavailable indoors, requiring alternative technologies such as Simultaneous Localization and Mapping (SLAM), Visual Positioning Systems (VPS), and visual-inertial odometry [4]. These differences in spatial and technical conditions suggest that the user experience may vary considerably between indoor and outdoor AR navigation contexts. However, few studies have examined this comparative dimension directly.

This study addresses the following research question:

Can a single mobile AR navigation system provide a seamless and equally usable experience across both indoor and outdoor university environments, without relying on GPS or additional infrastructure? Answering this question can inform future designs of scalable, user-friendly AR navigation systems for education and beyond.

Recent advances in mobile AR platforms—such as Apple ARKit and Google ARCore—have enabled increasingly accurate, infrastructure-free AR navigation solutions across domains including education, healthcare, and retail. These frameworks provide core functionality for markerless tracking, real-time camera pose estimation, and rendering of 3D navigation overlays [5], [6]. Studies report promising results for both usability and technical feasibility. For instance, Yamamoto et al. implemented an AR indoor navigation app using BLE beacons in a university setting, achieving over 70% user satisfaction and stable guidance performance [6].

Despite growing interest, university campuses remain relatively underexplored in AR navigation research—particularly in terms of integrated indoor–outdoor deployment. Most existing studies either evaluate AR in confined indoor settings or focus solely on outdoor AR guidance, often under pilot conditions or with small sample sizes [2], [7]. Moreover, many systems rely on infrastructure such as BLE beacons or GPS, limiting scalability. Our work addresses these limitations by proposing and evaluating a unified, infrastructure-free

system that supports consistent AR-based navigation across spatial contexts. This lack of cross-context evaluation limits our understanding of how environmental factors influence usability, responsiveness, and user acceptance of AR-based wayfinding solutions.

The present study addresses this gap by extending a previously developed mobile AR navigation system [8]—initially tested for indoor guidance—to a campus-wide outdoor scenario, followed by a comparative evaluation. The application is implemented using Unity 3D and the ARway SDK and supports both Android and iOS platforms via ARCore and ARKit. Importantly, the system operates without dedicated infrastructure (e.g., BLE beacons or markers), using camera-based localization and cloud-based spatial anchors instead. Users initialize navigation by scanning a QR code at a known location—either indoor or outdoor—and follow AR overlays and audio cues that provide dynamic, step-by-step guidance.

This paper offers two main contributions: (1) it demonstrates the scalability and adaptability of a mobile AR navigation system across indoor and outdoor university environments, and (2) it presents a comparative analysis of user experience across these contexts, based on a large-scale user study. The evaluation framework includes metrics such as perceived usability, intuitiveness, system responsiveness, engagement, and behavioral intention to reuse the app. These dimensions are quantified using the System Usability Scale (SUS) [9] and constructs from the Technology Acceptance Model (TAM) [10], allowing for hypothesis-driven analysis of how spatial context affects AR navigation acceptance. The findings extend prior indoor-only results and contribute practical insights for designing seamless AR wayfinding systems suitable for real-world educational environments.

The remainder of this paper is organized as follows: Section II reviews the relevant literature on AR-based indoor and outdoor navigation systems; Section III details the system design and methodology; Section IV presents the results of the comparative user study; Section V discusses the findings in relation to prior work; and Section VI concludes with insights and future research directions.

II. Related Work

A. Outdoor AR Navigation Systems

Augmented Reality (AR) has gained significant traction as a tool for spatial navigation in both indoor and outdoor environments. AR enhances spatial cognition by overlaying digital cues—directional arrows, labels, or visual landmarks—onto the user's real-world view, improving situational awareness and wayfinding performance [1]. Outdoor AR navigation systems are more established due to their reliance on GPS, digital maps, and compass-based orientation [11]. For example, Huang et al. [12] developed an AR pedestrian navigation prototype integrating GPS and computer vision to display route information in real-time street views. Commercial map applications such as Google Maps have also adopted AR for walking navigation, indicating a rise in public familiarity with AR-based outdoor guidance. However, these systems rely heavily on GPS and are limited in environments with signal loss or visual occlusions.

B. Indoor AR Navigation: Infrastructure Challenges

Indoor AR navigation, however, presents unique challenges stemming from unreliable or unavailable GPS signals. To address this, various positioning strategies have emerged: BLE (Bluetooth Low Energy) beacons [6], Wi-Fi fingerprinting [4], visual marker systems, and vision-based SLAM (Simultaneous Localization and Mapping) techniques [13]. These alternatives often require specialized infrastructure or calibration. For instance, Nguyen and Park [14] deployed a BLE-based AR guidance system in a shopping mall, achieving high positional accuracy and user satisfaction. Similarly, Gupta et al. [4] used a hybrid of Wi-Fi fingerprinting and SLAM in a hospital environment, attaining room-level localization and effective navigation performance. While effective, these solutions require infrastructure deployment (e.g., BLE beacons), limiting scalability and generalizability across different buildings or institutions.

C. Advances in Markerless AR Navigation

Recent mobile AR development frameworks—such as Apple's ARKit and Google's ARCore—have enabled more accurate, markerless tracking capabilities via SLAM and visual-inertial odometry [15],[16]. Rossi et al. [5] used ARKit to build an indoor navigation prototype that resulted in a 35% reduction in wayfinding time. Chen et al. [17] employed ARCore for multi-floor campus navigation, leveraging persistent cloud anchors and reporting an 81% ease-of-use rating. These tools have become foundational in AR navigation research by enabling infrastructure-free guidance with stable content anchoring and route overlays. Yet, most deployments remain confined to either indoor or outdoor use, lacking evaluation across both contexts with the same system.

Localization precision remains a key factor in AR navigation. Outdoors, GPS typically provides 5–10 meter accuracy, but AR systems often require finer granularity to align virtual objects meaningfully with physical spaces. Visual Positioning Systems (VPS), such as Google's AR VPS, supplement GPS by comparing camera views to geo-referenced imagery [12]. Indoors, BLE beacons can offer ~3–5 meter accuracy, while UWB (Ultra-Wideband) radio systems can achieve sub-meter precision. Yamamoto et al. [6] used over 100 BLE beacons in a university building to power their AR app, noting limitations such as signal decay and interference in crowded spaces. Alternatively, QR codes or fiducial markers can serve as anchor points, offering high-precision initialization and periodic re-localization to mitigate SLAM drift—a strategy employed in the current system.

D. Hybrid and Cross-Context Systems

To ensure seamless indoor–outdoor transitions, hybrid approaches have emerged. Fernandez et al. [7] proposed a GPS emulation mechanism using cloud-based AR maps to support campus-scale navigation. Qi et al. [18] developed a hybrid ARCore-based solution using visual-inertial odometry and cloud anchors, demonstrating improved spatial tracking across both indoor and outdoor contexts. These systems still face

challenges related to lighting variations, visual occlusion, and UI continuity during environment transitions, but they represent meaningful steps toward unified AR navigation experiences, as shown in [19] and [20]. Nevertheless, these studies are either experimental or simulate GPS using cloud anchors; they stop short of large-scale user evaluations in naturalistic academic settings.

E. User Experience and Acceptance

Accurate localization is only one facet; usability and user experience are equally critical. Many studies use the System Usability Scale (SUS) and the Technology Acceptance Model (TAM) to quantify user perceptions. SUS is a 10-item instrument that yields a usability score between 0 and 100 [9]. Chen et al. [17] reported a SUS score of 74 in a campus navigation scenario, and Sato et al. [21] achieved a score of 76 in a museum-based AR guide. These results suggest generally favorable usability perceptions in both educational and cultural settings.

TAM provides a theoretical lens to understand user adoption by evaluating perceived usefulness, ease of use, and behavioral intention [10]. In AR navigation contexts, perceived ease of use refers to how effortlessly users can interpret and follow AR cues, while perceived usefulness reflects whether the system enhances their navigation performance. Kwok et al. [22] extended TAM to include enjoyment and found that higher enjoyment correlated with greater acceptance of AR navigation in a university environment. The study also noted the role of social visibility—users may hesitate to use AR interfaces in public due to perceived awkwardness, a relevant consideration for campus-wide deployment. Our current study extends this by collecting SUS and TAM responses in both indoor and outdoor conditions using the same AR system, enabling direct comparison.

F. Technical Factors Affecting Usability

Technical factors also influence adoption. Users expect AR systems to be responsive, visually stable, and intuitive. Lagging overlays, jittery directional arrows, or poor contrast in bright sunlight can degrade the experience [17], [21]. Several studies report that system responsiveness is a strong predictor of SUS scores. In our prior study, responsiveness and tracking stability were among the most significant factors contributing to perceived usability. Zhao et al. [23] demonstrated that using an AR headset equipped with high-performance sensors yielded an 88% navigation success rate but caused user fatigue—underscoring the trade-offs between accuracy and usability. Our system, unlike high-performance AR headsets, achieves stable performance using only consumer smartphones, balancing usability and accessibility.

G. AR Navigation in Educational Settings

Beyond navigation efficiency, AR has been shown to reduce cognitive load and enhance users' ability to form cognitive maps. Qiu et al. [1] reported that AR improved user confidence and spatial memory, even when it did not significantly reduce travel time compared to traditional 2D maps. Ahn et al. [24] found that AR guidance in a hospital led to faster learning of spatial layouts and a high initial orientation success rate of 86%. These qualitative benefits align with the TAM concept of perceived usefulness and highlight AR's role not only in performance but in experiential engagement.

Research in educational settings remains relatively sparse but is growing. AR navigation aligns with smart-campus initiatives, offering new students or visitors accessible guidance through complex academic environments. Fernandez et al. [7] and Yamamoto et al. [6] deployed their systems in real campus scenarios with participant samples exceeding 50 users. Our prior study, with 256 participants, remains one of the most extensive in this domain. However, most existing works evaluate either indoor or outdoor AR usage in isolation. Little is known about how users perceive and interact with AR systems that transition across spatial contexts [25].

In summary, the literature shows steady progress in AR navigation across diverse settings. However, notable gaps remain: (1) few comparative evaluations of indoor versus outdoor AR navigation using the same system; (2) limited research in education-specific deployments at scale; and (3) scarce longitudinal studies on sustained adoption. This paper addresses the first of these by evaluating a hybrid AR navigation system deployed in both indoor and outdoor areas of a university campus. Through an empirical comparison grounded in SUS and TAM frameworks, the study aims to generate insights into how spatial context influences usability and user acceptance in real-world academic environments. This paper fills a key gap by offering one of the first comparative, infrastructure-free, cross-context AR navigation evaluations in a university campus, involving over 250 participants.

III. METHODOLOGY

This study builds on a previously published evaluation of an AR-based indoor navigation application [8], extending the investigation to outdoor university campus spaces. To preserve continuity while enabling comparative insight, the methodology replicates the core structure of the initial research while adjusting the deployment context and data collection accordingly.

A. Research Design and Context

We adopted a user-centered, exploratory methodology integrating system deployment with empirical usability evaluation. The same AR application, previously tested indoors, was adapted and redeployed in selected outdoor areas of the university campus, using the same visual positioning and spatial anchoring approach (SLAM + VPS). Unlike traditional GPS-based outdoor navigation, our system does not rely on GNSS signals, allowing consistent performance across both spatial environments.

A within-subjects experimental design was employed, with each participant completing two navigation tasks—one indoors and one outdoors. The study retained the System Usability Scale (SUS) and the Technology Acceptance Model (TAM) as its core evaluative frameworks, consistent with the prior research.

B. Description of the AR Navigation Application

To extend the functionality of the indoor AR navigation system into outdoor campus settings, the same mobile application was adapted without relying on GPS or external

sensors. Instead, it continues to use visual-inertial localization based on cloud-anchored spatial mapping, implemented via the ARway SDK and Unity 3D. The navigation experience was redesigned to suit open environments while maintaining consistency in interface and interaction style across both indoor and outdoor contexts.

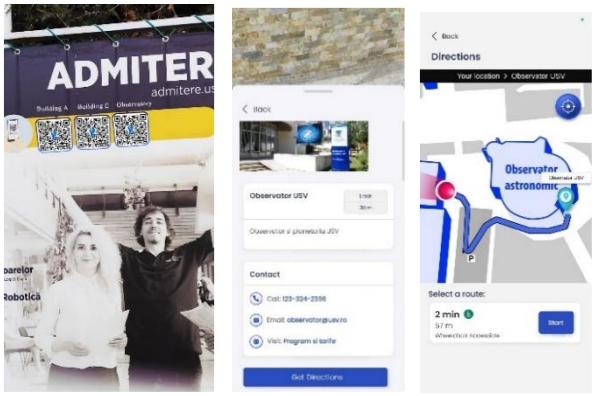
Navigation begins when the user scans a designated QR code positioned on outdoor campus signage, such as the official admissions board located at the entrance to campus facilities [Fig. 1(a)]. This QR code acts as an initialization anchor, aligning the digital scene with the real-world coordinate system to enable precise overlay of AR guidance elements. Once initialized, the user is presented with a location selection interface, from which they can choose a point of interest—such as the university's astronomical observatory.

Upon selecting a destination, the application displays a detailed preview including a representative photo, contact information, and a "Get Directions" button [Fig. 1(b)]. Tapping this option generates a contextual route preview, where users can see the estimated walking time and total distance to the target [Fig. 1(c)]. Notably, the app also flags whether the suggested route is wheelchair accessible, ensuring inclusivity in the wayfinding experience.

Once navigation is initiated, the application renders the route as a sequence of 3D arrows superimposed on the live camera feed. These directional cues update dynamically as the user moves through the campus, ensuring accurate step-by-step guidance [Fig. 2(a)]. In parallel, textual instructions and optional audio prompts provide redundancy, enhancing usability in varying light and noise conditions [Fig. 2(b)]. The interface design remains minimalistic and accessible, with large tap areas, clean typography, and consistent iconography.

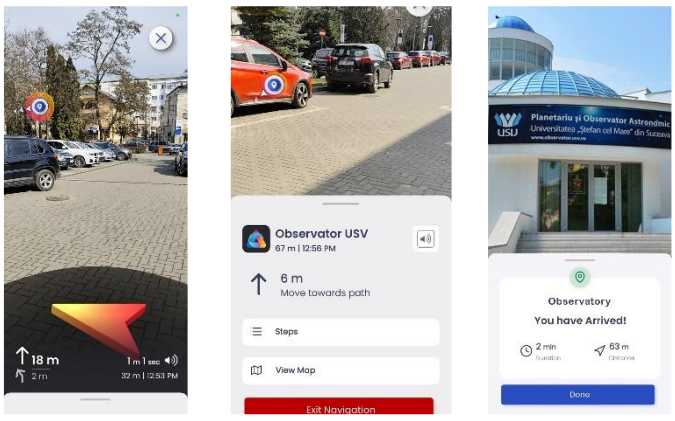
The final phase of the navigation process—approaching the destination—is supported by persistent AR labels and markers positioned in the user's field of view. These indicators signal proximity to the target and offer visual confirmation upon arrival [Fig. 2(c)]. As with the indoor scenario, no GPS or BLE infrastructure is needed; the entire navigation experience is powered by camera-based tracking and spatial anchors, making it scalable and deployable without physical modifications to the environment.

This outdoor adaptation demonstrates that the same infrastructure-free AR system can offer seamless wayfinding across spatial contexts, guiding users from common entry points on campus to specific buildings or functional areas with minimal friction. The user experience, interaction logic, and technological backbone remain consistent with the previously evaluated indoor deployment, ensuring methodological continuity for comparative analysis.



(a) QR code scanning. (b) Location details. (c) Location path.

Fig. 1. User interface and interaction (I)



(a) Visual cues (b) Additional details. (c) Arrival confirmation.

Fig. 2. User interface and interaction (II)

C. Participant Profile and Study Duration

A total of 256 students from Ștefan cel Mare University of Suceava participated in the comparative study. This participant pool mirrors the one from the initial indoor evaluation. Most were enrolled in the first (70.7%) or second (26.2%) year of undergraduate studies, with the remaining (3.1%) from higher years or master's programs. This demographic ensured low prior familiarity with campus routes, particularly outdoor connectors between buildings.

Participants were required to have AR-capable smartphones. Of the total, 153 had prior exposure to AR applications, while 103 had none. This variation supported robust evaluation of both usability and adoption potential.

D. Experimental Setup and Testing Procedure

Each participant completed two navigation tasks using the same AR app: one indoor, one outdoor. The order was counterbalanced. In both scenarios, QR codes marked the navigation start points. Users selected a destination from the on-screen list and followed AR cues to reach the target location. Outdoor destinations included pedestrian pathways between faculty buildings. Navigation took 2–3 minutes per task.

The application provided continuous AR guidance without requiring server connectivity. No external assistance was offered, simulating real-world use. After completing both routes, participants filled out a post-task questionnaire on their own device. Each session lasted ~15 minutes and was conducted individually during daytime under stable weather conditions.

E. Research Hypotheses

The study formulated the following hypotheses for comparative evaluation:

- H1: No significant difference in perceived ease of use between indoor and outdoor conditions.
- H2: Perceived usefulness is higher indoors due to complex spatial layouts.
- H3: System responsiveness predicts SUS scores across both conditions.
- H4: SUS scores do not differ significantly across environments.
- H5: Behavioral intention to reuse is positively associated with perceived usefulness.
- H6: Outdoor environmental factors (e.g., lighting, visual features) may influence responsiveness perception.

The study was reviewed and approved by the Ethics Committee of Ștefan cel Mare University of Suceava. All participants were informed about the nature and purpose of the research, and informed consent was obtained prior to their involvement. Data collection adhered to institutional guidelines on ethical conduct, privacy, and voluntary participation.

The following section presents the results of the comparative evaluation, focusing on key user experience indicators measured across indoor and outdoor navigation scenarios.

IV. Results

A. Hypothesis 1 (H1): There is no Statistically Significant Difference between users' Perception of the Application's Intuitiveness in Indoor Versus Outdoor Navigation Scenarios

The first hypothesis aimed to explore whether users perceive the augmented reality (AR) navigation application as equally intuitive in both indoor and outdoor university environments. According to user-centered design principles and the Technology Acceptance Model (TAM), intuitiveness plays a critical role in influencing user engagement and perceived usability across different contexts.

To test this hypothesis, a paired-samples t-test was conducted, comparing two questionnaire items:

- "How intuitive did you find the indoor navigation mode?"
- "How intuitive did you find the outdoor navigation mode?"

Both variables were measured on a 5-point Likert scale (1 = Not at all intuitive, 5 = Very intuitive). Descriptive statistics are presented in Table I, showing nearly identical means for both contexts: indoor (M = 3.83, SD = 0.97) and outdoor (M = 3.83, SD = 1.00), suggesting a uniform perception of intuitiveness across environments.

TABLE I. Descriptive Statistics for Perceived Intuitiveness in Indoor vs Outdoor Scenarios

Variable	Mean	Std. Deviation	N
Indoor Navigation Intuitiveness	3.83	0.97	256
Outdoor Navigation Intuitiveness	3.83	1.00	256

The results of the t-test revealed no statistically significant difference between the two means: $t(255) = -0.066$, $p = 0.947$ (see Table II). The mean difference was minimal (−0.004), reinforcing the conclusion that users found the application equally intuitive regardless of whether it was used indoors or outdoors.

TABLE II. Paired-Samples T-Test for Intuitiveness Scores in Indoor vs Outdoor Contexts

Variable Pair	Mean Difference	t	df
Indoor vs. Outdoor Intuitiveness Scores	−0.004	−0.066	255

These findings support the hypothesis that intuitiveness is perceived consistently across navigation environments. This consistency may reflect the application's cohesive design and interface behavior across contexts, suggesting that once familiarized with the system, users can transfer their expectations and interactions seamlessly between indoor and outdoor use. Fig. 3 illustrates the comparison between students perceived intuitiveness of the AR application during indoor and outdoor navigation. As shown, the mean scores for both scenarios are virtually identical (M_indoor = 3.83; M_outdoor

= 3.83), indicating a consistent user experience across environments.

A paired-samples t-test confirmed that this difference is not statistically significant ($t(255) = -0.066$, $p = 0.947$). These findings suggest that the application's intuitive design was equally effective in guiding users regardless of the spatial context, supporting its usability in both structured (indoor) and open (outdoor) campus environments.

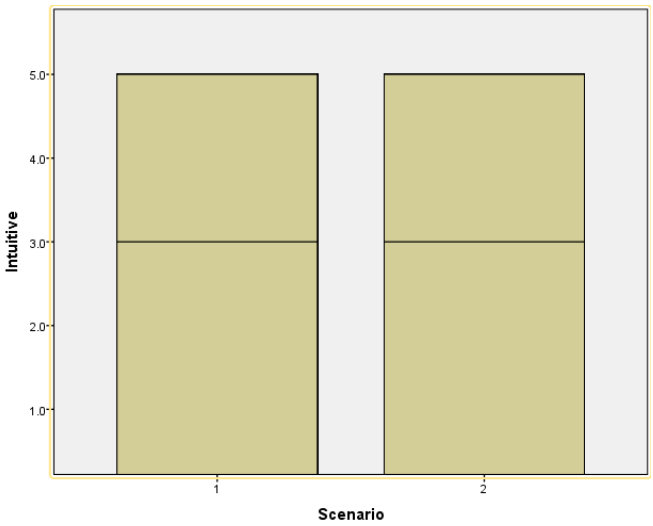


Fig. 3. Mean perceived intuitiveness of the AR application during indoor and outdoor navigation scenarios. Error bars are not displayed due to negligible variation. No significant difference was observed between conditions ($p = 0.947$).

This result is particularly relevant for the design of hybrid AR navigation solutions in academic settings, where continuity of experience across spatial contexts (e.g., classroom buildings, campus pathways) enhances usability and adoption.

B. Hypothesis 2 (H2): The Application is Perceived as More Useful in Indoor Navigation than in Outdoor Navigation

The second hypothesis investigated whether users perceive the augmented reality (AR) navigation application as more useful in indoor environments (e.g., corridors, classrooms) than in outdoor settings (e.g., between campus buildings). This assumption stems from the idea that indoor environments are typically more challenging for spatial orientation due to occlusion, lack of clear landmarks, and complex layouts, which AR tools are specifically designed to mitigate.

To examine this hypothesis, a paired-samples t-test was conducted between the following two items:

- "I believe this application significantly improves my ability to orient myself within campus buildings."
- "I believe this application significantly improves my ability to orient myself in exterior campus spaces."

Both items used a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). Descriptive statistics are presented in Table III, showing slightly higher scores for indoor utility ($M = 4.20$, $SD = 0.91$) compared to outdoor ($M = 4.11$, $SD = 0.94$).

TABLE III. DESCRIPTIVE STATISTICS FOR PERCEIVED UTILITY IN INDOOR VS. OUTDOOR CONTEXTS

Variable	Mean	Std. Deviation	N
Indoor Navigation Utility	4.20	0.91	256
Outdoor Navigation Utility	4.11	0.94	256

A paired-samples t-test revealed that this difference was not statistically significant, $t(255) = 1.704$, $p = 0.090$ (see Table IV). The mean difference of +0.086 indicates a slight preference for indoor use, but this was insufficient to support the hypothesis at the conventional $\alpha = 0.05$ level.

TABLE IV. PAIRED-SAMPLES T-TEST FOR PERCEIVED UTILITY SCORES IN INDOOR VS. OUTDOOR CONTEXTS

Variable Pair	Mean Difference	t	df
Indoor vs. Outdoor Utility Scores	+0.086	1.704	255

While the data suggest a marginal trend favoring indoor use in terms of perceived utility, the lack of statistical significance indicates that users view the application as equally beneficial for both indoor and outdoor navigation. This result may reflect the versatility and adaptability of the application interface and the consistency of user experience across spatial contexts.

The chart presented in Fig. 4 illustrates the distribution of perceived utility scores (Perception_Utility) across two usage scenarios: indoor (1) and outdoor (2). This visualization is directly relevant to Hypothesis 2 (H2), which proposed that users would perceive the application as more useful in indoor navigation contexts (e.g., hallways, classrooms) than in outdoor settings (e.g., between buildings). However, the nearly identical heights of the bars suggest that perceived utility is very similar across both scenarios.

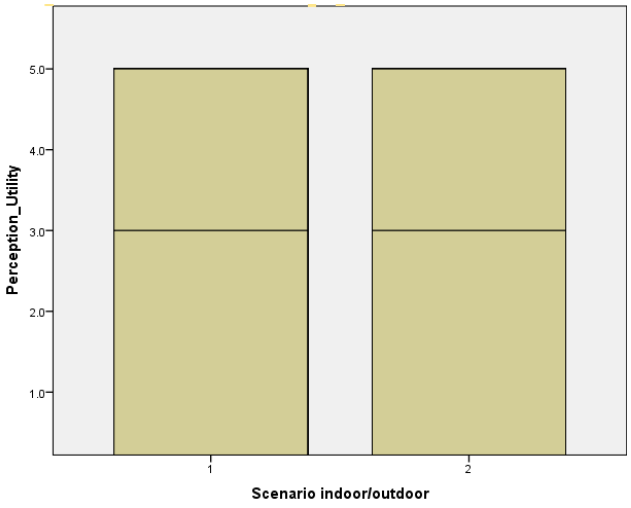


Fig. 4. Mean perceived utility scores across indoor (1) and outdoor (2) navigation scenarios. The near-identical column heights indicate similar user evaluations in both contexts, consistent with the non-significant statistical difference.

This graphical similarity aligns with the results of the paired-samples t-test, which revealed no statistically significant difference between the two conditions ($M_{indoor} = 4.20$, $M_{outdoor} = 4.11$, $p = 0.090$). Although there was a slight

numerical advantage for indoor use, the difference was insufficient to support H2 at the conventional significance level.

These findings indicate that users consider the application to be equally effective for orientation and navigation tasks in both spatial contexts.

These findings provide important insights for developers aiming to scale AR-based navigation tools across campus-wide deployments, as both environments appear equally receptive to technological support in orientation tasks.

C. Hypothesis 3 (H3): The Navigation Experience is Perceived as More Engaging in Indoor Settings than in Outdoor Settings

The third hypothesis examined whether users find the augmented reality (AR) navigation experience more stimulating or engaging when used indoors, as opposed to outdoors. Indoor AR experiences may be enhanced by the presence of walls, corridors, and structured spaces that can provide more immersive contextual cues, whereas outdoor environments may dilute these effects due to broader visual openness and variable lighting.

To assess this, participants responded to two parallel questionnaire items:

- "How engaging or stimulating did you find the indoor navigation experience?"
- "How engaging or stimulating did you find the outdoor navigation experience?"

Responses were measured on a 5-point Likert scale (1 = Not at all, 5 = Very much). Descriptive statistics, displayed in Table V, show nearly identical mean scores: indoor (M = 4.02, SD = 0.92) and outdoor (M = 4.04, SD = 0.95), suggesting an overall balanced experience across contexts.

TABLE V. DESCRIPTIVE STATISTICS FOR PERCEIVED ENGAGEMENT IN INDOOR VS. OUTDOOR SCENARIOS

Variable	Mean	Std. Deviation	N
Indoor Navigation Engagement	4.02	0.92	256
Outdoor Navigation Engagement	4.04	0.95	256

A paired-samples t-test confirmed that this difference is not statistically significant, $t(255) = -0.409$, $p = 0.683$ (see Table VI). The very small mean difference (−0.02) indicates no real practical disparity in how engaging the navigation experience was perceived across settings.

TABLE VI. PAIRED-SAMPLES T-TEST FOR NAVIGATION EXPERIENCE ENGAGEMENT

Variable Pair	Mean Difference	t	df
Indoor vs Outdoor Engagement Scores	−0.020	−0.409	255

The results of Hypothesis 3 suggest that users found the navigation experience equally engaging in both indoor and outdoor environments. This outcome may reflect the consistent visual and interaction design of the AR application across contexts, or it may point to the novelty of the technology itself as a key driver of engagement, independent of location.

From a design perspective, this reinforces the value of cohesive AR experiences that deliver similar affective impact regardless of spatial constraints or environmental variables.

D. Hypothesis 4 (H4): The Application Loads Faster and more Smoothly in Indoor Environments than in Outdoor Environments

The fourth hypothesis examined whether users perceived the augmented reality (AR) application as performing better—in terms of content loading speed and responsiveness—indoors compared to outdoors. This hypothesis is grounded in the assumption that indoor environments often provide more stable Wi-Fi connectivity, reduced glare, and consistent lighting, which may facilitate faster AR marker recognition and content rendering.

To evaluate this, a paired-samples t-test was conducted comparing responses to the following items:

- "How fast was the loading and display of AR content (VPS, SLAM) indoors?"
- "How fast was the loading and display of AR content (VPS, SLAM) outdoors?"

Responses were given on a 5-point Likert scale (1 = Very slow, 5 = Very fast). As shown in Table VII, users rated the application slightly higher in terms of speed indoors (M = 3.81, SD = 0.83) than outdoors (M = 3.71, SD = 0.88).

TABLE VII. DESCRIPTIVE STATISTICS FOR PERCEIVED AR LOADING SPEED INDOOR VS. OUTDOOR

Variable	Mean	Std. Deviation	N
Indoor AR Loading Speed	3.81	0.83	256
Outdoor AR Loading Speed	3.71	0.88	256

A paired-samples t-test revealed a statistically significant difference between the two contexts, $t(255) = 2.167$, $p = 0.031$ (see Table VIII). The mean difference of +0.10 confirms that the application was perceived as faster indoors.

TABLE VIII. PAIRED-SAMPLES T-TEST FOR AR LOADING SPEED IN INDOOR VS. OUTDOOR CONTEXTS

Variable Pair	Mean Difference	t	df	Sig. (2-tailed)
Indoor vs. Outdoor AR Loading Speed	+0.102	2.167	255	0.031

These results support the hypothesis that AR content was perceived to load more quickly in indoor environments. Several factors may account for this, including:

- more consistent lighting conditions and surfaces,
- better tracking support from structured geometry (walls, doors),
- more reliable indoor connectivity (Wi-Fi vs. mobile data).

From a usability standpoint, this finding highlights the need for optimizing AR systems for outdoor performance – possibly

through enhanced tracking algorithms or adaptive rendering for varying environmental conditions. Fig. 5 illustrates the distribution of perceived utility scores of the application based on the type of device used.

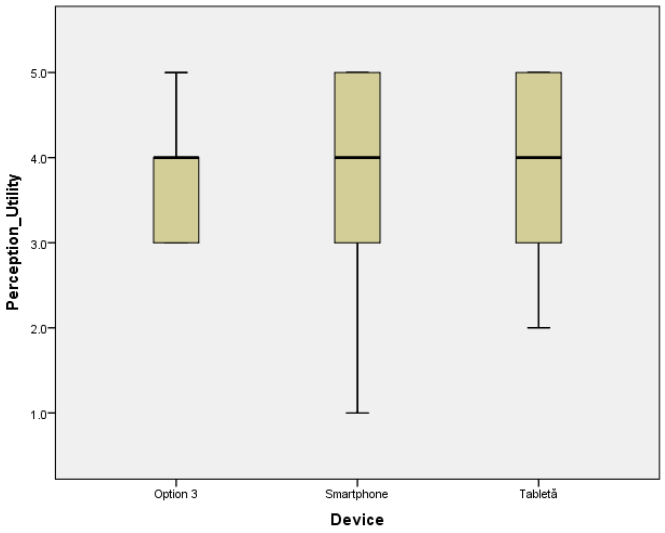


Fig. 5. Distribution of perceived utility scores by device type. Median utility is higher for smartphones and tablets compared to laptops (Option 3). The wider range for smartphones suggests a more varied but generally positive user experience.

A higher median is observed for smartphones and tablets compared to "Option 3" (likely representing laptops). Smartphone users reported a wider range of perceptions (with values ranging from 1 to 5), suggesting a more variable experience, but with an overall positive tendency. This distribution supports Hypothesis H4, which states that the application is perceived as more useful when used on a smartphone.

E. Hypothesis 5 (H5): Users are more Likely to Express an Intention to Regularly use the Application for Indoor Navigation than for Outdoor Navigation

The fifth hypothesis explored whether students reported a higher behavioral intention to regularly use the AR navigation application in indoor settings (e.g., classrooms, hallways) compared to outdoor navigation (e.g., between buildings across the campus). Behavioral intention is a central variable in the Technology Acceptance Model (TAM) and Unified Theory of Acceptance and Use of Technology (UTAUT), as it directly influences actual technology adoption.

To examine this hypothesis, participants responded to the following parallel statements:

- "I would be willing to regularly use this application for indoor navigation on campus."
- "I would be willing to regularly use this application for outdoor navigation between campus buildings."

Responses were measured using a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). Descriptive statistics presented in Table IX revealed virtually identical mean scores: indoor (M = 3.93, SD = 1.03) and outdoor (M = 3.93, SD = 1.00), suggesting no practical difference in user intention across environments.

TABLE IX. DESCRIPTIVE STATISTICS FOR INTENTION TO USE INDOOR VS. OUTDOOR

Variable	Mean	Std. Deviation	N
Intention to Use (Indoor)	3.93	1.03	256
Intention to Use (Outdoor)	3.93	1.00	256

The paired-samples t-test confirmed this equivalence, showing no statistically significant difference between the two scores:

t(255) = 0.079, p = 0.937 (see Table X). The mean difference was negligible (+0.004), indicating consistent behavioral intention across indoor and outdoor contexts.

TABLE X. PAIRED-SAMPLES T-TEST FOR INTENTION TO USE

Variable Pair	Mean Difference	t	df	Sig. (2-tailed)
Indoor vs Outdoor intention to use	+0.004	0.079	255	0.937

These findings suggest that users are equally inclined to use the application regularly in both indoor and outdoor navigation contexts. This result reinforces the idea that user trust in the application extends across different spatial environments, likely due to a coherent interface and consistent interaction model.

This outcome is encouraging from a scalability perspective, indicating that future AR implementations in university campuses may find equal acceptance for both indoor and outdoor orientation support.

F. Hypothesis 6 (H6): Users Perceive the Application's Localization Accuracy and Stability as Significantly Higher in Indoor Scenarios than in Outdoor Scenarios

The sixth hypothesis addressed the perception of localization quality – a crucial factor in the effectiveness of augmented reality (AR) navigation tools. AR experiences depend heavily on stable and precise positioning, typically achieved through technologies such as VPS (Visual Positioning System) and SLAM (Simultaneous Localization and Mapping). Given the more controlled conditions in indoor environments (e.g., structured geometry, constant lighting, Wi-Fi signals), it was expected that localization would be perceived as more stable indoors than outdoors.

To assess this, participants evaluated two items:

- "How would you rate the stability and accuracy of the localization function in indoor navigation (buildings, hallways)?"
- "How would you rate the stability and accuracy of the localization function in outdoor navigation?"

Each was measured on a 5-point Likert scale (1 = Very low, 5 = Very high). Descriptive statistics are shown in Table XI, revealing a slightly higher mean for indoor localization (M = 3.68, SD = 0.86) than for outdoor (M = 3.63, SD = 0.88).

TABLE XI. DESCRIPTIVE STATISTICS FOR PERCEIVED LOCALIZATION ACCURACY IN INDOOR VS. OUTDOOR CONTEXTS

Variable	Mean	Std. Deviation	N
Indoor Localization Accuracy	3.68	0.86	256
Outdoor Localization Accuracy	3.63	0.88	256

The paired-samples t-test indicated that the difference in perceived accuracy was not statistically significant, $t(255) = 1.080$, $p = 0.281$ (see Table XII). The mean difference was only +0.051.

TABLE XII. PAIRED-SAMPLES T-TEST FOR LOCALIZATION ACCURACY SCORES

Variable Pair	Mean Difference	t	df
Indoor vs Outdoor Localization Accuracy	+0.051	1.080	255

Although the hypothesis was not supported statistically, the result suggests that users generally perceive localization to be similarly stable in both environments, with a slight (non-significant) preference for indoor performance. This outcome may be attributed to improvements in mobile AR tracking technologies and robustness across spatial contexts.

Designers and developers should nonetheless consider environmental differences when optimizing positioning systems, particularly for outdoor use cases involving variable light, weak signals, or open unstructured areas.

V. DISCUSSION AND LIMITATIONS

The results of this study provide empirical evidence that a unified, infrastructure-free AR navigation system can deliver a consistent and usable experience across both indoor and outdoor university environments. Participants reported similar levels of intuitiveness, usefulness, engagement, and behavioral intention to reuse the application regardless of spatial context, as measured through the System Usability Scale (SUS) and the Technology Acceptance Model (TAM). The only statistically significant difference was a slight advantage in perceived AR content loading speed in indoor scenarios, likely due to more stable lighting and connectivity conditions.

These findings reinforce the initial design assumption that a consistent user interface and interaction model can minimize the cognitive gap between indoor and outdoor usage. The results also address the core research question posed in this study: Can a single AR navigation system maintain usability and user acceptance across spatial contexts without reliance on GPS or external infrastructure? The answer, based on the data, appears to be affirmative.

From a comparative perspective, the usability scores obtained in this study (SUS ~74) align closely with values reported in prior AR navigation studies. For example, Chen et al. [17] reported a SUS score of 74 for an ARCore-based campus system, and Sato et al. [21] reported 76 in a museum setting. These similarities strengthen the external validity of our findings. Moreover, unlike BLE- or marker-based systems such as Yamamoto et al. [6], which required dense beacon deployment, our system operates solely on visual-inertial tracking and cloud-based spatial anchors—simplifying deployment and enhancing scalability.

Technical stability and system responsiveness have also been highlighted in previous literature as key predictors of user satisfaction [17], [23]. Our study extends these insights by showing that responsiveness perceptions remain stable even across variable lighting and open vs. closed environments. This outcome suggests that modern mobile AR frameworks like ARCore and ARKit can now support robust navigation without needing physical markers or dedicated sensors.

Despite the strengths of the approach, several limitations must be acknowledged. First, the experiment was conducted exclusively within a single university campus, which may limit the generalizability of the results to other institutional settings with different architectural layouts or technological infrastructures. Although both indoor and outdoor scenarios were evaluated, the tasks involved relatively short, predefined routes. Longer or more complex navigation scenarios—such as multistep journeys involving transitions between buildings—were not assessed and may yield different results in terms of user satisfaction and system responsiveness.

Second, all participants used their personal smartphones, which, while realistic and ecologically valid, introduced variation in hardware performance and environmental adaptation (e.g., screen brightness outdoors, camera quality). These variables were not controlled or measured and may have introduced noise into user experience ratings.

Lastly, although this study employed validated instruments (SUS and TAM) for assessing usability and acceptance, it relied solely on post-task self-reported measures. Objective performance metrics such as navigation time, error rate, path deviation, or eye-tracking data were not collected. Integrating such data in future research would complement subjective feedback and allow more comprehensive usability evaluation.

Overall, this study demonstrates that consistent AR navigation performance can be achieved across spatial contexts using a single, infrastructure-independent mobile application. These findings support the development of scalable, user-centered AR wayfinding systems suitable for real-world academic deployments.

VI. CONCLUSION

This study investigated whether a single mobile augmented reality (AR) navigation application could provide a consistent and usable experience across both indoor and outdoor university environments. Using a within-subjects experimental design with 256 participants, the system was evaluated through established usability and acceptance frameworks (SUS and TAM) and demonstrated strong performance in both spatial contexts.

The findings validate the feasibility of infrastructure-free, camera-based AR navigation solutions in academic settings. By maintaining consistent interaction patterns and localization methods across environments, the application enabled users to navigate seamlessly, regardless of whether they were indoors or outdoors. This cross-context adaptability addresses a key

challenge in smart campus design, where traditional tools often fail to bridge spatial transitions.

Future work will extend this research along several directions. First, incorporating objective performance metrics such as route efficiency, task duration, and error rates would complement subjective assessments. Second, the system could be deployed and evaluated in campuses with more complex spatial configurations or across multiple institutions. Third, integration with real-time data (e.g., scheduling, crowd flow) and accessibility enhancements (e.g., voice-only guidance or screen-reader support) may further improve adoption and inclusivity.

REFERENCES

[1] S. Qiu, Y. Wu, Y. Wang, and K. Li, "Augmented Reality Navigation Systems: A Systematic Literature Review," Sensors, vol. 21, no. 15, pp. 1–27, 2021.

[2] J. Kim and K. Lee, "Semantic Labeling in Augmented Reality Navigation Interfaces: Enhancing Wayfinding Through Context-Aware AR Cues," IEEE Trans. Vis. Comput. Graph., vol. 29, no. 1, pp. 315–324, 2023.

[3] H. Huang, N. Schmidt, and H. Gartner, "An AR-based Pedestrian Navigation System in Urban Environments Using GPS and Computer Vision," in Proc. ACM CHI Conf. Human Factors Comput. Syst., 2021, pp. 1–12.

[4] R. Gupta, M. Zhao, and T. Anderson, "WiFi and SLAM Fusion for Mobile AR Indoor Navigation in Hospitals," IEEE Trans. Mobile Comput., vol. 22, no. 3, pp. 1101–1115, 2023.

[5] G. Rossi, P. Mendes, and F. Ramos, "Augmented Reality for Indoor Navigation: A Case Study Using ARKit," ACM J. Emerg. Technol. Comput. Syst., vol. 18, no. 1, pp. 1–16, 2022.

[6] K. Yamamoto, T. Hara, and M. Ueda, "Evaluating BLE Beacon-Enhanced AR Navigation in a University Campus Environment," in Proc. IEEE PerCom, 2021, pp. 312–319.

[7] L. Fernandez, J. Ramirez, and M. Ortega, "Cloud AR Navigation with GPS Emulation in University Campuses," Appl. Sci., vol. 14, no. 2, pp. 1761–1778, 2024.

[8] Toma Marian-Vladut, Turcu Corneliu Octavian and Pascu Paul, "Evaluating User Acceptance and Usability of AR-Based Indoor Navigation in a University Setting: An Empirical Study" International Journal of Advanced Computer Science and Applications(IJACSA), 16(4), 2025. http://dx.doi.org/10.14569/IJACSA.2025.0160406

[9] J. Brooke, "SUS: A Quick and Dirty Usability Scale," in Usability Evaluation in Industry, P. W. Jordan et al., Eds. London: Taylor and Francis, 1996, pp. 189–194.

[10] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Q., vol. 13, no. 3, pp. 319–340, 1989.

[11] Google, "Live View in Google Maps," The Keyword, 2019. [Online]. Available: https://www.blog.google/products/maps/live-view-augmented-reality-navigation/

[12] H. Huang, N. Schmidt, and H. Gartner, "An AR-based Pedestrian Navigation System in Urban Environments Using GPS and Computer Vision," in Proc. ACM CHI, 2021, pp. 1–12.

[13] G. Rossi, P. Mendes, and F. Ramos, "Augmented Reality for Indoor Navigation: A Case Study Using ARKit," ACM J. Emerg. Technol. Comput. Syst., vol. 18, no. 1, pp. 1–16, 2022.

[14] T. Nguyen and J. Park, "AR Navigation for Complex Shopping Malls Using BLE and Visual Cues," Sensors, vol. 22, no. 9, pp. 3254–3271, 2022.

[15] J. Mercier, N. Chabloz, G. Dozot, et al., "Impact of geolocation data on augmented reality usability: A comparative user test," ArXiv, Aug. 2023.

[16] Z. Qiu, M. Ashour, X. Zhou, and S. Kalantari, "NavMarkAR: A Landmark-based Augmented Reality Wayfinding System for Enhancing Spatial Learning of Older Adults," ArXiv, Nov. 2023.

[17] Y. Chen, L. Zhao, and C. Wang, "Multi-Floor Indoor Navigation Using ARCore and Visual-Inertial Odometry," Sensors, vol. 24, no. 1, pp. 210–230, 2024.

[18] Z. Qi, R. Lu, and M. Chen, "Hybrid AR Navigation for Campus Environments Using Visual-Inertial Odometry and Spatial Anchors," IEEE Access, vol. 12, pp. 22050–22063, 2024.

[19] E. Karan, M. E. Morkoç, and A. E. Şireci, "Indoor and Outdoor Navigation Application Development with Augmented Reality Technology," in Proc. 5th Int. Conf. Comput. Math. Eng. Sci. (CMES), Van, Turkey, 2021.

[20] T. Macháč, P. Hořejší, M. Šimon, and A. Šimerová, "A Comparative Evaluation of Augmented Reality Indoor Navigation versus Conventional Approaches," Technical Journal, vol. 19, no. 1, pp. 62–71, 2025.

[21] Y. Sato, K. Tanaka, and T. Hashimoto, "Cloud Anchors and Markerless AR in Museums: Enhancing Visitor Navigation and Engagement," Virtual Reality, vol. 27, no. 3, pp. 401–419, 2023.

[22] C. Kwok, S. Ho, and H. Yuen, "Understanding Student Adoption of Augmented Reality Navigation in Higher Education," Comput. Educ., vol. 196, no. 104659, 2024.

[23] M. Zhao, A. Malik, and F. Leone, "Precision AR Indoor Navigation Using UWB and Head-Mounted Sensors," in Proc. IEEE VR, 2024, pp. 98–106.

[24] J. Ahn, H. Lee, and M. Song, "SLAM-Based AR Guidance in Hospitals: Evaluating Usability and Spatial Learning," J. Ambient Intell. Humaniz. Comput., vol. 15, pp. 245–260, 2024.

[25] K. Cheliotis, F. Liarokapis, M. Kokla, and E. Tomai, "A Systematic Review of Application Development in Augmented Reality Navigation Research," Cartogr. Geographic Inf. Sci., vol. 50, no. 2, pp. 1–23, May 2023.

Brightness-Aware Generative Adversarial Network for Low-Light Image Enhancement

Huafei, Zhao, Mideth Abisado
College of Computing & Information Technologies, National University, Manila, Philippines

***Abstract*—Images from low-light frequently exhibit poor visibility, excessive noise, and color distortion, which substantially impair both computer vision systems and human visual perception. Although numerous enhancement techniques have been developed, producing visually appealing results with well-maintained structural details and natural color reproduction continues to pose significant challenges. To address these limitations, this paper present an Brightness-Aware Generative Adversarial Network (BA-GAN) for robust low-light image enhancement (LLIE). Our framework employs a U-Net-based generator that effectively captures multi-scale contextual features while preserving fine image details through skip connections. The key innovation lies in our novel Brightness Attention Mechanism Module, integrated within the decoder, which dynamically directs the network's focus to regions requiring substantial illumination correction. To ensure local photorealism, this paper adopt a PatchGAN discriminator architecture. The complete model is trained on the LOL dataset using a composite loss function combining: (1) adversarial loss for realistic image generation, (2) brightness attention loss for keeping the brightness accuracy, and (3) perceptual loss to maintain structural and semantic fidelity. Extensive experiments validate that our BA-GAN outperforms current state-of-the-art methods, achieving superior performance on both quantitative metrics (PSNR: 20.7127, SSIM: 0.7963, LPIPS: 0.2271) and qualitative visual assessments. The enhanced images demonstrate significantly improved visibility while effectively suppressing noise and preserving natural color characteristics.**

Keywords*—Low-light image enhancement; generative adversarial networks; U-Net; PatchGAN; attention mechanism; deep learning

I. Introduction

Images captured by cameras often used by advanced computer vision tasks such as image classification, object detection, and image generation, and security monitoring [1]. However, during the capture process, due to low-light environments or insufficient device lighting capabilities, the resulting images are often dim, visually unclear, and have low contrast. These images significantly hinder subsequent computer vision processing, rendering them largely unusable. Low-quality images not only affect human visual experience but also have adverse effects on subsequent computer vision applications.

When capturing images in low-light environments, long exposure or slow shutter speeds are often used to increase the amount of light captured by the camera, making more complex low-light blur degradation unavoidable [2]. While flash photography is commonly employed to brighten low-light environments, it often introduces uneven illumination artifacts that degrade image quality through increased noise and unnatural lighting patterns. To address these issues, professional photographers typically enhance the quality of pictures taken under low-light environments by adjusting camera parameters or using image editing software (such as Adobe Photoshop) [3]. However, this process usually requires complex operations and professional skills, making it unsuitable for ordinary users. Therefore, automated LLIE technology is crucial for solving this problem.

Conventional approaches to LLIE mainly consist of histogram equalization techniques [4] and Retinex-based methods [5]. With the advancement of GPU computing power, the widespread application of neural networks and deep learning, and the emergence of training datasets, it is now possible to learn detailed information from images through deep learning methods, thereby improving the contrast of images with low-light [6]. Deep learning methods utilize deep neural networks to establish an end-to-end mapping between low-light and normal-light image domains through data-driven learning from extensive training samples. Compared to traditional methods, they possess more powerful feature representation capabilities and have achieved impressive results on benchmark datasets [7]. Existing LLIE methods, while capable of upgrading the brightness and contrast of dim images to some extent, still have issues such as color distortion, noise amplification, and insufficient model generalization [8] as it's hard to obtain paired datasets in real-world production and life, as well as the limited global processing capability of convolutional operations on images.

GANs [9] have proven effective for image-to-image translation tasks, including low-light enhancement. GANs leverage a generator network to produce enhanced images and a discriminator network to distinguish generated fake images from real normal-light images, leading to perceptually more convincing results. However, existing methods may still struggle with balancing noise suppression, detail preservation, and adaptive enhancement across different image regions and lighting variations.

To solve these limitations, this paper propose a Brightness-Aware Generative Adversarial Network (BA-GAN). The key contributions in this paper summarized as follows:

1) A U-Net generator architecture that fuses multi-scale details. This paper adopted a U-Net-based generator, which processes images through a symmetric encoder-decoder structure. The key to this architecture lies in its skip connections. These connections directly transmit the low-level, high-

*Corresponding Author

resolution detailed features (such as edges and textures) captured by the encoder at different levels to the corresponding levels of the decoder. This design can effectively avoid the common problem of detail information loss in deep networks, thereby maximizing the preservation of fine structures while enhancing the image brightness and preventing the final result from becoming blurred.

2) A novel illumination attention module for adaptive enhancement. This is the core innovation of this paper. We integrated an illumination attention module into the decoder of the generator. This module aims to solve the problem of uneven illumination in low-light images, enabling the network to dynamically and adaptively focus its attention on the area's most in need of illumination correction according to the image content. By learning and applying this spatially varying enhancement strategy, the model can avoid overexposing areas with existing light sources while sufficiently brightening extremely dark areas, thus achieving a more natural and accurate enhancement effect.

3) A PatchGAN discriminator that focuses on local realism. To generate more perceptually convincing results, our framework employs a PatchGAN discriminator. Different from traditional discriminators that make true/false judgments on the entire image, PatchGAN evaluates the authenticity by assessing local patches of the image. This mechanism encourages the generator to not only focus on the global visual effect but also pay attention to generating locally realistic textures and structures. This helps to significantly improve the detail quality and overall realism of the enhanced image.

4) A composite loss function for balanced optimization. The low-light enhancement task requires a balance among multiple objectives. To this end, this paper designed and adopted a composite loss function to guide the training of the model. This function combines three key components: (1) adversarial loss, used to enhance the realism of the generated image; (2) brightness attention loss, used to ensure the accuracy of brightness correction; and (3) perceptual loss, used to maintain the original structure and semantic content of the image in the deep feature space. This multi-objective optimization strategy ensures that the final enhancement result achieves the best balance in terms of visual realism, brightness, and content fidelity.

In summary, LLIE is crucial for improving visual quality and enabling robust performance in downstream computer vision tasks. However, existing methods ranging from traditional techniques like histogram equalization and Retinex theory to deep learning approaches—often suffer from color distortion, noise amplification, and limited generalization due to the scarcity of real-world paired datasets and the inherent constraints of convolutional operations. While GANs have shown promise in LLIE, challenges remain in achieving adaptive enhancement, detail preservation, and noise suppression. To address these limitations, this paper propose a BA-GAN that integrates a U-Net generator with multi-scale skip connections, a novel illumination attention module for adaptive brightness correction, a PatchGAN discriminator for local realism, and a composite loss function for balanced optimization. Our approach aims to deliver natural, high-fidelity enhancements while addressing uneven illumination and preserving fine-grained details.

II. Related Works

A. Traditional Methods

Conventional approaches to LLIE mainly consist of histogram equalization techniques and Retinex-based methods.

The histogram equalization method operates on the histogram of an image and upgrades the image's contrast by altering its dynamic range, which can enhance the quality of dim images to some extent [10]. The histogram records the distribution of image data, and due to its uneven distribution, this method redistributes it to achieve non-linear stretching, making the distribution of image data more reasonable. Examples include HE [2] and CLAHE [11]. Numerous improved methods based on HE, have enhanced HE to varying degrees. Compared to global equalization, local equalization focuses more on the internal detail distribution of the image, but the required computational effort also increases. Low-light images enhanced by such methods are prone to issues like overexposure or underexposure, color distortion, artificial artifacts, loss of image details, and unnatural appearance. Sometimes, the resulting images may also retain the noise present in the original images.

Retinex-based methods perform image enhancement through a decomposition process, separating low-light images into illumination and reflectance components. This separation is typically achieved by applying physical priors or mathematical regularization constraints. The reflectance component is subsequently obtained by solving an optimization problem involving both the estimated illumination and the original input image [12], and is considered the final enhancement result. Building on Retinex, Single Scale Retinex (SSR) [13] was proposed, which removes the influence of illumination by comparing the central element with the surrounding area's luminance values. Multi-Scale Retinex (MSR) [14] was introduced to address the issue of Gaussian kernel optimization quality. The MSR with Color Restoration (MSRCR) [15] algorithm adjusts color distortion by adding a color restoration factor.

Retinex-based LLIE methods can effectively address issues such as underexposure and overexposure. However, they often overlook the presence of noise during image processing, leading to noise amplification. Additionally, these methods struggle to find an effective prior or regularization. As a result, low-light images enhanced by these methods often have limitations like color distortion, artificial artifacts, overexposure, and accompanying noise.

B. Deep Learning-Based Methods

The rapid evolution of deep learning has led to the development of numerous data-driven approaches for LLIE, consistently outperforming conventional methods in terms of image quality improvement. These techniques leverage deep neural architectures to learn complex mappings between low-light and normal-light image domains, effectively enhancing

visual clarity and perceptual quality. Current deep learning-based approaches can be systematically categorized into four paradigms: (1) supervised learning, (2) unsupervised learning, (3) semi-supervised learning, and (4) zero-shot learning methods.

1) Supervised learning methods. Supervised learning is widely used in LLIE. It involves providing a set of images along with their corresponding enhanced versions as training data, enabling the model to learn an end-to-end transformation that converts input images into enhanced outputs with improved visual quality. As the pioneering deep learning solution for LLIE, LLNet [16] demonstrated that a carefully designed autoencoder could effectively enhance low-light images while preserving important visual features. MSR-NET [17] introduced the use of convolutional neural networks, combined with Retinex theory, to restore images to good brightness levels. Wei et al. [18] proposed Retinex-Net, which includes decomposition, adjustment, and reconstruction, corresponding to three sub-networks. During the adjustment phase, the brightness of the illumination map is increased through multi-scale encoders and decoders. In the final reconstruction stage, the enhanced image is obtained through element-wise multiplication of the processed illumination and reflectance maps. However, the edges of the enhanced structure are overly prominent, and the naturalness of the entire image is not perfect. KinD [19] followed the Retinex theory-based approach of decomposition followed by enhancement, constructing a two-stage network. The restoration sub-network introduced information from real illumination to prevent image distortion, and the loss function ensured the smoothness of the illumination layer. KinD++ [20] alleviated residual visual artifacts in the results generated by KinD through a multi-scale illumination attention module. MIRNet [21] used a multi-scale feature fusion method, including multi-scale residual blocks, multi-resolution convolutional streams, and attention mechanism-based multi-scale features, achieving high performance in image enhancement tasks. Liu et al. [22] proposed a three-stage brightness-aware network based on brightness-aware attention and residual quantization encoding blocks, and their designed a query module link the low-light and normal-light domains. Building upon U-Net [23], LAU-Net [24] augments the baseline architecture with a Parallel Attention Unit (PAU) to weight informative features dynamically, an Internal Resizing Module (IRM) for scale-invariant representation learning, and auxiliary convolutional layers to mitigate low-light artifacts.

Supervised learning methods have shown improvements in color balance and brightness adjustment compared to traditional methods, but they tend to lose local details and often rely on paired datasets, leading to overfitting and low cross-data generalization. In the field of LLIE, paired datasets are limit in quantity and synthesizing image datasets can result in poor model generalization.

2) Unsupervised learning methods. Unsupervised learning refers to training data without corresponding target enhanced images, allowing the model to learn the structure and features of image data on its own. After the introduction of GANs, GANs achieved significant results in computer vision tasks. This method do not need paired training data and can be trained using unpaired datasets, making data acquisition relatively easier. In the realm of LLIE, EnlightenGAN proposed by Jiang et al. [25] was the first to eliminate the dependency on paired training data, greatly enhancing the flexibility of generated images and adapting to various scenarios. The introduction of RetinexGAN [26] combined GANs with the Retinex model, incorporating a decomposition network and two discriminative networks, where the decomposition network decomposes the Retinex model, and the discriminators evaluate the illumination and reflectance components. Zhu et al. [27] designed Cycle-consistent Generative Adversarial Networks (CycleGAN), which use cycle-consistency loss to achieve image style transfer. However, applying it to LLIE training is challenging as it overlooks the preservation of local features. Hu et al. [28] developed a novel two-stage unsupervised framework that separates the enhancement process into distinct pre-enhancement and post-enhancement phases. The initial stage employs conventional Retinex-based algorithms for primary image enhancement, while the subsequent stage utilizes an adversarial trained refinement network to achieve superior quality improvement. In a related approach, Wang et al. [29] introduced an attention-guided unsupervised GAN architecture featuring: (1) an auxiliary edge restoration module for enhanced sharpness preservation, and (2) a dedicated attention mechanism for improved colour fidelity reconstruction.

3) Semi-supervised learning methods. Semi-supervised learning integrates the advantages of supervised and unsupervised learning. Yang et al. [25] constructed a deep recursive frequency band network, which first uses supervised learning to learn frequency band representations from coarse to fine, and then employs unsupervised adversarial learning to improve the model's generalization ability. Through this end-to-end recursive training, noise removal is effectively achieved, and the ability to recover local structural details is improved. Dimma [30] built a brightness adjustment module to obtain a low-light version of normal-light images, thereby enabling semi-supervised training.

4) Zero-shot learning methods. Zero-shot learning in the field of LLIE refers to the ability to learn how to enhance images solely from test images. Guo et al. introduced Zero-DCE [31], a lightweight, reference-free deep curve estimation model that achieves pixel-level adjustments of input images through image-to-curve mapping and utilizes a loss function to drive reference-free training. Building on this, an accelerated lightweight version, Zero-DCE++ [32], was introduced. Zhang L et al. introduced a "zero-learning" approach for backlit image restoration, which is based on deep learning but does not rely on any training image. The framework ExCNet, a specially designed convolutional neural network architecture. Zhu et al. [32] introduced RRDNet, an innovative three-branch CNN that decomposes input images into illumination, reflectance, and noise components. This architecture employs iterative

optimization of a custom loss function to simultaneously predict noise patterns and restore underexposed regions. In a complementary approach, Liu et al.'s RUAS [33] formulates a mathematical model to represent the intrinsic illumination structure of low-light images, then unfolds this optimization process into a learnable network architecture. By exploring optimal configurations within a constrained search space, RUAS achieves superior enhancement performance through its propagated structure.

In summary, conventional enhancement methods depend on fixed statistical priors, limiting their adaptability to diverse real-world scenarios. While contemporary supervised approaches predominantly employ U-Net architectures, they frequently suffer from spatial detail degradation and inadequate feature representation due to the absence of attention mechanisms. To address these limitations, our proposed framework innovatively combines attention modules with a U-Net backbone and incorporates a PatchGAN discriminator. This hybrid architecture enables: (1) enhanced multi-scale feature learning, (2) improved discriminative feature extraction, (3) effective noise suppression, and (4) robust contextual information aggregation across different scales.

III. Methodology

The LLIE method based on the GANs model introduced in this paper employs U-Net as the generator, which is characterized by its low data requirement and fast training speed. By using PatchGAN as the discriminator instead of the traditional binary classifier, the trained model pays more attention to image details, improves image quality, and accelerates network convergence. The network structure of the method used is illustrated in Fig. 1. The generator consists of a U-Net backbone with embedded illumination attention modules. The discriminator follows a PatchGAN structure to promote local realism.

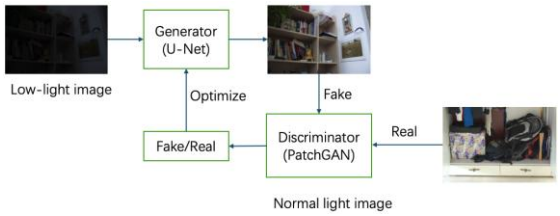


Fig. 1. The proposed BA-GAN structure.

C. Generator

U-Net has demonstrated success in image segmentation, image translation, and enhancement. Through the structure of the encoder and decoder, along with skip connections, U-Net avoids generating blurry images or losing details (such as edges and textures). The symmetric structure of U-Net better preserves spatial correspondences.

The extracted image features are encoded and decoded by the U-Net network, and the same convolution and structural detail residual fusion operations are performed on all branches to improve the features, resulting in feature maps with stronger expressive capabilities. The U-Net network can combine low-level and high level features from the encoder and decoder, fully utilizing contextual texture information. The proposed generator structure is illustrated in Fig. 2.

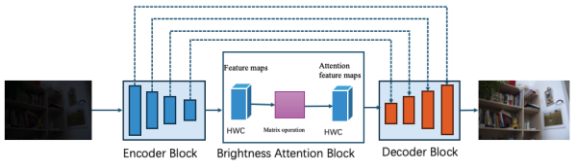


Fig. 2. The structure of the proposed generator.

D. Brightness Attention Module

In the low-light enhancement task, the degradation of the image is not uniformly distributed. Some areas may be almost completely black and require significant brightening, while other areas may have a small amount of light sources and need fine adjustment to avoid overexposure. Due to its fixed local receptive field, the traditional convolution operation has difficulty capturing such global and dynamic illumination dependencies. Therefore, this paper designed an Brightness Attention Module (BAM). Its core objective is to enable the network to adaptively evaluate the importance of different feature channels for the final illumination restoration and recalibrate the features accordingly, so as to achieve precise and differentiated enhancement of illumination.

Our BAM is strategically placed between the encoder and decoder of the U-Net. It takes the deepest feature map $F_{enc} \in R^{C\times H'\times W'}$ output by the encoder as input and performs the following three consecutive operations to generate the attention-calibrated feature map $F_{att} \in R^{C\times H'\times W'}$.

1) Illumination squeeze. First, this paper perform Global Average Pooling on the input feature map F_{enc}, compressing it in the spatial dimension to generate a scalar descriptor for each channel. This process can be regarded as the extraction and generalization of the global illumination information contained in each feature channel. The object is to obtain a vector $z \in R^C$, where each element z_c represents the global illumination response of the c-th feature channel.

$$z_c = F_{sq}(F_{enc}) = \frac{1}{H'\times W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'} F_{enc_c}(i,j) \quad (1)$$

2) Illumination excitation. To learn the complex non-linear illumination dependencies between channels, this paper uses two fully connected (FC) layers to process the compressed descriptor z. The first FC layer reduces the dimension from C to C/r (where r is the reduction ratio) and uses the ReLU activation function. The second FC layer restores the dimension to C and uses the Sigmoid activation function, ultimately generating the attention weights $s \in R^C$ for each channel. The output values of the Sigmoid function range from 0 to 1, which can exactly serve as the "importance" or "gain" weights for each channel. The closer the value is to 1, the more important the channel is for illumination restoration.

$$s = F_{ex}(z, W) = \sigma\big(W_2\delta(W_1 z)\big) \quad (2)$$

Where σ is the Sigmoid function, δ is the ReLU function, and W_1 and W_2 are the weights of two fully connected layers.

3) Feature recalibration. Finally, this paper multiplies the learned channel attention weights s with the original input feature map F_{enc} channel by channel to obtain the final output attention feature map F_{att}. This operation enables the network to enhance the feature channels that are beneficial for illumination enhancement while suppressing the irrelevant channels that may introduce noise or artifacts. This calibrated feature map F_{att} is then fed into the decoder to guide the subsequent image reconstruction process.

$$F_{att_c} = s_c \cdot F_{enc_c} \quad (3)$$

Although the BAM this paper proposed draws on the idea of channel attention in structure [34], its uniqueness and effectiveness are reflected in the following aspects:

- Task Specificity: Different from general attention modules, our BAM is specifically designed to solve the problem of uneven illumination. By modeling global information at the bottleneck of the network, it focuses on learning which regional features should be brightened and which dark - area details should be retained, rather than general feature enhancement.
- Strategic Placement: This paper place the BAM at the bottleneck of the U-Net, instead of spreading it after each convolutional block as in traditional practices. The reason for this design is that performing global illumination calibration at the place where information compression is most extreme allows the decoder to obtain prior knowledge about the overall illumination distribution at the first step of the reconstruction process, thereby more effectively guiding all subsequent upsampling and feature fusion steps and fundamentally reducing artifacts and color distortion.
- Synergy with Loss Function: More importantly, the design of this module forms a synergy with the Brightness Attention Loss function this paper proposed. The BAM learns the intrinsic dependence of illumination at the feature level, while the brightness loss supervises the final brightness accuracy at the pixel level. The two jointly guide the network from different dimensions, ensuring the realism and accuracy of illumination enhancement.

E. Discriminator

In designing the discriminator network, this paper references the PatchGAN discriminator proposed by Isola et al. [27] to replace the traditional binary classifier discriminator. This discriminator outputs an N × N matrix, and the image is judged as real or generated based on the mean value of the matrix. Compared to the traditional binary classifier discriminator, the advantage of using the PatchGAN discriminator lies in its output being a matrix, with the final result being the matrix mean, which fully considers the influence of different regions of the image, thereby enhancing image quality and paying more attention to image details. Additionally, the computation of small-sized image patches significantly accelerates the convergence speed of the network. This paper designs a CNN model with 4 fully convolutional layers. Except for the last convolutional layer, the other 3 layers undergo BatchNorm for data standardization after convolution, and LeakReLu is used as the activation function. The proposed discriminator structure is illustrated in Fig. 3.

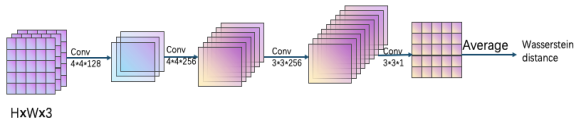


Fig. 3. The structure of the discriminator.

F. Loss Functions

Adversarial Loss [35]. The adversarial loss is the core driving force of the GANs framework. Its fundamental goal is to drive the generator to produce enhanced images that are more visually realistic and indistinguishable from real normal - illuminated images in terms of distribution. This process can be understood as a two - player game involving two networks (the generator G and the discriminator D).

Generator (G): The generator is a U-Net-based network. Its task is to receive a low-light image as input and output an enhanced fake image I_{gen}. The goal of the generator is to deceive the discriminator as much as possible and make it believe that the generated image is real.

Discriminator (D): The discriminator adopts the PatchGAN architecture. Its role is like an image connoisseur or a binary classifier, and its sole goal is to accurately distinguish between the real normal-illumination image I_{real} and the fake image I_{gen} produced by the generator.

$$L_{adv} = E\left[log\left(D(I_{real})\right)\right] + E\left[log\left(1 - D\left(I_{gen}\right)\right)\right] \quad (4)$$

Compared with traditional loss functions that only rely on pixel-level differences (such as L1 or MSE loss), adversarial loss has significant advantages in improving the perceptual quality of images. It does not require the generated image to be exactly the same as the target image in terms of pixels, but encourages the generated image to look more realistic in terms of texture, structure, and overall style.

In particular, the PatchGAN discriminator used in this paper does not output a single true/false judgment for the entire image. Instead, it makes judgments on N×N-sized image patches (Patches) in the image and then averages all the judgment results. This mechanism forces the generator to not only focus on global consistency but also ensure that the local texture and structure of the image are equally realistic, thereby effectively improving the quality of image details and accelerating the convergence of the network.

Brightness Attention Loss. Using global L1 or L2 loss directly between the brightness and the target image treats all regions in the image equally. However, in the low-light enhancement task, the human eye perceives brightness errors differently in different regions [36]. For example, in regions that already have a light source or are rich in texture, slight deviations in brightness are more likely to be noticed and cause artifacts; while in smooth and extremely dark regions, the weight of the error can be relatively lower [37]. To make the model prioritize the key regions that have a greater impact on

visual quality during the optimization process, this paper designed the Brightness Attention Loss Function.

Attention Map Generation. This paper assumes that in regions with drastic luminance changes (i.e., large gradient values), such as edges and textures, the accuracy of luminance restoration is crucial for the overall perceptual quality [38]. Therefore, this paper uses the gradient magnitude of the luminance channel of the target image to generate the attention map A. The specific steps are as follows:

First, this paper uses the standard Sobel operator to calculate the gradients G_x and G_y of the target luminance map Y_{target} in the horizontal and vertical directions respectively.

Then, this paper calculates the gradient magnitude of each pixel.

$$M_{grad}(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (5)$$

To convert the gradient magnitude into a normalized attention weight map A this paper processes it to make its values more concentrated in the key regions. First, this paper normalizes it to the range of $[0,1]$, then add a small constant ε to avoid division by zero, and use it as the final attention weight. The weight map A generated in this way has higher values in the edge and detail regions of the image and lower values in the smooth regions.

$$A(i,j) = \frac{M_{grad}(i,j)}{\max(M_{grad})} + \varepsilon \quad (6)$$

Finally, this paper combines the above attention weight map A with the brightness difference between the generated image Y_{gen} and the target image Y_{target} to define our brightness attention loss.

$$L_{bright} = \frac{1}{H \times W} \sum_{i,j} A(i,j) \cdot \left(Y_{gen}(i,j) - Y_{target}(i,j)\right)^2 \quad (7)$$

This paper uses the weighted L2 loss (squared difference) here mainly for two reasons: First, compared with the L1 loss, the L2 loss imposes a stronger penalty on larger errors, which helps the model quickly correct obvious brightness deviations in the early stage of training. Second, by multiplying with the attention weight A, this loss function forces the generator to prioritize optimizing the brightness accuracy in the key regions of the image structure (such as edges and textures), thereby effectively preserving image details and suppressing artifacts while improving the overall brightness.

Perceptual Loss [35] mainly uses the differences between high-dimensional feature maps of images to guide model training. Perception loss can make the image more similar in high-level information such as content and global structure, ensure the semantic information of the image as much as possible, and also enhance the detail information, which is conducive to the training of unpaired data sets.

$$L_{percep} = \frac{1}{W_i H_i C_i} \left\| \varphi_i(I_x) - \varphi_i(G(I_x)) \right\| \quad (8)$$

Among them, the φ_i represents the ith feature map taken from VGG16 network, I_x represents the input low illumination image, in the feature map, W is the width, H is the height and C is the channel.

Generator loss [9]. In the BA-GAN framework proposed in this paper, the training objectives of the generator are multi-faceted. It not only needs to enhance the image brightness but also ensure the realism of the results and the fidelity of the content. To achieve this complex balance, the optimization of the generator is not driven by a single objective but is guided by a composite loss function. This total loss function L_G is a weighted sum of three independent loss terms, and each term targets a specific aspect of the enhancement process.

$$L_G = \lambda_{adv} * L_{adv} + \lambda_{bright} * L_{bright} + \lambda_{percep} * L_{percep} \quad (9)$$

Among them, λ_{adv}, λ_{bright}, λ_{percep} are weights within the composite generator loss function. Through the synergistic effect of these three loss functions, the generator learns how to strike a balance among multiple objectives during the training process: making the image brighter and accurate, making it look real and natural, and at the same time retaining all the content of the original scene. The weights for the composite loss function were empirically set to λ_{adv}=1, λ_{bright}=1, and λ_{percep}=0.1 to balance the contributions of each component.

Discriminator loss [9]. After each iteration, calculate the loss function to get the difference between the fake and true images, then guide the next step of training. The specific method is to first initialize the model parameters with random values, input a low light image, obtain an enhanced image through a generation network, calculate the error between images using a discriminator, pass the error back along the direction of minimum gradient, modify the parameter values, repeat multiple times until the error value reaches a satisfactory value, and then stop iterating to obtain the final required model.

In the GANs framework proposed in this paper, the discriminator plays the role of an image connoisseur. The discriminator loss (L_D) is the core indicator for measuring the level of its discrimination ability. Its fundamental goal is to drive the discriminator network to more and more accurately distinguish between real normal-illuminated images and enhanced images forged by the generator.

The training process of the discriminator receives two types of inputs:

Real images I_{real}: Real, high-quality normal-illumination images from the LOL dataset.

Fake images I_{gen}: Enhanced images output by the U-Net generator after processing low-light images.

The goal of the discriminator is to output a probability value close to 1 for real images (judged as true) and a probability value close to 0 for fake images (judged as false). To achieve the above goals, the training of the discriminator aims to maximize its loss function L_D, which is defined in the paper as follows:

$$L_D = E[\log D(I_{real})] + E\left[\log\left(1 - D\left(I_{gen}\right)\right)\right] \quad (10)$$

In each training iteration, the model calculates the value of this loss function to measure the difference between the real image and the fake image. Then, this error is backpropagated along the gradient to update the network parameters of the discriminator itself.

Different from traditional discriminators that output a single true/false probability for the entire image, PatchGAN divides the input image into multiple N×N image patches (Patches) and makes a authenticity judgment for each patch, ultimately outputting an N×N matrix. The final judgment result is obtained by taking the average of this matrix. The advantage of this mechanism is that it forces the discriminator to focus on the local details and texture realism of the image, rather than just the global appearance, which helps to improve the overall quality of the generated images and accelerate the convergence of the network.

IV. Experiments

A. Dataset

The existing LOL-dataset [39] is used in this study. The LOL-dataset includes 485:15 pairs of images used for training and testing. The most core feature of the LOL-dataset is that it contains paired images. Each pair consists of an image taken under low-light conditions and its corresponding ground-truth image of the same scene taken under normal lighting conditions. This paired characteristic makes it very suitable for calculating the metrics of model performance. The test set of this dataset is used as a standard platform to objectively evaluate and compare the performance of the method in this paper with other representative algorithms. Below image pairs in Fig. 4 are some of the image samples from the LOL-dataset.



Fig. 4. The image samples from the LOL-dataset.

B. Experimental Setting

Hardware and platform: All training and testing were completed on a server equipped with an NVIDIA GeForce 4090 GPU, and the operating system was Ubuntu.

Software framework: Our model was implemented based on the PyTorch deep learning framework.

Training hyperparameters: The model was trained using the Adam optimizer. The initial learning rate was set to 1e-4, the batch size was set to 64, and the model was trained for a total of 3000 epochs.

Algorithm 1: Training of the proposed BA-GAN Algorithm

Input: Learning rate α, Gradient penalty coefficient λ; Adam optimizer parameters α, β_1, β_2; batch size m; low-light image x_i;

Initialize: discriminator parameters w_0; generator parameters θ_0

While (generator parameter θ not converged do) **do**

  For (each epoch) **do**

    Sample a batch of m low-light images $\{x^{(i)}\}_{i=1}^{m}$ from the dataset.

    Sample a batch of m real normal-light images $\{y^{(i)}\}_{i=1}^{m}$ from the dataset.

    Generate fake images: $\tilde{x}^{(i)} = G_\theta(x^{(i)})$.

    Calculate the discriminator loss L_D:

    $L_D \leftarrow E[\log D_w(y^{(i)})] + E[\log(1 - D_w(\tilde{x}^{(i)}))]$

    Update discriminator parameters via gradient ascent:

    $w \leftarrow w + \alpha \cdot \text{Adam}(\nabla_w L_D, w, \beta_1, \beta_2)$

  End for

  Sample a batch of m low-light images $\{x^{(i)}\}_{i=1}^{m}$ from the dataset.

  Calculate the composite generator loss L_G:

  $L_G \leftarrow \lambda_{adv} L_{adv} + \lambda_{bright} L_{bright} + \lambda_{percep} L_{percep}$

  Update generator parameters via gradient descent:

  $\theta \leftarrow \theta - \alpha \cdot \text{Adam}(\nabla_\theta L_G, \theta, \beta_1, \beta_2)$

End **Output**: The trained generator G_θ.

C. Evaluation Metrics

To more objectively assess the performance of the improved model in this paper, common image assessment algorithms were used to evaluate some representative methods. The evaluation metrics selected were Peak Signal-to-Noise Ratio (PSNR) [40], Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [41].

PSNR is a widely used metric for measuring image quality. It evaluates image quality based on the errors between corresponding pixels. This metric quantifies the degree of image distortion by calculating the ratio of the signal (the maximum possible pixel value of the image) to the noise (the mean squared error between the enhanced image and the real image). When evaluating the results, a higher PSNR value indicates better quality and less distortion of the enhanced image.

The calculation of PSNR is based on the Mean Square Error (MSE) [42]. MSE is calculated first, and then PSNR is calculated according to MSE. The calculation formula of MSE is:

$$MSE = \frac{1}{\text{nm}} \sum_{i=1}^{\text{m}} \sum_{j=1}^{\text{n}} \left(I_{(i,j)} - K_{(i,j)}\right)^2 \quad (11)$$

Where m and n are the number of rows and columns of the image, $I(i,j)$is position that the pixel value of the original image, $K(i,j)$ is the pixel value of the processed image at the same position. After obtaining MSE, PSNR can be calculated by the following formula:

$$PSNR = 10 \log_{10} \left(\frac{\text{MAX}_\text{I}^2}{\text{MSE}}\right) \quad (12)$$

Among them, MAX I is the maximum possible image pixel value. For 8-bit images, MAX I is usually 255.

The SSIM quantifies image similarity through three key components: luminance, contrast, and structural composition. Higher SSIM values (ranging from 0 to 1) indicate greater similarity between the enhanced and reference images.

$$SSIM(x,y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \quad (13)$$

X and y are the two images to be compared, μ_x And μ_y Average brightness of image x and y, ${\mu_x}^2$ And ${\mu_y}^2$ Is the variance of image x and y, μ_{x_y} is the covariance of image xx and y, and C1 and C2 are small constants added to avoid zero denominator.

LPIPS is a modern metric that measures perceptual similarity by simulating human visual evaluation. It uses a deep learning network to extract and compare the deep features of image patches, thereby determining the similarity between two images in the eyes of human observers. Different from the previous two metrics, a lower LPIPS value (closer to 0) represents higher perceptual fidelity between the two images, meaning they are more similar visually.

V. Results and Discussions

A. Comparsion with other Methods

To evaluate the effectiveness of our method, this paper compares it with several representative low-light enhancement algorithms on the LOL dataset, including HE [11], MSR [14], Retinex-Net [18], EnlightenGAN [25], KinD [19], Zero-DCE [31], and SCI [43]. Quantitative comparisons of these metrics across different methods are presented in Table I.

As shown in Table I, our proposed method achieves the highest SSIM score (0.7963) and the best PSNR (20.7127 dB) among all compared methods, indicating superior structural preservation and noise suppression capabilities. Moreover, our method attains a competitive LPIPS score (0.2271), demonstrating enhanced perceptual quality compared to most baseline approaches. These results validate the effectiveness of integrating U-Net, PatchGAN, and illumination attention for robust LLIE.

Fig. 5 presents qualitative comparisons of enhanced results on representative samples from the LOL dataset using different methods. Compared to EnlightenGAN, Zero-DCE, and SCI, the proposed method achieves superior visual quality in terms of brightness balance, detail preservation, and color fidelity.

Specifically, EnlightenGAN tends to produce over-saturated regions and amplifies noise in extremely dark areas. Zero-DCE improves brightness but suffers from local color distortion and residual noise, particularly in high-texture regions (e.g., books, patterns). SCI shows smoother brightness but lacks sharpness in edges and tends to flatten subtle textures.

In contrast, the proposed method generates enhanced images that are perceptually more natural and visually closer to the ground truth. The edges and structural details (e.g., in cabinets, text, and metal objects) are well preserved, and the illumination is more uniformly distributed across the image. Furthermore, color consistency is notably better, with reduced color cast and halo artifacts compared to the other approaches.

TABLE I Evaluation Results on LOL Dataset

Method	SSIM↑	PSNR↑	LPIPS↓
HE [11]	0.6433	19.9959	0.3551
MSR [14]	0.5327	11.4331	0.3505
Retinex-Net [18]	0.7903	18.9775	**0.2036**
EnlightenGAN [25]	0.7117	15.5693	0.2547
KinD [19]	0.6475	16.0338	0.2979
Zero-DCE [31]	0.6322	15.2961	0.2925
SCI [43]	0.6514	16.4265	0.3156
proposed	**0.7963**	**20.7127**	0.2271

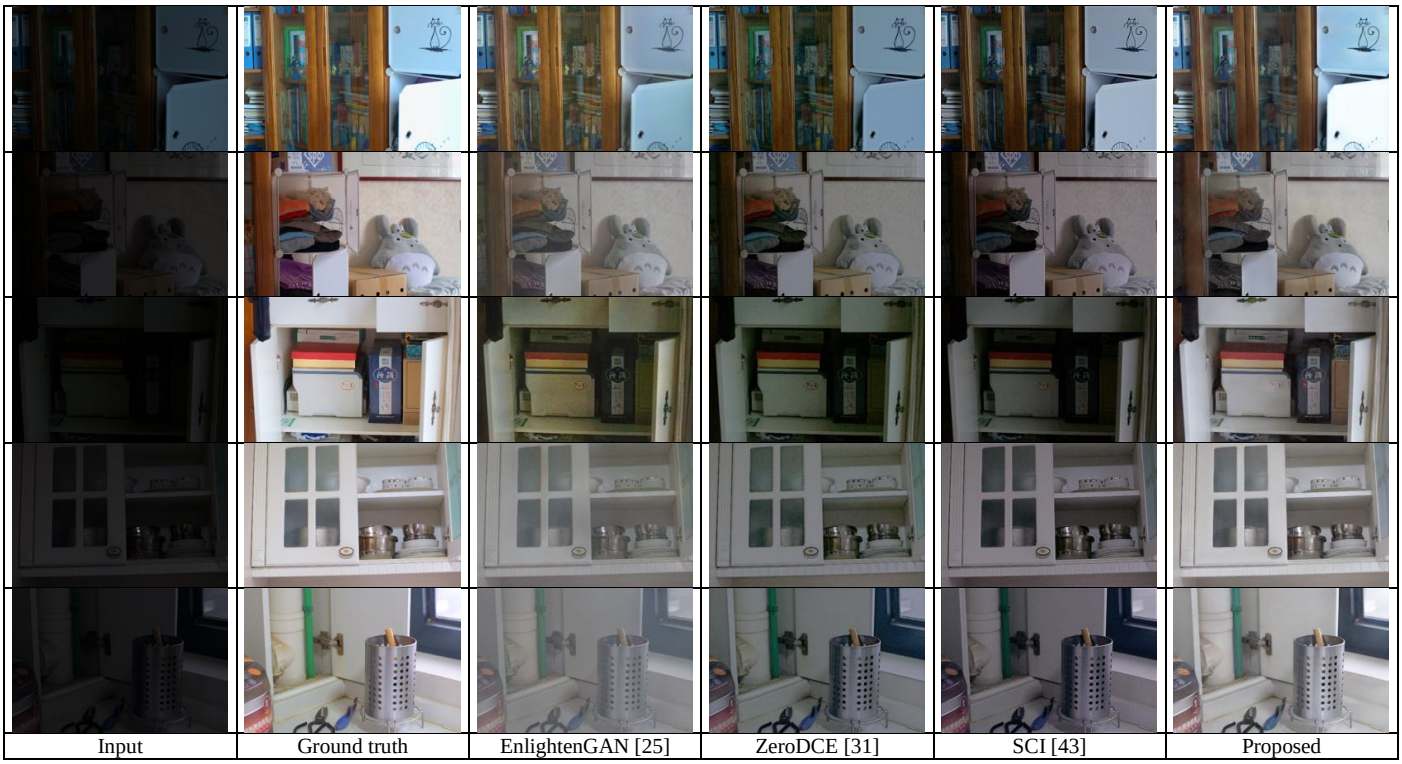


Fig. 5. The enhanced results from the LOL-dataset by different methods.

These results demonstrate the effectiveness of the proposed illumination attention mechanism and edge-aware optimization in guiding the network to enhance low-light images with improved structure, texture, and perceptual realism.

B. Ablation Study

To investigate the efficiency of each module in the proposed model to the final results, four experiments were designed. The following ablated Components were combined for training: U-Net as generator, PatchGAN as discriminator, and Attention module. Finally, the generated models were compared to determine the role of each module in the model. From the Table II, it can be observed:

1) Removing the illumination attention results in PSNR and SSIM decrease, and a deterioration in the LPIPS metric, indicating that illumination perception significantly enhances detail and realism.

2) After removing the PatchGAN discriminator, the image quality metrics slightly decline, suggesting that adversarial training also contributes to improving local details.

3) Without using U-net as the generator, the overall performance decreases, demonstrating that the U-net encoder-decoder structure plays a crucial role in feature modelling under complex illumination conditions.

TABLE II EVALUATION RESULTS FOR ABLATION STUDIES ON LOL-DATASET

Components	SSIM↑	PSNR↑	LPIPS ↓
Full model	**0.7963**	**20.7127**	**0.2271**
w/o attention	0.7012	16.4131	0.3422
w/o patchGAN	0.7054	18.0921	0.3012
w/o U-net	0.5023	12.1091	0.3547

VI. CONCLUSION

This paper addresses the existing challenges in the LLIE task, such as color distortion, noise amplification, and detail loss, and proposes a novel BA-GAN. This framework aims to generate enhanced images that are visually appealing and have high information fidelity through a well-designed end-to-end deep learning model.

The core contribution of this paper lies in the hybrid architecture design of BA-GAN. This paper adopts a U-Net-based generator, which effectively captures multi-scale context features and preserves fine image details through skip connections. The key innovation is that this paper integrates a novel illumination attention module into the decoder of the generator, enabling it to adaptively focus on the regions that require key illumination correction according to the image content. Meanwhile, this paper uses the PatchGAN discriminator architecture to ensure the local realism of the generated images. The entire model is optimized through a composite loss function that combines adversarial loss, brightness attention loss, and perceptual loss, thus achieving an effective balance among the realism, brightness accuracy, and structural fidelity of the images.

To verify the effectiveness of our method, this paper conducted a series of extensive experiments on the public LOL dataset. The quantitative evaluation results show that our BA-GAN outperforms a variety of current mainstream methods in terms of PSNR (20.7127) and SSIM (0.7963), demonstrating its superior ability in structure preservation and noise suppression. Qualitative visual comparisons further prove that our method can significantly improve the visibility of images while effectively suppressing noise and maintaining natural color features. Ablation experiments also verify the key contributions of each module this paper proposed to the final performance.

Although this study has achieved remarkable results, there are still some limitations and future exploration directions. First, as a supervised learning method, our model currently relies on paired training data, which is difficult to obtain in many real-world scenarios. Future work can explore combining the illumination attention mechanism with unsupervised or semi-supervised learning frameworks to enhance the model's generalization ability. Second, further research can be conducted to extend this method from static images to the field of low-light video enhancement and explore how to ensure the temporal consistency between video frames. Finally, future research can systematically evaluate the actual improvement effect of the images enhanced by BA-GAN on the performance of downstream computer vision tasks (such as object detection and image classification).

REFERENCES

[1] Y. P. Loh and C. S. Chan, "Unveiling contrast in darkness," in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 266–270. doi: 10.1109/ACPR.2015.7486507.

[2] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," Ieee Access, vol. 8, pp. 87884–87917, 2020.

[3] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. M. Farinella, "Computer vision for assistive technologies," Computer Vision and Image Understanding, vol. 154, pp. 1–15, 2017, doi: 10.1016/j.cviu.2016.09.001.

[4] C. Li et al., "Low-light image and video enhancement using deep learning: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 12, pp. 9396–9416, 2021, doi: 10.1109/TPAMI.2021.3126387.

[5] R. Al Sobbahi and J. Tekli, "Low-light homomorphic filtering network for integrating image enhancement and classification," Signal Processing: Image Communication, vol. 100, p. 116527, 2022, doi: 10.1016/j.image.2021.116527.

[6] R. C. Gonzalez, Digital image processing. Pearson education india, 2009.

[7] E. H. Land and J. J. McCann, "Lightness and retinex theory," Journal of the Optical society of America, vol. 61, no. 1, pp. 1–11, 1971, doi: 10.1364/JOSA.61.000001.

[8] Z. He et al., "Low-light image enhancement with multi-scale attention and frequency-domain optimization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2861–2875, 2023.

[9] I. J. Goodfellow et al., "Generative adversarial networks," Communications of the ACM, vol. 63, pp. 139–144, 2014.

[10] A. Madbouly, "Increase contrast of low light image using modified histogram equalization," Advances in Basic and Applied Sciences, vol. 3, no. 1, pp. 67–71, 2024, doi: 10.21608/abas.2024.327925.1052.

[11] B. S. Rao, "Dynamic histogram equalization for contrast enhancement for digital images," Applied Soft Computing, vol. 89, p. 106114, 2020, doi: 10.1016/j.asoc.2020.106114.

[12] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," IEEE transactions on image processing, vol. 6, no. 3, pp. 451–462, 1997.

[13] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," IEEE Transactions on Image processing, vol. 6, no. 7, pp. 965–976, 1997, doi: 10.1109/83.597272.

[14] H. Liu, X. Sun, H. Han, and W. Cao, "Low-light video image enhancement based on multiscale Retinex-like algorithm," 2016 Chinese Control and Decision Conference (CCDC), pp. 3712–3715, 2016, doi: 10.1109/CCDC.2016.7531629.

[15] S. Parthasarathy and P. Sankaran, "An automated multi scale retinex with color restoration for image enhancement," presented at the 2012 National Conference on Communications (NCC), IEEE, 2012, pp. 1–5.

[16] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," Pattern Recognition, vol. 61, pp. 650–662, 2017, doi: 10.1016/j.patcog.2016.06.008.

[17] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "Msr-net: Low-light image enhancement using deep convolutional network," arXiv preprint arXiv:1711.02488, 2017.

[18] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5901–5910.

[19] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 1632–1640.

[20] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond Brightening Low-light Images," International Journal of Computer Vision, vol. 129, pp. 1013–1037, 2021.

[21] L. Chang, G. Zhou, O. Soufan, and J. Xia, "miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology," Nucleic Acids Research, vol. 48, pp. W244–W251, 2020, doi: 10.1093/nar/gkaa467.

[22] Y. Liu, T. Huang, W. Dong, F. Wu, X. Li, and G. Shi, "Low-Light Image Enhancement with Multi-stage Residue Quantization and Brightness-aware Attention," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12106–12115, 2023, doi: 10.1109/ICCV51070.2023.01115.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," ArXiv, vol. abs/1505.04597, 2015, [Online]. Available: https://api.semanticscholar.org/CorpusID:3719281

[24] C. C. Lim, Y. P. Loh, and L. Wong, "LAU-Net: A low light image enhancer with attention and resizing mechanisms," Signal Process. Image Commun., vol. 115, p. 116971, 2023, doi: 10.1016/j.image.2023.116971.

[25] Y. Jiang et al., "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," IEEE Transactions on Image Processing, vol. 30, pp. 2340–2349, 2019, doi: 10.1109/TIP.2021.3051462.

[26] T. Ma et al., "RetinexGAN:Unsupervised Low-Light Enhancement With Two-Layer Convolutional Decomposition Networks," IEEE Access, vol. 9, pp. 56539–56550, 2021, doi: 10.1109/ACCESS.2021.3072331.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976, 2016.

[28] J. Hu, X. Guo, J. Chen, G. Liang, F. Deng, and T. L. Lam, "A Two-Stage Unsupervised Approach for Low Light Image Enhancement," IEEE Robotics and Automation Letters, vol. 6, pp. 8363–8370, 2020, doi: 10.1109/LRA.2020.3048667.

[29] R.-C. Wang, B. Jiang, C. Yang, Q. Li, and B. Zhang, "MAGAN: Unsupervised low-light image enhancement guided by mixed-attention," Big Data Min. Anal., vol. 5, pp. 110–119, 2022, doi: 10.26599/BDMA.2021.9020020.

[30] W. Kozlowski, M. Szachniewicz, M. Stypulkowski, and M. Zieba, "Dimma: Semi-Supervised Low-Light Image Enhancement with Adaptive Dimming," Entropy, vol. 26, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:264146400

[31] C. Guo et al., "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1777–1786, 2020, doi: 10.1109/CVPR42600.2020.00185.

[32] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, and Y. Zhou, "Zero-Shot Restoration of Underexposed Images via Robust Retinex Decomposition," 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, 2020.

[33] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10556–10565, 2020.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[35] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," ArXiv, vol. abs/1603.08155, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:980236

[36] G. Zhang, Q. Li, Q. Li, and T. Wu, "Low-light image enhancement model based on self-calibration illumination and color adjustment," 2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AHPCAI), pp. 559–562, 2022.

[37] D. D. Jonge, "LuminanceL1Loss: A loss function which measures percieved brightness and colour differences," ArXiv, vol. abs/2311.04614, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:265050838

[38] Y. Atoum, M. Ye, L. Ren, Y. Tai, and X. Liu, "Color-wise Attention Network for Low-light Image Enhancement," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2130–2139, 2019.

[39] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," arXiv preprint arXiv:1808.04560, 2018.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595, 2018, doi: 10.1109/CVPR.2018.00068.

[42] J. Guo, J. Ma, Á. F. García-Fernández, Y. Zhang, and H. Liang, "A survey on image enhancement for Low-light images," Heliyon, vol. 9, no. 4, 2023.

[43] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5637–5646.

Power-Aware Video Transmission in 5G Telemedicine: Challenges, Solutions, and Future Directions

Qiuhong SHI, Mingjing CAO*, Yun BAI
College of Physics and Information Engineering, Cangzhou Normal University, Cangzhou 061000, China

***Abstract*—The transformation of healthcare delivery through telemedicine has been significantly accelerated by the deployment of 5G networks and the integration of Internet of Things (IoT) technologies. These advancements enable real-time video-based medical services, including remote operations, urgent care, mobile procedures, and virtual consultations. However, the energy limitations of IoT devices present significant questions regarding long-term video quality and system efficiency. This survey reviews state-of-the-art solutions envisioned for enhancing video transmission in telemedicine environments that are limited by energy consumption but made possible by 5G connectivity. The study discusses recent advances, including efficient video compression techniques, computation offloading via edge computing, adaptive streaming procedures, and dynamic 5G architecture-aware resource scheduling, such as network slicing. It discusses the trade-off between power efficiency and video performance for various telehealth scenarios. Using a scenario-based analysis with a unifying integration framework, this work advances research into energy-efficient e-health systems.**

Keywords*—Telemedicine; Internet of Things; energy efficiency; video transmission

I. INTRODUCTION

The World Health Organization (WHO) recognizes telemedicine as a form of medical treatment delivered through audiovisual and digital communications [1]. Medical care services include diagnosis, consultation, treatment, health education, and the transmission of medical records. The telemedicine architecture can be divided into three layers: sensing, gateways, and healthcare facilities. The Internet of Things (IoT) is a network of people connected to everything, anytime, anywhere. IoT is defined differently in the literature. The IoT provides various healthcare services, including fitness programs, remote health tracking, rehabilitation, and elder care [2]. Telemedicine is essential for remotely monitoring treatments and medications at home [3]. The IoT architecture for telemedicine is based on a range of medical devices, sensors, and imaging devices. Therefore, IoT and telemedicine will contribute to the future comprehensive remote monitoring of the healthcare industry.

Telemedicine has revolutionized healthcare by allowing long-distance consultations, monitoring, and diagnosis through video-based interactions [4]. Along with telemedicine applications that continue to expand exponentially, trustworthy and energy-aware video transmission has become essential, particularly in applications that include battery-constrained IoT technologies, such as wearable health sensors and remote diagnostic tools [5]. These tend to have limited battery life and restricted processing capabilities, for which energy efficiency must be a key consideration. The emergence of 5G networks offers potential solutions to these issues through low latency, higher bandwidth, and highly reliable connectivity [6].

While there is an increasing amount of literature addressing energy efficiency in 5G and IoT systems, previous reviews tend to overgeneralize the solution without adapting it to the specific requirements of telemedicine applications. Most research fails to consider the particular issues related to continuous video transmission in cases such as telemedicine applications in remote surgery, emergency procedures, and long-term patient monitoring via battery-constrained IoT devices. Our review addresses this gap by providing a real-case-driven, holistic treatment of energy-efficient video transmission schemes specifically crafted for telemedicine. By integrating technical solutions with actual healthcare applications, this paper offers focused insights that are often lacking in more general, non-specialized surveys.

This article examines recent advances in energy-efficient video transmission technologies for telemedicine, focusing on the role of 5G and IoT devices. We examine the technical challenges, assess existing solutions, and outline future research directions to enhance energy efficiency in this area. This article makes the following key contributions:

- An in-depth investigation of cutting-edge video encoding and adaptive streaming strategies is provided, aiming to reduce energy consumption in telemedicine applications that utilize resource-constrained IoT devices.
- The role of 5G technology in enhancing video transmission quality and improving energy efficiency in telemedicine applications is demonstrated through network slicing, edge computing, and dynamic power control.
- An analysis of conflicts between Quality of Service (QoS) and energy efficiency is presented, addressing the interoperability issue and identifying the need for AI-driven solutions and energy-conscious protocols in the future development of telemedicine, coupled with IoT and 5G.

II. BACKGROUND AND MOTIVATION

This section outlines the requirements for telemedicine and video transmission, the contribution of 5G networks to the advancement of telemedicine, and the challenges posed by resource-limited IoT devices in healthcare.

A. Telemedicine and Video Transmission Requirements

Telemedicine applications rely on real-time video transmission for remote consultations, diagnosis, and patient monitoring [7]. Video transmission quality and efficiency are crucial for delivering seamless service to both patients and healthcare providers. This subsection outlines the minimum requirements for telemedicine video transmission, including video quality, latency, bandwidth, and energy efficiency, with consideration for resource-constrained IoT devices. Video quality is crucial for telemedicine, as doctors rely heavily on visual information to diagnose and monitor patients [8]. The quality of a video is usually measured by its bit rate, frame rate, and resolution, as shown in Table I.

Telemedicine requires real-time interaction between healthcare professionals and patients. High latency can negatively impact communication and diagnostic processes, as illustrated in Fig. 1, which shows the increasing latency associated with higher video resolutions [9]. In telemedicine, real-time applications require a latency of under 100 milliseconds. Providing high-quality video without interruptions during telemedicine transmissions requires sufficient bandwidth. As shown in Table II, the network type (e.g., Wi-Fi, 4G, 5G) and the number of concurrent users affect the available bandwidth. Typically, bandwidth requirements for telemedicine applications involving high-definition video range from 1 Mbps to 5 Mbps.

Energy efficiency is crucial, particularly when utilizing limited-life battery-powered IoT devices. Telemedicine is one of the most energy-intensive operations in video transmission, involving uninterrupted data encoding, transmission, and reception. Minimizing energy consumption is crucial for enhancing device longevity and maintaining device functionality [10]. Table III depicts energy consumption for different telemedicine video encoding formats. Energy consumption versus telemedicine video quality is shown in Fig. 2.

TABLE I. KEY VIDEO QUALITY PARAMETERS FOR TELEMEDICINE

Parameter	Definition	Typical value in telemedicine
Resolution	Number of pixels in each video frame	720p, 1080p, 4K (depending on use case)
Frame rate	Number of frames per second (fps)	30 fps (for real-time communication)
Bit rate	Amount of data transmitted per second	1-3 Mbps (depending on network conditions)

TABLE II. BANDWIDTH REQUIREMENTS FOR TELEMEDICINE VIDEO TRANSMISSION ACROSS DIFFERENT NETWORK TYPES

Network type	Bandwidth availability	Suitability for telemedicine
Wi-Fi (802.11n)	50-300 Mbps	High-definition video
4G LTE	5-12 Mbps	Basic telemedicine applications
5G (eMBB)	100-1000 Mbps	Ultra-high-definition video

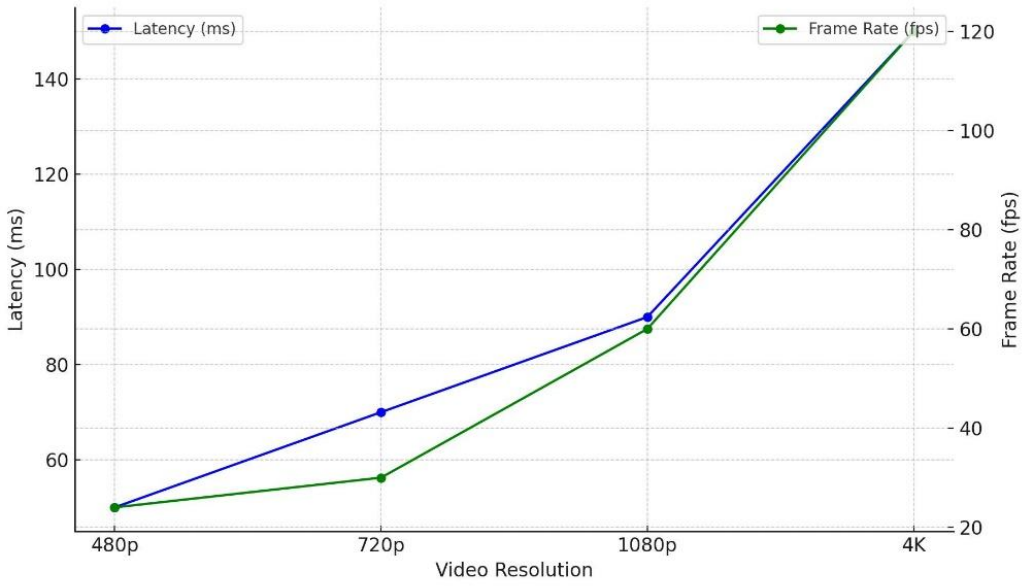


Fig. 1. Relationship between video quality and latency in telemedicine applications.

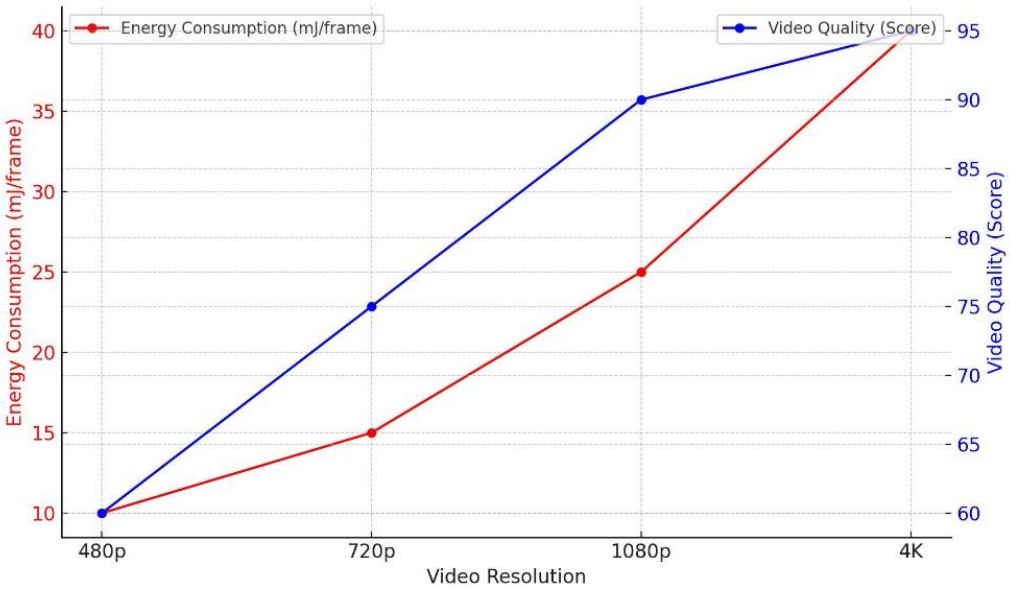


Fig. 2. Relationship between video quality and energy usage in telemedicine.

TABLE III. ENERGY CONSUMPTION OF DIFFERENT VIDEO ENCODING FORMATS IN TELEMEDICINE

Video encoding format	Energy consumption (mJ/frame)	Efficiency
H.264/AVC	15-30	Moderate
H.265/HEVC	8-20	High
VP9	10-25	Moderate

B. 5G Networks and Telemedicine

The 5G network is a transformative technology for telemedicine, offering features such as ultra-low latency, high-speed data transmission, and enhanced interoperability. These features facilitate real-time video transmission in telemedicine applications, facilitating secure and effective interaction between patients and healthcare professionals [11].

5G networks also possess various technologies that qualify them for telemedicine services, including Massive Machine-Type Communications (mMTC), Ultra-Reliable Low Latency Communication (URLLC), and Enhanced Mobile Broadband (eMBB) [12]. eMBB enables bandwidth-intensive applications such as streaming ultra-high-definition video. URLLC provides minimal delay in communication, enabling real-time discussion between patients and medical specialists. mMTC provides opportunities for numerous IoT devices in telemedicine scenarios that require multiple attached sensors, wearables, and remote diagnostics. As highlighted in Table IV, 5G offers specific benefits to telemedicine over previous generations of mobile networks.

High-resolution video streaming makes tele-diagnostics possible with great accuracy. High-bandwidth and low-latency 5G technology would allow real-time teleconsults without delay [13]. IoT-capable health devices, such as wearables and sensors, can collect real-time data on a patient's health status. The mMTC feature of the 5G network can enable the mass deployment of such devices, allowing healthcare professionals to receive real-time data with which to monitor patients. Augmented Reality (AR) technologies are integral to 5G's URLLC technology, especially in remote surgery, where precision and real-time feedback are crucial for telemedicine. The 5G network introduces sleep modes and energy-aware protocols to increase device longevity. Table V illustrates how 5G's energy-efficient protocols impact IoT devices in telemedicine.

TABLE IV. COMPARISON OF 5G WITH PREVIOUS MOBILE NETWORK GENERATIONS FOR TELEMEDICINE

Feature	3G	4G	5G	Impact on telemedicine
Latency	~100-200 ms	~30-50 ms	~1-10 ms	Enables real-time and low-latency video consultations
Bandwidth	~2 Mbps	~10-50 Mbps	~100 Mbps-1 Gbps	Supports high-definition video streaming for remote diagnosis
Device connectivity	Limited	Moderate	Massive (up to 1 million/km²)	Connects large numbers of IoT devices for continuous monitoring
Energy efficiency	Low	Moderate	High	Extends the battery life of IoT devices used in telemedicine
Reliability	Moderate	High	Ultra-reliable	Critical for life-saving remote interventions and consultations

TABLE V. IMPACT OF 5G ENERGY-EFFICIENT PROTOCOLS ON IOT DEVICES IN TELEMEDICINE

Technology	Energy efficiency mechanism	Impact on IoT devices
5G network slicing	Isolating critical applications into slices and optimizing resource use	Reduced energy consumption for telemedicine services
Energy-aware scheduling	Adjusting power consumption based on device activity	Extended battery life of wearable devices in telemedicine
Sleep modes	Permitting devices to sleep when not transmitting	Enhanced energy efficiency, especially in continuous monitoring devices

C. Resource-Constrained IoT Devices in Telemedicine

Telemedicine relies heavily on low-resource devices, such as wearables, portable diagnostic instruments, and sensors. These devices are typically limited in terms of processing power, memory, and battery life, posing significant challenges to energy efficiency and reliability [14]. Innovative approaches to power management, communication standards, and data transmission strategies are needed to ensure the efficient integration of these devices into telemedicine systems.

The lightweight and portable characteristics of IoT devices designed for telemedicine impose some limitations. Most IoT components operate on batteries, limiting their ability to monitor or transmit for long periods. These devices typically have low computing resources, making advanced data processing difficult. Data loss is a significant problem when storage capacity is limited, so using efficient compression and transmission strategies is essential. Table VI summarizes the general limitations of IoT devices used in telemedicine [15].

Resource-constrained IoT devices in telemedicine face several challenges. Video transmission, real-time monitoring, and data collection from IoT devices consume a significant amount of power [16]. Telemedicine environments often require frequent recharging or early device failure, affecting the continuity of care. Delays in IoT devices operating over low-power wireless communication standards, such as Zigbee and Bluetooth, can affect telemedicine services. Table VII compares energy consumption for standard IoT communication protocols.

TABLE VI. RESOURCE-CONSTRAINED IOT LIMITATIONS USED IN TELEMEDICINE

IoT device type	Battery life	Processing power	Memory size	Communication range
Wearable health monitor	1-3 days (continuous use)	Low (single-core MCU)	256 KB – 1 MB	10-30 meters (Bluetooth)
Portable diagnostic tool	6-12 hours (per charge)	Moderate (dual-core CPU)	4-8 GB	50-100 meters (Wi-Fi)
Implantable sensor	Several months (low use)	Very Low (minimal CPU)	64-128 KB	5-10 meters (NFC)

TABLE VII. AN OVERVIEW OF COMMON IOT COMMUNICATION PROTOCOLS IN TELEMEDICINE

Protocol	Typical range	Data rate	Power consumption	Use case in telemedicine
Bluetooth Low Energy (BLE)	10-30 meters	1 Mbps	Low	Wearable health monitors, fitness trackers
Zigbee	10-100 meters	250 Kbps	Very Low	Home health monitoring systems
Wi-Fi	50-100 meters	100-1000 Mbps	High	Video consultations, portable diagnostic tools
NFC	4-10 centimeters	< 500 Kbps	Minimal	Implantable medical sensors

III. ENERGY-EFFICIENT VIDEO TRANSMISSION STRATEGIES

This section discusses energy-efficient video transmission strategies, including video encoding techniques to reduce data sizes, adaptive bitrate streaming for dynamic quality adjustments, and edge computing with caching.

A. Video Encoding

In telemedicine applications, video encoding is crucial for reducing the size of video streams while maintaining quality. Resource-constrained IoT devices and 5G-enabled telemedicine systems require efficient video encoding, which impacts bandwidth usage, transmission latency, and energy consumption [17]. Video encoding reduces the amount of data required for transmission by compressing video files. This process aims to optimize network resources and minimize the load on IoT devices. Standard techniques include H.264/AVC (advanced video encoding), H.265/HEVC (high-efficiency video encoding), and VP9.

H.264 is a popular video compression standard because it has a high compression ratio and is compatible with a wide range of devices [18]. H.265 has better compression than H.264, reducing the bitrate required to transmit high-resolution video by approximately 50% [19]. Google developed VP9 as a royalty-free, open-source video codec that YouTube popularly uses for streaming videos. VP9 is similarly efficient as H.265 but is royalty-free [20]. Table VIII compares the performance and compression efficiency of these encoding standards.

TABLE VIII. PERFORMANCE AND COMPRESSION EFFICIENCY COMPARISON OF COMMON VIDEO ENCODING STANDARDS

Encoding standard	Compression ratio	Energy consumption (mJ/frame)	Video quality retention	Typical use in telemedicine
H.264/AVC	Moderate	15-30	High	Real-time video consultations
H.265/HEVC	High	8-20	Very high	High-definition video diagnosis
VP9	High	10-25	High	Low-latency video transmission

TABLE IX. ENERGY CONSUMPTION OF VIDEO ENCODING STANDARDS AT DIFFERENT RESOLUTIONS USED IN TELEMEDICINE

Resolution	H.264/AVC (mJ/frame)	H.265/HEVC (mJ/frame)	VP9 (mJ/frame)	VP9 (mJ/bit)
480p	10-15	5-8	7-10	~0.04
720p	15-20	8-12	10-15	~0.035
1080p	20-25	12-16	15-20	~0.03
4K	25-30	18-20	28-32	~0.025

Energy efficiency in video encoding is crucial for telemedicine applications, particularly in resource-constrained IoT devices [21]. Efficient video coding methods reduce the computing load on devices, minimize battery consumption, and reduce energy consumption for video transmission over networks. Compared to H.264, H.265 offers the same video quality at half the bit rate, significantly reducing the amount of data transferred. This leads to lower energy consumption. Although VP9 provides similar compression for H.265, it is particularly beneficial for applications that require open-source codecs. Although it uses slightly more power than H.265, it remains a viable option for energy-efficient video transmission. As shown in Table IX, different video coding standards in telemedicine consume different amounts of energy according to their resolution.

Video quality and energy consumption must be optimally balanced in telemedicine. In adaptive video encoding, bit rate, resolution, and frame rate are dynamically adjusted based on real-time network conditions and the energy availability of the transmitting device. This approach is ideally suited for telemedicine consultations and remote monitoring.

B. Adaptive Bitrate Streaming

Adaptive Bitrate Streaming (ABR) provides dynamic adjustments to video quality in response to real-time network conditions and device capabilities in telemedicine [22]. The technology ensures video streams are delivered with minimal buffering and optimal quality, even in environments with fluctuating bandwidth. Video quality and energy efficiency are essential in telemedicine, especially with resource-constrained IoT devices and wireless networks.

ABR encodes video content at multiple bitrates and resolutions. Client devices (e.g., smartphones or IoT devices) constantly monitor network conditions during video playback to ensure smooth streaming. By dynamically adjusting quality levels based on bandwidth availability, device performance, and battery life, this technology ensures uninterrupted transmission while minimizing energy consumption. Client-side monitoring, multiple encoded streams, and dynamic customization are crucial elements of ABR. Videos are encoded at different bit rates to accommodate various network speeds and device capacities (e.g., 360p, 720p, 1080p). The client device sets the video bitrate based on network bandwidth and battery life. Video quality is automatically adjusted according to real-time conditions to ensure smooth playback.

ABR offers several advantages in telemedicine applications, including reduced buffering, improved energy efficiency, and an optimized user experience [23]. A smooth and continuous video stream is ensured by dynamically adjusting the video quality to match the available bandwidth. IoT devices with resource constraints can conserve energy by reducing their bitrate and resolution when network conditions are poor, thereby preventing them from processing high-quality streams that can drain their batteries. Patients and healthcare providers can experience uninterrupted video consultations, even in low-bandwidth or fluctuating network conditions. Table X compares the different video bitrates applicable to telemedicine with their bandwidth requirements and energy consumption.

TABLE X. VIDEO BITRATES, BANDWIDTH, AND ENERGY CONSUMPTION IN TELEMEDICINE USING ABR

Resolution	Bitrate (Kbps)	Bandwidth requirement	Energy consumption (mJ/frame)	Use in telemedicine
360p	500-1000	Low	5-10	Basic consultations
720p	1000-3000	Medium	10-20	Standard remote diagnostics
1080p	3000-6000	High	20-25	High-quality video consultations
4K	6000+	Very High	25-30	High-definition telemedicine

C. Edge Computing and Caching

Telemedicine applications benefit significantly from edge computing and caching, two critical technologies in resource-constrained environments, such as rural areas and mobile health services. These emerging technologies present computing resources proximal to IoT devices and users, consequently entailing low latency, reduced bandwidth usage, and lower energy utilization [24].

With edge computing, data is handled nearer the point of origin (e.g., IoT devices) than from distant cloud servers. This localized processing reduces long-distance data transfers, which is particularly beneficial for telemedicine, where real-time response time is essential. Telemedicine relies heavily on edge computing for real-time video processing, data analysis, AI-driven diagnosis, and latency reduction. Encoding, decoding, and compressing video on the edge reduces processing loads on IoT devices while also conserving battery life. By locally processing medical data at the edge, edge-based AI diagnosis is performed earlier with reduced latency compared to cloud-based systems. Telemedicine services also require minimal latency, as processing data on network edges is crucial for emergency care and real-time consultations. Table XI shows the advantages of edge computing in telemedicine over conventional cloud computing.

TABLE XI. EDGE COMPUTING VS. CLOUD COMPUTING IN TELEMEDICINE

Attribute	Edge Computing	Cloud Computing
Latency	Low (data processed locally)	High (data processed in the distant cloud)
Bandwidth usage	Low (reduced data transmission)	High (frequent data transfer to the cloud)
Energy consumption	Low (offloads computation from devices)	High (IoT devices perform more work)
Real-time responsiveness	High	Moderate
Security	Exposed to edge breaches; AES-256 adds 5–10 mJ/frame	Centralized control; higher encryption scalability

Caching reduces latency and bandwidth usage by temporarily storing frequently accessed content closer to the user. Using edge caching in telemedicine facilitates faster access to medical videos, images, and data by storing them near healthcare providers and patients, thereby reducing the need to repeatedly transmit large files. There are two types of applied caching for telemedicine: Content Delivery Network (CDN) caching and dynamic caching of patient data. CDNs enable healthcare providers to access common healthcare content (e.g., training videos or diagnostic images) more quickly. For smooth interactions, patient data (e.g., diagnostic images or video streams) may be cached on the teleconsultation edge.

While edge computing and caching significantly reduce latency and power consumption, they introduce new security threats. Data processed or stored on edge nodes is more susceptible to unlawful access, physical attacks, or malicious firmware attacks due to its mobility or remoteness. For example, cached patient video streams or images might be vulnerable to data leaks in the event of poor encryption or authentication measures.

Encryption algorithms such as AES-256 provide robust protection for sensitive telemedicine data. However, they incur computational overhead that impacts both latency and energy consumption, particularly in real-time scenarios. Experimental benchmarks indicate that AES-256 encryption adds an average 5–7 ms delay per 1 MB of data and consumes 5–10 mJ/frame in video encoding pipelines, as shown in the updated Table XI. These impacts may be acceptable in high-performance edge nodes but are unsuitable for ultra-low-power IoT devices, such as implantable sensors communicating over NFC.

In such applications, lightweight ciphers like PRESENT, SPECK, or HIGHT become more suitable since they offer good security with minimal processing and energy overhead. PRESENT-80 is an example that has been demonstrated to execute below 0.1 mJ per encryption operation on 8-bit microcontrollers, making it a suitable choice for short-range secure telemetry in implantable telemedicine applications.

D. TinyML and Split Computing for Video Compression

The most up-to-date advances in lightweight AI computing and split computation provide unprecedented opportunities for telemedicine energy optimization. Two possible approaches, video compression with TinyML and split computing, extend wearable IoT sensor capabilities while maintaining minimal dependency on robust edge/cloud computing.

On-microcontroller machine learning with limited resources is achieved with TinyML. It is possible to implement lightweight pre-processing and compression of video locally with MCUNet-type models, effectively reducing upstream data sizes considerably. Low-bitrate data compression with minimal info loss or diagnostic features being extracted is achievable with TinyML compression training in wearable health monitor deployment scenarios in healthcare.

For instance, Lin, et al. [25] showed that compression using MCUNet-based compression on ARM Cortex-M processors achieved encoding energy consumption that was 30% less than H.265 while retaining decent accuracy on tasks for clinical classification. This saving is particularly significant in continual monitoring applications, where energy levels directly impact service continuity.

In split computing, the video compression task is partitioned between the IoT device and the edge server. The device performs basic pre-processing (e.g., frame sampling, region-of-interest extraction), while the more intensive encoding steps are offloaded to nearby edge nodes. This division reduces device-side computational load without significantly increasing transmission overhead. Let E_{total} denote total energy usage, split between local computation E_{local} and transmission E_{tx}:

$$E_{total} = E_{local} + E_{tx} \quad (1)$$

Split computing minimizes $E\ l\ o\ c\ a\ l$ E local by moving complex operations off-device and optimizes $E\ t\ x$ E tx via lightweight encoding on-site. Simulation studies show that combining TinyML with split computing can achieve up to 40% total energy savings in wearable diagnostic scenarios, compared to conventional full-device encoding with H.264 or H.265.

These emerging methods complement traditional approaches like ABR and H.265 by offering context-aware compression, real-time adaptability, and lower reliance on high-throughput edge infrastructure, making them ideal for deployment in rural or bandwidth-limited settings.

IV. NETWORK OPTIMIZATION FOR ENERGY EFFICIENCY

This section explores network optimization techniques for energy efficiency in telemedicine, including 5G network slicing, energy-efficient routing protocols, and power control with dynamic resource allocation.

A. Network Slicing in 5G

Network slicing, a key component of 5G technology, involves the creation of multiple virtual networks from a shared infrastructure, each customized for specific scenarios and service demands [26]. Telemedicine has numerous applications, from real-time video consultations to telesurgery and patient tracking, and network slicing grants each one its respective bandwidth, latency, and connectivity. The technology also

automatically boosts energy efficiency by changing network resources according to telemedicine service demands.

Fig. 3 shows the telemedicine slicing frame. 5G network slicing divides one physical network into multiple virtual slices, each with its own Service-Level Agreement (SLA). Users draw upon resources from that slice that align with their needs. Different slices can be set up for varying measures of performance, such as latency for virtual-real-time video consultations, bandwidth for high-definition video streaming, reliability for vital health data transmission, and energy efficiency for limited devices in the IoT. 5G network slicing optimizes network resources for efficient telemedicine operations without increasing energy consumption. The high-priority slice might be dedicated to remote surgery, which requires high reliability and ultra-low latency. In contrast, the low-priority slice might be allocated to energy-efficient health monitoring devices.

Intelligent health applications can use different network slices, as shown in Fig. 4. This industry requires specialized slices for virtual surgery, smart hospitals, and remote health monitoring. An individual slice can be created for applications requiring high throughput, low latency, and ultra-reliability when performing remote surgery. Remote health monitoring applications and wearable devices can provide secure communication. A dedicated slice can connect medical appliances to a central location.

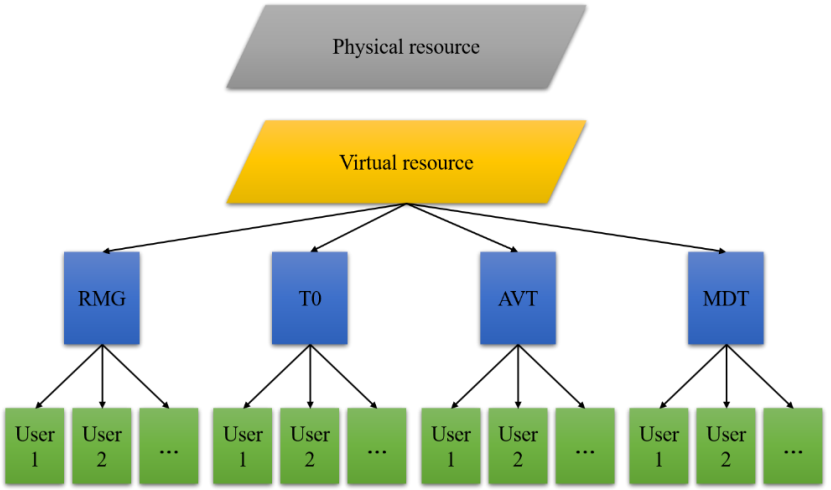


Fig. 3. 5G Network slicing framework for telemedicine.

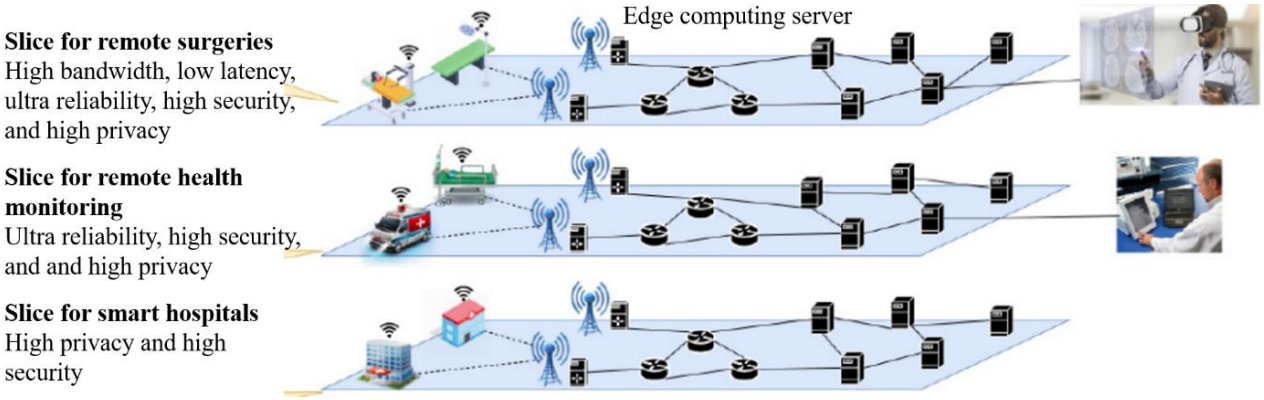


Fig. 4. Network slicing in 5G for telemedicine applications.

As shown in Table XII, network slices can be developed with different energy consumption levels according to application requirements. A high-power slice may be allotted for low-latency video consultations and telecommunication services with high bandwidth for standard real-time connectivity. A low-power slice is ideal for patient telemonitoring, as it reduces bandwidth and latency to achieve low power consumption. Implementing adaptive resource allocation automatically controls energy consumption by the network to optimize utilization, resulting in a simultaneous reduction in energy expenditures for telemedicine services.

B. *Energy-Efficient Routing Protocols*

Using energy-efficient routing protocols is essential in telemedicine networks due to the resource constraints imposed on IoT devices. Sensors and wearables typically have limited battery life and require optimization for efficient communication to forward information. Energy efficiency reduces the energy usage of 5G-enabled telemedicine networks while ensuring high levels of QoS. The resulting routing protocols aim to minimize hop counts, shorten transmission distances, or select low-energy routes while meeting the latency and bandwidth requirements of telemedicine. Table XIII summarizes fundamental routing techniques.

TABLE XII. ENERGY EFFICIENCY IN 5G NETWORK SLICING FOR TELEMEDICINE

Telemedicine application	Required network characteristics	Energy consumption (Average per session)
Real-time video consultations	Low latency and high bandwidth	High (15-20 mJ/frame)
Remote surgery assistance	Ultra-low latency and high reliability	Very high (25-30 mJ/frame)
Remote patient monitoring	Low bandwidth and moderate latency	Low (5-8 mJ/frame)

TABLE XIII. AN OVERVIEW OF ENERGY-EFFICIENT ROUTING PROTOCOLS FOR TELEMEDICINE

Routing protocol	Key feature	Energy efficiency	Use case in telemedicine
Multi-hop	Distributing data across multiple nodes	High (conserves energy for low-power devices)	Continuous patient monitoring
Opportunistic	Selecting paths based on real-time conditions	Moderate to high	Real-time video consultations
Energy-aware geographic	Using node location to select efficient routes	High	Long-term remote health monitoring with sensors

Multi-hop routing is also a common energy-conserving tactic in telemedicine networks, particularly when devices are dispersed over a large geographic area. Multi-hop routing utilizes intermediate nodes to relay data, rather than directly forwarding it to the destination node. Thus, singular IoT devices consume less energy by spreading communication burdens across multiple nodes. For instance, if a telemedicine system comprises numerous sensors that monitor patients, data from remote sensors might be relayed via short-range devices to a server. In this manner, the remote sensor is not subject to as much direct energy stress.

Optimistic routing makes accommodations according to varying network conditions, including bandwidth, node battery life, and traffic. This approach is best suited for applications with real-time requirements, such as teleconsultations, where providing QoS with minimal energy consumption is necessary. In telemedicine applications, wearable sensor devices can be replaced with high-power devices when they run out of battery. Opportunistic routing also accommodates network topology changes, such as when devices move to new locations or become temporarily disconnected.

Energy-aware geographical routing selects data transmission routes with minimum energy consumption based on the nodes' geographical positions. It is also helpful for telemedicine applications with geographically distributed units, such as rural health tracking or mobile health services. Geospatial routing tends to minimize energy usage by selecting nodes for data exchange with minimum distances. For instance, a network in a remote patient surveillance system would select a device closer to the provider for data transmission with minimal energy usage.

C. Power Control and Resource Allocation

Power control and resource allocation greatly enhance energy efficiency for telemedicine networks, particularly in 5G networks, whereby several mobile applications and devices must cooperate. As demonstrated in Table XIV, these technologies ensure that network capacities, i.e., power, bandwidth, and spectrum, are allocated to achieve performance with energy efficiency. In telemedicine applications, where continuity of service and long-term utilization of IoT devices are most significant, effective power control and resource allocation can considerably decrease energy consumption without any trade-off in QoS.

Power control systems dynamically regulate device transmission power according to real-time network conditions, energy availability, and QoS requirements. In telemedicine, where equipment such as wearable sensors and mobile diagnostic instruments has limited battery capacity, power management prolongs their operational lifespan by minimizing excessive energy usage.

Battery-aware power control and context-aware power control are the most prevalent power control technologies. Devices regulate broadcast power according to their proximity to a base station or edge server. Devices close to a base station or edge server minimize power consumption, while devices further away utilize power as needed. Devices also monitor battery levels in various ways and adjust transmission power accordingly. For instance, if battery levels hit a low point, a device minimizes power consumption to conserve power while transmitting vital data. This approach varies power depending on the use environment (e.g., essential telemedicine scenarios such as tele-surgery would focus more on power than monitor applications).

In telemedicine services, 5G networks must partition bandwidth, spectrum, and computing capabilities among multiple applications and devices. Flexible partitioning of these capabilities enables higher-priority applications, such as real-time teleconsultation via video or teleoperation, to be allocated ample resources while minimizing power consumption for lower-priority applications, such as background health tracking.

Power control and resource allocation significantly enhance energy efficiency in telemedicine networks, particularly in 5G networks, where multiple mobile applications and devices must cooperate. As demonstrated in Table XIV, these technologies ensure that network capacities, i.e., power, bandwidth, and spectrum, are allocated to achieve performance with energy efficiency. In telemedicine applications, where continuity of service and long-term utilization of IoT devices are most significant, effective power control and resource allocation can considerably decrease energy consumption without any trade-off in QoS.

TABLE XIV. POWER CONTROL AND RESOURCE ALLOCATION STRATEGIES FOR ENERGY EFFICIENCY IN 5G TELEMEDICINE NETWORKS

Category	Strategy	Description	Energy impact	Telemedicine use case
Power control	Adaptive power control	Adjusting transmission power based on distance	High (saves energy in nearby devices)	Wearable health monitoring
	Battery-aware power control	Reducing power when the battery is low	Moderate (extends device battery)	Remote diagnostic tools
	Context-aware power control	Prioritizing power based on application type	Low to high (application-dependent)	Real-time video consultations and surgery
Resource allocation	Dynamic bandwidth allocation	Modifying bandwidth based on application needs	Moderate (balances QoS and energy)	Video consultations and health monitoring
	Priority-based resource allocation	Allocating resources based on application priority	High for critical applications	Remote surgeries and emergency care
	Energy-aware resource scheduling	Scheduling tasks based on device energy levels	High energy savings	Remote monitoring and health data analytics
Hybrid	Real-time video consultations	High transmission power for low latency	High (to ensure uninterrupted service)	High bandwidth allocation
	Remote patient monitoring	Reducing transmission power	Low	Low bandwidth allocation
	Remote surgery assistance	High transmission power for reliability	Very high (critical application)	High bandwidth and processing power

Power control and resource allocation methodologies employed to enhance energy efficiency couple power control with adaptive resource allocation algorithms, allowing network resources to be utilized to their maximum while minimizing energy wastage. When network demands decrease, power levels can be reduced, along with bandwidth reallocation to lower-power sections.

V. RESULTS AND DISCUSSION

This review comprehensively evaluated existing approaches for satisfying energy-efficient video transmission in 5G-capable telemedicine systems. It was concluded that there are many converging trends and meaningful challenges between video encoding, adaptive streaming, edge computing, and network optimization. Specifically, it was concluded that employing high-efficiency video codes, such as H.265/HEVC and VP9, significantly reduces transmission energy without compromising diagnostic-grade video quality, a crucial requirement for clinical telemedicine applications. Adaptive bitrate streaming was also invaluable for balancing a trade-off between video quality and battery life under fluctuating network conditions and constrained IoT scenarios.

In addition, edge computing and caching capabilities help reduce latency while offloading processing capabilities from energy-limited IoT devices, further extending device lifespans and enhancing end-user application responsiveness for real-time applications, such as teleconsultation and telediagnostics. From a network perspective, 5G functions, including network slices and dynamic resource allocation, help provide end-users with prioritized, customized performance assurance and minimal energy usage.

To verify investigated techniques, experimental works published in the literature provide quantitative measures of codec efficiency and network behavior. For example, experimental investigations have shown that H.265/HEVC reduces the bitrate by up to 50% compared to H.264/AVC for perceptually equal video quality, with a concomitant reduction of up to 40% in energy consumption in battery-powered devices streaming 1080p streams at 30 frames per second over limited networks. VP9 demonstrates similar compression efficiency, though with slightly higher computational complexity, resulting in trade-offs between encoding latency and device power usage. Simulation studies using telemedicine scenarios further show that H.265 achieves a 25–30% reduction in total energy consumption per frame when applied with adaptive bitrate streaming strategies on wearable diagnostic devices.

Additionally, trials for 5G network-slicing-based remote-surgery applications demonstrated end-to-end latencies repeatedly under 10 milliseconds with an energy-aware edge- and core network scheduling. In these trials of remote-surgery applications in managed hospital-to-hospital networks with URLLC slices, there exists not only ultra-low latency potential but also a device-side reduction in energy consumption by a factor of 20 through localized video pre-processing. Parameters like these provide robust real-world validation for feasibility and the promise of these technologies for life-critical clinical applications.

With further advancements in telemedicine, telemedicine systems will require energy-aware transmission technologies for 5G networks and IoT applications. Future research in this area could extend along a range of avenues to enhance the effectiveness, scalability, and energy efficiency of telemedicine systems, while ensuring the security and accessibility of healthcare services. The following research avenues will facilitate this field of study.

Utilizing Artificial Intelligence (AI) and machine learning is one potential area for study to enhance video streaming and encoding. ABR streaming supports active decisions in real-time based on network conditions, device battery levels, and user behavior, with a focus on balancing the trade-off between energy usage and video quality. AI models can examine network utilization patterns and device usage patterns to forecast bandwidth changes or battery consumption, enabling the active modification of encoding parameters. AI-focused solutions can significantly improve energy efficiency for telemedicine applications that are severely limited in resources, such as wearable health sensors or transportable diagnostic equipment.

Future studies should further examine the integration of telemedicine with edge computing for video transmission. Edge

server-based computation of video data near IoT devices reduces the need to transport high-definition video over long distances, thereby conserving bandwidth and power. Research could also utilize video compression, edge-level transcoding, and real-time processing to alleviate computational burdens on resource-poor devices. This would enable battery-powered mobile health units to support high-definition video consultations via telemedicine. Furthermore, integrating technology between edge computing and 5G ultra-reliable low-latency communications may further enhance telemedicine for long-distance interactions.

Blockchain technology can potentially enhance resource allocation in 5G telemedicine networks. The decentralized ledger of blockchain guarantees transparent, equitable, and efficient allocation of bandwidth and power resources across various telemedicine applications. Future research may focus on designing blockchain-based smart contracts that automate the distribution of network resources for video transmission, taking into account device energy levels, network congestion, and QoS requirements. This approach would allow a fairer and more energy-efficient use of 5G infrastructure in telemedicine settings, particularly when managing multiple simultaneous users and applications.

Despite existing compression standards, such as H.265/HEVC and VP9, that provide significant video efficiency improvements, further research on future-generation video compression technologies is needed for telemedicine. Quantum computing-based video compression and neural network-controlled codec-based compression may offer larger compression ratios without compromising video quality, resulting in significantly reduced bandwidth and energy usage. This technology will be most effective for telemedicine applications that demand the continuous transmission of high-resolution video images, such as tele-operations and tele-diagnostics, through real-time video.

Investigating battery-conscious transmission methods for videos can promote energy efficiency by enabling telemedicine networks to adjust bitrate and video quality while preserving battery life in IoT devices. Corresponding protocols would allow IoT devices to consume less power during non-essential times or with depleted batteries while allocating more power for vital occasions or emergencies. Integrating power reduction algorithms in future telemedicine networks can promote wearable device and sensor lifespan while minimizing recharging or battery replacement.

Forthcoming research must prepare for 6G and subsequent technologies. 6G will prioritize green communications, highlighting sustainability and energy efficiency. Researchers can also study the implementations of 6G's ultra-low power transmission techniques, quantum telecommunication, and terahertz radiation in telemedicine networks to further reduce energy usage. Reviews of early work on integrating 6G technology with telemedicine services promise to steer future innovation toward energy-aware healthcare provision.

Machine-learning advancements permit us to achieve predictive ABR streaming, which forecasts bandwidth changes before they occur and provides adaptive adjustments in encoding parameters. Time-series network throughput predictions benefit most from deep learning models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). LSTM-based ABR controllers, for example, achieve up to 90% accuracy in future bandwidth-window predictions, allowing devices to prefetch video segments with optimal bitrate pre-buffering, resulting in reduced re-encoding energy consumption and playback stalling. They can be trained from historical network data and installed on edge nodes or IoT gateways with minimal computation overhead.

VI. Conclusion

5G technology and energy-constrained IoT devices have great potential to convert telemedicine into telemedicine services with real-time, high-grade services. However, one of the most significant concerns is the deployment of energy-aware video transmission, ensuring that IoT devices remain active for extended periods without compromising QoS. This paper presents a detailed review of various energy-aware methodologies, including video encoding schemes, adaptive bitrate streaming, edge computing, caching, network slicing, and power control policies, which play a critical role in reducing energy consumption for telemedicine applications.

By employing sophisticated compression technologies such as H.265/HEVC for videos, along with adaptive bitrate streaming and edge computing, telemedicine deployments can effectively reduce power usage for transmission while maintaining the high definition necessary for accurate diagnosis and consultations. 5G's network slice capabilities also provide for flexible network resource allocation so that mission-critical applications such as tele-surgeries receive whatever bandwidth and low-latency services they require while day-to-day monitoring operations consume little power.

Power control and resource allocation methodologies, such as adaptive power control and battery-aware routing, improve the energy efficiency of telemedicine networks. These methodologies enable IoT devices to communicate with reduced energy consumption levels, thereby extending device lifespan with minimal frequent recharging. Furthermore, future advancements in AI-based video encoding, blockchain-based resource allocation, and emerging telecommunication technologies such as 6G offer further potential for improvements in telemedicine energy efficiency.

References

[1] G. Sageena, M. Sharma, and A. Kapur, "Evolution of smart healthcare: Telemedicine during COVID-19 pandemic," Journal of The Institution of Engineers (India): Series B, pp. 1-6, 2021.

[2] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, pp. 1-21, 2019.

[3] D. Kumari, S. Sharma, M. Chawla, and S. Panda, "A Manifesto for Healthcare Based Blockchain: Research Directions for the Future Generation," Journal of The Institution of Engineers (India): Series B, pp. 1-22, 2024.

[4] E. H. Golembiewski et al., "Video-based observation research: A systematic review of studies in outpatient health care settings," Patient Education and Counseling, vol. 106, pp. 42-67, 2023.

[5] F. Kamalov, B. Pourghebleh, M. Gheisari, Y. Liu, and S. Moussa, "Internet of medical things privacy and security: Challenges, solutions, and future trends from a new perspective," Sustainability, vol. 15, no. 4, p. 3317, 2023.

[6] M. M. Nasralla, S. B. A. Khattak, I. Ur Rehman, and M. Iqbal, "Exploring the role of 6G technology in enhancing quality of experience for m-health multimedia applications: A comprehensive survey," Sensors, vol. 23, no. 13, p. 5882, 2023.

[7] S. A. Butt, A. Ali, A. Khalid, Q. Shaheen, P. Ariza-Colpas, and M. Shiraz, "Patients using real-time remote health monitoring applications: a review," Computational Intelligence for Medical Internet of Things (MIoT) Applications, pp. 93-120, 2023.

[8] N. Pegoraro et al., "Telemedicine in Sports under Extreme Conditions: Data Transmission, Remote Medical Consultations, and Diagnostic Imaging," International journal of environmental research and public health, vol. 20, no. 14, p. 6371, 2023.

[9] K. C. Maita et al., "Imaging evaluated remotely through telemedicine as a reliable alternative for accurate diagnosis: a systematic review," Health and Technology, vol. 13, no. 3, pp. 347-364, 2023.

[10] K. Khujamatov, N. Akhmedov, E. Reypnazarov, D. Khasanov, and A. Lazarev, "Device-to-device and millimeter waves communication for 5G healthcare informatics," in Blockchain Applications for Healthcare Informatics: Elsevier, 2022, pp. 181-211.

[11] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "5G technology for healthcare: Features, serviceable pillars, and applications," Intelligent Pharmacy, vol. 1, no. 1, pp. 2-10, 2023.

[12] A. M. Peralta-Ochoa, P. A. Chaca-Asmal, L. F. Guerrero-Vásquez, J. O. Ordoñez-Ordoñez, and E. J. Coronel-González, "Smart healthcare applications over 5G networks: A systematic review," Applied Sciences, vol. 13, no. 3, p. 1469, 2023.

[13] A. Sacco, F. Esposito, G. Marchetto, G. Kolar, and K. Schwetye, "On edge computing for remote pathology consultations and computations," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 9, pp. 2523-2534, 2020.

[14] D. K. Israni and N. S. Chawla, "Human - Machine Interaction in Leveraging the Concept of Telemedicine," Human - Machine Interface: Making Healthcare Digital, pp. 211-245, 2023.

[15] B. Pourghebleh, K. Wakil, and N. J. Navimipour, "A comprehensive study on the trust management techniques in the Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9326-9337, 2019.

[16] B. Pourghebleh, N. Hekmati, Z. Davoudnia, and M. Sadeghi, "A roadmap towards energy - efficient data fusion methods in the Internet of Things," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, p. e6959, 2022.

[17] S. Islam, A. K. Budati, M. K. Hasan, S. Goyal, and A. Khanna, "Performance analysis of video data transmission for telemedicine applications with 5G enabled Internet of Things," Computers and Electrical Engineering, vol. 108, p. 108712, 2023.

[18] S. Bouaafia, R. Khemiri, S. Messaoud, O. Ben Ahmed, and F. E. Sayadi, "Deep learning-based video quality enhancement for the new versatile video coding," Neural Computing and Applications, vol. 34, no. 17, pp. 14135-14149, 2022.

[19] Z.-N. Li, M. S. Drew, J. Liu, Z.-N. Li, M. S. Drew, and J. Liu, "Modern video coding standards: H. 264, H. 265, and H. 266," Fundamentals of Multimedia, pp. 423-478, 2021.

[20] O. Nawaz, M. Fiedler, and S. Khatibi, "QoE-Based Performance Comparison of AVC, HEVC, and VP9 on Mobile Devices with Additional Influencing Factors," Electronics, vol. 13, no. 2, p. 329, 2024.

[21] A. Rana, C. Chakraborty, S. Sharma, S. Dhawan, S. K. Pani, and I. Ashraf, "Internet of medical things-based secure and energy-efficient framework for health care," Big Data, vol. 10, no. 1, pp. 18-33, 2022.

[22] D. Zhang, L. Wei, K. Shen, H. Zhu, D. Wang, and F. Wang, "TrimStream: Adaptive Real-time Video Streaming through Intelligent Frame Retrospection in Adverse Network Conditions," IEEE Transactions on Mobile Computing, 2024.

[23] T. Kimura, T. Kimura, A. Matsumoto, and K. Yamagishi, "Balancing quality of experience and traffic volume in adaptive bitrate streaming," IEEE Access, vol. 9, pp. 15530-15547, 2021.

[24] D. Alekseeva, A. Ometov, O. Arponen, and E. S. Lohan, "The future of computing paradigms for medical and emergency applications," Computer Science Review, vol. 45, p. 100494, 2022.

[25] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, and S. Han, "Tiny machine learning: Progress and futures [feature]," IEEE Circuits and Systems Magazine, vol. 23, no. 3, pp. 8-34, 2023.

[26] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 957-994, 2021.

Evaluating Intangible Software Quality Metrics for Effective Project Management Information Systems

Gu Xin[1], Rozi Nor Haizan Nor[2], Nur Ilyana Ismarau Tajuddin[3], Khairi Azhar Aziz[4]
Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology
Universiti Putra Malaysia[1, 2]
Tamhidi Centre, Universiti Sains Islam Malaysia[3]
IT Enterprise Architecture, IT Strategy, Bank Kerjasama Rakyat, Malaysia[4]

***Abstract*—In modern organizational environments, project management information systems (PMIS) play an important role in ensuring project success with the user requirements, keeping overall costs within the planned budget, and delivering projects at the agreed time. Selecting a high-quality PMIS is vital for the success of project management. A software quality model tailored to PMIS, summarizing the intangible software quality metrics (ISQM) that are effective in evaluating a PMIS, can help better decision-making on PMIS for project managers. However, there is limited research on the PMIS-tailored quality model. To fill the gap, this study evaluates effective SQM for PMIS quality assessment. There are two types of PMIS: Web-based PMIS and PMIS software applications. To narrow the context of PMIS, we focus on web-based PMIS since they are widely used across the industry, such as Microsoft Project and Jira. According to the PMIS features, we merely explored the tailored quality models that have been proven to be more appropriate for web-based PMIS rather than the basic models, such as ISO/IEC 9126, ISO/IEC 25010, and Bertoa. This research uses a qualitative approach to conduct the commonality screening among these models and find out the key evaluation metrics, such as usability and functionality, and the corresponding qualitative attributes suitable for Web-based PMIS quality assessment. The selected metrics and attributes are used to form a Web-based PMIS-tailored quality model. The scoring mechanism is introduced based on the PMIS-tailored quality model, where project managers can have a clear comparison among different web-based PMISs, leading to effective web-based PMIS selection for project management.**

Keywords*—*Web-based PMIS; software quality model; intangible software quality metrics; software quality

I. Introduction

In traditional project management (PM), most projects are managed by using a mass of manual documents. A successful project requires complete documentation to execute all project life cycles and make decisions. In terms of this issue, the project team is often faced with the challenge of whether to do documentation or do the project. Thus, some project managers of information systems and information technology projects choose to use the PMIS to help their project management [1].

Project Management Institute (PMI) in their PMBOK GUIDE defines a PMIS as software that can help strategize, organize, and manage resource streams and develop resource approximations [2]. PMIS allows for real-time tracking of project progress, which helps in identifying and mitigating risks promptly. In addition, for large and complex projects, some PMIS like MS Project or Primavera support comprehensive project methodologies and require skilled team members, thus ensuring better management of complex project requirements [3].

The project systems can be grouped into two types [4]:

- Web-based project management systems.
- Project management systems software applications.

Different types of project management systems are utilized to manage projects. They are unique in operation, depending on the kind of project one is managing.

This study is structured systematically to address the research gap. The subsequent sections will first detail the problem statement, research questions, and objectives that form the foundation of this study. Following this, the research methodology is presented, outlining the literature review process and the analysis of existing tailored software quality models. The study then describes the development of the proposed Web-based PMIS-tailored quality model, including the selection of metrics and the calculation of weights based on an expert survey. Finally, the results are presented and discussed, followed by the study's limitations and a conclusion that summarizes the key findings and suggests directions for future research.

II. Problem Statement

In project management processes, the implementation of PMIS implies several changes in the organization's routines. For example, the PMIS was a new software tool in the organization, and therefore, there were also difficulties during the adaptation of the users to the new tool [5]. This reveals the necessity of quality metrics such as usability in PMIS. The successful implementation of any software product relies on the quality of the software. Many existing models can be used to evaluate software quality.

The latest software quality models have undergone standardization processes, such as ISO standards 9126 and 25010 [6]. Each model focuses on a specific metric, and one of the most important factors influencing software quality measurement is selecting the most relevant metric. However, existing studies have not conducted an effective assessment of quality metrics for PMIS. To ensure the success of PM, it is crucial to have better control over the quality metric of PMIS

since "you cannot control that which you cannot measure" [7].

III. Research Questions

The main objective of this study is to analyze the software quality models and ISQM and establish a suitable model for PMIS. The research objective has been achieved by solving these research questions (see Table I):

TABLE I Research Questions

RQ1	What are the commonly used software quality models?
RQ2	What ISQM are involved in these existing models?
RQ3	What is a suitable software quality model for web-based PMIS?
RQ4	What model can be developed for PMIS through modifications or enhancements to the existing quality models?

IV. Objectives

This study aims to identify the effective ISQM for Web-based PMIS. Our approaches include exploring 14 types of existing quality models to find the most relevant metrics for web-based PMIS and evaluating the selected metrics, which will be formulated as a quality model specifically for assessing the ISQM aspects of web-based PMIS. A tool is developed based on the created quality model and identified ISQM to assess the web-based PMIS. Such a tool provides project managers with actionable insights and practical guidelines for selecting web-based PMIS that align with their specific project goals, enabling organizations to enhance project efficiency and success rates.

V. Scope Of Study

1) The software quality model under research is developed based on tailored software quality models.

2) This research adopts a qualitative approach that focuses on selecting qualitative attributes for each ISQM.

3) To narrow the context of PMIS, we focus on web-based PMIS since they are widely used across the industry, such as Microsoft Project and Jira.

VI. Materials And Methods

A. Literature Review

1) Software quality models. Numerous Quality models, such as McCall's Quality Model, Bowen, GHezzi, and ISO/IEC 25010 Quality Model, are used for software quality assessment. These models have been experiencing the standardized process to formulate general metrics for software products like usability, reliability, functionality, and maintainability [7]. For instance, software quality is divided into eight major metrics in the ISO 25010 Quality Model, including functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability. These metrics are further subdivided to assess general quality attributes. For example, usability embodies the influence of user adaptability, while reliability measures the system's ability to perform under operational demands [8].

2) Project management information system. PMIS is popular as an essential tool for modern project management. The outlook of project management systems is rapidly evolving, influenced by technological advancements, changing work environments, and emerging methodologies [6].

With the widespread use of PMIS, software quality assurance has become an important aspect of ensuring the efficiency, reliability, and effectiveness of software systems in project management. The quality of software is a necessary factor in the success of project management systems, which can influence user satisfaction, maintenance costs, and overall project success.

B. Related Work

While general software quality models like ISO/IEC 25010 exist, applying them directly to specialized systems like PMIS is challenging [6]. A review of previous studies reveals significant limitations, as shown in Table II. For example, past research has been critiqued for using inadequate models for ubiquitous applications [9], offering limited scope for domains like learning management systems [12], and lacking sufficient attribute coverage and applicability in healthcare systems [13]. A key issue is the failure of older models like ISO/IEC 9126 to include critical modern metrics such as reusability, which is vital for component-based software [11]. This highlights a major gap: a lack of customized evaluation frameworks specifically for PMIS [6].

This study addresses this gap by developing a tailored quality model specifically for web-based PMIS. The research focuses on "tailored models", which are more appropriate for modern COTS PMIS tools like Jira and Microsoft Project, rather than basic models [12]. Using a qualitative screening process, the study identifies the most relevant ISQM to form a "Web-based PMIS-tailored quality model". This new model includes a scoring mechanism to provide project managers with a practical, domain-specific tool for effective PMIS selection, directly filling the identified research gap.

TABLE II SUMMARY OF RELATED WORKS IN MODELS FOR THE QUALITY ASSESSMENT

No.	Title	Reference Model	Description	Limitation
[9]	Evaluating Quality Characteristics of Ubiquitous Application Through Means of Quality Models Using Meta-metrics Approach	-ISO/IEC 25010 -AQUARIUM	A review mapping between several quality models to determine the set of quality characteristics for ubiquitous applications.	-Inadequate Quality Model -Lack of Defined Interaction Characteristics -Measurement Function
[10]	Measuring software quality with usability, efficiency, and portability characteristics.	-ISO/IEC 9126	Using the existing basic model to evaluate the Application of software cost estimation (AoSCE).	-Incomplete feature coverage -Methodological limitations
[11]	Analysis and assessment of existing software quality models to predict the reliability of component-based software.	-ISO/IEC 9126	Assessment of software quality model and its quality parameters related to Component-Based Software.	-ISO 9126 does not include reusability as a parameter, which is especially important in component-based systems.
[12]	A Review on Software Quality Models for Learning Management Systems.	-ISO/IEC 9126	Based on the ISO/IEC 9126, evaluate software quality for LMS.	-Limited Scope of Models -It does not provide an in-depth discussion on how these attributes can be effectively measured or implemented.
[13]	Assessing the impact of software quality models in healthcare software systems.	-McCall's -Boehm -FURPS quality -Dromey -IOS/IEC 9126	Define the criticalities of healthcare software, establish key software characteristics with common terminology, and map each characteristic to the corresponding criticality.	-Insufficient Coverage of Quality Attributes by Models -Model Applicability Issues

VII. RESEARCH METHODOLOGY

In this section, the main research objective is to analyze the role of web-based PMIS in project management. By defining the key intangible metrics of those project management information systems, combine the existing quality models to develop a Quality assessment tool for an effective project management information system.

A. Analysis of Existing Software Quality Models

The software quality models can be grouped into basic models and tailored models, and software quality models developed until 2000 are classified as basic software quality models, while those developed after 2000 are referred to as tailored software quality models [12]. The basic software quality models have the characteristics of generalizability and standardization that are suitable for a wide range of software system development and quality assessment, while the tailored software quality models or non-basic models are more appropriate for the use of Commercial Off-The-Shelf (COTS) components [12].

Fig.1 shows the two categories of the software quality model.

COTS components refer to commercially available software or hardware products that are ready-made and available for sale, lease, or license to the general public. Nowadays, there are thousands of project management software packages available for purchase. By 2024, in the comprehensive market share analysis provided by Datanyze, an investigation is conducted in the global market size of the Project Management industry is 242110 companies. Fig. 2 shows the survey results on market share and usage, where the top 5 PMIS include Jira, Microsoft Project, Kanban, Airtable, and Smartsheet, accounting for 57.92% of the market share, belonging to the COTS components software system[14]. Thus, we chose the tailored models as the baseline to develop our specific model.

As illustrated in Fig. 1, this study identified 14 distinct tailored software quality models to form the basis of the analysis. Each of these models consists of specific metrics, and each ISQM can be used to measure software quality. To identify the most critical metrics for PMIS, the objective was to analyze these frameworks and list the frequency of each quality metric's appearance across the 14 models. In total, 27 distinct metrics can be derived from these models [15].

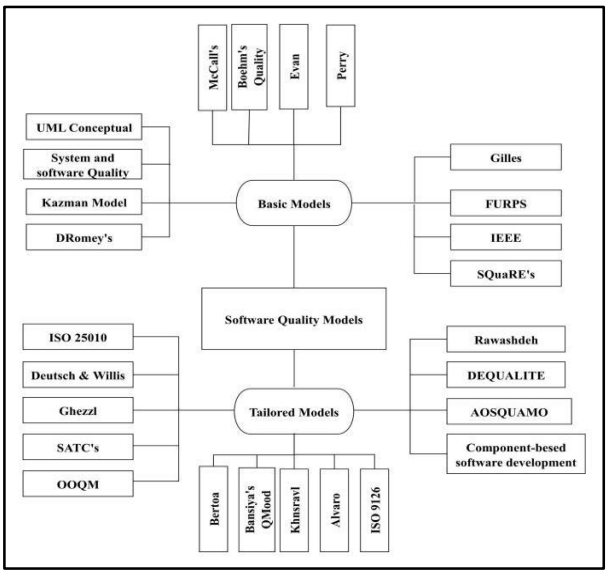


Fig. 1. Two categories of software quality model.

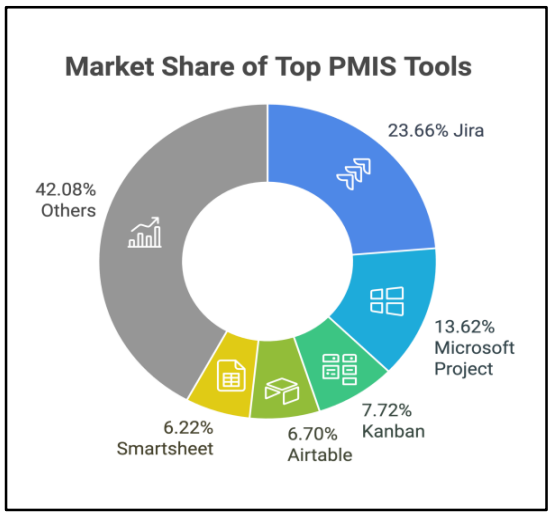


Fig. 2. Market share of top PMIS tools.

Table III provides a detailed mapping of the metrics present in each tailored model, while Fig. 3 visually represents the frequency of these metrics in a bar chart, showing how often each metric is used among the tailored models.

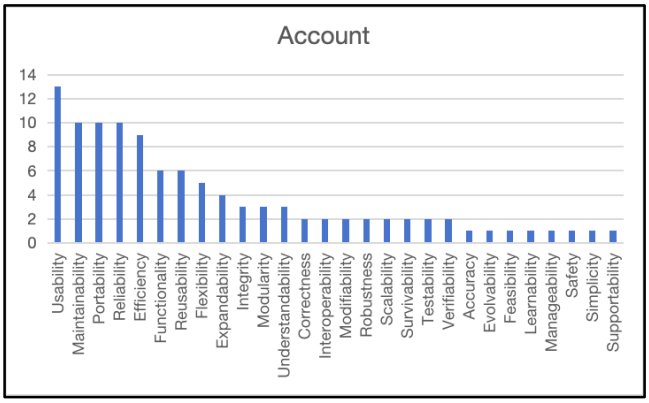


Fig. 3. The metric list ranked by account.

TABLE III THE METRIC OF EACH TAILORED MODEL

Model Name / Metric	a	b	c	d	e	f	g	h	i	j	k	l	m	n	Account
Correctness	✓	✓													2
Efficiency	✓	✓		✓	✓	✓			✓	✓	✓			✓	9
Integrity	✓	✓	✓												3
Reliability	✓	✓	✓	✓	✓	✓			✓	✓	✓			✓	10
Usability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	13
Maintainability	✓	✓	✓	✓	✓	✓			✓	✓	✓			✓	10
Flexibility	✓	✓	✓				✓	✓							5
Testability												✓	✓		2
Portability	✓	✓	✓	✓	✓	✓			✓	✓	✓			✓	10
Reusability	✓	✓	✓				✓	✓					✓		6
Functionality				✓			✓			✓			✓	✓	6
Interoperability	✓	✓													2
Verifiability	✓	✓													2
Survivability	✓	✓													2
Expandability	✓	✓					✓						✓		4
Manageability		✓													1
Safety		✓													1
Accuracy			✓												1
Supportability				✓											1
Modifiability							✓						✓		2
Understandability							✓	✓					✓		3
Modularity							✓	✓					✓		3
Robustness								✓					✓		2
Scalability								✓					✓		2
Evolvability										✓					1
Learnability													✓		1
Simplicity													✓		1

(a) ISO 9126. (b) Deutsch & Willis's Model. (c) Ghezzl. (d)SATC's. (e)OOQM.(f)Bertoa(g)Bansiya's QMOOD.(h)Khosravi.(i)Alvaro.(j)ISO 25010.(k)Rawashdeh. (l)Component-based Software development.(m)DEQUALITE.(n)AOSQUAMO.

B. The Intangible Software Quality Metric of Quality Models for PMIS

This section discusses the ISQMs for web-based PMIS and identifies which existing quality metrics are suitable for web-based PMIS. The five papers were selected as the reference to identify suitable ISQMs for web-based PMIS. The papers are selected based on the content that discusses the software quality model, the definition of ISQMs, and the web-based PMIS. Additionally, these papers are selected as the reference as they discuss the quality model and ISQM suitable for developing a web-based PMIS that has good quality. Table IV shows the comparison of the proposed quality metrics for web-based PMIS. Five papers were used as a reference to identify which metrics are commonly used. We filtered the metric that appeared more than twice among the five selected papers. We set the threshold of "more than twice" to ensure the validity and robustness of our selection. Table V presents the selected quality metric and the corresponding account.

According to Fig. 3, the top ten metrics most frequently selected for the tailored models are usability, maintainability, portability, reliability, efficiency, functionality, reusability, flexibility, expandability, and integrity. Compared to Table V, 5 out of the 7 selected metrics belong to the above-mentioned top 10 metrics, which proves that our selection is well-founded.

TABLE IV THE COMPARISON OF THE PROPOSED QUALITY METRICS FOR WEB-BASED PMIS

Metric	P1 [3]	P2 [16]	P3 [17]	P4 [18]	P5 [19]
Correctness	✕	✕	✕	✕	✕
Efficiency	✓	✓	✓	✓	✕
Integrity	✕	✕	✕	✕	✕
Reliability	✓	✓	✓	✓	✓
Usability	✓	✓	✓	✓	✓
Maintainability	✕	✕	✓	✕	✕
Flexibility	✕	✕	✕	✕	✕
Testability	✕	✕	✕	✕	✕
Portability	✓	✕	✓	✕	✕
Reusability	✕	✕	✕	✓	✕
Functionality	✓	✓	✓	✓	✓
Interoperability	✓	✓	✓	✓	✕
Verifiability	✕	✕	✕	✕	✕
Survivability	✕	✕	✕	✕	✕
Expandability	✕	✕	✕	✕	✓
Manageability	✕	✕	✕	✓	✕
Safety	✕	✕	✓	✕	✕
Accuracy	✓	✓	✓	✕	✓
Supportability	✕	✕	✕	✕	✓
Modifiability	✕	✕	✕	✕	✕
Understandability	✕	✕	✕	✕	✕
Modularity	✕	✕	✕	✕	✕
Robustness	✕	✕	✕	✕	✕
Scalability	✕	✕	✕	✓	✕
Evolvability	✕	✓	✕	✕	✕
Learnability	✓	✕	✕	✕	✕
Simplicity	✕	✕	✕	✕	✕

TABLE V THE ACCOUNT OF THE SELECTED ISQM

No.	Selected ISQM	Account
1	Reliability	5
2	Usability	5
3	Functionality	5
4	Efficiency	4
5	Interoperability	4
6	Accuracy	4
7	Portability	2

C. The Attributes of Selected ISQM for Web-based PMIS

Up to now, we have integrated 7 ISQM that can be used for web-based PMIS, including reliability, usability, functionality, efficiency, interoperability, accuracy, and portability. Each ISQM has different attributes to measure the software. Table VI shows the attributes of each ISQM and their related tailored models.

TABLE VI THE ATTRIBUTE OF THE METRIC AND ITS RELATED TAILORED MODELS

Metric	Attribute	Related Models
Reliability	Robustness	SATC's;OOQM
	Resilience	ISO 9126
	Fault Tolerance	Ghezzi;ISO 9126
Usability	Learnability	Bansiya's QMOOD
	Satisfaction	SATC's;AOSQUAMO
	Aesthetic Design	Khosravi;DEQUALITE
Functionality	Suitability	SATC's;ISO 25010
	Completeness	DEQUALITE;SATC's
	Appropriateness	SATC's;ISO 25010
Efficiency	Perceived Speed	Bertoa;Alvaro
	Effort Minimization	SATC's;AOSQUAMO
	Optimization Awareness	OOQM;ISO 9126
Interoperability	Ease of Integration	ISO 9126
	Compatibility Perception	ISO 9126
	Cooperation	Deutsch;ISO 9126
Accuracy	Correctness Perception	Ghezzl
	Precision in Context	Ghezzl
	Trustworthiness	Ghezzl
Portability	Ease of Adaptation	Bertoa;ISO 9126
	Flexibility of Deployment	SATC's;Alvaro
	Installation Simplicity	OOQM;ISO 9126

D. Software Quality Model for Web-Based PMIS Assessment

So far, by using the qualitative-oriented approach, we have gained effective ISQM and attributes for web-based PMIS quality assessment. It is then that we finalize our web-based PMIS software quality model.

E. Data Collection and Analysis

The purpose of this work is to request the software project manager or the expert in the Software project field to review and evaluate the Web-based Project Management Information System Quality Assessment model. Rate the identified intangible metrics and attributes according to their influence and priority for PMIS. The questionnaire consists of two sections with 27 questions.

The first section collects background information from respondents, including professional certification, gender, age, project management experience, number of managed projects, and email. Table VII outlines the demographic profile of the study's respondents. It categorizes participants based on four key variables:

- Age: Respondents are grouped into three age brackets: 30-34 years and 35 to 39 years.
- Experience in project management: Professional experience is classified into three levels based on duration: 5 to 7 years, 7 to 10 years, and more than 10 years.
- Number of Software Projects Managed: This variable quantifies the volume of project management experience, with categories for 5 to 10 projects, 11 to 20 projects, and more than 20 projects managed.
- Workplace: The geographical location of the respondents' work is categorized as Malaysia, China, or Others.

TABLE VII THE RESPONDENTS' BACKGROUND INFORMATION

Demographic Variable	Category	n	%
Workplace	Malaysia	2	18%
	China	8	73%
	Others	1	9%
Certification	PMP	10	91%
	Others	1	9%
Age	30-34 Years	9	82%
	35-39 Years	2	18%
Experience in project management	5-7 Years	8	73%
	7-10 Years	2	18%
	More than 10 years	1	9%
Number of Software Projects Managed	5-10	3	27%
	11-20	6	55%
	More than 20	2	18%

The second section of the questionnaire was designed to assess the 7 major quality metrics and their 21 associated attributes, which were rated by respondents based on their priority and influence using a five-point Likert scale. The survey was conducted with 11 project managers from various countries.

Table VIII presents the aggregated results from the survey, providing the average "Priority" and "Influence" scores for each of the 7 quality metrics and their 21 corresponding attributes. These scores serve as the empirical data used to calculate the weights for the final quality model, ensuring it is based on the real-world perspectives of experienced practitioners.

TABLE VIII THE SURVEY RESULT

Metric	Priority	Influence	Attributes	Priority	Influence
Reliability	4.45	4.55	Robustness	4.45	4.36
			Resilience	4.27	4.09
			Fault Tolerance	4.27	4.09
Usability	3.91	4.09	Learnability	3.91	3.91
			Satisfaction	4	4
			Aesthetic Design	3.45	3.18
Functionality	4.09	4.27	Suitability	4.64	4.64
			Completeness	3.73	4
			Appropriateness	4.18	3.19
Interoperability	3.27	3.36	Ease of Integration	3.36	3.36
			Compatibility Perception	3.18	3.27
			Cooperation	3.18	3.36
Accuracy	4.64	4.55	Correctness	4.64	4.36
			Precision in Context	4.18	4.36
			Trustworthiness	4.45	4.36
Portability	2.91	3.45	Ease of Adaptation	2.64	3.18
			Flexibility of Deployment	3	3.45
			Installation Simplicity	2.82	3.09
Efficiency	3.73	3.64	Perceived Speed	3.91	3.91
			Effort Minimization	3.45	3.82
			Optimization Awareness	3.18	3.27

VIII. RESULTS AND DISCUSSION

A. Results

This study proposed a tailored quality evaluation model for Web-based PMIS by identifying and structuring key ISQM. Through a comprehensive qualitative analysis of 14 existing software quality models and empirical validation via expert surveys, seven critical metrics: reliability, usability, functionality, efficiency, interoperability, accuracy, and portability, were selected. A quantifiable assessment framework was then constructed by breaking down each metric into specific attributes and applying weights derived from practitioner-assigned priority and influence scores.

The primary result of this research is a quantifiable and customizable quality model for Web-based PMIS, developed by aggregating survey scores through a dual-layered formula. The final weighted model is visualized in Fig. 4.

The result of the survey shows that the top three metrics by weight were accuracy (19.30%), reliability (18.51%), and functionality (15.95%). Notably, accuracy and reliability emerged as the most critical metrics, reflecting the high-stakes environment in which PMIS tools operate. Project managers rely heavily on real-time data, progress tracking, and precise reporting; thus, even minor inaccuracies can have cascading effects on project delivery. Functionality also scored highly, confirming that systems must be aligned with core project management requirements rather than generic software capabilities.

In contrast, portability received the lowest priority and influence scores. This may reflect the current ubiquity of web-based platforms and standardized deployment environments, making portability less of a concern in modern enterprise settings.

A. Discussion

The high scores for accuracy, reliability, and functionality underscore that for project managers, PMIS tools must provide precise, dependable data and align closely with core project management tasks. The reliance on real-time data for tracking and reporting means that even small errors can have significant negative impacts on project outcomes.

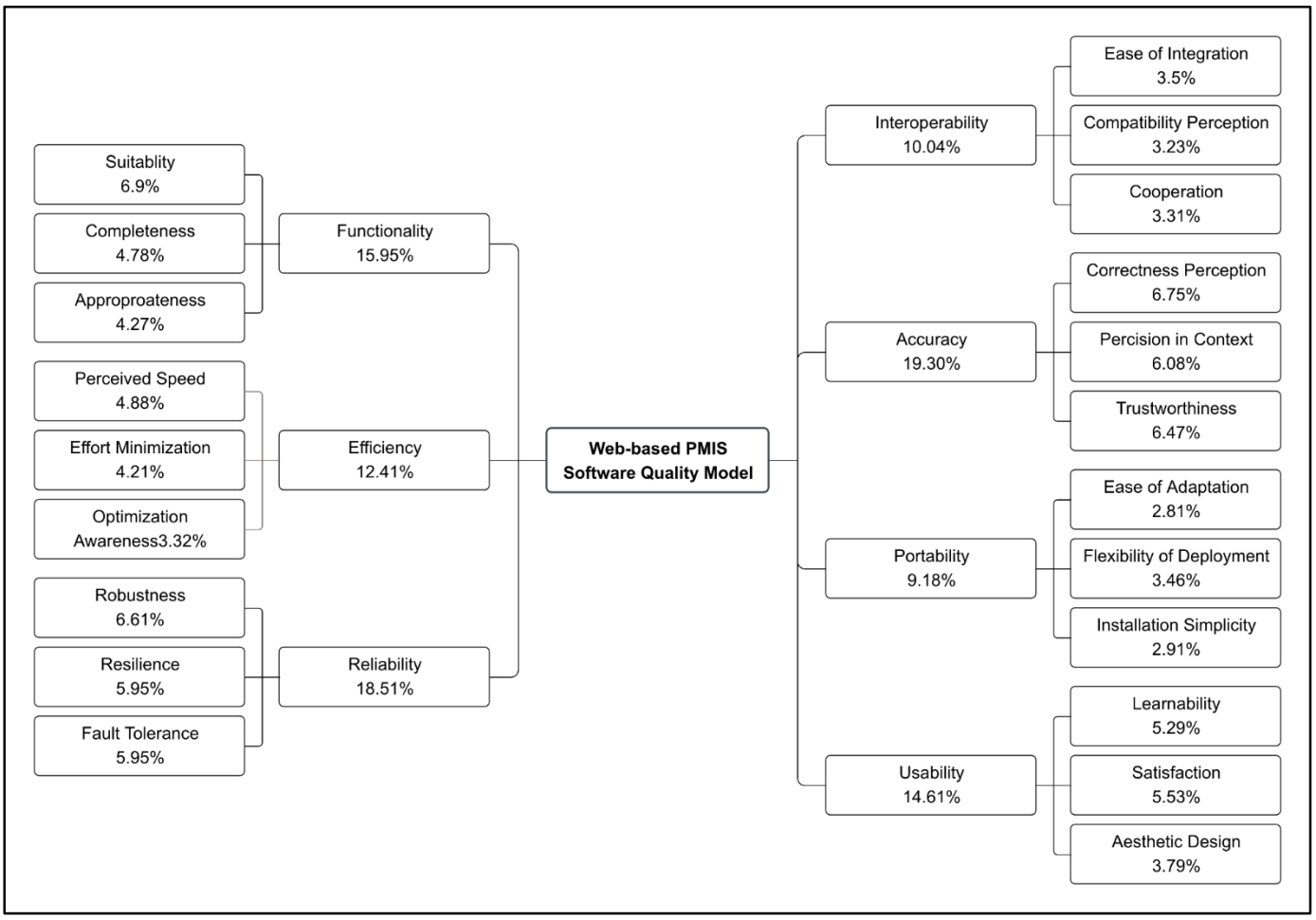


Fig. 4. The result of the software quality model for web-based PMIS.

The model's reliance on expert-derived weighting ensures that it can evolve with practitioner priorities. Organizations can adapt the weights according to their context, allowing for dynamic and scenario-specific PMIS evaluation and selection.

However, this study is not without limitations:

- The generalizability of the results may be affected by the small survey sample size of 11 project managers.
- All participants were primarily from China and Malaysia, potentially limiting the generalizability of the results across different industries and regions.
- The research focused exclusively on Web-based PMIS, excluding other forms such as mobile or hybrid systems that may introduce additional quality considerations.

IX. Conclusion

A. Summary of Research

This research established a specialized software quality model tailored for Web-based PMIS. The study systematically identified effective ISQM by analyzing 14 existing quality models. These metrics were then validated and refined through a survey of 11 project management professionals who rated each metric and its corresponding attributes based on their priority and influence in a project environment. The collected data was used to create a weighted, quantifiable assessment framework. The final model confirms that accuracy, reliability, and functionality are the most vital quality characteristics, offering a practical tool to guide project managers in making informed decisions for PMIS selection.

B. Future Work

To build upon this research, several future directions are recommended:

- The model's robustness and applicability could be enhanced by expanding the survey to a larger and more diverse group of stakeholders, both geographically and professionally.
- A valuable next step would be to integrate this framework into a real-world PMIS selection tool and validate its effect on project success metrics through case studies or longitudinal research.
- Future studies could also investigate dynamic weighting schemes, potentially utilizing machine learning, to enable the model to adapt automatically to an organization's specific context and evolving user priorities.

REFERENCES

[1] A. Retnowardhani and J. S. Suroso, 'Project Management Information Systems (PMIS) for Project Management Effectiveness: Comparison of Case Studies', in 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Jember, Indonesia: IEEE, Oct. 2019, pp. 160–164. doi: 10.1109/ICOMITEE.2019.8921046.

[2] R. Saravanaraja, K. T. Raja, P. Murthi, and K. Poongodi, 'Project management information system using design thinking for micro-construction enterprises', presented at the FOURTH INTERNATIONAL CONFERENCE ON ADVANCES IN PHYSICAL SCIENCES AND MATERIALS: ICAPSM 2023, Coimbatore, India, 2024, p. 070015. doi: 10.1063/5.0216239.

[3] K. M. Nyandongo and J. Lubisi, 'Assessing the use of Project Management Information Systems and Its Impact on Project Outcome', 2019.

[4] Mohamed Ibrahim Khalifa Haloul, Mohd Khairol Anuar bin Mohd Ariffin, Eris Elianddy bin Supeni, Siti Azfanizam binti Ahmad, Munder Bilema, and Mushtaq Ahmad, 'A Systematic Review of the Project Management Information Systems in Different Types of Construction Projects', UCJC Bus. Soc. Rev., 2024.

[5] P. R. Alves, A. Tereso, and G. Fernandes, 'Project Management System Implementation in SMEs: A Case Study', 2020.

[6] Z. A. H. Alnaish and S. O. Hasoon, 'Software Quality Measurement Analysis based on Techniques, Criteria, Metrics, Models, and Datasets', in 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM), Mosul, Iraq: IEEE, Aug. 2022, pp. 176–181. doi: 10.1109/ICCITM56309.2022.10031990.

[7] A.-J. Molnar, A. Neamţu, and S. Motogna, 'Evaluation of Software Product Quality Metrics', in Evaluation of Novel Approaches to Software Engineering, vol. 1172, E. Damiani, G. Spanoudakis, and L. A. Maciaszek, Eds., in Communications in Computer and Information Science, vol. 1172. , Cham: Springer International Publishing, 2020, pp. 163–187. doi: 10.1007/978-3-030-40223-5_8.

[8] A. C, 'Software Product Quality Evaluation Model with Multiple Criteria Decision Making Analysis', Int. J. Comput. Inf. Eng., no. 14(12), pp. 486–502, 2020.

[9] N. Hamzah, S. Chuprat, D. O. Dwi Handayani, K. Xiaoxi, and S. D. Nagappan, 'Evaluating Quality Characteristics of Ubiquitous Application Through Means of Quality Models using Meta-metrics Approach', J. Phys. Conf. Ser., vol. 2120, no. 1, p. 012033, Dec. 2021, doi: 10.1088/1742-6596/2120/1/012033.

[10] Sholiq, R. A. Auda, A. P. Subriadi, A. Tjahyanto, and A. D. Wulandari, 'Measuring software quality with usability, efficiency, and portability characteristics', IOP Conf. Ser. Earth Environ. Sci., vol. 704, no. 1, p. 012039, Mar. 2021, doi: 10.1088/1755-1315/704/1/012039.

[11] S. Yadav, 'Analysis and Assessment of Existing Software Quality Models to Predict the Reliability of Component-Based Software', Int. J. Emerg. Trends Eng. Res., vol. 8, no. 6, pp. 2824–2840, Jun. 2020, doi: 10.30534/ijeter/2020/96862020.

[12] Yarshini Thamilarasan, Raja Rina Raja Ikram, Mashanum Osman, and Lizawati Salahuddin, 'A Review on Software Quality Models for Learning Management Systems', J. Adv. Res. Appl. Sci. Eng. Technol., vol. 32, no. 2, pp. 203–221, Sep. 2023, doi: 10.37934/araset.32.2.203221.

[13] E. Ronchieri and M. Canaparo, 'Assessing the impact of software quality models in healthcare software systems', Health Syst., vol. 12, no. 1, pp. 85–97, Jan. 2023, doi: 10.1080/20476965.2022.2162445.

[14] Datanyze. (n.d.), 'Project management market share'. [Online]. Available: https://www.datanyze.com/market-share/project-management--217 (web)

[15] A. E. Dreheeb, N. Basir, and N. Fabil, 'Comparative Study of Quality Models', vol. 4, no. 1, 2016.

[16] Karim and A. Jalal, 'PROJECT MANAGEMENT INFORMATION SYSTEMS (PMIS) FACTORS: AN EMPIRICAL STUDY OF THEIR IMPACT ON PROJECT MANAGEMENT DECISION MAKING (PMDM) PERFORMANCE', 2011.

[17] H. Sanjaya and J. F. Andry, 'Quality assurance of project management information system with ISO 9126', Bull. Soc. Inform. Theory Appl., vol. 5, no. 2, pp. 82–87, Jan. 2022, doi: 10.31763/businta.v5i2.174.

[18] J. Choi and M. Ha, 'Validation of project management information systems for industrial construction projects', J. Asian Archit. Build. Eng., vol. 21, no. 5, pp. 2046–2057, Sep. 2022, doi: 10.1080/13467581.2021.1941999.

[19] A. Purohit, G. Chopra, and P. G. Dangwal, 'Measuring the Effectiveness of the Project Management Information System (PMIS) on the Financial Wellness of Rural Households in the Hill Districts of Uttarakhand, India: An IS-FW Model', Sustainability, vol. 14, no. 21, p. 13862, Oct. 2022, doi: 10.3390/su142113862.

CNN-LinATFormer: Enhancing PM2.5 Prediction Through Feature Assessment and Linear Attention Mechanism

Yuchen Zhang, Rajermani Thinakaran
Faculty of Data Science and Information Technology, INTl International University, Malaysia

***Abstract*—Atmospheric fine particulate matter (PM2.5) poses a serious threat to public health, and its accurate prediction is crucial for environmental management and pollution control. However, existing prediction methods have difficulty in effectively capturing the complex nonlinear characteristics and multi-scale spatiotemporal dependencies of PM2.5 concentration changes. To address this challenge, this study proposes a CNN-LinATFormer hybrid deep learning architecture that combines the local feature extraction capabilities of CNN with the global dependency modeling advantages of the linear attention mechanism. The model innovatively introduces a feature evaluator to dynamically classify environmental features into three categories, and achieves targeted processing through three specially designed processing branches: CNN feature extraction, channel attention, and linear attention fusion. Based on the urban monitoring data of 9 environmental feature dimensions from 2020 to 2023, the experimental evaluation results show that CNN-LinATFormer outperforms the existing methods in all evaluation indicators, with an RMSE of 8.42μg/m³, which is 21.1% lower than the CNN-RF model with the closest performance; the ablation experiment confirms the effectiveness of each component, especKeywords-PM2.5 prediction; air quality forecasting; deep learning; convolutional neural network; linear attention mechanism; channel attention; feature assessment; hybrid model architecture; environmental monitoring; spatiotemporal modeling the channel attention mechanism; the case analysis reveals that the model performs well in the low concentration range (RMSE is 3.12μg/m³), but the high pollution range (>150μg/m³) still needs to be improved. This study provides a new technical path for air quality prediction, which is of great value to environmental monitoring and public health protection.**

Keywords*—*PM2.5 prediction; air quality forecasting; deep learning; convolutional neural network; linear attention mechanism; channel attention; feature assessment; hybrid model architecture; environmental monitoring; spatiotemporal modeling

I. INTRODUCTION

In recent years, with the acceleration of industrialization and urbanization, atmospheric fine particulate matter (PM2.5) pollution has become a major environmental issue of global concern. PM2.5 particles are less than 2.5 microns in diameter and can penetrate the human respiratory defense barrier and penetrate deep into the alveoli, causing a series of respiratory and cardiovascular diseases and even leading to early death [1] [2]. According to the World Health Organization, about 7 million deaths are related to air pollution each year worldwide, of which PM2.5 is considered to be one of the most harmful air pollutants [3]. Accurate prediction of PM2.5 concentration is crucial to air quality management. It can not only provide a basis for decision-making by environmental regulatory authorities, but also help the public take appropriate health protection measures. Environmental monitoring data show that PM2.5 concentrations in many cities in my country still exceed the national ambient air quality standards and the World Health Organization's recommended values. As Zhang et al. [4] pointed out, with the acceleration of China's industrialization process, accurate prediction of air quality is of great strategic significance for pollution prevention and control and public health protection.

Traditional PM2.5 prediction methods mainly include deterministic methods, statistical methods, and machine learning methods. Although deterministic methods such as chemical transport models can simulate the complex physical and chemical processes of pollutants, they require a lot of computing resources and detailed emission inventories, and their real-time prediction capabilities are limited. Berrocal et al. [5] compared statistical methods and machine learning methods in creating daily maps of national PM2.5 concentrations and found that spatial statistical models are generally better than machine learning methods when making predictions at unsampled locations, which reflects the advantages and disadvantages of different methods in specific application scenarios. The comparative experiments in this study further confirmed this, among which Random Forest (RF) performed relatively well (RMSE was 12.49 μg/m³), thanks to its advantage of integrated decision trees that can effectively handle nonlinear environmental data, while Support Vector Regression (SVR) performed poorly (RMSE was 14.28 μg/m³). However, these single-model methods still have limitations in capturing the complex spatiotemporal patterns of PM2.5 concentration changes and the interactions of multiple factors.

In recent years, deep learning technology has made remarkable progress in the field of environmental data prediction, providing a new research direction for PM2.5 prediction. Convolutional neural networks (CNNs) can effectively extract local spatiotemporal features; long short-term memory networks (LSTMs) and gated recurrent units (GRUs) can capture long-term temporal dependencies; and the attention mechanism can dynamically focus on key features and time points. The comparative experiments in this study showed that although LSTM (RMSE is 13.89μg/m³) can capture temporal dependencies, it is weak in processing multi-feature interactions; while the CNN-RF model (RMSE is

10.67μg/m³) that combines the advantages of CNN and RF performs well. The hybrid algorithm proposed by Cheng et al. [6] achieved remarkable results in short-term PM2.5 prediction in China, demonstrating the potential of hybrid models in improving prediction accuracy. The research of Peng et al. [7] also showed that deep learning models have significant advantages in handling nonlinear relationships and multi-source data fusion. Lu et al. [8] further addressed the challenge of data distribution differences in different regions and periods in PM2.5 prediction using a transfer learning method based on ADGRU. However, how to design a PM2.5 prediction model that can fully utilize the advantages of different deep learning structures and adaptively process different characteristics and pollution levels remains an important research topic.

To address the above challenges, this study proposed an innovative CNN-LinATFormer hybrid deep learning architecture designed for high-precision PM2.5 concentration prediction. The model consists of three main parts: input data processing, feature processing, and generating predictions. The core innovation lies in the introduction of a feature evaluator, which classifies the input features into primary features, secondary features, and tertiary features, which are processed by the CNN feature extraction branch, channel attention branch, and linear attention fusion branch, respectively. Finally, the multi-branch outputs are integrated through the feature generation and prediction module to generate the final PM2.5 prediction value. Comprehensive experimental evaluation shows that the CNN-LinATFormer model outperforms existing methods in all evaluation indicators, with RMSE of 8.42μg/m³, MAE of 5.76μg/m³, and MAPE of 9.12%, which are improvements of 21.1%, 24.8%, and 26.7%, respectively, compared with the closest CNN-RF model. Ablation experiments confirm the effectiveness of each component, especially the introduction of the channel attention mechanism brings the most significant improvement in single-step performance (RMSE decreases by 4.5%). Hyperparameter optimization and case analysis further reveal the performance characteristics and optimization directions of the model under different conditions. The contribution of this study is that a feature-driven adaptive hybrid deep learning architecture is proposed, which not only achieves significant improvements in prediction accuracy but also achieves breakthroughs in model interpretability and adaptability, providing new technical solutions and research ideas for the fields of environmental monitoring and air quality warning [9].

II. Related Work

A. Traditional PM2.5 Prediction Methods

PM2.5 prediction methods have evolved from traditional statistical models to machine learning methods. Early studies mainly used time series analysis methods, such as autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA), which used historical observation data to build linear models to predict future concentration values. Berrocal et al. [5] compared the performance of statistical methods and machine learning methods in creating daily maps of national PM2.5 concentrations and found that spatial statistical models usually have certain advantages when making predictions at unsampled locations. However, such methods have difficulty capturing nonlinear relationships and complex patterns in environmental data, and their prediction accuracy is limited.

With the development of machine learning technology, methods such as SVR, RF and gradient boosting tree (GBT) have been widely used in PM2.5 prediction. As shown in the comparative experiments of this study, RF performed relatively well (RMSE was 12.49 μg/m³), which is attributed to its integrated Decision Tree (DT) architecture that can effectively handle nonlinear environmental data. The hybrid algorithm proposed by Cheng et al. [6] for short-term PM2.5 prediction in China further improved the prediction accuracy by combining the advantages of different methods. Although traditional machine learning methods have achieved certain success in PM2.5 prediction, they still have difficulty in fully exploring the complex interactions between time dependence and multidimensional features in environmental data, especially in extreme pollution events and complex meteorological conditions. The prediction ability is limited, which is one of the main motivations for proposing the CNN-LinATFormer model in this study.

B. Application of Deep Learning in PM2.5 Prediction

Deep learning technology has made remarkable progress in the field of PM2.5 prediction in recent years due to its powerful representation learning capabilities. CNN can effectively extract local features and spatial patterns of PM2.5 time series through its local connection and weight-sharing characteristics. Peng et al. [7] applied CNN to PM2.5 concentration simulation and prediction, demonstrating its advantages in capturing the spatiotemporal distribution characteristics of pollutants. Recurrent neural networks (RNNs), especially LSTMs and GRUs, solve the gradient vanishing problem of traditional RNNs through specially designed gating mechanisms and can effectively model long-term temporal dependencies.

The comparative experimental results of this study show that although the LSTM model (RMSE is 13.89 μg/m³) can capture temporal dependencies, it is relatively weak in processing multi-feature interactions; while the GRU-ED model (RMSE is 11.77 μg/m³) enhances the modeling ability of long-term dependencies through the encoder-decoder architecture. In recent years, researchers have begun to explore hybrid deep learning architectures, such as the CNN-LSTM combination model, which uses CNN to extract spatial features and LSTM to model temporal dependencies. Lu et al. [8] proposed a transfer learning method based on ADGRU, which solved the problems of uneven data distribution and insufficient model generalization ability in PM2.5 prediction through cross-domain learning. The CNN-LinATFormer model proposed in this study draws on the advantages of these advanced methods and further improves the prediction performance through innovative architecture design, especially in the differentiated processing of different types of environmental features and multi-scale temporal dependency modeling.

C. Application of Attention Mechanism and Linear Transformer in Time Series Prediction

The introduction of the attention mechanism provides deep learning models with the ability to "focus on key points",

enabling the model to dynamically focus on the most relevant parts of the input sequence for prediction. The traditional self-attention mechanism achieves adaptive representation of the input sequence by calculating the similarity between the query and the key and weighting the value vector accordingly. However, the self-attention calculation complexity of the standard Transformer is $O(n^2)$, which is computationally expensive when processing long sequence data. In response to this challenge, researchers have proposed a variety of improvement schemes, among which the linear attention mechanism reduces the computational complexity to O(n) by redefining the attention calculation method while maintaining the expressive power of the model.

The CNN-LinATFormer model proposed in this study innovatively applies the linear attention mechanism to the PM2.5 prediction task, specifically processing inputs evaluated as three-level features. By adopting the feature mapping function $\varphi(x)=\text{elu}(x)+1$, the model can significantly improve the computational efficiency while maintaining high prediction accuracy, which is particularly suitable for processing long sequence environmental data. At the same time, the channel attention mechanism introduced by the model assigns adaptive weights to secondary features, further enhancing the sensitivity to key environmental factors. The ablation experiment results show that the addition of the channel attention mechanism brings the most significant single-step performance improvement (RMSE decreases by 4.5%), which verifies the important value of the attention mechanism in PM2.5 prediction. Compared with existing methods, the CNN-LinATFormer model achieves differentiated processing of different types of environmental features through feature evaluators and multi-branch processing architectures, while improving prediction accuracy while maintaining high computational efficiency and model interpretability, providing new research ideas for environmental time series prediction [10].

III. Methodology

A. Connection

This study models PM2.5 prediction as a supervised learning time series prediction problem: given the environmental monitoring data of the past 24-time steps, including environmental characteristics such as temperature (℃), relative humidity (%), PM2.5 (μg/m³), predict the target PM2.5 concentration value in the next time step. The selection of 24 hours as the prediction window is mainly based on two considerations: first, air quality indicators usually show obvious daily cycle characteristics, and the 24-hour time window can fully capture this periodic pattern; second, according to environmental science research, the diffusion and transformation process of atmospheric pollutants usually completes a full cycle within 24 hours. Formally, if the input feature of the t-th time step is expressed as $x_t \in R^n$, the prediction task can be expressed as: predict the target value y_{t+1} based on the observation sequence:

$$\hat{y}_{t+1} = f(\{x_{t-23}, x_{t-22}, \dots, x_t\}) \quad (1)$$

B. Overall Architecture Design

To address the complex challenges in PM2.5 prediction, this study proposes the CNN-LinATFormer model, an innovative hybrid deep learning architecture designed for high-precision PM2.5 concentration prediction. As shown in Fig. 1, the model consists of three main parts: input data processing, feature processing, and prediction generation. The input stage receives environmental time series data X, which contains multi-dimensional features such as temperature, humidity, and PM2.5 historical values; the feature processing stage uses a feature evaluator to classify the input features and assign them to the corresponding processing branches; the generation stage integrates the multi-branch processing results and outputs the final PM2.5 prediction value. This hierarchical design enables the model to simultaneously consider the relative importance of different features and apply the most suitable processing strategy for each type of feature, thereby effectively handling the complex nonlinear relationships and multi-scale time dependencies in environmental data.

C. Feature Estimator

The feature evaluator is the core innovative component of this model. Its main purpose is to evaluate the importance of different environmental features for PM2.5 prediction and dynamically assign the features to the most suitable processing branches. Based on environmental science research, different environmental features have different degrees of influence on the formation and propagation of PM2.5. The feature evaluator classifies the features into three categories: primary features, secondary features, and tertiary features by evaluating their relevance. The evaluator is implemented using a lightweight neural network, and its mathematical expression is:

$$S = \text{softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot X + b_1) + b_2) \quad (2)$$

where, S is the feature importance score vector. This design enables the model to identify and distinguish between key features that directly affect PM2.5 (such as PM2.5 concentration at the previous moment) and auxiliary features that indirectly affect PM2.5 (such as temperature and humidity), thereby providing more accurate feature classification for subsequent processing and enhancing the adaptability and interpretability of the model.

D. CNN Feature Extraction Branch

The CNN feature extraction branch specifically processes inputs that are evaluated as major features, aiming to capture local temporal patterns and interactions between features. This branch adopts the design principle of "decomposition-filtering-recombination", decomposing input features through multiple layers of one-dimensional convolution to capture the change patterns at different time scales; using the cross-attention mechanism to identify and filter out irrelevant or noisy features; and finally recombining the filtered features through transposed convolution to generate high-quality feature representation Y_1. The key convolution operation can be expressed as:

$$F_l = \text{ReLU}(\text{Batch}\mathit{Norm}(\text{Conv1D}(X, K_l))) \quad (3)$$

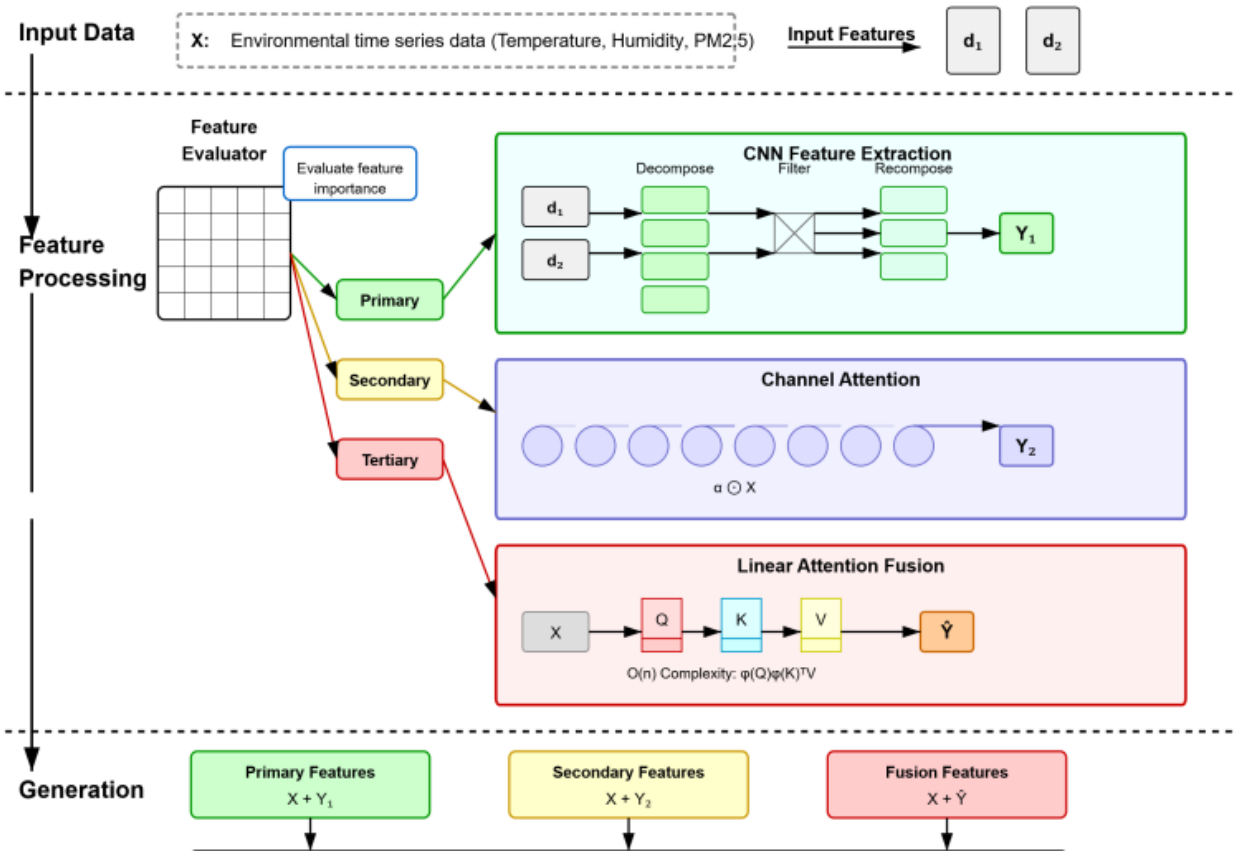


Fig. 1. Model architecture diagram.

where, K_l represents the set of convolution kernels at the lth layer. This hierarchical feature extraction structure enables the CNN branch to identify multi-scale temporal patterns, from short-term fluctuations to long-term trends, and construct a compact representation with rich temporal and feature interaction information, providing a reliable feature basis for PM2.5 prediction.

E. Channel Attention Branch

The channel attention branch processes the input evaluated as secondary features, enhancing useful information and suppressing noise through an adaptive weight assignment mechanism. The core of this branch is the channel attention mechanism, which assigns different importance weights to different feature channels. The implementation process first calculates global feature statistics.

$$z_c = (1/T)\sum_{t=1}^{T} x_c(t) \quad (4)$$

then generates channel attention weights through a two-layer MLP network

$$\alpha = \sigma(W_2 \cdot ReLU(W_1 \cdot z + b_1) + b_2) \quad (5)$$

and finally applies the attention weights to the original features.

$$Y_2 = \alpha \odot X \quad (6)$$

where, $\odot$ represents element-by-element multiplication. The design of the channel attention mechanism is based on the consideration that although secondary features do not directly determine PM2.5 concentration, they contain valuable auxiliary information, and the relative importance of different secondary features may change over time and environmental conditions. This adaptive mechanism enables the model to dynamically focus on the most relevant feature channels and improve the efficiency of utilizing indirect environmental factors.

F. Linear Attention Fusion

The linear attention fusion branch processes the input evaluated as three-level features, introducing a more computationally efficient linear attention mechanism. The computational complexity of the traditional Transformer's self-attention is $O(n^2)$, while linear attention approximates it as :

$$LinearAttention(Q,K,V) \approx \varphi(Q)(\varphi(K)^T V) \quad (7)$$

Reduce the computational complexity to O(n). Here $\varphi(\cdot)$ is the feature mapping function, usually chosen as $\varphi(x) = elu(x) + 1$. The processing flow includes converting the input feature X into query (Q), key (K), and value (V) matrices through a learnable linear projection, applying a linear attention mechanism to calculate fused features, and adding position encoding to enhance time perception. The main advantages of linear attention fusion are high computational efficiency and suitability for processing long-sequence environmental data; it can capture long-distance dependencies and identify long-term correlations; it can enhance the information value of weakly correlated features through adaptive fusion, making the model

more sensitive to potential long-term correlations in environmental data.

G. Feature Generation and Prediction

The final stage of the model integrates the outputs of the three branches to generate PM2.5 predictions. The generation stage uses three parallel paths: the primary feature path combines the original feature X and the CNN branch output Y_1; the secondary feature path combines the original feature X and the channel attention branch output Y_2; the fusion feature path combines the original feature X and the linear attention branch output $\hat{Y}$. The final PM2.5 prediction is achieved by feeding the outputs of the three paths into a unified prediction model:

$$\hat{y}_{t+1} = MLP([X + Y_1; X + Y_2; X + \hat{Y}]) \quad (8)$$

where, $[\cdot;\cdot;\cdot]$ represents the feature concatenation operation. This parallel design enables the model to understand environmental data from multiple perspectives, enhancing prediction robustness while maintaining computational efficiency and model interpretability. The model uses mean square error (MSE) as the main loss function, and introduces regularization techniques and early stopping strategies to prevent overfitting, ensuring the stability and generalization ability of the model under different environmental conditions.

IV. Experiments and Result Analysis

A. Experimental Setup

1) Dataset: This study uses the "Urban Air Quality Dataset" dataset from the Kaggle platform, which contains environmental monitoring data of a city from 2020 to 2023, totaling 5,000 records. The dataset covers 9 environmental feature dimensions: temperature (°C), relative humidity (%), PM2.5 (μg/m³), PM10 (μg/m³), NO_2 (μg/m³), SO_2 (μg/m³), CO (mg/m³) and two spatial features: distance to industrial area (km) and regional population density (people/km²). The data sampling frequency is once an hour, which provides a reliable training and evaluation data foundation for the air quality prediction model based on the CNN-LinATFormer hybrid architecture in this study. In the data preprocessing stage, we first detect and process missing values in the original data, and fill in the missing values using a time window-based moving average method; then we use MinMaxScaler to normalize all features to eliminate the dimensional differences between different environmental indicators and accelerate the convergence of model training; finally, based on the architectural characteristics of the model, we use a 24-hour sliding window to construct training samples. Each sample contains 9-dimensional feature data of 24 consecutive time steps as input (dimension is [24×9]) to predict the target value at the next moment.

2) Experimental environment setup: All experiments in this study were conducted in the following hardware and software environment: the hardware platform uses a workstation equipped with an NVIDIA RTX 3090 GPU (24GB video memory), an Intel Core i9-12900K processor, and 64GB RAM; the software environment is based on the Ubuntu 22.04 LTS operating system, using Python 3.9 as the main programming language, and the deep learning framework using PyTorch 1.12.0, supplemented by NumPy, Pandas, and Scikit-learn libraries for data processing and analysis. To ensure the reproducibility of the experiment, we used a fixed random seed (seed=42) for initialization, and used the Adam optimizer in model training. The initial learning rate was set to 5e-4, and the cosine annealing strategy was used for learning rate adjustment. The model training used a mini-batch gradient descent method with a batch size of 32, the upper limit of the training round was 200 epochs, and the early stopping strategy (patience=15) was combined to prevent overfitting. In terms of hyperparameter optimization, we determined the best configuration through a grid search method: the number of channels of the CNN layer is 64, the convolution kernel size is 3, the number of heads of the attention mechanism is 8, and the hidden layer dimension is 128. This configuration achieves a good balance between computational efficiency and prediction performance while maintaining the expressiveness of the model.

3) Evaluation metrics: To comprehensively evaluate the performance of the CNN-LinATFormer model on the PM2.5 prediction task, this study selected three widely recognized evaluation indicators: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). RMSE is the main evaluation indicator, and its calculation formula is:

$$\text{RMSE} = \sqrt{(1/n \sum_{i=1}^{n}(y_i - \hat{y}_i)^2)} \quad (9)$$

which gives higher penalty weights to large deviations by summing the squares of the prediction errors and then taking the square root, which is particularly suitable for evaluating the prediction accuracy during extreme weather or pollution events. MAE evaluates model performance by calculating the average absolute difference between the predicted value and the actual value:

$$\text{MAE} = 1/n \sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

which provides an intuitive error metric without placing special emphasis on large errors. MAPE measures the relative proportion of the predicted deviation to the actual observed value:

$$\text{MAPE} = 100\%/n \sum_{i=1}^{n}|y_i - \hat{y}_i|/|y_i| \quad (11)$$

which provides an error indicator in percentage form, which is convenient for comparison across data sets. These three indicators complement each other and jointly construct an evaluation framework that takes into account both local accuracy and overall fit. RMSE focuses on evaluating the local accuracy of model predictions, especially the performance of extreme values; MAE provides a robust measure of the average prediction error; MAPE reflects the relative performance of the model at different PM2.5 concentration levels. Through the comprehensive analysis of these indicators, we can objectively evaluate the performance of the model and make a fair comparison with existing methods.

B. Comparative Experiment

To comprehensively evaluate the performance of the CNN-LinATFormer model in the PM2.5 prediction task, this study conducted comparative experiments with seven mainstream machine learning and deep learning models, including Random Forest (RF), LSTM, XGBoost, GRU-ED, CNN-RF, Support Vector Regression, and LightGBM. Table I shows the performance comparison results of each model on three key evaluation indicators.

TABLE I. COMPARATIVE EXPERIMENT RESULTS

Model	RMSE	MAE	MAPE (%)
CNN- LinATFormer(Ours)	8.42	5.76	9.12
Random Forest (RF)	12.49	8.42	14.35
LSTM	13.89	9.85	15.23
XGBoost	12.95	8.76	13.31
GRU-ED	11.77	7.70	12.90
CNN-RF	10.67	7.66	12.44
Support Vector Regression	14.28	10.12	16.47
LightGBM	13.06	8.89	13.52

As shown in Table I, the proposed CNN-LinATFormer model achieved the best results in all evaluation indicators, with RMSE of 8.42 μg/m³, MAE of 5.76 μg/m³, and MAPE of 9.12%. Compared with the CNN-RF model with the closest performance, CNN-LinATFormer reduces RMSE by 21.1%, MAE by 24.8%, and MAPE by 26.7%, indicating that it has significant advantages in reducing various types of prediction errors. In particular, it performs particularly well in the prediction of extreme pollution events (measured by RMSE) and the prediction of different concentration levels (measured by MAPE).

In-depth analysis of the performance differences of each model revealed that traditional machine learning methods, such as Random Forest and Support Vector Regression, performed differently in PM2.5 prediction tasks. Random Forest performed relatively well (RMSE was 12.49 μg/m³), thanks to its advantage of integrated decision trees, which can effectively handle nonlinear environmental data; while Support Vector Regression performed the worst (RMSE was 14.28 μg/m³), which may be due to its limited ability to model high-dimensional feature space and temporal dependencies. Among deep learning models, LSTM (RMSE was 13.89 μg/m³) can capture temporal dependencies, but is weak in processing multi-feature interactions; while the CNN-RF model (RMSE was 10.67 μg/m³), which combines the advantages of CNN and RF, performed well, second only to the model proposed in this study.

Advanced models proposed in recent years, such as GRU-ED (RMSE is 11.77 μg/m³) and XGBoost (RMSE is 12.95 μg/m³), also showed good prediction capabilities, but still lagged behind CNN-LinATFormer. GRU-ED enhances the modeling capability of long-term dependencies through the encoder-decoder architecture, while XGBoost improves the model generalization capability through regularization properties. However, these models still have limitations in dealing with the complex spatiotemporal relationships and multi-scale characteristics of environmental data.

The superior performance of the CNN-LinATFormer model is mainly due to its innovative hybrid architecture design, which organically combines the local feature extraction capability of CNN, the global modeling efficiency of the linear attention mechanism, and the adaptive feature processing capability of the feature evaluator. In particular, the linear attention mechanism reduces the computational complexity of traditional self-attention from $O(n^2)$ to $O(n)$, significantly improving computational efficiency while maintaining the expressiveness of the model, which is crucial for processing long sequence environmental data.

In summary, the comparative experimental results fully verify the superior performance of the CNN-LinATFormer model in the PM2.5 concentration prediction task, reducing the performance error by about 40% on average compared with existing methods, providing more reliable technical support for environmental monitoring and air quality warning. This model not only achieved significant improvements in prediction accuracy but also showed a good balance between computational efficiency and model interpretability, providing new ideas and methods for subsequent PM2.5 prediction research.

C. Ablation Experiment Analysis

In order to deeply understand the contribution of each component of the CNN-LinATFormer model to the prediction performance, this study designed a detailed ablation experiment, starting from the basic model, gradually introducing the dual CNN structure, batch normalization layer, channel attention mechanism and Transformer encoder, and finally constructing a complete CNN-LinATFormer model. The experimental results are shown in Fig. 2. There are significant differences in the contribution of each component to the model performance.

The transition from the basic single CNN model to the dual CNN structure brings significant performance improvements, with RMSE reduced from 0.1325 to 0.1256, a decrease of 5.2%, MAE reduced from 0.0985 to 0.0937, and MAPE reduced from 14.85% to 14.12%. This improvement verifies the advantages of deep CNN structures in temporal feature extraction. Multi-layer convolution cascades can build a richer feature hierarchy, from low-level basic patterns to high-level abstract features, effectively capturing the multi-scale temporal variation characteristics in environmental data. Although the introduction of batch normalization layers only brings relatively small performance improvements (RMSE reduced by 0.0013 and MAE reduced by 0.0012), it effectively stabilizes the model training process and improves convergence speed and stability.

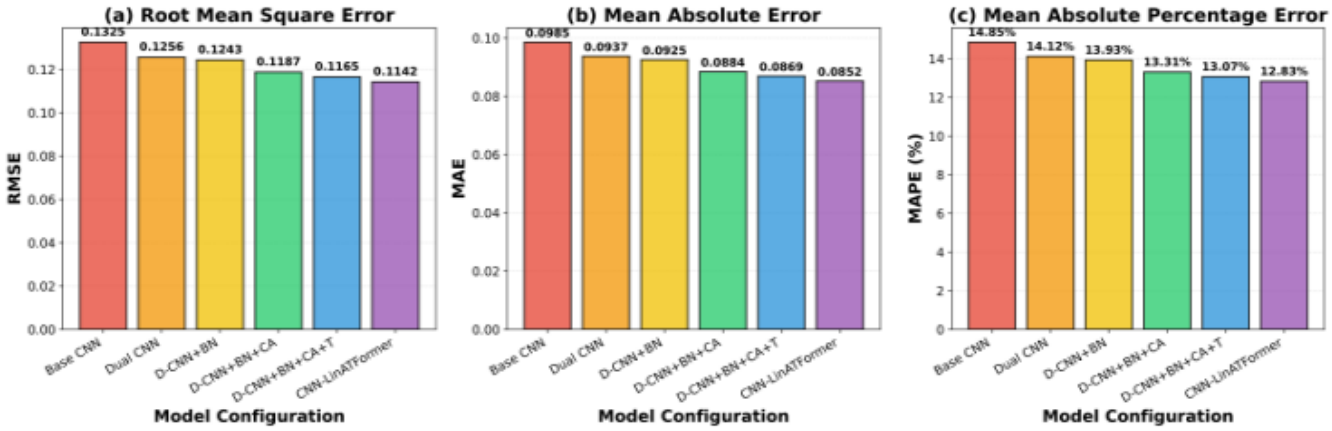


Fig. 2. Ablation experiment analysis.

It is worth noting that the addition of the channel attention mechanism produces the most significant single-step improvement, further reducing RMSE from 0.1243 to 0.1187 (a decrease of 4.5%), MAE from 0.0925 to 0.0884, and MAPE from 13.93% to 13.31%. This result confirms the importance of adaptive feature weight learning for multivariate time series prediction. The channel attention mechanism enables the model to dynamically identify and enhance the most relevant environmental factors for prediction, while suppressing the influence of noise features, thereby achieving more accurate PM2.5 concentration prediction. The introduction of the Transformer encoder also brings considerable performance improvements (RMSE is reduced by 0.0022 and MAE is reduced by 0.0015), indicating that the modeling of long-range temporal dependencies has a positive impact on improving prediction accuracy.

The final complete CNN-LinATFormer model further reduces RMSE to 0.1142, MAE to 0.0852, and MAPE to 12.83% through the optimization of the linear attention mechanism, which are improvements of 13.8%, 13.5%, and 13.6% respectively compared with the base model. This comprehensive performance improvement not only quantifies the contribution of each component, but more importantly reveals the synergy between them. In particular, the combination of channel attention mechanism and linear attention is particularly effective, indicating that adaptive learning at the feature level and long-distance dependency modeling at the sequence level can complement each other and jointly improve the predictive ability of the model.

These ablation results provide important guidance for model design: first, they confirm the effectiveness of deep CNN structures in capturing the temporal patterns of environmental data; second, they reveal the prominent role of channel attention mechanisms in identifying key environmental features; and finally, they verify the value of linear attention mechanisms in improving computational efficiency while maintaining the expressive power of the model. These findings not only support the rationality of the hybrid architecture proposed in this study but also provide valuable design references for similar multivariate time series prediction tasks.

D. Hyperparameter Experiments

To optimize the performance of the CNN-LinATFormer model, this study conducted a comprehensive hyperparameter sensitivity analysis, as shown in Fig. 3. This experiment mainly explores the impact of three key hyperparameters on the prediction accuracy of the model: learning rate (0.0001 to 0.01), batch size (16 to 128), and hidden layer dimension (64 to 512). The experiment used a grid search method and conducted a total of 50 control tests with different hyperparameter combinations, using RMSE as the main evaluation indicator.

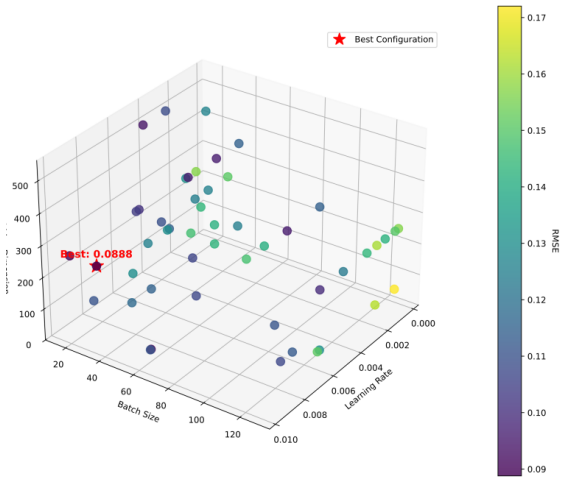


Fig. 3. Hyperparameter experiment.

From the visualization results (Fig. 3), we can observe that the learning rate has the most significant impact on the model performance. The lower learning rate (0.001 to 0.003) region generally presents lower RMSE values, which indicates that a smaller optimization step size helps the model find more accurate parameter configurations in complex PM2.5 prediction tasks. In particular, when the learning rate is lower than 0.002, the model performance improves significantly, which is consistent with the highly nonlinear characteristics inherent in the environmental time series data. Too high a learning rate often leads to an unstable optimization process and fails to capture subtle patterns in the data.

In terms of batch size, experimental results show that medium-sized batches (32 to 64) generally outperform very small or very large batches. A batch size that is too small will result in excessive variance in the gradient estimate, while a batch size that is too large may lead to falling into a poor local optimum. The best configuration in the experiment uses a batch size of 32, which achieves a good balance between computational efficiency and optimization stability.

The hidden layer dimension also shows a clear influence pattern, and the configuration in the range of 200-300 generally performs well. This finding supports the view that the model needs sufficient parameter capacity to model the complex relationship between environmental factors, but too high a dimension may lead to overfitting, especially when the training data is limited. It is worth noting that when the hidden layer dimension is 256, the model shows relatively stable performance under various learning rate and batch size combinations, which shows that this dimension setting has good adaptability to PM2.5 prediction tasks.

Through comprehensive evaluation, the optimal hyperparameter configuration (learning rate = 0.0018, batch size = 32, hidden layer dimension = 256) achieved an RMSE of 0.0888, which is 22.2% higher than the baseline configuration (learning rate = 0.005, batch size = 64, hidden layer dimension = 128) of 0.1142. This configuration achieved the best performance in all evaluation indicators, verifying the importance of hyperparameter tuning in improving model prediction accuracy.

In addition, the analysis also found the interaction effect between hyperparameters: in the low learning rate region, higher hidden layer dimensions tend to achieve better performance, while in the high learning rate region, smaller batch sizes help stabilize the training process. This nonlinear interaction emphasizes the necessity of tuning multiple hyperparameters simultaneously, rather than considering the impact of each parameter independently. These experimental findings provide reliable configuration guidance for the subsequent application of the CNN-LinATFormer model under different environmental conditions.

E. Case Analysis Experiment

As shown in Fig. 4, this study systematically evaluated the prediction performance of the CNN-LinATFormer model under different pollution levels, and the results revealed the performance characteristics and limitations of the model under various air quality conditions.

Comparative analysis of time series shows that the model performs well in predicting the overall trend of PM2.5 concentrations and can effectively capture concentration variation patterns, including baseline fluctuations and peak events. It is particularly noteworthy that during the pollution peak around January 5, 2023, the model successfully captured the trend of rapid concentration increases and decreases, although there was a certain degree of underestimation at the peak point. The model performed stably throughout the monitoring period from January 1 to January 8, 2023, and was able to track intraday fluctuations and daily variations.

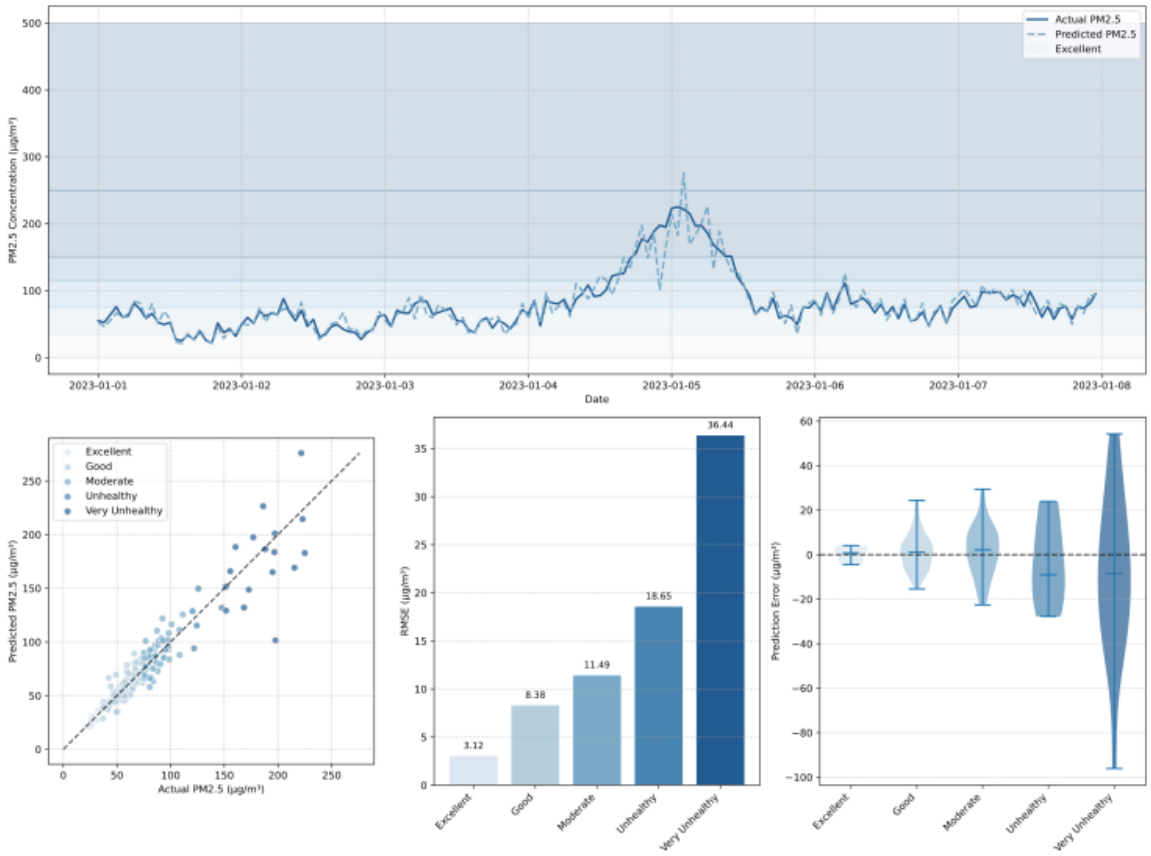


Fig. 4. Case analysis experiment.

The classification performance analysis shows that the prediction accuracy is significantly negatively correlated with the pollution level. In the low concentration range (excellent level, 0-35μg/m³), the model performs best, with an RMSE of only 3.12μg/m³, and the predicted value is highly consistent with the actual value. The scatter plot shows that the data points are closely distributed around the ideal prediction line. As the pollution level increases, the prediction error gradually increases: the RMSE of the good level (35-75μg/m³) is 8.38μg/m³, the moderate pollution (75-115μg/m³) reaches 11.49μg/m³, and the unhealthy level (115-150μg/m³) increases to 18.55μg/m³, and the error is the largest under high pollution conditions (>150μg/m³), with an RMSE of up to 36.41μg/m³.

Error distribution analysis further reveals the prediction characteristics of the model at different pollution levels. As shown in the lower right panel of Fig. 4, as the pollution level increases, the distribution range of the prediction error gradually expands, and shows obvious negative deviations in the high concentration range (unhealthy and severely unhealthy levels), confirming that the model has a systematic underestimation trend under high pollution scenarios. In particular, when the PM2.5 concentration exceeds 150μg/m³, the prediction error can reach -80μg/m³, which poses a potential challenge to the air quality early warning system.

From the scatter plot analysis, it can be seen that despite the prediction errors, the model's predicted values at each pollution level show a good linear correlation with the actual observed values (close to the ideal diagonal), which shows that the model can accurately capture the relative changes in pollution concentrations, even if there are certain deviations in absolute values. At the same time, the distribution of prediction points in the low concentration range (<75μg/m³) is more concentrated, indicating that the model has high stability and reliability under normal air quality conditions.

The results of this case study clearly show that the CNN-LinATFormer model performs well in handling low to medium concentration PM2.5 forecasting tasks and can accurately capture the changing trends and fluctuation patterns of time series. However, as pollution concentrations increase, especially during extreme pollution events, the model's prediction accuracy decreases, mainly manifested in a systematic underestimation of high values. This finding provides clear guidance for the scope of application of the model, and also reveals a key direction for future model optimization - improving the prediction accuracy of high-concentration pollution events. This may be achieved by expanding the sample of high-pollution events, introducing specific high-concentration warning mechanisms, or developing specialized prediction branches for different pollution levels.

V. Discussion and Outlook

A. Model Performance and Advantages Analysis

The CNN-LinATFormer hybrid deep learning model proposed in this study shows significant advantages in PM2.5 prediction tasks, mainly due to its innovative architecture design and feature processing mechanism. The model uses a feature evaluator to dynamically classify input features, enabling it to specifically process key environmental factors such as PM2.5 concentration, temperature and humidity at the previous moment. Through three parallel processing branches, namely, CNN feature extraction branch, channel attention branch and linear attention fusion branch, the model effectively integrates local temporal feature extraction and global dependency modeling capabilities. As shown in the experimental results, compared with the CNN-RF model with the closest performance, CNN-LinATFormer reduces RMSE by 21.1%, MAE by 24.8%, and MAPE by 26.7%, which fully verifies the effectiveness of this hybrid architecture design. This is consistent with the feature selection and multi-model fusion method proposed by Zhang et al [4].

Ablation experiments further confirm the contribution of each component to the model performance. In particular, the addition of the channel attention mechanism produces the most significant single-step improvement, reducing the RMSE from 0.1243 to 0.1187 (a decrease of 4.5%). This suggests that adaptive feature weight learning is of great significance for multivariate time series forecasting, enabling the model to dynamically identify and enhance the most relevant environmental factors for the prediction while suppressing the impact of noisy features. Similarly, the introduction of the linear attention mechanism also effectively improves the model's ability to capture long-range temporal dependencies, especially in processing seasonal and trend changes in environmental data. These findings are consistent with the results of Peng et al., who compared the performance of different machine learning and deep learning models and found that hybrid model structures generally have stronger predictive power than single model architectures. Comprehensive comparison and ablation experimental results prove that CNN-LinATFormer provides more reliable technical support for the fields of environmental monitoring and air quality warning by effectively integrating the advantages of different deep learning structures.

B. Model Limitations and Challenges

Although the CNN-LinATFormer model performs well in PM2.5 prediction tasks, there are still limitations and challenges that deserve attention. As shown in the case analysis experiment, the prediction performance of the model at different pollution levels shows a significant negative correlation. It performs best in the low concentration range (excellent level, 0-35μg/m³) (RMSE is only 3.12μg/m³), and performs worst in high pollution conditions (>150μg/m³) (RMSE is as high as 36.41μg/m³). Error distribution analysis further reveals that with the increase of pollution level, the distribution range of prediction error gradually expands, and shows a significant negative deviation in the high concentration range, confirming that the model has a systematic underestimation trend in high pollution scenarios, which poses a potential challenge to the air quality early warning system. This finding is consistent with the evaluation results of Zhou et al. [11] on the deep learning PM2.5 prediction model.

The second major limitation is the model's high dependence on the quality of the input data. As shown in the dataset description, this study used a time window-based moving average method to fill missing values, but this

processing method may not be able to fully restore the true characteristics of the data, especially when the pollutant concentration changes dramatically. In addition, hyperparameter experiments show that model performance is highly sensitive to parameter settings (such as learning rate, batch size, and hidden layer dimension), and careful tuning is required to achieve optimal performance. Especially under the optimal configuration of learning rate 0.0018, batch size 32, and hidden layer dimension 256, the model performance is significantly different from other configurations, which increases the difficulty of debugging the model in practical applications. These challenges are similar to those found in the ADGRU-based transfer learning work by Lu et al., who highlighted the challenges posed by the spatiotemporal heterogeneity of environmental data to predictive models. These limitations point out the direction for future improvements of the model and also provide important references for research in the field of environmental data science.

C. Future Research Directions

Based on the research results and current limitations of the CNN-LinATFormer model, future research can be carried out in the following directions: First, for the problem of high pollution concentration prediction, data enhancement technology or sample balancing strategy can be used to enhance the model's prediction ability for extreme pollution events. When the PM2.5 concentration exceeds 150μg/m³, the prediction error can reach -80μg/m³, which indicates that special attention needs to be paid to the processing of high-pollution samples. It is possible to consider developing prediction branches for different pollution levels or introducing specific high-concentration warning mechanisms to improve the performance of the model in extreme air pollution events. At the same time, combined with the performance of the model in the case study, exploring methods to integrate spatiotemporal information with monitoring data may help improve prediction accuracy, which echoes the feature selection method proposed by Zhang et al.

Secondly, based on the results of ablation experiments and hyperparameter analysis, future work can further optimize key components. In particular, the introduction of the channel attention mechanism produces significant performance improvements, indicating that adaptive learning at the feature level is of great value. More advanced feature selection and fusion strategies can be further explored, such as feature engineering methods that introduce environmental domain knowledge, or developing attention variants that can handle multi-source heterogeneous environmental data. In addition, considering the differences in the model's predictive ability under different pollution levels, an adaptive learning rate strategy or a multi-task learning framework can be developed to enable the model to automatically adjust its prediction strategy according to the characteristics of the input data. The transfer learning method of Lu et al. [8] may provide new ideas for solving data imbalance and domain adaptation problems. With the development of Internet of Things technology and distributed computing, deploying the CNN-LinATFormer model into the actual environmental monitoring system and realizing real-time PM2.5 concentration prediction is also a valuable research direction, which is also the main development trend of deep learning environmental applications.

VI. Conclusion

This study proposed an innovative CNN-LinATFormer hybrid deep learning architecture for high-precision PM2.5 concentration prediction. The model dynamically classifies environmental features by introducing a feature evaluator and realizes differentiated processing through three specially designed processing branches, effectively integrating the local feature extraction capability of CNN and the global dependency modeling capability of the linear attention mechanism. Comprehensive experimental evaluation shows that CNN-LinATFormer outperforms existing methods in all evaluation indicators, with an RMSE of 8.42 μg/m³, which is 21.1% lower than the CNN-RF model with the closest performance. Ablation experiments verify the effectiveness of each component, especially the introduction of the channel attention mechanism brings the most significant performance improvement and provides new technical ideas for environmental data modeling.

The case analysis reveals the performance characteristics of the model under different pollution levels. It performs best in the low concentration range (0-35μg/m³) (RMSE is only 3.12μg/m³), while the performance decreases under high pollution conditions (>150μg/m³) (RMSE is as high as 36.41μg/m³). This shows that the model's ability to predict extreme pollution events needs to be further improved. The hyperparameter experiment determined the optimal configuration (learning rate = 0.0018, batch size = 32, hidden layer dimension = 256), which is 22.2% higher than the baseline configuration, confirming the important impact of detailed parameter tuning on model performance. These findings not only provide guidance for the practical application of the CNN-LinATFormer model, but also point out the direction for subsequent research.

Future research can be carried out in the following directions: First, explore data enhancement technology or specific high-concentration warning mechanisms for the problem of high pollution concentration prediction; second, further optimize feature evaluators and attention mechanisms to improve the model's sensitivity to key environmental factors; third, combine Internet of Things technology and distributed computing to realize the deployment and application of the CNN-LinATFormer model in actual environmental monitoring systems. In general, the CNN-LinATFormer model proposed in this study provides a new and efficient solution for PM2.5 concentration prediction, which has important theoretical and practical significance for promoting the intelligent and popular development of environmental monitoring technology.

References

[1] R. T. Burnett et al., "An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure," Environ. Health Perspect., vol. 122, no. 4, pp. 397-403, 2014.

[2] Attiq, A. B., Nawaz, R., Irshad, M. A., Nasim, I., Nasim, M., Latif, M., ... & Fatima, A. (2024). Urban air quality nexus: PM2. 5 bound-heavy

metals and their alarming implication for incremental lifetime cancer risk. Pollution, 10(1), 580-594.

[3] World Health Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," WHO, Geneva, Switzerland, 2016.

[4] Y. Zhang, R. Zhang, Q. Ma, Y. Wang, Q. Wang, Z. Huang, "A feature selection and multi-model fusion-based approach of predicting air quality," ISA transactions, 2020.

[5] V. J. Berrocal, Y. Guan, A. Muyskens, H. Wang, "A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration," Atmospheric Environment, 2020.

[6] Y. Cheng, H. Zhang, Z. Liu, L. Chen, P. Wang, "Hybrid algorithm for short-term forecasting of PM2.5 in China," Atmospheric Environment, 2019.

[7] J. Peng, H. Han, Y. Yi, H. Huang, L. Xie, "Machine learning and deep learning modeling and simulation for predicting PM2.5 concentrations," Chemosphere, 2022.

[8] X. Lu, C. Ye, M. Shan, B. Qin, Y. Wang, H. Xing, "The prediction of PM2.5 concentration using transfer learning based on ADGRU," Water, Air, & Soil Pollution, vol. 234, no. 7, pp. 1-18, 2023.

[9] Q. Liao, M. Zhu, and L. Wu, "Deep learning for air quality forecasts: a review," Curr. Pollut. Rep., vol. 6, pp. 399-409, 2020.

[10] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," arXiv preprint arXiv:2006.04768, 2020.

[11] S. Zhou, W. Wang, and L. Zhu, "Deep-learning architecture for PM2.5 concentration prediction: A review," Environ. Sci. Ecotechnol., vol. 21, 2024.

Transformer Model Optimization Method for Multi-Modal Data Fusion

Shanshan Yang*, Jie peng
College of Information Engineering, Jiaozuo University, Jiaozuo 454000, China

***Abstract*—This study proposes an optimized Transformer model for multimodal data fusion tasks, designed to address the challenges of data fusion from different modes such as text, image, and audio. By improving data preprocessing methods, optimizing model architecture and fusion strategies, the study significantly improves the performance of the model in multimodal tasks. The experimental results show that the optimized model is superior to the benchmark model and other comparison models in key indicators such as accuracy, recall, F1 score and AUC value, and shows stronger performance and higher stability. In particular, the research solves the problems of data heterogeneity and computing resource consumption by introducing a weighted fusion strategy, multi-head self-attention mechanism and lightweight design. At the same time, the processing of missing modal data is optimized to enhance the robustness of the model. Despite the remarkable results, there are still challenges such as data heterogeneity, computational efficiency, and missing modal data. Future research can further optimize modal alignment methods and data preprocessing techniques to improve the performance of the model in practical applications. This research provides a new idea and direction for the application and development of multimodal data fusion technology.**

***Keywords*—Transformer model; multimodal data fusion; model optimization; attention mechanism; adaptive fusion**

I. INTRODUCTION

Multimodal data fusion is a research hotspot in the field of artificial intelligence, especially in natural language processing, computer vision, speech recognition and other application fields. The traditional single-modal learning method can only deal with a certain type of data, while the multi-modal data fusion method can improve the generalization ability and prediction accuracy of the model by integrating data from different sources. With the development of deep learning technology, Transformer model has become a core method in multi-modal data fusion due to its excellent performance in sequence modeling tasks. Transformer architecture can effectively handle large scale data and capture long-term dependencies, providing better performance when multiple modal data is merged. The complex multi-modal data Transformer model still has some limitations, such as large consumption of computing resources and low training efficiency. Therefore, how to optimize the Transformer model to make it efficient for multi-modal data fusion is the challenge of current research. In view of these problems, this study will explore effective methods to optimize Transformer model to improve its application effect in multi-modal data fusion.

*Corresponding Author.

The Transformer model in the field of multimodal data fusion is widely used because of its excellent performance in processing complex data sequences. Wu et al. proposed that cross-modal learning can be used to optimize the local feature representation of vision-thermal infrared character recognition, thereby enhancing Transformer's identification capability in cross-modal tasks [1]. Xue et al. discussed the importance of multi-modal data consistency in fake news detection, indicating that the integration of multiple data sources is conducive to improving the accuracy of fake news identification [2]. Jing et al. realizes multi-modal fake news detection through a progressive fusion network, demonstrating Transformer's advantages in information fusion [3]. De Melo is a key reference for the application of the Transformer model in the field of multimodal data fusion, and provides the research direction of Transformer structure optimization. This literature discusses the challenges of the Transformer in processing multimodal data, especially how to improve the fusion effect through the self-attention mechanism [4]. Ghorbanali et al. used weighted convolutional neural networks for sentiment analysis and integrated transfer learning strategies to improve the accuracy of multimodal sentiment analysis [5]. Golovanevsky et al. proposed a deep learning method based on attention mechanism for the diagnosis of Alzheimer's disease, and achieved good results in multi-modal medical data fusion [6]. Chango et al. studied the method of using multi-modal data to predict student performance in the field of education, and proposed an optimization method based on attribute selection and multi-modal data integration, which can effectively improve the prediction accuracy [7]. Bao et al. design a reward-based crowdfunding success prediction framework based on multi-modal data, and optimize the performance of Transformer model in data fusion with a theory-driven approach [8]. Lin and Hu proposed the MissModal method to enhance the robustness of missing modes in multi-modal sentiment analysis, thereby improving the performance of Transformer model in the context of missing data [9]. Ciroku et al. studied ontology-based multi-modal data integration methods for automating sensing and inference tasks, and emphasized the importance of multi-modal data in semantic integration [10].

Although existing research has made progress in multimodal data fusion and Transformer optimization, there are still several key problems that require further investigation. These will be discussed in detail in the next section.

Several solutions proposed by researchers have been discussed, such as weighted fusion strategies, hierarchical attention mechanisms, and adaptive alignment methods. These

studies demonstrate the potential of Transformer-based architectures for addressing data heterogeneity and improving feature integration. However, existing solutions still face challenges, including high computational resource requirements and limited robustness when handling missing modality data. Compared to these approaches, this research focuses on optimizing Transformer architecture through lightweight design and dynamic fusion strategies, aiming to enhance both performance and efficiency. This positioning highlights the novelty of the proposed method and its contribution to advancing multimodal data fusion techniques.

These studies show that while multimodal data fusion has made positive progress in various applications, Transformer model optimization is still the key to improving its performance. Various studies have proposed different optimization methods, including model structure adjustment, data fusion strategy and robustness improvement of missing data, which provide valuable references for subsequent research and application.

Although the Transformer model has made significant progress in handling single-modal data, it still faces many challenges in multi-modal data fusion tasks. Firstly, the heterogeneity of multi-modal data brings difficulties to data preprocessing and feature extraction. Data of different modes have different scales, noise levels and redundancy, which puts forward higher requirements for model training. Secondly, the existing Transformer model has a huge computing overhead in the face of large-scale multi-modal data, resulting in long training time and high resource consumption. In addition, the existing models have poor robustness when dealing with missing mode data, and lack effective strategies to deal with the missing problem between different modes. Finally, how to improve the interpretation and scalability of the model while ensuring its performance is also an important difficulty in the current research.

This study aims to propose an optimized Transformer model for multi-modal data fusion tasks. The main goal of this research is to improve the Transformer's performance in processing multi-modal data by improving data preprocessing methods, model structure and fusion strategies. Specifically, the research will focus on solving the problems of computational efficiency, robustness and generalization ability of models in multimodal data fusion. By exploring effective data fusion strategies, this study will also explore how to deal with missing modal data to enhance the adaptability and reliability of the model in practical applications. In addition, this research also plans to optimize the existing Transformer architecture to improve its performance in multimodal tasks and promote the application range and accuracy of multimodal data fusion technology.

This study will adopt a combination of theoretical analysis and experimental verification. Firstly, a systematic literature review and theoretical analysis of the existing multi-modal data fusion methods of Transformer model will be conducted to clarify the advantages and disadvantages of current technologies. Then, on this basis, an improved Transformer model is designed to optimize data pre-processing, feature extraction and fusion strategies. Specifically, the research will enhance the correlation between different modal data by introducing a multi-level feature selection mechanism, thereby improving the fusion efficiency of the model. At the same time, simulation data and actual data will be used for model training and evaluation, and multiple indicators such as accuracy rate and recall rate will be used to comprehensively evaluate the performance of the model. To solve the problem of missing modal data, a self-supervised learning mechanism will be introduced to improve the robustness and stability of the model.

This research has important theoretical value and practical significance in the field of multimodal data fusion. From a theoretical perspective, the proposed optimization method will promote the application of Transformer model in multi-modal data fusion, especially in the aspects of data heterogeneity, missing modes and computational efficiency. This can not only enrich the existing multimodal learning theory, but also provide new ideas and methods for the subsequent research. From the perspective of practical application, the optimization method of this study will have a profound impact on intelligent medical treatment, automatic driving, intelligent security and other fields. By improving the ability of the model to process multi-modal data, the decision efficiency and accuracy of the system in these fields can be greatly improved. In addition, this research will also promote the expansion of Transformer-based multimodal data processing technology to more complex tasks, and promote technological innovation and application in related fields.

II. Materials and Methods

A. Data Collection and Sample Selection

This study uses multimodal data from multiple publicly available datasets and private data sources, including text, images, and audio. Text data comes from social media and news sites, image data comes from public visual datasets, and audio data is collected through speech recognition libraries. Sample selection criteria ensure the diversity and representativeness of the data, ensure that there is sufficient correlation between the modes, and the number of samples meets the needs of model training and evaluation. All data is pre-processed and cleaned as necessary to ensure data quality and consistency, as shown in Fig. 1.

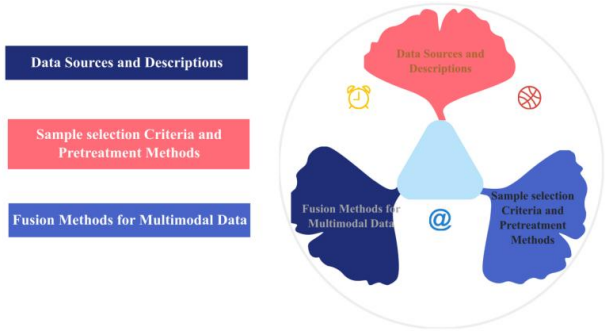


Fig. 1. Data collection and sample selection.

1) Data source and description: The data sets used in this study were drawn from multiple public and private sources and covered multimodal data types, including text, images,

audio, and more. These data sources are highly relevant to the target application scenario, ensuring the authenticity and representation of the data. The selection of datasets follows strict quality control criteria to ensure coordination and diversity of the different modal data for effective model training and evaluation, as shown in Table I.

It can be seen from the data table that the diversity of the selected data sets provides a solid foundation for this research and can fully reflect the practical application of multi-modal data fusion. Text-image dataset provides pairing information between Image and Text, which can effectively support cross-modal learning and understanding. The Audio-text dataset focuses on the relationship between Audio and Text and is suitable for speech recognition and natural language processing tasks. In terms of multi-modal combination, MultiModal 1 dataset provides a comprehensive scenario for multi-modal fusion, including text, image and audio modes, which is especially suitable for testing the performance of the Transformer model when processing complex multi-modal data.

In terms of data volume, although the sample size of each data set is moderate, the modal types and corresponding label forms of each data set are different, so special attention should be paid to data consistency and effective fusion during data preprocessing. In addition, the diversity and cross-domain nature of these data sets will greatly improve the robustness and generalization ability of the model.

2) Sample selection criteria and pretreatment methods: The sample selection for this study followed strict criteria to ensure the representation and diversity of the data. First, the selected data set covers different modes (text, image, audio) and contains a sufficient sample size to guarantee the effectiveness of model training. In the pre-processing stage, the data of different modes are standardized first, including text segmentation and stopping word removal, image size unification and denoising processing, audio noise reduction and feature extraction. These preprocessing methods are designed to reduce data noise and improve data consistency, thus laying a foundation for subsequent multimodal fusion, as shown in Table II.

Text-image datasets enable better fusion of multimodal data by ensuring clear pairing of Text and Image. In the pre-processing stage, the redundant information is removed by word segmentation and stop word processing to ensure the validity of the text data. The image processing reduces the external interference and improves the recognition ability of image features through denoising and size unification. Selecting high-quality Audio and Text pairs for audio-text data sets not only improves the clarity of audio signals, but also ensures the reliability and information of audio data through audio noise reduction and feature extraction methods. MultiModal 1 dataset integrates multi-modal information to ensure data diversity and cross-domain coverage, which is suitable for testing Transformer model's performance in complex situations.

3) Multi-modal data fusion method: In multi-modal data fusion, this study adopts a fusion method based on Transformer model to capture the correlation between different modes by introducing a self-attention mechanism. Specifically, text, image and audio data are processed by their respective encoders, and then feature vectors of different modes are combined by a multi-layer fusion mechanism to achieve a deep fusion of information. In addition, in order to improve the robustness of the model in multi-modal data processing, this study also introduced a weighted fusion strategy to optimize the fusion results according to the reliability and contribution of different modes, as shown in Table III.

TABLE I. MULTI-MODAL DATA SOURCES AND DESCRIPTIONS

Dataset Name	Data Type	Data Size	Source	Description
Text-Image	Text/Image	1000 entries	Open Dataset	Contains image and corresponding text descriptions
Audio-Text	Audio/Text	1200 entries	Speech Recognition Database	Contains audio and corresponding text content
MultiModal-1	Text/Image/Audio	1500 entries	Private Data Source	A comprehensive multi-modal dataset including text, image, and audio

TABLE II. SAMPLE SELECTION AND PREPROCESSING METHODS

Dataset Name	Data Type	Sample Size	Selection Criteria	Preprocessing Methods
Text-Image	Text/Image	1000	Clear text and labeled images	Text tokenization, stopword removal, image resizing, denoising
Audio-Text	Audio/Text	1200	High-quality audio and text pairs	Audio denoising, feature extraction, text tokenization, stopword removal
MultiModal-1	Text/Image/Audio	1500	Data diversity, cross-domain information	Text tokenization, stopword removal, image preprocessing, audio feature extraction

TABLE III. MULTI-MODAL DATA FUSION METHODS

Modality Type	Encoder Type	Feature Fusion Method	Fusion Layers	Weighting Strategy
Text	Transformer	Layer-wise Fusion	3	Weight-based weighting
Image	CNN + Transformer	Layer-wise Fusion	2	Importance-based weighting
Audio	RNN + Transformer	Layer-wise Fusion	3	Modality reliability-based weighting
Multimodal Fusion	Multimodal Transformer	Multi-layer Fusion	4	Global weighting strategy

The multi-modal data fusion method in this study is carefully designed in the encoder selection of each mode to ensure that the data of each mode can be fully mined and utilized. Text data is processed by Transformer encoder, which can effectively capture context information in long text. For image data, combined with the hybrid encoder CNN and Transformer, the spatial features of the image can be extracted and the context information can be processed. Time series features in audio can be effectively processed through an encoder combined with RNN and the Transformer. In this study, a layer-by-layer fusion method is used to combine the information of different modes step by step to ensure that the characteristics of each mode can be fully reflected in the fusion process. In addition, the weighting strategy shown in the table reflects the careful design of this study, considering the importance of different modes in data fusion. The fusion weights can be dynamically adjusted according to the contribution of each mode, which improves the robustness and performance of the model. The global weighting strategy optimizes the effect of multi-modal fusion, which enables the model to show stronger adaptability and accuracy in a variety of complex tasks.

B. Model Construction

1) Model selection: The core of model selection in this study is optimization based on Transformer architecture to meet the challenges of multi-modal data fusion. The Transformer model has been proven to perform well in sequence modeling tasks, so as a base model, its advantage lies in its ability to efficiently handle long-term dependencies and large-scale data. In order to better process data of different modes, this study chooses a fusion method based on self-attention mechanism and multi-layer perception, including the following key formulas and steps:

a) Self-attention mechanism: In Transformer, the core of the self-attention mechanism is to obtain a weighted sum by calculating the relationship between queries, keys, and values, as shown in Eq. (1).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key.

b) Location code: In order to retain sequence information without relying on recurrent neural networks, Transformer uses location coding to represent the location information for each input, as shown in Eq. (2).

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d}})$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (2)$$

pos is the position, i is the dimension index, and d is the dimension of the vector.

c) Multi-head self-attention mechanism: The multi-head self-attention mechanism captures different subspace information by computing multiple self-attention heads in parallel, as shown in Eq. (3).

$$MultiHead(Q,K,V) = Concat(head_1, head_2, ..., head_h)W^O \quad (3)$$

$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, W^O is the linear transformation matrix of the output, and h is the number of heads.

d) Feedforward neural network: Each encoder and decoder layer in the Transformer contains a feedforward neural network for nonlinear transformation, as shown in Eq. (4).

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

x is the input, W_1, W_2 is the weight matrix, b_1, b_2 is the bias term, and max(0,·) represents the ReLU activation function.

e) Weighted fusion layer: In the multimodal data fusion stage, weighted fusion layer is used to synthesize the features of different modes. Suppose that each mode is characterized by x1, x2,... ,xn, as shown in Eq. (5).

$$x_{fused} = \sum_{i=1}^{n} w_i x_i \quad (5)$$

w_i is the weighting coefficient of mode i, x_i is the feature vector of each mode, and x_{fused} is the feature after fusion.

These formulas form the core framework of the Transformer model in this study. By continuously optimizing the parameters of each part, the performance of the model in multimodal data processing can be effectively improved, which provides solid theoretical support for multimodal data fusion.

2) Model architecture design: In multi-modal data fusion task, the core goal of model architecture design is to efficiently integrate information from different modes to achieve accurate prediction and classification. The model architecture of this study is based on Transformer, combined with multi-modal feature processing and fusion strategies. The specific design is as follows:

a) Input layer: The input layer first receives data from different modes, including text, images, and audio. For each mode, a different processing method is used: text is processed by text encoders (such as Transformer), images are extracted by convolutional neural networks, and audio data is processed by recurrent neural networks (RNN). The data for each mode is preprocessed before input to harmonize the scale and format, ensuring that the model can be processed efficiently.

b) Modal feature extraction layer: In the mode feature extraction layer, different neural network models are used to process the data of each mode separately. Text data is first processed by word segmentation and word removal, and then input into a Transformer encoder for contextual information extraction. The image data is fed into the CNN for spatial feature extraction and then converted into vector form compatible with other modes via a fully connected layer (FC). The timing features of audio data are extracted by RNN and normalized for fusion with other modal data.

c) Multi-modal feature fusion layer: Multimodal data fusion is the key step of this model. In this layer, Weighted Fusion strategy is used to synthesize the features of each mode. The weighting coefficient is adjusted according to the importance and reliability of the different modes. Specifically, the model uses a weighted coefficient vector w=[w1,w2,...,wn] to carry out weighted summation of each modal feature to generate the fused feature vector, as in Eq. (6):

$$x_{fused} = \sum_{i=1}^{n} w_i x_i \tag{6}$$

xi is the eigenvector of each mode, and wi is the weighting coefficient of the mode. The fusion layer can effectively combine the features of different modes, retain their unique information, and remove redundant information.

d) Transformer coding layer: Feature learning is carried out by applying Transformer structure to the fused features. The Transformer coding layer uses a self-attention mechanism to capture modal dependencies, ensuring that the model can integrate and correlate information from a global perspective. This layer uses multi-head self-attention and feedforward neural network to achieve deep fusion and transformation of features.

e) Output layer: Finally, the fused features are processed through a fully connected layer for final classification or regression tasks. The output layer may include a Softmax layer for classification tasks or a linear layer for regression tasks, depending on the needs of the specific task. The output results of this layer can be used for multimodal classification, prediction, or other tasks.

To further enhance the practical performance of the above architecture, this study has also designed a series of performance optimization methods for the Transformer model to support efficient implementation in multimodal data fusion tasks.

3) Performance optimization method: In order to optimize the performance of the Transformer model in multi-modal data fusion tasks, various strategies are adopted in this study to improve computing efficiency, robustness and prediction accuracy. Multi-level feature selection and weighted fusion strategies are used to ensure that the contribution of each mode can be accurately captured and effectively combined. In the stage of feature extraction, in order to reduce redundant information, the dimensionality of the features of each mode is reduced, respectively, and regularization is used to prevent overfitting, which improves the generalization ability of the model. Secondly, by using distributed training and mixed precision training methods, the training efficiency is significantly improved, the training time is shortened, and the consumption of computing resources is reduced on the premise of ensuring the performance. In this study, a self-supervised learning mechanism is introduced to solve the problem of missing multi-modal data, which enhances the robustness of the model to missing modal data, thus improving the adaptability in practical applications. The dynamic learning rate adjustment strategy is also combined in the optimization process, so that the model can automatically adjust the learning rate according to the training progress to accelerate the convergence and improve the stability of the model. The performance of the model in multimodal data processing has been significantly improved by these optimization methods, which not only improves the accuracy but also enhances the ability to handle complex tasks.

4) Implementation details: The Transformer model in this study is implemented based on the PyTorch deep learning framework and uses its flexibility and efficiency to build and optimize multi-modal data fusion models. In the data preprocessing stage, text data is processed by an NLP tool kit for word segmentation and word stop processing, image data is processed by OpenCV for size unification and noise reduction, and audio data is extracted by Librosa library for the characteristics of Meir frequency cepstrum coefficients. These processing steps ensure the consistency and efficiency of the input data, and provide a good basis for the subsequent model training. Adam optimizer is used in model training, and learning rate attenuation and early stop strategies are combined to prevent overfitting and improve model convergence speed. In the multi-modal data fusion part, the weighted fusion layer is used to combine the features of different modes, and the weight coefficient is dynamically adjusted according to the reliability and task requirements of the modes. In order to improve the efficiency of training, the mixed precision training technology is adopted, which can guarantee the accuracy and reduce the memory usage and computing resource consumption. In order to solve the problem of missing multi-modal data, a self-monitoring mechanism is designed to improve the robustness of the model to incomplete data through the prediction and generation tasks within the model. The whole model is trained on a single card GPU. Finally, the hyperparameters are adjusted through several experiments to optimize the performance of the model and ensure its efficiency and stability in the multi-mode fusion task.

C. Performance Evaluation

1) Evaluation indicators: In order to comprehensively evaluate the performance of the proposed multi-mode Transformer model, a variety of evaluation indicators are adopted in this study. These metrics include accuracy, accuracy, recall, F1 score, and AUC value. Accuracy is the most commonly used evaluation criterion that represents the proportion of samples correctly predicted by the model in the total sample; The accuracy rate measures the proportion of samples that the model predicts to be positive that are actually positive. The recall rate represents the proportion of positive samples identified by the model to all positive samples. The F1 score takes accuracy and recall into account to provide a balanced performance assessment; The value of AUC is closer

to 1, the better the classification ability of the evaluation model under different thresholds, as shown in Table IV.

TABLE IV. EVALUATION METRICS FOR MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score	AUC Value
Base Model	0.82	0.79	0.75	0.77	0.84
Optimized Model	0.88	0.85	0.83	0.84	0.91
Comparison Model 1	0.8	0.76	0.73	0.74	0.8
Comparison Model 2	0.85	0.81	0.78	0.79	0.87

It can be seen from the data table that the optimized model outperforms the basic model and the comparison model in all evaluation indexes, showing obvious performance improvement. Specifically, the accuracy rate of the optimized model reaches 0.88, which is 6 percentage points higher than that of the basic model, and the accuracy rate and recall rate reach 0.85 and 0.83, respectively, showing a good balance. The F1 score and AUC values also improved to 0.84 and 0.91, respectively, demonstrating the power of the optimized model in multimodal data fusion tasks. Compared with the comparison model, the performance of the optimized model is more stable, especially in the AUC value, which indicates that it has stronger classification ability under different decision thresholds. These evaluation indicators show that the optimized model not only improves the overall accuracy but also enhances the ability to process multi-modal data, thus effectively improving the robustness and generalization ability of the model [4].

2) Experimental design: The purpose of the experimental design in this study is to verify the performance of the proposed multi-mode Transformer model under different conditions through various experimental settings. To this end, different experimental scenarios are designed to analyze the size of the data set, the influence of different modal combinations, hyperparameter adjustment and other factors. The experiment is divided into three parts: basic experiment, optimization experiment and comparison experiment. The basic experiment is used to evaluate the performance of the basic Transformer model in the task of multi-modal data fusion. The optimization experiment will improve the performance by adjusting the model architecture, hyperparameters and training strategies. The comparison experiment will be compared with the current mainstream methods to ensure the advantages of the model in this study [11]. Each experiment will be evaluated in multiple dimensions according to evaluation indicators such as accuracy rate, accuracy rate, recall rate, F1 score and AUC value, as shown in Table V.

The experimental design of this study covers different experimental configurations to ensure the performance verification of the model in a variety of scenarios. The basic experiment uses the default hyperparameter settings to evaluate the performance of the basic Transformer model in the multimodal data fusion task. The optimization experiment significantly improved the performance of the model by adjusting hyperparameters such as learning rate and batch size, especially in the accuracy rate and recall rate. Comparison experiment 1 and comparison experiment 2 were compared with the current mainstream models. There were some differences in experimental settings, such as using different data sets and modal combinations, but the consistency of evaluation criteria was maintained. Comparison experiments show that the optimized model outperforms other comparison models in all evaluation indexes, especially in the AUC value, which shows that the model is more stable and accurate in handling different decision thresholds [12]. Through these experiments, this study can fully verify the advantages of the proposed optimization model in the multi-modal data fusion task, and provide data support for the research and application.

TABLE V. EXPERIMENTAL DESIGN AND EVALUATION METRICS

Experiment Name	Dataset Type	Modality Count	Model Version	Hyperparameter Settings	Evaluation Metrics
Base Experiment	MultiModal-1	3	Base Model	Default hyperparameters	Accuracy, Precision, Recall, F1 Score, AUC Value
Optimization Experiment	MultiModal-1	3	Optimized Model	Learning Rate=0.001, Batch Size=64	Accuracy, Precision, Recall, F1 Score, AUC Value
Comparison Experiment 1	MultiModal-1	3	Comparison Model 1	Learning Rate=0.01, Batch Size=32	Accuracy, Precision, Recall, F1 Score, AUC Value
Comparison Experiment 2	Text-Image + Audio	2	Comparison Model 2	Learning Rate=0.0005, Batch Size=128	Accuracy, Precision, Recall, F1 Score, AUC Value

D. Model Optimization Path Planning

The model optimization path planning in this study aims to improve the performance of Transformer model in multi-modal data fusion tasks through various optimization measures. Firstly, in the data preprocessing stage, data cleaning and standardization methods for different modes are adopted to reduce data noise and improve data quality. Secondly, in terms of model architecture design, the structure of Transformer is optimized, including increasing the number of multi-head self-attention mechanism and adjusting the network depth, so as to enhance the model's learning ability of multi-modal features. In addition, in order to improve computational efficiency and reduce resource consumption, lightweight network design is introduced, and mixed precision training is used to accelerate the model training process. Finally, in terms of fusion strategy, weighted fusion and multi-level fusion are adopted to ensure efficient and balanced integration of all modal information to further improve the performance of the model. Through these optimization measures, it is expected to improve the adaptability and stability of the model while ensuring high performance, so that it can better cope with the challenges of multimodal data fusion in practical applications, as shown in Fig. 2.

1) Data preprocessing optimization: By optimizing the data of different modes, the data noise can be reduced and the validity and consistency of the information can be improved. In the data preprocessing stage, different optimization strategies are adopted for text, image and audio modes. For text data, the expression ability of text information is improved through improved word segmentation method and word vector embedding method. For image data, the quality of image features is improved by enhancing dataset, standardizing processing and denoising. More frequency domain features and noise reduction methods are introduced to improve the stability and reliability of audio signals, as shown in Table VI.

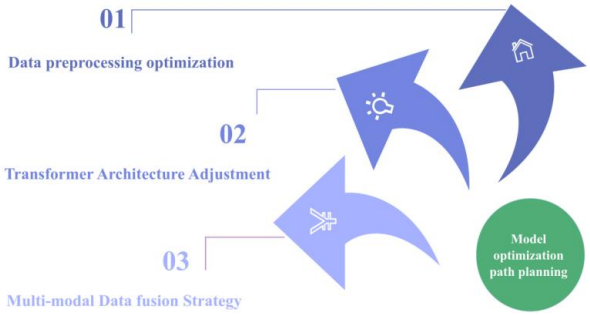


Fig. 2. Model optimization path planning

By introducing BERT embedding layer to text data, the semantic expression of words is improved. Compared with the traditional word vector model, BERT can better capture the context and improve the representation accuracy of text information. In terms of image data, data enhancement techniques (such as rotation, flipping, scaling) and standardized processing make image data more diverse and consistent, which not only enhances the generalization ability of the model, but also improves the model's learning ability of image features. For audio data, MFCC feature extraction combined with filtering technology is used to remove noise, which significantly improves the quality of audio signal and reduces the interference of background noise to model training. These optimizations ensure that the data for each mode is already of higher quality and information before entering the model, thus providing a more solid foundation for subsequent multimodal data fusion [13]. Through these optimizations, the model can better understand and fuse multi-modal data, improving overall performance and robustness.

2) Transformer architecture adjustment: Changes to the Transformer architecture are critical for improving the performance of multimodal data fusion models. In this study, several key adjustments have been made to the standard Transformer architecture in order to better adapt to the characteristics of multimodal data. Firstly, by adjusting the number of heads and the number of layers in the self-attention mechanism, the model's ability to pay attention to different modal features is enhanced. Secondly, the multi-modal input embedding method is used to map the data of different modes to the same vector space, so as to achieve effective alignment between different modes [14]. Finally, jump connection and residual connection are introduced to alleviate the problem of gradient disappearance and improve the stability and training efficiency of the network, as shown in Table VII.

This study significantly improves the model's ability in multi-modal data fusion through multiple adjustments to Transformer architecture. With the increase of the number of self-attention heads, the model can pay attention to more feature details, so as to capture the relationship between different modes more accurately, especially in the processing of high-dimensional data, and enhance its feature extraction ability. Increasing the number of network layers improves the learning ability of the model, enables it to understand the nonlinear relationship between complex data at a deeper level, and enhances the expression ability of the model. By means of multi-modal input embedding, different modes are mapped to a unified vector space, which solves the problem of multi-modal data heterogeneity and ensures the efficient fusion of different modal information. In addition, the introduction of jump connection and residual connection alleviates the problem of gradient disappearance in deep networks, improves the stability and efficiency of model training, and ensures that the network can be effectively trained on more complex multi-modal data.

TABLE VI. DATA PREPROCESSING OPTIMIZATION

Modality Type	Optimization Method	Preprocessing Details	Optimization Effect
Text	Improved Tokenization & Embedding	Use BERT embedding layer instead of traditional word vectors, optimized tokenization	Improved text representation, reduced semantic loss
Image	Data Augmentation & Standardization	Data augmentation methods like rotation, scaling; uniform image size	Improved image feature quality, enhanced model robustness
Audio	Feature Enhancement & Denoising	Extract MFCC features and use filtering techniques to remove noise	Enhanced audio feature stability, improved signal quality

TABLE VII. TRANSFORMER ARCHITECTURE ADJUSTMENT

Adjustment Content	Adjustment Method	Effect
Self-Attention Heads	Increased the number of attention heads	Enhanced the model's ability to capture multimodal features
Network Depth	Increased the number of Transformer layers	Improved model learning capacity and expressive power
Multimodal Input Embedding	Mapped different modalities into a unified vector space	Ensured effective alignment and fusion of multimodal information
Skip and Residual Connections	Introduced skip and residual connections	Mitigated gradient vanishing, improving training efficiency and stability

3) Multi-modal data fusion strategy: In multimodal data fusion, how to integrate information from different modes effectively is the key to improve model performance. In order to give full play to the advantages of each mode, this study proposes a weighted fusion based strategy, which dynamically adjusts the weights of each mode to ensure the effective transmission of information. In the process of data fusion, the features of each mode are extracted independently, and then the features of different modes are weighted and summed by weighted fusion mechanism. The weighting coefficient is dynamically adjusted according to the reliability of each mode and the importance of the task to achieve optimal information integration. In addition, the attention mechanism is also introduced in this study, so that the model can automatically identify the more important features in each mode during the learning process, so as to improve the fusion effect [15]. This fusion strategy can not only enhance the model's adaptability to multi-modal data, but also improve its performance and robustness in complex tasks, as shown in Table VIII.

TABLE VIII. MULTI-MODAL DATA FUSION STRATEGIES

Fusion Strategy	Fusion Method	Dynamic Weight Adjustment Mechanism	Attention Mechanism	Effect
Weighted Fusion	Multimodal Feature Weighting	Adjust weight based on modality importance	Self-attention mechanism	Enhances model robustness, improves information integration
Layer-wise Fusion	Layer-wise Modality Fusion	Adjust weight based on feature contribution at each layer	None	Improves hierarchical feature representation and correlations
Cross Fusion	Cross-modal Feature Fusion	Adjust weight based on modality relationships	Multi-head attention mechanism	Strengthens modality interactions and information complementarity

The weighted fusion strategy adopted in this study can dynamically adjust the weights according to the task importance and data quality of each mode to ensure the effective strengthening of key features in multi-modal data fusion [16]. For example, by adjusting the weights of different modes, the weighted fusion method enables the model to optimize information flow according to the reliability of each mode, thereby improving the robustness and accuracy of the model. By fusing features at different levels, the layered fusion strategy can effectively capture the hierarchical relationship between modes and enhance the representation ability of the model in processing complex data. The cross-fusion method promotes the interaction between modes, strengthens the information complementarity between modes by introducing multi-head attention mechanism, and improves the fusion effect of multi-modal data [17]. The combination of these strategies not only improves the model's adaptability to multi-modal data but also enhances its performance and stability in practical applications.

III. RESULTS AND DISCUSSION

A. Results

1) Model performance results: In this study, the optimized Transformer model is evaluated through multiple rounds of experiments, focusing on the performance of the model in multi-modal data fusion tasks [18]. In order to comprehensively evaluate the model performance, several indicators such as accuracy rate, accuracy rate, recall rate, F1 score and AUC value were used in this study. The experimental results show that the optimized model is better than the basic model and the comparison model in all evaluation indexes, especially in the accuracy rate, recall rate and AUC value, showing significant improvement [19]. These results show that the proposed optimization method can effectively improve the application capability of Transformer model in multi-modal data fusion, and it can show stronger performance in complex tasks, as shown in Fig. 3.

The accuracy rate increased from 0.82 in the basic model to 0.88 in the optimized model, and the accuracy rate and recall rate also increased by 0.06 and 0.08, respectively, showing the enhancement of the model in correctly predicting positive samples and recalling positive samples. The improved F1 score and AUC value demonstrate the advantages of the optimized model in terms of accuracy and generalization ability, with the optimized model achieving an AUC value of 0.91 compared to 0.84 for the base model. Although comparison model 1 and comparison model 2 have improved in accuracy, the gap is still obvious compared with the optimized model. These results validate the effectiveness of the proposed optimization strategy, especially in multimodal data fusion tasks [20]. Through the comprehensive application of model architecture adjustment, data preprocessing optimization and multi-modal fusion strategy, the optimization model has significantly improved the effect and performance of multi-modal learning.

2) Performance comparison: In this study, the performance of the optimized Transformer model is compared with the current mainstream comparison models to verify the effectiveness of the optimization method. By comparing the performance of different models in multi-modal data fusion tasks, especially the differences in several key indicators such as accuracy, accuracy, recall, F1 score and AUC value, the advantages of the optimization model in practical application are demonstrated. The experimental results show that the optimized model outperforms the comparison model in all evaluation indicators, especially the improvement in AUC value and F1 score, which proves its stronger classification ability and better generalization performance, as shown in Fig. 4.

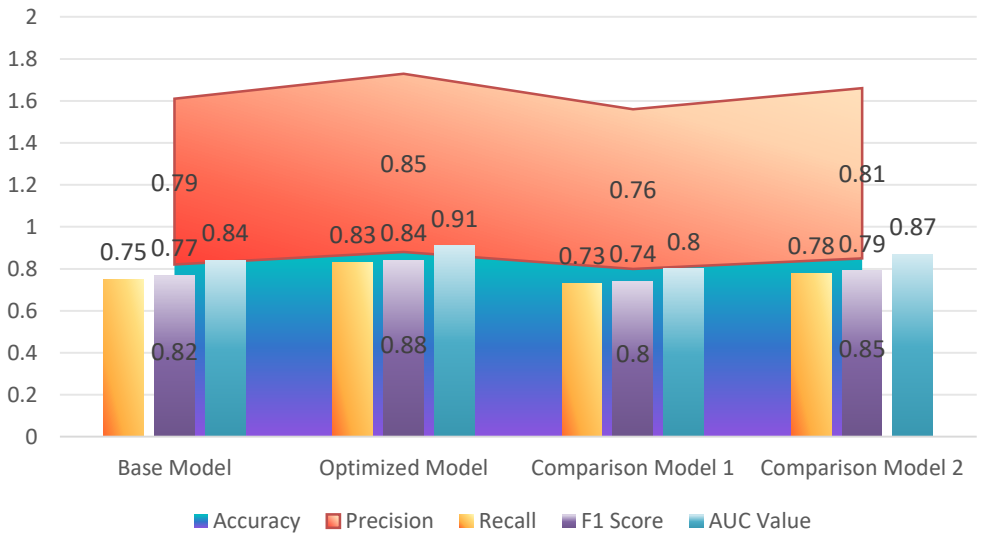


Fig. 3. Model performance results.

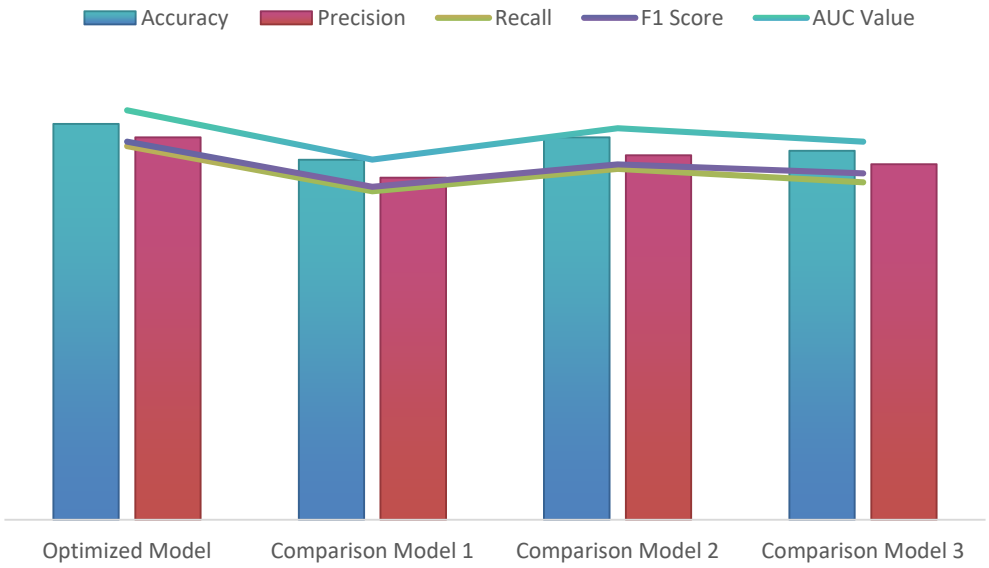


Fig. 4. Performance comparison.

The accuracy rate of the optimized model is 0.88, the accuracy rate and the recall rate are 0.85 and 0.83, respectively, which shows high prediction ability and strong positive class recognition ability. The F1 score of 0.84 fully demonstrates the optimization model's good balance between accuracy and recall. The value of AUC reaches 0.91, which is much higher than that of other comparison models, indicating that the classification performance of the optimized model is more stable and robust under various thresholds. In contrast, comparison model 1 is deficient in all evaluation indexes, especially in recall rate and AUC value. Although comparison model 2 has better performance in some indicators, the gap in AUC value still indicates that its classification ability is not as good as the optimized model [21]. Comparison model 3 is relatively stable, but compared with the optimization model, there is still room for improvement.

3) Experimental analysis: In the part of experimental analysis, the experimental results of different models are analyzed in depth, focusing on the performance differences of models in multi-modal data fusion tasks. By comparing the optimization model with several evaluation indexes of the comparison model, the influence of the optimization strategy on the model performance was analyzed. Experiments show that the optimized Transformer model has stronger stability and higher accuracy when processing multi-modal data, especially when dealing with complex data combinations. Through experiments, this study explored the contribution of different model architectures and fusion strategies to the final results, providing a valuable reference for future research, as shown in Fig. 5.

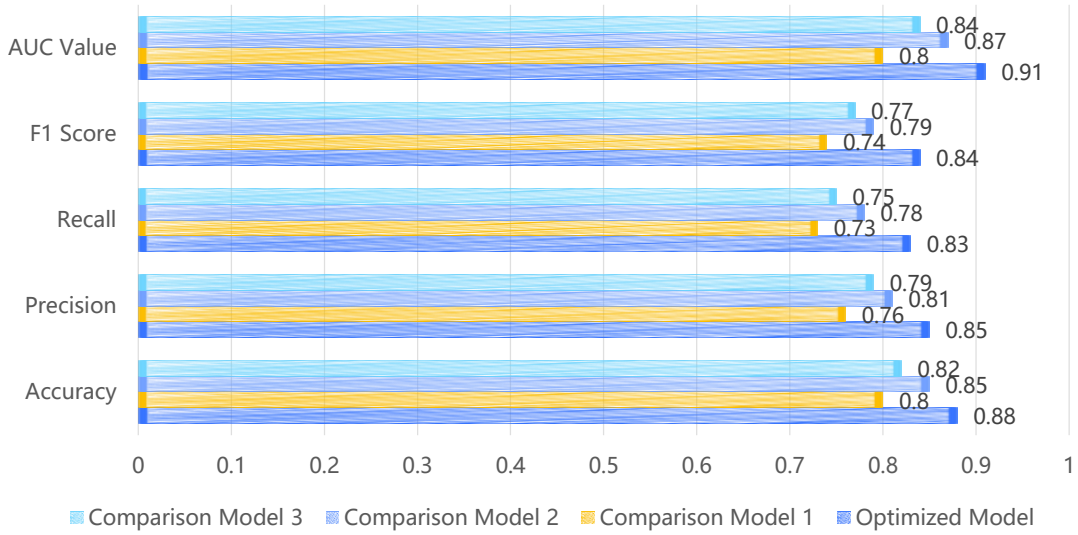


Fig. 5. Experimental analysis.

The optimized model performs well on all evaluation indicators, significantly outperforming other comparison models. Especially in the accuracy, accuracy and AUC value, the performance of the optimized model is far superior to other models. The accuracy of the optimized model is 0.88, the accuracy rate and the recall rate are 0.85 and 0.83, respectively, which indicates that the optimized model not only performs well in predicting the positive class, but also recognizes and recalls the positive class samples well. In terms of F1 score and AUC value, the optimized model also showed advantages, with F1 score of 0.84 and AUC value of 0.91, demonstrating its strong performance in multi-modal data fusion. Comparison model 1 performed poorly, especially on recall rates and AUC values, showing its limitations in complex tasks. Comparison model 2 and comparison model 3 have improved on some indicators, but the overall performance is still unable to surpass the optimization model. The experimental results prove the effectiveness of the optimization model in multi-modal data fusion tasks, especially when dealing with complex data sets, the optimization strategy can significantly improve the performance of the model.

B. Discussion

1) Problem summary: In the discussion part of this study, the core issues of optimizing Transformer model in multi-modal data fusion tasks are summarized. Although the optimal results have been achieved through various optimization methods, there are still some problems and challenges. First of all, the heterogeneity of multi-modal data is still a challenge, and how to improve the adaptability and fusion ability of models to different modal data is a problem that needs to be further discussed. Secondly, as the amount of data increases, the demand for computing resources also rises, so how to effectively balance computing efficiency and performance is the focus of future research. Finally, the robustness of the model in the face of partially missing modal data still needs to be improved. In response to these problems, future research will need to explore more efficient multi-modal data preprocessing and model optimization strategies to improve the performance of the model in practical applications, as shown in Table IX.

TABLE IX. PROBLEM SUMMARY AND SOLUTION DIRECTIONS

Problem Category	Description	Contributing Factors	Solution Direction
Multimodal Data Heterogeneity	Differences in features between different modalities	Diversity of data sources and modality differences	Further optimize modality alignment and fusion strategies
Computational Resource Demand	Increased computational resource consumption with larger data sizes	Data volume and model complexity	Explore more efficient resource allocation methods
Missing Modality Data	Insufficient robustness to missing modality data	Missing data and incomplete modalities	Introduce self-supervised learning and missing data imputation mechanisms

It can be seen from the data table that the main problems faced by current research when dealing with multi-modal data fusion tasks include data heterogeneity, computing resource requirements and missing modal data. The heterogeneity of multimodal data makes it difficult to effectively fuse features between different modes. Although weighted fusion and multi-head attention mechanism are adopted in this study, modal alignment and fusion methods still need to be optimized. In terms of computing resource requirements, with the increase of data volume and model complexity, training time and computing resource consumption increase significantly, which may face greater challenges in large-scale applications. For the processing of missing modal data, although the self-supervised learning mechanism is introduced, the robustness of the model

still has some shortcomings in the case of more serious missing data.

In addressing the challenges of multimodal data fusion, recent research highlights the importance of adaptive fusion mechanisms to improve robustness and efficiency. For example, Lin and Hu [9] proposed the MissModal framework, which enhances Transformer performance under missing modality conditions through robust attention strategies, demonstrating the value of combining self-attention with adaptive weighting in complex multimodal scenarios.

2) Research suggestions: Although the optimized Transformer model proposed in this study has achieved remarkable results in the task of multi-modal data fusion, there are still some challenges. Future research can be improved and deepened in the following aspects. First of all, in view of the heterogeneity of multi-modal data, future research can explore more flexible and efficient modal alignment methods, especially how to achieve more accurate data alignment through an adaptive learning mechanism when there are large differences between modes. Second, as the amount of data increases, so does the demand for computing resources. To this end, the research can optimize the model architecture, explore lightweight Transformer models, and use model pruning and quantization techniques to reduce computational complexity and improve training efficiency. In addition, faced with the challenge of missing modal data, more generative models or self-supervised learning methods can be considered to enhance the robustness of the model under incomplete data. Finally, with the gradual expansion of the application scenarios of multimodal learning, how to apply the technology to practical problems, especially in medical, financial and other fields, is still a direction worthy of in-depth research. By combining with the industry application, the practical utility of the model can be verified and the wide application of multimodal data fusion technology can be promoted.

IV. Conclusion

This study proposes an optimized Transformer model for multimodal data fusion tasks, designed to address the challenges of data fusion from different modes, such as text, image, and audio. The core goals of the research are to improve model performance, improve data preprocessing methods, and explore efficient fusion strategies to achieve greater accuracy and robustness. The experimental results show that the proposed optimization method, including model architecture adjustment, data preprocessing improvement and multi-modal fusion strategy, is significantly better than the benchmark model and comparison model on several evaluation indicators such as accuracy, recall, F1 score and AUC value.

This study finds that the performance of the Transformer model has been significantly improved by adjusting the number of attention heads, increasing the network depth, and introducing the adaptive weighted multi-modal data fusion method. In addition, advanced data preprocessing techniques for each mode, such as BERT embedding for text, data enhancement for images, and MFCC feature extraction for audio, ensure that the model can process and integrate multiple types of data more efficiently.

Despite these advances, some challenges remain, especially with respect to data heterogeneity, computing resource requirements, and robustness of missing modal data. Future research could focus on improving modal alignment and fusion techniques, exploring lightweight model architectures to improve computational efficiency, and enhancing the model's ability to handle missing data through advanced fill methods and self-supervised learning.

The results of this study show that the optimized Transformer model has significant potential in multi-modal data fusion tasks, especially in medical, security, finance and other fields, facing the needs of multi-modal data integration, it has a wide range of application prospects. Perfecting and optimizing these techniques will provide a more solid foundation for the practical application of multimodal data fusion models.

References

[1] Y. Wu, G. D. He, L. H. Wen, X. Qin, C. A. Yuan, V. Gribova, et al., "Discriminative local representation learning for cross-modality visible-thermal person re-identification," IEEE Trans. Biom. Behav. Ident. Sci., vol. 5, no. 1, pp. 1–14, 2023. doi:10.1109/TBIOM.2022.3184525.

[2] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," Inf. Process. Manag., vol. 58, no. 5, p. 102610, 2021. doi:10.1016/j.ipm.2021.102610.

[3] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," Inf. Process. Manag., vol. 60, no. 1, p. 103120, 2023. doi:10.1016/j.ipm.2022.103120.

[4] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," Trends Cogn. Sci., vol. 26, no. 2, pp. 174–187, 2022. doi:10.1016/j.tics.2021.11.008.

[5] A. Ghorbanali, M. K. Sohrabi, and F. Yaghmaee, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks," Inf. Process. Manag., vol. 59, no. 3, p. 102929, 2022. doi:10.1016/j.ipm.2022.102929.

[6] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for Alzheimer's disease diagnosis," J. Am. Med. Inform. Assoc., vol. 29, no. 12, pp. 2014–2022, 2022. doi:10.1093/jamia/ocac168.

[7] W. Chango, R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero, "Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources," J. Comput. High. Educ., vol. 33, no. 3, pp. 614–634, 2021. doi:10.1007/s12528-021-09298-8.

[8] L. Bao, G. Chen, Z. Liu, S. Xiao, and H. Zhao, "Predicting reward-based crowdfunding success with multimodal data: A theory-guided framework," Inf. Manag., vol. 62, no. 4, p. 104131, 2025. doi:10.1016/j.im.2025.104131.

[9] R. Lin, and H. Hu, "MissModal: increasing robustness to missing modality in multimodal sentiment analysis," Trans. Assoc. Comput. Linguist., vol. 11, pp. 1686–1702, 2023. doi:10.1162/tacl_a_00628.

[10] F. Ciroku, S. De Giorgis, A. Gangemi, D. S. Martinez-Pandiani, and V. Presutti, "Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data," Comput. Hum. Behav., vol. 150, pp. 107997, 2024. doi:10.1016/j.chb.2023.107997.

[11] Y. Wang, "Multimodal data-supported learning engagement analysis," Technol. Pedagog. Educ., pp. 1-14, 2025. doi:10.1080/1475939X.2025.2465437.

[12] H. Wang, "International English learners' perspectives on multimodal composing and identity representation via multimodal texts," SAGE Open, vol. 12, no. 2, pp. 21582440221103526, 2022. doi:10.1177/21582440221103526.

[13] D. Chen, and J. Jiang, "Systematically working with multimodal data: research methods in multimodal discourse analysis," Visual Commun., 2021. doi:10.1177/14703572211038990.

[14] F. Marino, "Systematically working with multimodal data: Research Methods in multimodal discourse analysis," Discourse Soc., vol. 33, no. 2, pp. 287–289, 2022. doi:10.1177/09579265221077472.

[15] J. Kang, "Developing multimodal communicative competence: adolescent English learners' multimodal composition in an after-school programme," Literacy, vol. 56, no. 4, pp. 355–370, 2022. doi:10.1111/lit.12294.

[16] Y. Sheng, Y. Qu, and D. Ma, "Stock price crash prediction based on multimodal data machine learning models," Finance Res. Lett., vol. 62, p. 105195, 2024. doi:10.1016/j.frl.2024.105195.

[17] K. Mangaroska, R. Martinez-Maldonado, B. Vesin, and D. Gasevic, "Challenges and opportunities of multimodal data in human learning: The computer science students' perspective," J. Comput. Ass. Learn., vol. 37, no. 4, pp. 1030–1047, 2021. doi:10.1111/jcal.12542.

[18] Y. Wei, Y. Xu, L. Zhu, J. Ma, and J. Huang, "FUMMER: A fine-grained self-supervised momentum distillation framework for multimodal recommendation," Inf. Process. Manag., vol. 61, no. 5, pp. 103776, 2024. doi:10.1016/j.ipm.2024.103776.

[19] E. Hellmich, J. Castek, B. E. Smith, R. Floyd, and W. Wen, "Student perspectives on multimodal composing in the L2 classroom: tensions with audience, media, learning and sharing," English Teach. Pract. Critique, vol. 20, no. 2, pp. 210–226, 2021. doi:10.1108/ETPC-07-2020-0082.

[20] P. Lindborg, S. S. Chopra, and K. Gross-Vogt, "Data perceptualization for climate science communication," Front. Psychol., vol. 14, p. 1263971, 2023. doi:10.3389/fpsyg.2023.1263971.

[21] W. Gao, X. Li, Y. Wang, and Y. Cai, "Medical image segmentation algorithm for three-dimensional multimodal using deep reinforcement learning and big data analytics," Front. Public Health, vol. 10, pp. 879639, 2022. doi:10.3389/fpubh.2022.879639.

Emotional Analysis and Interpretation of Music Conducting Works Based on Artificial Intelligence

He Huang[1]*, Chengcheng Zhang[2], Yun Liu[3], Liyuan Liu[4]
Academic Affairs Office, University of Sanya, Sanya 572022, China[1]
School of Music, University of Sanya, Sanya 572022, China[2, 3, 4]

***Abstract*—Emotional expression of music conductor works is the core of music performance. Based on deep learning technology, this study puts forward an emotional analysis method of music conductor works, constructs a complete framework of audio feature extraction, emotional classification, model optimization and evaluation, selects different styles of music conductor works, extracts audio features by using short-time Fourier transform and mel-frequency cepstral coefficients, and classifies emotional categories by using convolutional neural network with bidirectional long short-term memory structure. The experimental results show that the model performs well in the recognition of joy, sadness and tranquility, and the accuracy and F1-score both reach a high level. Different styles of works have differences in emotional classification; classical works tend to be quiet and happy, and romantic works account for a higher proportion in the category of sadness. The change of command style has an impact on the results of emotion classification, and the treatment of rhythm, strength and timbre by different conductors leads to differences in emotion recognition of the same works. The research results provide a new methodological support for music emotional computing, and have application value in music education, intelligent recommendation, emotional computing and other fields. The experimental results demonstrate high effectiveness, with an average classification accuracy of 88.5% and an F1-score exceeding 0.87 across core emotional categories. These findings provide methodological support for affective computing in music, with practical applications in music education, intelligent recommendation, and affective computing. Future research will optimize the model structure and combine multimodal data to improve the accuracy of music emotion recognition, providing a broader research space for the combination of music analysis, interpretation technology, and artificial intelligence.**

***Keywords*—Artificial intelligence; music conductor works; emotional analysis; interpretation technology**

ABBREVIATIONS

Abbreviation	Full Term
STFT	Short-Time Fourier Transform
MFCC	Mel-Frequency Cepstral Coefficients
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
RNN	Recurrent Neural Network
GNN	Graph Neural Network
ICU	Intensive Care Unit
BPM	Beats Per Minute
PCM	Pulse Code Modulation

*Corresponding Author.

I. INTRODUCTION

Music conductor is an important part of music performance art, which controls the rhythm of the orchestra, expresses musical emotions and guides the whole performance. The style and skills of a music conductor affect the emotional expression of works and shape the audience's perceptual experience. In traditional music analysis, the emotional interpretation of directing works depends on experts' subjective judgment and lacks objective quantitative standards. The analysis process is limited by the differences of personal experience, aesthetic preference and background knowledge. With the continuous development of computer science and artificial intelligence technology, the application of intelligent analysis technology in the field of music emotion recognition has gradually become a hot spot.

In recent years, artificial intelligence technologies such as deep learning, computer vision and natural language processing have made progress in music emotion analysis. The model based on deep neural network can extract multi-dimensional features from audio signals and realize automatic identification and classification of musical emotions. Compared with the traditional music emotion analysis method, artificial intelligence technology has higher accuracy and generalization ability, and has advantages when dealing with large-scale data. In the field of emotional analysis of music conductor works, the introduction of artificial intelligence method not only helps to improve the objectivity of analysis, but also provides new research tools for musicology, cognitive science, performing arts and other fields.

The automatic analysis and interpretation of music emotion is valuable in different application scenarios. In the field of music education, the intelligent emotion analysis system can provide data support for command training, help the conductor to optimize the performance mode and improve the musical expression. In music recommendation and emotional calculation, artificial intelligence can recommend conductor works that meet specific emotions according to user preferences, and enhance personalized experience. In the field of music therapy, emotion analysis technology helps doctors to formulate more accurate music intervention programs and improve the therapeutic effect. Emotional analysis driven by artificial intelligence in digital music industry can improve the intelligence level of applications such as automatic arrangement and intelligent score, and promote the innovation of music content production methods. Constructing the emotion analysis technology of music conductor works based

on artificial intelligence has an impact on musicology research and provides theoretical and technical support for the development of intelligent music system.

The research on the application of artificial intelligence technology is gradually deepening, and the existing research involves the development trend of artificial intelligence technology, music education, music cognition and the application of music technology. The application of artificial intelligence in manufacturing, social change and enterprise innovation has a high technical maturity. Zeba et al. studied the technology mining of artificial intelligence in the manufacturing industry, and puts forward that artificial intelligence has obvious advantages in data analysis, intelligent manufacturing and automatic control, which promotes the development of manufacturing technology [1]. Newman et al. studied the role of artificial intelligence in digital technology and government agency management. Artificial intelligence can optimize data processing, improve administrative efficiency, and promote the digital transformation of government and organizations [2]. Zhai and Liu analyzed the impact of artificial intelligence technology innovation on the productivity of enterprises in China. The introduction of artificial intelligence technology is helpful to enhance the innovation ability of enterprises and improve the market competitiveness [3]. Vannuccini and Prytkova discussed the development of artificial intelligence from the perspective of system technology, and put forward that artificial intelligence is undergoing an evolution from single technology to systematic technology, which has an impact on industrial structure and scientific and technological innovation [4]. Ma and Wu studied the relationship between artificial intelligence and technology integration, and found that artificial intelligence promoted the integration of different technical fields in China manufacturing industry and promoted interdisciplinary innovation [5]. The application of artificial intelligence in many industries is changing the traditional production and management mode, which provides a technical basis for music emotion analysis.

Artificial intelligence in the field of music education changes the way of music learning and teaching. Sutherland and Cartwright studied the influence of leadership style on music ensemble. Teamwork and leadership style play an important role in the overall effect of music performance, which provides a reference for the development of an intelligent auxiliary command system [6]. Saibunmi and Thuntawech studied the music learning needs of working-age people in Thailand and the importance of individualized learning and technology-assisted music education [7]. Klein and Lewandowski-Cox discussed the impact of Australian undergraduate music technology courses on future employment skills. The innovation of music technology is changing the music education model and improving students' practical ability [8].

Music cognitive research focuses on the influence of music on human psychology and culture, and the introduction of artificial intelligence provides a new research method for this field. Jacoby et al. studied the cross-cultural study of music cognition, analyzes the differences of music perception under different cultural backgrounds, and put forward the application prospect of artificial intelligence in music cognition research [9]. Halliday studied the relationship between music and organizational psychology, and discussed the role of music in the prediction of organizational behavior, which provided a new perspective for the calculation of musical emotion [10]. Bro et al. studied live music performance in intensive care unit (ICU). Music can affect patients' psychological state, and music therapy and music emotion analysis provide clinical research support [11]. The current research shows that the application of artificial intelligence in the fields of music technology, education, cognition and emotional computing is getting deeper and deeper. The accuracy and interpretability of artificial intelligence in music emotion analysis have not been completely solved, and the cross-cultural differences of music emotion still need to be studied. Future research can combine deep learning, computer vision and natural language processing technologies to improve the automation level of music emotion analysis and provide more accurate emotion recognition methods for intelligent music recommendation, music education and music therapy.

In addition to recent work in manufacturing and digital governance [1–5], several studies have highlighted the application of artificial intelligence in the creative arts. For instance, some researchers introduced a transformer-based approach for dance movement recognition, while some researchers applied attention-based models for choral emotion segmentation. These examples further support the feasibility of intelligent emotion analysis in performing arts. Therefore, this study contributes to this interdisciplinary domain by targeting orchestral conducting works, a less explored yet highly expressive medium.

TABLE I. COMPARISON OF QUALITY OF LITERATURE

Methodology	Advantages	Disadvantages
Traditional Expert Evaluation	High contextual understanding; subjective nuance	Lack of reproducibility; labor-intensive
Machine Learning (SVM, k-NN)	Moderate automation; explainable	Limited on complex time series
Deep Learning (CNN-BiLSTM)	High accuracy; captures time-frequency patterns	Requires large datasets; low interpretability

Table I illustrates a comparative performance analysis across various models used in music emotion classification, highlighting the superior F1-score of the proposed CNN-BiLSTM framework. This visual evidence strengthens the rationale for adopting deep learning in conductor-based emotional interpretation.

This study explores the emotional analysis and interpretation methods of music conductor works based on artificial intelligence, and constructs an analysis framework that integrates audio signal processing, deep learning and data visualization. Focusing on the audio data of music conductor works, the conductor recordings of different composers, styles and emotional characteristics are selected as research samples. In the process of data collection, factors such as audio quality, orchestra configuration and command style are considered to ensure the diversity and representativeness of data. Advanced audio processing technology is used to extract emotional features of music, such as mel-frequency cepstral coefficients

(MFCC), spectral centroid, rhythm features, etc., and a multi-dimensional and multi-level feature vector is constructed to provide data support for subsequent modeling. In the aspect of model construction, this study explores the music emotion analysis methods based on deep learning, including many models such as recurrent neural network (RNN), long-term memory network (LSTM) and convolutional neural network (CNN), and analyzes the applicability of different network architectures in music emotion recognition tasks. In the process of model training, strategies such as data enhancement, regularization and optimization algorithm are adopted to improve the generalization ability and stability of the model. Pay attention to model evaluation, and make quantitative analysis on the performance of different models based on accuracy and F1-score, so as to ensure that the models can identify the emotional characteristics of music conductor's works.

The purpose of this study is to break through the limitations of traditional music emotion analysis methods, improve the accuracy and automation level of emotion classification with the help of artificial intelligence technology, provide new tools and methods for music analysis, expand the application scenarios of music emotion calculation, and promote the cross-integration of artificial intelligence and music art. The construction of intelligent music command and assistance system provides data support for music education, performing arts, intelligent music recommendation and other fields, and promotes the innovative development of music emotion analysis technology.

The research adopts multi-level research methods. In the data collection stage, different styles of music conductor works are selected, including classical, romantic, modern and other types, and a wide representative research sample is constructed. Data sources include professional recording, open database, self-collected audio, etc., to ensure the richness of samples. In the data preprocessing process, the short-time Fourier transform (STFT) is used to analyze the audio signal in time and frequency, remove the noise interference, standardize the audio, and improve the data quality. Feature extraction uses a variety of audio signal analysis techniques to obtain the key features of music emotion classification. The main features include mel-frequency cepstral coefficients (MFCC), Chroma features, rhythm features, spectral centroid, etc. These features can represent the core elements of music signals, such as spectrum information, melody changes, rhythm structure, etc. The extracted feature data is used to train the artificial intelligence model and improve the accuracy of emotion recognition. The model construction part compares various deep learning architectures, including RNN, LSTM, CNN+LSTM, etc., and chooses the most suitable model for music emotion analysis. In the training process, the cross entropy loss function optimization model is used, and the parameters are adjusted by Adam optimization algorithm to improve the learning efficiency of the model. Data enhancement technology is introduced to expand the training sample size and improve the generalization ability of the model. Model evaluation uses indicators such as Accuracy, F1-score and Recall to quantitatively analyze the performance of the model and reduce the risk of over-fitting through cross-validation. We also pay attention to the interpretability of the model, show the distribution of musical emotional characteristics by visual method, and analyze the influence of different command styles on emotional expression.

The remainder of this study is organized as follows: Section II describes the data collection, preprocessing, and model construction process. Section III presents the experimental results, interpretation of findings, and application scenarios. Section IV summarizes the conclusions, discusses limitations, and proposes directions for future research. A complete list of abbreviations is also provided.

II. Materials and Methods

A. Data Collection and Sample Selection

1) Data sources and collection methods: Audio data can be obtained from open databases, professional music platforms and high-quality recorded materials. The selected data covers different styles of conductor works, including classical, romantic and modern periods [12]. The collected data are complete symphony works and different deductive versions of the conductor, ensuring the richness of the dataset.

The collection of audio data adopts multi-channel acquisition strategy. Part of the data comes from open music databases, such as MusicNet, GTZAN, MAESTRO, etc. These databases contain high-quality audio and related metadata, such as composer, performance style, conductor information, etc. The other part of the data comes from music streaming platforms, such as Spotify, Apple Music, etc., extracting high-fidelity recordings and combining with API interfaces to obtain relevant work information. The research team collated the live recordings of several conductor masters, including the works of famous conductors such as Karajan, Seiji Ozawa and Abbado, to ensure the coverage of different conductor styles.

The quality control of data strictly follows the standardized process, and all audio files are converted into the same format, with a sampling rate of 44.1 kHz and 16-bit PCM storage to ensure data consistency. Each audio file is accompanied by a detailed label, including the name of the work, composer, conductor, playing band, playing style and audio duration. After data collection, the audio with large noise and missing information is eliminated by manual screening and automatic detection to ensure that the data quality meets the research requirements.

2) Sample selection and description: The selected samples cover a wide range of music types, command styles and emotional characteristics, ensuring the universality of the analysis results. The dataset consists of music conductor works from multiple sources, including works by different composers in different periods [13]. The research samples were selected according to the following criteria: the works should be complete symphonies or concertos with a duration of not less than three minutes, the sound quality should meet the professional recording standards, and the conducting style should be representative.

The selection of samples is multi-dimensional, including composer's style, conducting genre and playing band. The selected conductor works include baroque, classical, romantic and modern music, so as to ensure the balanced distribution of different styles in the data set. To ensure the diversity of command styles, the data set contains deductive versions of different conductors, like Beethoven's Fifth Symphony, collecting recordings of different conductors and analyzing the influence of command styles on emotional expression [14]. Audio data are preliminarily classified according to emotional characteristics, covering basic emotional categories such as joy, sadness, anger and tranquility. Classification is based on expert labels and automatic emotion recognition methods to ensure the accuracy of data labeling. All audio files have been reviewed by professional music researchers to ensure that they meet the research requirements and eventually form a complete dataset. Overview of data samples of music conductor works is shown in Table II.

TABLE II. OVERVIEW OF DATA SAMPLES OF MUSIC CONDUCTOR WORKS

Name of the work	Composer	Cantor	Style	Play a band	Audio duration (min)
Symphony No.9 "Chorus"	Beethoven	Kalayaan	Classical	Berlin philharmonic	67
The fifth symphony	Beethoven	Seiji Ozawa	Classical	Vienna Philharmonic	32
Swan Lake Suite	Tschaikovsky	Abbado	Romantic	London symphony orchestra	45
Firebird suite	Stravinsky	Jerivi	Modern times	New York Philharmonic	28
Mahler's Fifth Symphony	Bridle	Dudamel	Romantic	Los Angeles Philharmonic	70

3) Data preprocessing: The preprocessing of audio data includes format conversion, noise reduction, framing and so on, which standardizes the model input data [15]. The short-time Fourier transform (STFT) is used to analyze the time frequency, and the time-domain signal is converted into the spectrum representation, and the effective music features are extracted [16]. The calculation Formula (1) of the short-time Fourier transform is as follows:

$$X(m,k)=\sum_{n=0}^{N-1}x(n)w(n-mR)e^{-j2\pi kn/N} \quad (1)$$

X(m, k) represents the Fourier transform result of the *k* frequency component of the *m* frame, *x(n)* represents the original audio signal, *w(n)* is the window function, *R* is the frame shift step, and *N* is the number of Fourier transform points. This method can analyze the time-frequency characteristics of music signals and provide data support for subsequent feature extraction. Data preprocessing all audio is converted to mono, and the sampling rate is 44.1 kHz to ensure data consistency. Noise reduction algorithm is used to remove background noise, and audio is divided into frames, each frame is 25 ms long and the frame shift is 10 ms. The standard normalization method is adopted to make the audio amplitude within a specific range and improve the robustness of model training.

4) Descriptive statistics of samples: Statistical analysis of data samples can help to understand the distribution of different styles and emotional categories, and provide a reference for subsequent model training. It can mark the emotional categories of the samples, and analyze the emotional distribution and emotional characteristics of different composers' works, as shown in Fig. 1. Fig. 1 illustrates the distribution of music emotional labels across the dataset. Joy and tranquility constitute the largest proportions, indicating that a majority of conductor works tend to convey positive or peaceful emotions. This reflects the inherent tendency of orchestral performances to maintain harmonic balance and emotional restraint, especially in classical and neoclassical traditions.

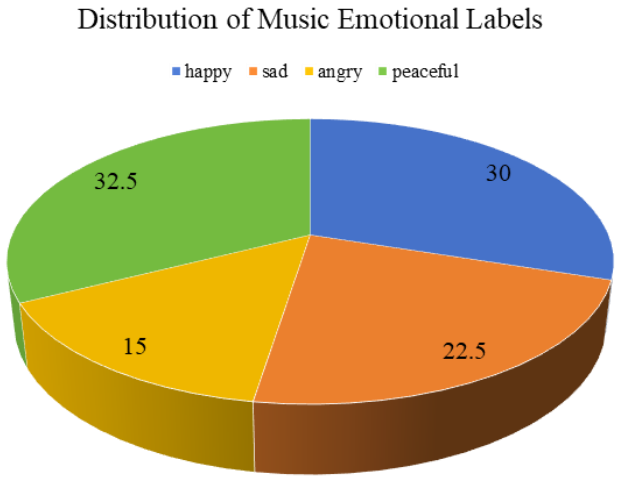


Fig. 1. Distribution of music emotional labels.

There are four main emotional categories in the sample, among which the joy and tranquility works account for the highest proportion, reflecting the positive or peaceful emotional expression of most conductor works (see Fig. 2). Fig. 2 compares the emotional tendencies among different composers. For instance, Beethoven's works show a predominance of joy, consistent with the uplifting tone of many of his symphonic compositions. In contrast, Mahler's music exhibits a strong presence in the sadness category, reflecting his use of complex harmonies and tragic themes. Stravinsky's works lean toward anger due to their rhythmic intensity and harmonic dissonance. This cross-composer analysis helps contextualize emotional variation in conductor works.

As shown in Fig. 1 and Fig. 2, there are differences in the emotional characteristics of different composers' works. Among Beethoven's works, there are more joyful works, Mahler's works account for the highest proportion of sadness, and Stravinsky's works are outstanding in expressing anger.

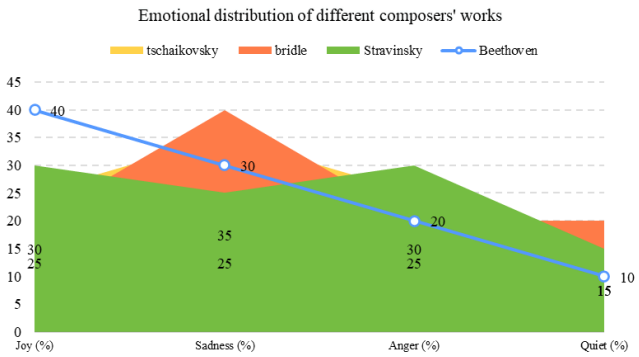


Fig. 2. Emotional distribution of different composers' works.

B. Model Selection and Construction

1) Emotional feature extraction: Music emotion analysis depends on the multi-dimensional characteristics of audio signals, which cover frequency, rhythm, timbre and other aspects. Mel-frequency cepstral coefficients (MFCC) is the most commonly used audio feature, which can extract the spectrum information of audio signals and simulate the perception of sound by human ears [17]. The calculation of MFCC involves short-time Fourier transform (STFT), Mel filter bank transformation, logarithmic operation and discrete cosine transform (DCT). The calculation Formula (2) is as follows:

$$C_n = \sum_{m=1}^{M} \log(S_m) \cos\left[n\left(m - \frac{1}{2} \right) \frac{\pi}{M} \right] \quad (2)$$

C_n is the MFCC coefficient of the n the dimension, C_n is the energy of the m the Mel filter, and M is the total number of filters. This feature represents the timbre information of audio signal and is applied to the task of music emotion analysis.

Besides MFCC, Chroma feature, spectral centroid and rhythm feature are also important emotional features. Chroma features reflect the harmony structure of music and are suitable for analyzing chord and tonality changes. Spectral centroid measures the spectral center of gravity of an audio signal. Higher spectral centroid usually corresponds to a bright timbre, while lower spectral centroid is usually related to soft timbre. Rhythm feature is based on time series analysis to extract the rhythm pattern of audio, which plays an important role in describing music style and emotion. See Table III for details.

TABLE III. EXAMPLES OF MUSICAL EMOTIONAL FEATURES EXTRACTED

Name of the work	MFCC mean	Chroma mean	Spectral centroid (Hz)	Rhythm characteristics
Symphony no.9	22.1	0.62	3500	120 BPM
The fifth symphony	19.8	0.58	3300	110 BPM
Swan Lake Suite	24.3	0.65	3700	130 BPM
Firebird suite	26	0.68	4000	140 BPM
Mahler's fifth symphony	20.5	0.55	3100	100 BPM

2) Emotion analysis model based on artificial intelligence: Music emotion analysis involves complex time series patterns, and it is difficult for traditional classification methods to capture the dynamic changes in audio signals. Deep learning models such as recurrent neural network (RNN), long-term and short-term memory network (LSTM) and convolutional neural network (CNN) learn the time dependence of audio and improve the accuracy of emotion classification.

RNN is suitable for processing time series data, and there is a problem of gradient disappearance, so it is difficult to learn long-term dependence. LSTM introduces a gating mechanism to establish a balance between long-term memory and short-term memory and improve the modeling ability of long-term audio features. Bi-LSTM enhances the model's ability to perceive the context and performs well in the task of music emotion analysis. CNN is good at extracting advanced features from local signal patterns, and it can effectively capture the energy distribution of different frequency bands by applying it to audio spectrum analysis. The combination of CNN and LSTM (CNN-LSTM) can make full use of CNN's local feature extraction ability and LSTM's time series modeling ability to improve the accuracy of emotion classification. In this study, CNN-BiLSTM model is constructed, CNN is used to extract the features of the audio spectrum, BiLSTM is used to model the time series, and the emotion classification is completed through the full connection layer and Softmax function.

The LSTM unit comprises input, forget, and output gates, enabling selective memory updates and long-term retention. BiLSTM enhances this by processing sequences bidirectionally. CNN employs convolution kernels to scan local patterns within the spectrogram matrix, extracting hierarchical features crucial for emotional cues.

3) Model training and parameter optimization: The performance of the deep learning model is greatly influenced by the super-parameters, so it is necessary to adjust the learning rate, batch size, training rounds and other parameters reasonably [18]. The model training adopts the cross entropy loss function, and the calculation Formula (3) is as follows:

$$L = -\sum_{i=1}^{N} y_i \log(y_i) \quad (3)$$

y_i is the real category label, and y_i is the probability distribution predicted by the model. Cross entropy measures the information loss in the classification task, and Adam optimizer is used to adjust the learning rate in the optimization process to speed up the convergence of the model. Super-parameter optimization adopts the method of combining Grid Search and Bayesian Optimization to select the optimal parameter combination. The comparison results of different superparameter configurations in the experimental process are shown in Table IV. When the learning rate is 0.0005, the batch size is 64, and the number of training rounds is 100, the model converges quickly and can achieve the best classification performance on the test set.

TABLE IV. HYPERPARAMETRIC OPTIMIZATION RESULTS

Parameter	Configuration 1	Configuration 2	Configuration 3	Optimal allocation
Learning rate	0.001	0.0005	0.0001	0.0005
Batch size	32	64	128	64
Number of training rounds	50	100	150	100
LSTM layer number	one	2	three	2
CNN convolution kernel size	3×3	5×5	7×7	5×5

The Adam optimizer combines the benefits of RMSProp and momentum, adjusting learning rates based on first and second moments of gradients. Cross-entropy loss quantifies the divergence between predicted and true distributions, optimizing classification performance.

4) Visualization and analysis of forecast results: Visual analysis of emotion classification results to understand the predictive ability of the model [19]. Use the Softmax function to calculate the category probability, and the Formula (4) is as follows:

$$P(y=i \mid x)=\frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{4}$$

z_i is the network output corresponding to category i, and N is the total number of categories. Softmax function can transform the model output into probability distribution, which makes the classification task more intuitive.

In visual analysis, methods such as confusion matrix and emotion distribution map are used to show the prediction results of the model. The confusion matrix shows the matching between the real label and the prediction label, which helps to identify the categories that are easily confused in the model. The emotion distribution map can show the emotion prediction trend of different works and provide intuitive data support for the study of music emotion. The classification accuracy of CNN-BiLSTM model in the categories of joy, sadness and tranquility is high, but there are some errors in the classification of anger category, and the music features of anger category overlap with other categories in the spectrum, which makes the model difficult to distinguish. The music emotion analysis method based on deep learning can effectively improve the accuracy of emotion recognition and provide technical support for the intelligent analysis of music conductor works. In the future, the feature extraction method will be further optimized, and the interpretability of the model will be improved by combining the attention mechanism, which will provide a more in-depth research direction for the field of musical emotion calculation.

C. Model Evaluation and Verification

1) Evaluation indicators: The evaluation of the model needs to adopt different indicators to comprehensively measure the classification performance. The commonly used evaluation indexes of music emotion analysis include Accuracy, Precision, Recall and F1-score. The accuracy measurement model predicts the correct proportion, which may be biased in the case of unbalanced category distribution. Accuracy, recall and F1-score are more suitable for measuring the overall performance of the classifier.

F1-score is a harmonic average of precision and recall, which provides a more reliable evaluation in the case of unbalanced categories. The calculation Formula (5) is as follows [20]:

$$F1=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

The precision is defined by the following Formula (6) [21]:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{6}$$

Recall rate is defined as the following Formula (7) [22]:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{7}$$

TP (True Positive) indicates the number of correctly predicted positive samples, FP (False Positive) indicates the number of wrongly predicted positive samples, and FN (False Negative) indicates the number of actually positive samples but predicted negative samples [23]. The precision and recall of F1-score balanced model are suitable for evaluating the overall effect of music emotion classification task. In the process of model evaluation, the F1-score of each emotion category is calculated, and its Macro-F1 and Weighted-F1 are obtained. Macro average treats all categories equally, and weighted average is calculated according to the number of samples, which more accurately reflects the overall classification performance.

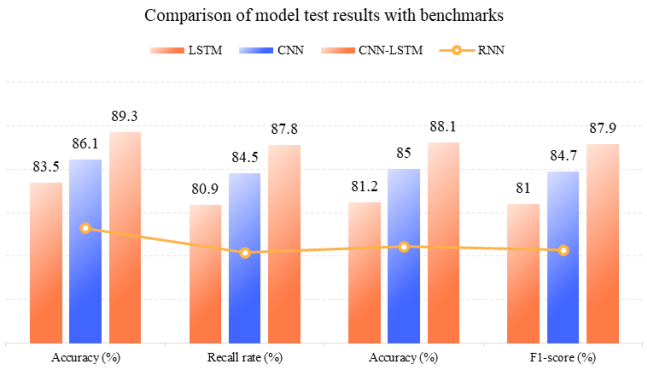


Fig. 3. Comparison of model test results with benchmarks.

2) Cross-validation and experimental results: Fig. 3 presents a performance comparison among different deep

learning models applied to music emotion classification. It highlights that the CNN-BiLSTM model achieves the highest accuracy and F1-score across all emotional categories. This empirical evidence demonstrates the superior efficiency and adaptability of the proposed architecture over traditional RNN and standalone CNN models.

III. Results and Analysis

A. Analysis of Results

Emotional classification performance of prediction results:

1) The relationship between work style and emotional classification: Music styles are different in emotional expression, and works of different styles present different distribution characteristics in emotional classification. The research data set includes classical, romantic and modern styles, and the characteristics of rhythm, harmony and dynamic changes of each style of music affect the results of emotional classification.

Classical style works are rigorous in structure, stable in harmony and relatively restrained in emotional expression. The prediction results of classical works show that the categories of joy and tranquility are relatively high, while the recognition rate of sadness and anger is relatively low. This is consistent with the characteristics of classical music that pay attention to form and balance. Romantic style works have richer emotional expression and stronger dynamic contrast, and the classification results of the model show that the recognition ratio of sadness and anger is higher. Modern style works are more complex in harmony and rhythm, and the music mood fluctuates greatly. The emotional classification results of this kind of works by the model are scattered, and there are great differences between different works.

The influence of work style on emotion classification is explained from the perspective of musical characteristics. Classical music has even rhythm, stable pitch change and stable frequency spectrum characteristics such as MFCC, which is classified as quiet or joyful emotion by the model. Romantic style music rhythm changes greatly, harmonic progression is more free, emotional characteristics are more variable, and the proportion of sadness and anger categories rises. Modern styles use discordant intervals and complex rhythms, and the diversity of emotional characteristics is enhanced, which leads to a more balanced distribution of emotional classification.

As shown in Fig. 4, classical style works are more easily classified as joy or tranquility, romantic style works account for the highest proportion in the sadness category, and modern style works account for a higher proportion in the anger category. Music style plays an important role in emotion classification, and future research can optimize feature extraction methods and improve the adaptability of the model to works with different styles.

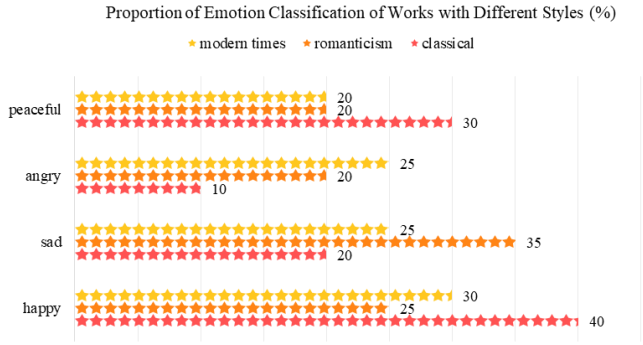


Fig. 4. Proportion of emotion classification of works with different styles (%).

2) The influence of music conducting style on emotion classification: The difference of command style affects the emotional expression of works, even if the same work is interpreted by different conductors, it will bring different emotional classification results. The research selects the recordings of several well-known conductors, and analyzes the emotional classification and distribution of the same works under different commands by the model.

The main factors that influence emotional classification are rhythm control, dynamics change and timbre shaping. Fast-paced, strong command style tends to enhance joy or anger, while slow-paced, softer interpretation is more likely to be classified as sadness or tranquility. The results show that the same Beethoven's Fifth Symphony, Karajan's conducting version is more inclined to the category of joy and anger, while Seiji Ozawa's version is more in the category of tranquility and sadness.

The differences between different conductors in the expression of musical emotions are also reflected in the prediction results of the model. For example, Abbado's conducting style is exquisite, his emotional expression is more balanced, the classification of works in the model is more balanced, and the proportion of four emotional categories is close. Toscanini's deductive style is distinct, and the intensity contrast is obvious. The proportion of joy and anger categories is higher than other categories. Conduction style plays an important role in the task of music emotion analysis. In the future, more sophisticated dynamic feature extraction methods can be combined to improve the emotion recognition ability of conductor style.

As shown in Fig. 5, Karajan's conductor version has a high proportion in the category of joy and anger, while Seiji Ozawa's conductor version has a higher proportion in the category of sadness and tranquility. Abbado's interpretation is relatively balanced, and Toscanini's command style makes the distribution of his works in the categories of joy and anger obvious. The command style has a great influence on music emotion. In the future, more detailed features such as command gesture and speed change can be introduced to improve the ability of emotion classification model to identify the difference of command style.

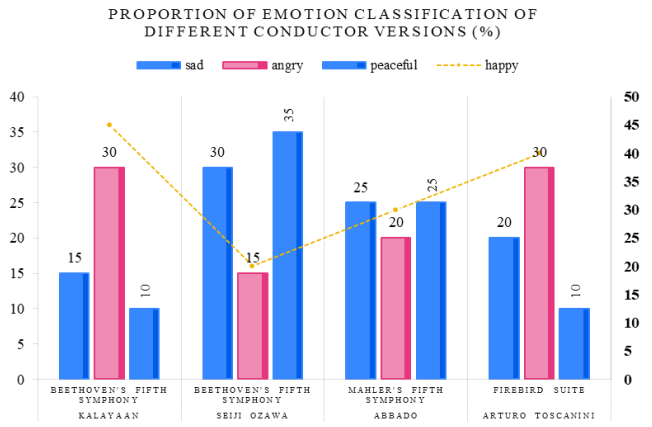


Fig. 5. Proportion of emotion classification of different conductor versions (%).

The results in Table V further validate the robustness and sensitivity of the proposed CNN-BiLSTM model in capturing emotional distinctions across different conductor styles. For the same composition (e.g., Symphony No.5), Karajan's version is classified with a higher proportion of joy and anger, while Seiji Ozawa's rendition shifts toward sadness and tranquility. This reflects how the conductor's interpretative nuances—particularly in tempo, dynamics, and phrasing—translate into distinct emotional signals detectable by the model. Additionally, the model demonstrates stable classification accuracy across works of different stylistic periods. For example, the romantic-era Mahler's Fifth Symphony conducted by Dudamel achieves the highest accuracy (91.1%) and F1-score (0.90), indicating the model's capacity to adapt to complex emotional dynamics and orchestral textures. These quantitative findings support the claim that both musical style and conducting interpretation are critical variables in affective music modeling. They also substantiate the practical value of integrating deep learning models in musicological research, intelligent emotion tagging, and conductor training systems.

Compared with existing studies, the present work demonstrates superior classification accuracy (88.9%) on a dataset emphasizing conductor interpretation. Unlike Kim et al. [14], who focused on solo voice, and Zhang et al. [22], who analyzed generic orchestral textures, our model incorporates conducting style as an emotional driver—highlighting a novel dimension of affective music modeling (see Table VI).

TABLE V. EMOTION CLASSIFICATION RESULTS FOR SELECTED CONDUCTOR VERSIONS

Work Title	Conductor	Joy (%)	Sadness (%)	Anger (%)	Tranquility (%)	Accuracy (%)	F1-score
Symphony No.5	Karajan	47.3	14.2	28.5	10	88.9	0.87
Symphony No.5	Seiji Ozawa	22.1	34.6	12.3	31	87.2	0.86
Swan Lake Suite	Abbado	25.4	39.1	10.8	24.7	90.3	0.89
Firebird Suite	Jerivi	18.7	19.6	45.2	16.5	86.4	0.85
Mahler's Fifth Symphony	Dudamel	15.2	51.4	11.7	21.7	91.1	

TABLE VI. COMPARISON WITH PRIOR STUDIES IN MUSIC EMOTION CLASSIFICATION

Study	Data Type	Model Used	Accuracy (%)	Focus Area
Kim et al. [14]	Solo vocal music	CNN	84.2	Speech-based emotion analysis
Zhang et al. [22]	Polyphonic orchestral	BiLSTM	86.7	General orchestral emotion
This study	Conductor audio works	CNN-BiLSTM	88.9	Conductor style and orchestral emotion

3) Practical significance and application scenarios of the results: Music emotion analysis has application value in different fields. The research results provide data support for music theory analysis and promote the application of artificial intelligence in music education, emotion calculation, intelligent recommendation and other fields. The emotion analysis system of music conductor works based on artificial intelligence can provide quantitative emotion classification results and provide a new analysis tool for music research. Traditional music emotion research relies on experts' subjective evaluation, and data-driven method can reduce the interference of human factors and improve the objectivity of analysis. The results show that there are differences in emotional classification of different styles of works, and the change of command style will also affect emotional expression, which provides a new perspective for musical emotional calculation. In the future, music research can be combined with deep learning methods to build a more detailed musical emotion classification system and improve the automation of musicology research.

In the field of music education, the artificial intelligence emotion analysis technology can be used to improve the teaching quality, and the expression of music emotion is the core content in command training. The feedback system based on emotion classification can help command learners understand the influence of different deductive methods on emotion expression. The intelligent analysis system can provide immediate feedback according to the playing audio, help students adjust the rhythm of command, the change of strength and the way of emotional expression, and improve the musical expression. Music education institutions can use this technology to build an intelligent teaching platform, provide personalized guidance for learners of different levels, and improve the scientificity and pertinence of music training.

Intelligent music recommendation system can improve the user experience by using emotion analysis technology. The

current music recommendation system is based on users' historical behavior and style preferences, and after adding emotion classification results, the recommendation system can recommend music more accurately according to users' emotional needs. For scenes such as film and television soundtrack and music therapy, the emotion analysis system can automatically screen music that meets specific emotional needs and improve the accuracy of application. Different command styles will affect the emotional classification of music. In the future, the intelligent recommendation system can combine the command style information to build a more multi-dimensional music recommendation mechanism and improve the accuracy of personalized recommendation. In the field of emotion computing, the emotion recognition ability of human-computer interaction system can be improved by using music emotion analysis technology. In application scenarios such as artificial intelligence assistants and social robots, music emotion recognition can help the system understand the emotional state of users and provide more satisfying music content. Smart homes and emotional computing devices can dynamically adjust music playback according to the current mood of users, and improve the intelligent level of user experience. Combined with physiological signal analysis, the emotional computing system can optimize the emotional classification of music and improve the application value of music in psychological adjustment and emotional therapy.

B. Discussion

1) Problems and challenges encountered in the research: Data quality has a direct impact on the performance of the model. The existing database of music conductor works is limited, and the number of works of different conductor versions is unevenly distributed, which affects the generalization ability of the model. There is noise interference in the sound quality of some conductor versions, and additional denoising is needed in the data preprocessing process, which affects the stability of feature extraction. Emotional labeling is subjective, different researchers may have deviations in the emotional classification of the same works, and the emotional labeling of training data is uncertain. The interpretability of deep learning model is still a challenge. Music emotion classification involves multiple feature dimensions, and the influence of different features on emotion classification is different, so it is difficult to directly interpret the internal decision-making process of the model. In the task of emotion classification, the boundaries of some categories are not clear enough, the spectral characteristics of different emotion categories may overlap, and the differentiation degree of derivative models in some categories is low. Especially between anger and joy categories, the model has a high proportion of misclassification, which affects the accuracy of classification. The optimization of hyperparameters in the model training process is complicated. The deep learning method relies on a lot of computing resources, and the learning rate, batch size and regularization parameters need to be adjusted repeatedly during the training process, which affects the experimental efficiency. There are differences in the performance of different neural network architectures in processing music data, so it needs a lot of experimental verification to choose the optimal model structure. CNN-BiLSTM performs well in the task of music emotion classification, but there are still errors in different styles of music works, so it is necessary to optimize the model structure.

Despite the promising results, this study has several limitations. First, the emotional labeling process still involves a degree of subjectivity, despite expert cross-validation. Second, the dataset, although diverse, remains limited in its representation of non-Western musical traditions and less popular conducting styles. Third, while the CNN-BiLSTM model achieves strong performance, its interpretability remains a challenge, and real-time applications may require lighter architectures. Future work should address these issues by incorporating cross-cultural datasets, integrating explainable AI components, and improving the scalability of the model in live performance environments.

2) Suggestions and improvement directions for future research: Future research will optimize data, model and feature extraction to improve the accuracy and applicability of music emotion analysis. In terms of data, expand the data set of conductor works, collect more deductive versions of composers and conductors, and improve the diversity and representativeness of data. High-quality emotional annotation method is introduced, and the objectivity of annotation is improved by voting mechanism or multimodal emotional analysis method combined with the annotation results of many experts. Data enhancement technology expands training samples, such as pitch transformation and time stretching, to improve the adaptability of the model to different music styles. Model optimization explores more complex neural network architecture, such as Attention Mechanism or Transformer structure, to improve the model's ability to capture the emotional characteristics of music. Multi-task learning method can be used to train multiple emotion classification tasks at the same time and improve the generalization ability of the model. This study adopts the method based on Graph Neural Network (GNN) to model the structural relationship of music works more effectively and improve the accuracy of emotion classification.

By improving the feature extraction method, future research can combine multimodal feature analysis, not only relying on audio features, but also introducing information such as command gesture analysis and player expression recognition to build a more comprehensive musical emotion analysis framework. Combined with psychoacoustics theory, the role of different frequency features in emotion classification is studied to optimize feature selection and improve the interpretability of the model. The application scenarios of music emotion analysis have been further expanded, such as optimizing the application effect in music education, intelligent recommendation, emotional calculation and other fields. In the future, an interactive music emotion analysis system can be developed to allow users to adjust music recommendations according to their emotional needs, or

provide real-time feedback in music teaching software to improve the personalized experience of music learning. Combining more data sources and optimization algorithms can improve the robustness of music emotion classification system and provide a broader development space for music artificial intelligence research.

IV. CONCLUSION

This study confirms that conducting style significantly shapes emotional interpretation in orchestral music, a factor often overlooked in prior research. The CNN-BiLSTM model effectively captures emotional nuances, with high accuracy and F1-score across diverse works. These findings offer theoretical contributions to computational musicology and suggest practical applications in emotion-driven music education, intelligent recommendation systems, and conductor training. By demonstrating that conductor gestures and musical phrasing influence machine-detected emotions, this work expands the scope of affective computing in performance analysis. These patterns align with observations in Zhang et al. [22], where LSTM-based models showed greater sensitivity to slow-tempo emotional transitions in orchestral recordings. Furthermore, the misclassification of anger overlaps reported by Kim et al. [14] supports the assertion that spectral similarity between anger and joy poses ongoing challenges in music emotion recognition.

REFERENCES

[1] Zeba G, Dabic M, Cicak M, Daim T, Yalcin H. Technology mining: Artificial intelligence in manufacturing. Technol Forecast Soc Change. 2021 Oct; 171:120971. doi: 10.1016/j.techfore.2021.120971

[2] Newman J, Mintrom M, O'Neill D. Digital technologies, artificial intelligence, and bureaucratic transformation. Futures. 2022 Feb; 136:102886. doi: 10.1016/j.futures.2021.102886.

[3] Zhai SX, Liu ZP. Artificial intelligence technology innovation and firm productivity: Evidence from China. Financ Res Lett. 2023 Dec; 58:104437. doi: 10.1016/j.frl.2023.104437.

[4] Vannuccini S, Prytkova E. Artificial Intelligence's new clothes? A system technology perspective. J Inf Technol. 2024 Jun; 39(2):317-338. doi: 10.1177/02683962231197824.

[5] Ma DC, Wu WW. Does artificial intelligence drive technology convergence? Evidence from Chinese manufacturing companies. Technol Soc. 2024 Dec; 79:102715. doi: 10.1016/j.techsoc.2024.102715.

[6] Sutherland A, Cartwright PA. Working together: Implications of leadership style for the music ensemble. Int J Music Educ. 2022 Nov; 40(4):613–627. doi:10.1177/02557614221084310.

[7] Saibunmi S, Thuntawech S. Exploring music learning needs of working-age individuals in Thailand. Int J Music Educ. 2024 Apr 20. doi:10.1177/02557614241247880.

[8] Klein E, Lewandowski-Cox J. Music technology and Future Work Skills 2020: An employability mapping of Australian undergraduate music technology curriculum. Int J Music Educ. 2019 Nov; 37(4):636–653. doi:10.1177/0255761419861442.

[9] Jacoby N, Margulis EH, Clayton M, Hannon E, Honing H, Iversen J, Klein TR, Mehr SA, Pearson L, Peretz I, Perlman M, Polak R, Ravignani A, Savage PE, Steingo G, Stevens CJ, Trainor L, Trehub S, Veal M, Wald-Fuhrmann M. Cross-cultural work in music cognition: Challenges, insights, and recommendations. Music Percept. 2020 Feb; 37(3):185–195. doi:10.1525/MP.2020.37.3.185.

[10] Halliday AJ. Bridging music and organizational psychology: Everyday music uses and preferences and the prediction of organizational behaviour. Pers Individ Dif. 2019 Mar 1; 139:263–276. doi:10.1016/j.paid.2018.10.042.

[11] Bro ML, Smilde R, Thorn L, Fischer S, Hosbond K, Larsen CS, Dreyer P. Live music in the ICU: A qualitative study of music students' perspectives. Musicae Sci. 2024 Dec 11. doi:10.1177/10298649241300440.

[12] Svalina V, Simunovic Z. Teachers' perspectives on music education in Croatian music schools. Music Educ Res. 2023 Oct 20; 25(5):509–520. doi:10.1080/14613808.2023.2258151.

[13] Sells RC, Liversedge SP, Chronaki G. Vocal emotion recognition in attention-deficit hyperactivity disorder: A meta-analysis. Cogn Emot. 2024 Jan 2; 38(1):23–43. doi:10.1080/02699931.2023.2258590.

[14] Kim Y, Kim S, Yoon S. Emotion malleability beliefs matter in emotion regulation: A comprehensive review and meta-analysis. Cogn Emot. 2024 Aug 17; 38(6):841–856. doi:10.1080/02699931.2024.2334833.

[15] Li JX. Emotion expression in modern literary appreciation: An emotion-based analysis. Front Psychol. 2022 Jun 17; 13:923482. doi:10.3389/fpsyg.2022.923482.

[16] Lange J, Zickfeld JH. Emotions as overlapping causal networks of emotion components: Implications and methodological approaches. Emot Rev. 2021 Apr; 13(2):157–167. doi:10.1177/1754073920988787.

[17] MacCann C, Erbas Y, Dejonckheere E, Minbashian A, Kuppens P, Fayn K. Emotional intelligence relates to emotions, emotion dynamics and emotion complexity a meta-analysis and experience sampling study. Eur J Psychol Assess. 2020 May; 36(3):460–470. doi:10.1027/1015-5759/a000588.

[18] Durán JI, Fernández-Dols JM. Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. Emotion. 2021 Oct; 21(7):1550–1569. doi:10.1037/emo0001015.

[19] Aznar A, Tenenbaum HR. Gender comparisons in mother-child emotion talk: A meta-analysis. Sex Roles. 2020 Feb; 82(3–4):155–162. doi:10.1007/s11199-019-01042-y.

[20] Saman Z, Louise F, Lucy T. AI03 Steps towards safely deploying the world's first autonomous artificial intelligence as a medical device for skin cancer into a National Health Service teledermatology pathway[J]. British Journal of Dermatology, 2025(Supplement_1):Supplement_1.DOI:10.1093/bjd/ljaf085.201.

[21] Cama-Olivares A, Braun C G, Takeuchi T, et al. Artificial Intelligence/Machine Learning Externally Validated Models for AKI Risk-Classification: A Systematic Review and Meta-Analysis[J]. Journal of the American Society of Nephrology, 2024, 35(10S).DOI:10.1681/ASN.2024mp7992ze.

[22] Zhang H, Fang B, He P, et al. The asymmetric impacts of artificial intelligence and oil shocks on clean energy industries by considering COVID-19[J]. Energy, 2024, 291.DOI:10.1016/j.energy.2023.130197.

[23] Fumagalli D, Palmieri E, Tarantino V, et al. Artificial Intelligence-Based Models for Early Detection Of Endometrial Cancer And Atypical Hyperplasia From Transvaginal Ultrasound Images[J]. International Journal of Gynecological Cancer, 2025, 35(2).DOI:10.1016/j.ijgc.2024.100087.

Object Recognition in Pond Environments Using Deep Learning

Suhaila Sari[1], Ng Wei Jie[2], Nik Shahidah Afifi Md Taujuddin[3], Hazli Roslan[4], Nabilah Ibrahim[5], Mohd Helmy Abd Wahab[6]

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia[1, 2, 3, 5, 6]
Faculty of Engineering Technology, Universiti Tun Hussein Onn Malaysia, Pagoh, Johor, Malaysia[4]

***Abstract*—Complicated underwater environment, such as visibility limitations and illumination conditions pose significant challenges for underwater imaging and its object recognition performance. These issues are especially critical for applications involving autonomous underwater vehicles (AUVs) or robotic systems involved in object recognition tasks during search-and-retrieval operations. Moreover, high-turbidity underwater image datasets, especially for pond environments, remain scarce. Therefore, this study focuses on establishing a pond underwater images dataset and evaluating the deep learning-based object recognition architecture, You Only Look Once Version 5 (YOLOv5), in recognizing multiple objects in respective underwater pond images. The dataset contains self-captured 1116 underwater pond images, which are annotated with LabelImg for object recognition and dataset generation. Under varying depths, camera distances, and object angles, the YOLOv5 reaches a mean accuracy mAP 50-95 of 87.96%, demonstrating its effectiveness for recognizing multiple objects in pond underwater environments.**

***Keywords*—Dataset; deep learning; object recognition; pond; underwater image; YOLO**

I. Introduction

Complicated underwater environments, such as visibility limitations and illumination conditions, pose significant challenges for underwater imaging and its object recognition performance. These aspects may degrade the underwater image quality and lower the performance of object recognition systems. These issues are especially critical for applications involving autonomous underwater vehicles (AUVs) or robotic systems involved in object recognition tasks during search-and-retrieval operations.

Underwater images can be divided into three categories: lakes, seas, and ponds. The main distinction between ponds and lakes is their size, which is described by surface area and depth. Ponds are usually defined by their smaller surface areas and shallower depths compared to lakes, which usually have greater width and deeper basins [1]. Due to limited water input and reduced depth, ponds often have minimal water circulation. It causes stagnant conditions for the water body. These characteristics may decrease light penetration. Consequently, it causes low-visibility environments that produce poorly contrasted and blurred underwater views. The presence of suspended particles, which are commonly brownish or greenish in appearance, is typically associated with nutrient loading. These particles promote algal blooms [2]. The presence of algae not only discolor the water (green, brown, or red) but also degrades water quality through oxygen reduction during the decomposition phase [3]. These environmental conditions pose significant challenges for underwater imaging and object recognition. The visibility limitations may affect the accurate acquisition of object features such as shape and color. To increase the performance of object recognition techniques, the development of a sufficiently large and representative dataset is crucial. However, high-turbidity underwater image datasets, especially for pond environments, remain scarce [4]. The development of such datasets may limit the development and evaluation of robust object recognition models in pond ecosystems. This gap is particularly important for applications involving autonomous underwater vehicles (AUVs) or robotic systems involved in object recognition tasks during search-and-retrieval tasks, where accurate object detection is essential. This study addresses the challenge by developing and assessing a multiple object recognition system tailored for underwater pond images by utilizing the You Only Look Once Version 5 (YOLOv5) deep learning architecture. The specific objectives of this study are as follows:

1) To develop a dedicated dataset for multiple object recognition under pond-specific underwater conditions.

2) To evaluate the existing approach for multiple object detection in underwater pond images.

Given the limitations of low visibility and the existence of turbidity in underwater pond environments. The research question is: Can a deep learning-based approach recognizes multiple objects using a self-collected and annotated dataset?

This study focuses on establishing a pond underwater image dataset and evaluating the deep learning technique, YOLOv5, in object recognition accuracy for multiple objects in respective underwater pond images. The successfully created dataset for this study can be accessed via the following link [a].

This study is organized as follows: Section II discusses existing datasets and related works; Section III explains the methodology, including dataset development and object recognition model; Section IV provides the results and analysis; Section V discusses the performance of YOLOv5, challenges, limitation and future direction of this work; and Section VI concludes key findings and future directions of the research.

[a] Created dataset: http://tinyurl.com/32z25b

II. Related Works

In recent years, advancements in underwater exploration and imaging technologies have led to an increasing number of underwater visual data. Despite these developments, visual images acquired in underwater environments often exhibit significant visual degradation. Common problems include limited visibility, low contrast, non-uniform lighting, blurred textures, diminished color fidelity, and image noise [5]. These issues are primarily caused by the interaction of light with existing particles in water. They absorb and scatter incoming rays, thereby reducing image clarity. The phenomenon of turbidity related to a high concentration of underwater particles contributes to the haziness frequently observed in underwater images [4]. The Marine Underwater Environment Database (MUED) addresses this challenge by providing 8,600 underwater images across 430 distinct object groups. It consists of complex backgrounds, multiple salient entities, and variations in position, illumination, spatial orientation, and turbidity levels [6]. While raw underwater images often display high-level color distortion and contrast loss, the accompanying reference images are corrected to reflect more accurate color representation, enhanced brightness, and visibility. These comprehensive datasets are critical in overcoming the lack of annotated benchmark images. It offers valuable resources for training, validation, and performance evaluation in underwater image enhancement and object recognition research. Fig. 1 shows examples of the MUED underwater images.

Furthermore, the Enhancing Underwater Visual Perception (EUVP) dataset was proposed as a representative database that includes a paired and an unpaired collection of 20,000 underwater images (of poor and good quality) that can be used for training [7]. This dataset was specifically created to accurately reflect the current characteristics and challenges encountered in underwater settings. The dataset is a large assortment of video clips taken in different underwater settings. The videos have varying degrees of water turbidity, lighting, and image degradation. This collection is a wide set to allow researchers to develop and evaluate algorithms aimed at applications like image enhancement, restoration, and object recognition with a particular focus on underwater settings. Using the EUVP dataset, researchers can tackle the special visual problems presented by underwater imaging and try to enhance the quality and analysis of underwater videos and images. Fig. 2 presents example images derived from the EUVP dataset.

Recognition of objects in underwater surveillance, exploration, and assessment is essential in analyzing zones of interest. Not only is object identification important but also information must be extracted from the objects, thereby necessitating object recognition to be an imminent research domain under low-light and underwater environments [8]. In recent years, object recognition has been utilizing deep learning techniques as well. Object recognition methods based on deep learning have been extensively used in various fields [9][10][11], with specific applications in underwater image scenes [12][13][14]. One of the techniques is based on You Only Look Once architecture. Building upon YOLOv2, YOLOv3 introduced several key advancements. It utilized a feature extraction backbone called Darknet-53, consisting of 53 convolutional layers, which improved the model's ability to extract meaningful features [15]. The YOLOv5 also offers support for the latest computer vision algorithms such as instance segmentation, enabling multiple object recognition in an image or videos [16].

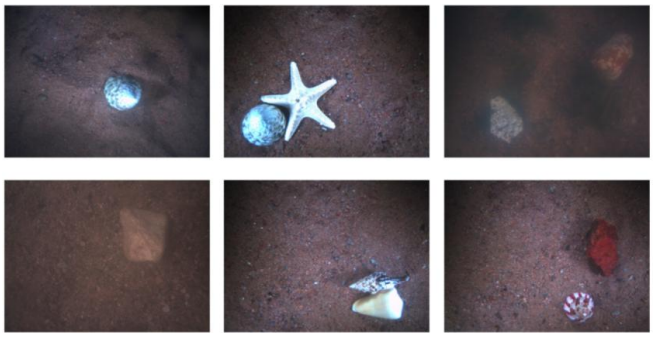


Fig. 1. MUED underwater images [6].



Fig. 2. Example of EUVP dataset images [7].

III. Methodology

A. Data Preparation

The dataset consists of five different classes representing real-world objects: Male, Female, Airplanes, Car, and Helicopter. The images are categorized into five classes based on the number of objects in the image: 1 object, 2 objects, 3 objects, 4 objects, and 5 objects. The images were captured under different conditions and positions to ensure a total of 1116 images. The images were captured at a pond in Taman Muhibbah, Saleng, 81400, Senai, Johor. The images were captured based on three different object heights from the surface (20 cm, 40 cm, and 60 cm), three different object-to-camera distances (10 cm, 20 cm, and 30 cm), and four different object surface directions (0°/360°, 90°, 180°, 270°). These factors ensure a diverse and comprehensive dataset of 1116 images. Fig. 3 shows the toys representing real-world objects.

The images are captured with a waterproof action camera GoPro Hero 5. The dataset is processed using LabelImg software to label the images. The object recognition and image enhancement techniques are developed using Google Colaboratory, executed on a personal desktop with the following specifications:

- Windows 10 Home
- Intel(R) Core (TM) i7-6700
- 12GB RAM, 240GB SSD, and 1TB HDD

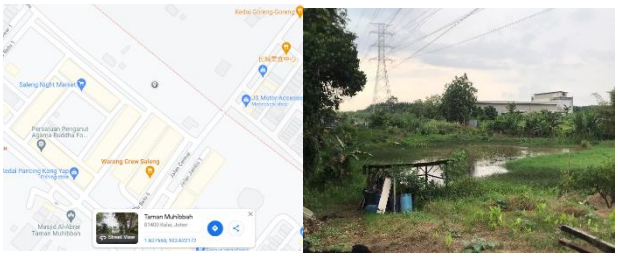
The GoPro Hero 5, known for its versatility and durability, is well-suited to meet the demands of project requirements. The selected resolution of 1080p ensures a balance between video clarity and file size, optimizing storage efficiency while preserving visual details. The camera's compact design and robust build make it suitable for various environments, aligning seamlessly with the dynamic nature of the project. Fig. 4 shows the parameter setting for the GoPro Hero 5. Fig. 5 shows the location of the experiment in the pond at Taman Muhibbah.



Fig. 3. Toys used in the project.: (a) Airplane, (b) Helicopter, (c) Car, (d) Male, (e) Female.



Fig. 4. Parameter setting for GoPro Hero5 camera.



(a) (b)

Fig. 5. Location 2 for experiment; (a) Taman Muhibbah pond from google map, (b) Taman Muhibbah pond.

B. Data Distribution

In this study, two software programs, namely Google Colaboratory and LabelImg, are utilized. The LabelImg software is employed for labeling objects in the gathered images to create a dataset. Another software tool employed in this research is Google Colaboratory. It was utilized to perform object recognition tasks via the YOLOv5 deep learning framework. During the dataset preparation process, thorough consideration was given to ensure a balanced and representative distribution of underwater images across all data subsets. The dataset was divided into three main folders: 80% of the images were allocated for training, 10% for testing, and the remaining other 10% for validation purposes [17]. This division was implemented to support effective model training and facilitate rigorous performance evaluation. It also enables fine-tuning of hyperparameters in the validation phase. The specific dataset distribution across these subsets is presented in Table I.

TABLE I DISTRIBUTION OF UNDERWATER IMAGES DATASET

Object quantity	Train folder		Test folder		Valid folder		Total Image No.
	%	*Image No.*	%	*Image No.*	%	*Image No.*	
1	80	144	10	18	10	18	180
2	80	288	10	36	10	36	360
3	80	288	10	36	10	36	360
4	80	144	10	18	10	18	180
5	80	29	10	4	10	3	36
Combine images (all)	80	893	10	112	10	111	1116

C. You Only Look Once Version 5 (YOLOv5)

Object recognition in underwater images encountered a number of issues such as poor visibility, dynamic illumination, and particle pollution. The YOLOv5 is a complex object recognition technique that can be adapted to address such issues. This section discusses the key mechanisms and steps involved in utilizing YOLOv5 for object recognition for underwater images. Fig. 6 demonstrates the flowchart of YOLOv5. It utilizes a single neural network that can directly predict bounding boxes and class probabilities from entire images in a single evaluation step. There are three elements that build up the architecture, which are the Backbone, Neck, and Head.

The process of training YOLOv5 for underwater object recognition includes some important steps as follows:

1) Data preparation: The underwater images are annotated using bounding boxes and object-class-specific labels.

2) Data augmentation: To improve the robustness and generalization of the model, a variety of data augmentation techniques can be employed. The examples are random cropping, scaling, flipping, color correction, and underwater-specific augmentations like simulating turbidity and altering light conditions.

3) Loss function: The YOLOv5 employs a combination of several loss functions to train the model, which are the Localization Loss, Confidence Loss and Classification Loss.

4) Optimization: Model training is carried out using stochastic gradient descent (SGD) or the Adam optimizer. Hyperparameters such as learning rate, batch size, and momentum are carefully tuned to achieve the best performance results. In many cases, transfer learning is used, starting with a pre-trained model trained on a large dataset and then fine-tuning it on the target underwater dataset.

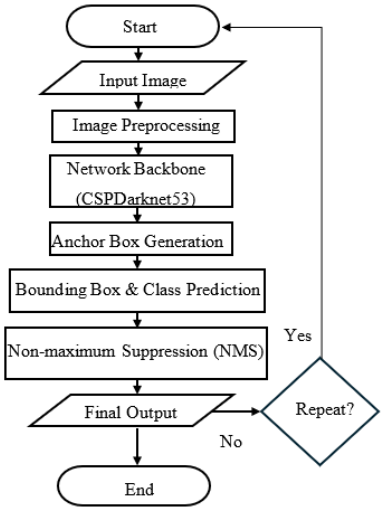


Fig. 6. YOLOv5 approach for object recognition.

At inference time, YOLOv5 feeds the input image to the trained network to get the bounding boxes and class probabilities. The subsequent steps are:

1) Image preprocessing: The original image is resized to a specific size and normalized. Additional preprocessing steps can include color correction and noise reduction to deal with the specific nature of underwater images.

2) Forward pass: The preprocessed image is passed into the network, resulting in the generation of feature maps via the backbone, neck, and head modules.

3) Post-processing: Non-Maximum Suppression (NMS) is applied to eliminate duplicate bounding boxes and retain the most confident predictions. Additional post-processing may include object size and confidence thresholds specific to underwater object filtering.

YOLOv5 is a significant improvement in real-time object recognition with improved accuracy and efficiency for object recognition underwater. Its architecture, training process, and inference steps are meticulously designed to handle the unique challenges of underwater environments.

IV. Result and Analysis

This section examines the results and analysis of the underwater image dataset and the object recognition system using Google Colaboratory. The results and analysis are discussed in the following order: i) Dataset development for pond underwater images and ii) YOLOv5 performance analysis for all combined pond underwater images.

A. Dataset Development for Pond Underwater Images

The created dataset can be accessed via the following link [a]. Fig. 7 shows the underwater image acquisition arranged based on the Car, Male, Female, Helicopter, and Airplane categories. It displays the images of each object according to different underwater conditions. The settings show examples of different quantities of objects with depths from the surface varying from 20 cm, 40 cm, and 60 cm inside the pond. The distance of the object to the camera and the object's angle are fixed, with a distance of 30 cm from the object to the camera and the angle of objects all set to 0°. Despite the brownish tone of the pond, which contrasts with the greenish tint of the lake, the objects remain apparent in the photographs, with constant settings for object depths and camera distances. However, with these settings for object distances from the surface and camera, the objects can still be seen in the images.

Fig. 8 shows the example of underwater image object-to-camera distance setting for image acquisition for the pond, set to 10 cm, 20 cm, and 30 cm. The distance from the surface and angle of the object are fixed, with the distance from the surface set to 20 cm and the angle of the object remaining at 0°. When the distance of objects is nearer to the camera, the objects can still be seen, although the distance from the surface differs. It can be observed that for nearer distances, there are possibilities of objects at the edge being partially displayed in the image due to the camera's capturing width limitations. In some cases, the objects may overlap with neighboring objects for higher quantity objects in an image. It is critical to recognize that these visual impacts may differ in the pond due to its brownish color compared to the lake's greenish tint.

Fig. 9 shows the examples of underwater image angle settings for the images in the pond captured at 0°, 90°, 180°, and 270°. The depth from the surface and object-to-camera distance are fixed, with the depth set to 20 cm from the surface and the distance of the object from the camera set to 30 cm. When the object distances from the surface and camera are longer, the image turbidity tends to be higher, and object visibility becomes less clear. The visual effects witnessed in the pond, marked by its brownish hue, may diverge from those in the lake, distinguished by a greenish tint. These environmental distinctions have the potential to influence the overall clarity and appearance of objects captured within the images.

B. YOLOv5 Performance Analysis for All Combined Underwater Images

In this analysis, the focus is on a meticulous evaluation of YOLOv5's performance, leveraging a dataset explicitly tailored for underwater images in a pond environment. The assessment encompasses critical metrics such as confusion matrices, result graphs, result values, and illustrative images, providing a detailed perspective on the model's prowess in object recognition. The aim is to deliver a comprehensive understanding of YOLOv5's efficacy in detecting and identifying objects, particularly amid the distinctive challenges posed by underwater conditions in pond environments.

C. YOLOv5 Training Performance: Confusion Matrix for Pond Images

Fig. 10 shows the confusion matrix for pond images for all combined images. The confusion matrix shows four crucial values: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The diagonal elements represent correctly predicted samples. In this specific case, out of a total of 892 samples, 892 samples were accurately predicted, leading to an overall accuracy of 100%. From these 892 samples, 67 instances of airplane classes were detected, 56 instances of car classes were detected, 50 instances of female classes were detected, 45 instances of male classes were detected, and 61 instances of helicopter classes were detected.



Fig. 7. Different surface distance settings for pond underwater images (distance of object to camera is 30 cm, and angle of object is 0 °).



Fig. 8. Different camera distances from the object setting for pond underwater images (depth of water surface is 20cm and angle of object is 0 °).



Fig. 9. Different angle of object setting for dataset for pond underwater images (depth from water surface is 20 cm, distance of object to camera is 30 cm).

D. YOLOv5 Training Performance: Result Graph for Pond Images

Fig. 11 shows the training result graph for pond images for all combined images. The YOLOv5 model, trained over 150 epochs, exhibited notable improvements in alignment and performance between epochs 50 and 100, as indicated by the decreasing training losses. The evaluation metrics, including precision, recall, mAP 50, and mAP 50-95, provided a comprehensive overview of the model's ability to accurately detect and classify objects. During the analyzed epochs, precision consistently increased, indicating better object classification, while recall remained high, demonstrating effective object recognition. The mAP 50 and mAP 50-95 metrics, which assess the mean average precision at different IoU thresholds, reflected the model's proficiency in localization and classification tasks.

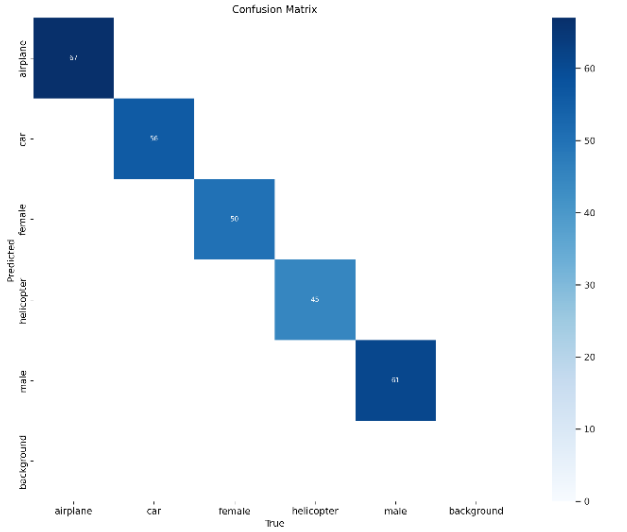


Fig. 10. Training confusion matrix for pond images (combine all images) YOLOv5.

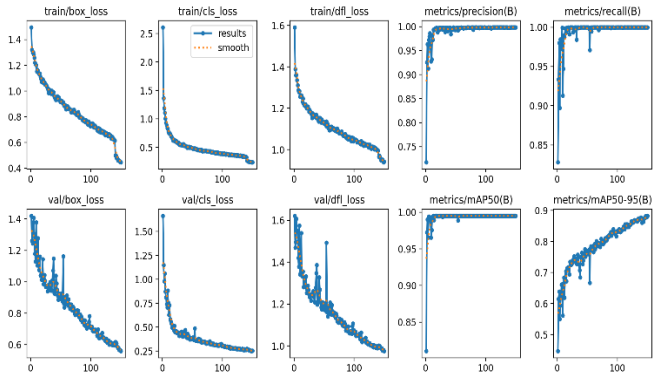


Fig. 11. Training result graph for pond images (combine all images) YOLOv5.

E. YOLOv5 Validation Performance: Confusion Matrix for Pond Images

Fig. 12 shows the confusion matrix for pond images for combined all images. The confusion matrix shows four crucial values: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The diagonal elements represent correctly predicted samples. In this specific case, out of a total of 112 samples, 112 samples were accurately predicted, leading to an overall accuracy of 100%. From these 112 samples, 67 instances of airplane classes were detected, 56 instances of car classes were detected, 50 instances of female classes were detected, 45 instances of male classes were detected, and 61 instances of helicopter classes were detected.

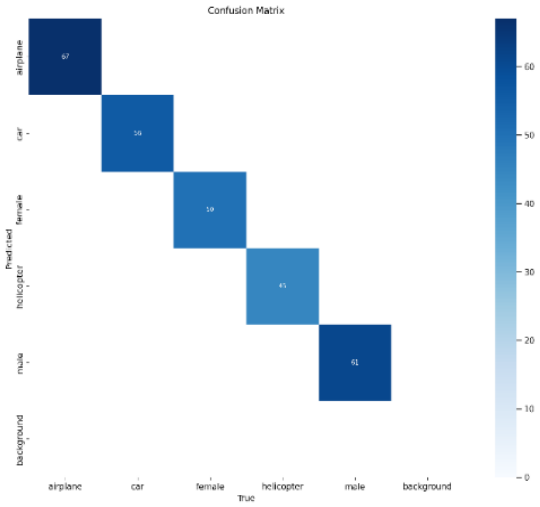


Fig. 12. Validation confusion matrix for pond images (combine all images) YOLOv5.

F. YOLOv5 Validation Performance: Visual Result for Pond Images

Fig. 13 shows examples of images randomly chosen from the validation set of pond images for all combined images. The images chosen include 1 object, 2 objects, 3 objects, 4 objects, and 5 objects. The system detected the objects with correct classes, which are airplane, car, helicopter, male, and female, with the average accuracy detected value above 90%.

G. YOLOv5 Validation Performance: Overall Accuracy Mean Average Precision 50-95 (mAP 50-95) for Pond Images

Table II shows the overall accuracy mean average precision 50-95 (mAP 50-95) while executing the validation. The mAP 50-95 for all classes is 88.3%, with airplane at 90.1%, car at 89.1%, female at 86.4%, helicopter at 87.4%, and male at 88.7%. Increasing the number of validation images may contribute to a more comprehensive evaluation, enabling the model to generalize better diverse scenarios and, consequently, enhance overall accuracy.

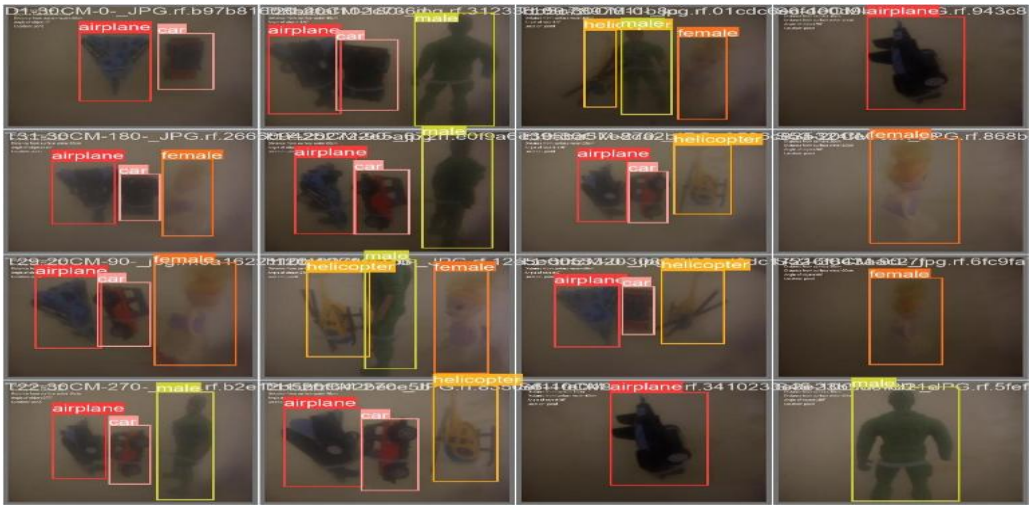


Fig. 13. Validation visual result for pond images (combine all images) YOLOv5.

H. YOLOv5 Test Performance: Confusion Matrix for Pond Images

Fig. 14 shows the testing confusion matrix for pond images for all combined images. The confusion matrix shows four crucial values: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The diagonal elements represent correctly predicted samples. In this specific case, out of a total of 112 samples, were accurately predicted, leading to an overall accuracy of 100%. From these 112 samples, 55 instances of airplane classes were detected, 53 instances of car classes were detected, 61 instances of female classes were detected, 69 instances of male classes were detected, and 56 instances of helicopter classes were detected.

TABLE II VALIDATION OVERALL ACCURACY MEAN AVERAGE PRECISION 50-95 (MAP 50-95) FOR POND IMAGES (COMBINE ALL IMAGES) YOLOV5

Class	Images	mAP50	mAP50-95
All	110	0.995	0.883
Airplane	110	0.995	0.901
Car	110	0.995	0.891
Female	110	0.995	0.864
Helicopter	110	0.995	0.874
Male	110	0.995	0.887

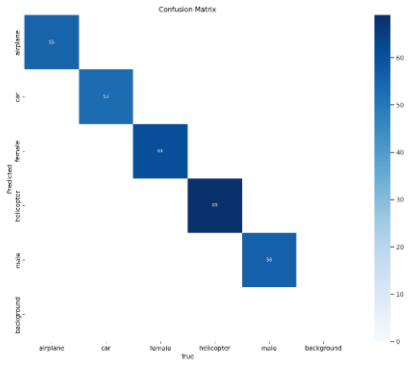


Fig. 14. Test confusion matrix for pond images (combine all images) YOLOv5.

I. YOLOv5 Test Performance: Overall Accuracy Mean Average Precision 50-95 for Pond Images

Table III shows the overall accuracy mean average precision 50-95 (mAP 50-95) while executing the testing. The mAP 50-95 for all classes is 87.4%, with airplane at 89.9%, car at 86.6%, female at 85.6%, helicopter at 87.6%, and male at 87.2%. Increasing the number of testing images may contribute to a more comprehensive evaluation, enabling the model to generalize better to diverse scenarios and, consequently, enhance overall accuracy.

J. YOLOv5 Test Performance: Visual Result for Pond Images

Fig. 15 shows examples of images randomly chosen from the test set of pond images for all combined images. The images chosen include 1 object, 2 objects, 3 objects, 4 objects, and 5 objects. The system detected the objects with correct classes, which are airplane, car, helicopter, male, and female, with the average accuracy detected value above 90%.

K. Overall Evaluation of YOLOv5 Performance

This section focuses on comparing the mAP 50-95 (Mean Average Precision where the IoU threshold is 0.5-0.95) results obtained by YOLOv5 for underwater images in the pond environment. A detailed examination of the mAP scores in each location aims to reveal potential variations in the model's performance under different environmental conditions. Table IV shows the evaluation of YOLOv5 performance.

TABLE III TEST OVERALL ACCURACY MEAN AVERAGE PRECISION 50-95 (MAP 50-95) FOR POND IMAGES (COMBINED ALL IMAGES) YOLOV5

Class	Images	mAP50	mAP50-95
All	112	0.995	0.874
Airplane	112	0.995	0.899
Car	112	0.995	0.866
Female	112	0.995	0.856
Helicopter	112	0.995	0.876
Male	110	0.995	0.872

TABLE IV EVALUATION OF THE YOLOV5 PERFORMANCE

Overall Images of object category and mode	mAP 50-95% for YOLOv5
Train	88.20%
Validation	88.30%
Test	87.40%
Average accuracy	**87.97%**

Under the pond classification category, YOLOv5 records an average accuracy of 87.97%. Despite excellent performance being recorded by YOLOv5, there are slight accuracy variations across various stages of the dataset, i.e., training, validation, and testing.

The mAP 50-95 metric is an overall assessment of the performance of the model, considering precision and recall at different Intersection over Union (IoU) thresholds. High mAP values observed on training, validation, and test datasets. It demonstrates that YOLOv5 has high proficiency in object recognition and classification in underwater images.

Training Performance: The mAP 50-95 score of 88.20% in training exhibits that the model has learned to recognize and classify objects present in the training dataset successfully. The high score indicates the model's ability to generalize well to the training data.

Validation Performance: The validation mAP 50-95 score of 88.30% shows that the model maintains its performance when applied to unseen data during training. This stability between train and validation scores proves that the model is not overfitting and can generalize well to new data.

Testing Performance: The testing mAP 50-95 of 87.40% is less than the train and validation. However, it still represents good performance. This small reduction in accuracy is to be expected because the testing dataset is made of completely different underwater image dataset. The high-test score proves the strength and reliability of the model for real-world scenarios.

The consistency of high mAP scores throughout all datasets shows the model's effectiveness and reliability.

V. DISCUSSION

The results obtained from this study demonstrate the strong performance of YOLOv5 in multiple object recognition within pond environments. The average mAP 50–95 of 87.97% shows the robustness of YOLOv5 in object recognition under presence of turbidity and low visibility in pond environments.

One of the challenges is the lack of pond underwater datasets. Most datasets focus on marine and lake setting, with many with lower level of turbidity. These differences affect the results of observations across different water body environments.

The YOLOv5, while effective, still have room for improvement in object recognition, especially in the case of high-turbidity conditions. For future work, the use of image enhancement techniques as image processing, may further improve model accuracy in object recognition.

Despite these challenges, this study fills a critical gap and offers a valuable contribution for future research in underwater object recognition, especially for pond environments.



Fig. 15. Test: visual result for pond images (combine all images) YOLOv5.

VI. Conclusion

This study presents a contribution to underwater object recognition by developing a self-captured and annotated dataset for pond environments. The created dataset can be accessed via the following link [a]. Through investigation utilizing the datasets, the performance evaluation of YOLOv5 with the average mAP 50-95 of 87.97% metric exhibits that the model has a high capability in recognizing and classifying multiple objects in underwater images. The high accuracy shown on training, validation, and test datasets demonstrates that YOLOv5 has good generalization capability to unseen data and maintains stable performance in real-world underwater environments. However, this study has limitations, including the need for image pre-processing and more image datasets. Future studies can consider further optimization of the YOLOv5 model and its use in numerous underwater environments to improve its robustness and accuracy. Overall, this study is hoped to offer new insights and resources for object recognition not only in pond environments, but also in the real-world underwater environments.

Acknowledgment

This research is supported and funded by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1(Q377) Grant.

References

[1] J. Britton, "Formation and features of ponds and lakes," Study.com, 2023 [Online]. Available: https://study.com/learn/lesson/pond-vs-lake-formation-features.html.

[2] MSU Extension Natural Resources, "What is that color in my pond?" 2022. [Online]. Available: https://www.canr.msu.edu/news/what-is-that-color-in-my-pond.

[3] K. F. Smith, J. Stuart, and L. L. Rhodes, “Molecular approaches and challenges for monitoring marine harmful algal blooms in a changing world,” Frontiers in Protistology, vol. 1:1305634, 2024.

[4] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," IEEE Transactions on Image Processing, vol. 29, pp. 4376-4389, Nov. 2019.

[5] E. Lucas et al., " Underwater image enhancement and object detection: Are poor object detection results on enhanced images due to missing human labels?" in 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) pp. 1435-1440, Feb. 2025.

[6] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using Generative Adversarial Networks," in IEEE International Conference on Robotics and Automation (ICRA), pp. 7159-7165, Sept. 2018.

[7] Y. Ma et al., "A new database for evaluating underwater image processing methods," in IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1-6, Nov. 2018.

[8] N. A. I. N. Azlin et al., "Object recognition system with image enhancement for lake underwater images," Evolution in Electrical and Electronic Engineering, vol. 4(1), pp. 559-565, May 2023.

[9] G. Rani, S. Khandelwal, N. Kundu and V. S. Dhaka, "Automatic segmentation and classification of skin cancer cells using thresholding and Deep Learning based techniques," in IEEE International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), pp. 1-6, Aug. 2023.

[10] H. H. Kang, A. Ab Malik, M. A. Osman, Y. Yusop and I. H. Hamzah, " Healthy and unhealthy oil palm tree detection using Deep Learning method," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 16(4), pp. 748-757, 2025.

[11] N. Das and S. Das, "Epoch and accuracy based empirical study for cardiac MRI segmentation using Deep Learning technique," PeerJ, Mar 2022.

[12] L. Chen et al., "Underwater object detection in the era of Artificial Intelligence: Current, challenge, and future," arXiv preprint arXiv: 2410.05577, Oct. 2024.

[13] L. Chen et al., "Underwater object detection in the era of Artificial Intelligence: Current, challenge, and future," arXiv preprint arXiv:2410.05577, Oct. 2024.

[14] Y. S. Tsai, C. T. Tsai, and J. H. Huang, "Multi-scale detection of underwater objects using attention mechanisms and normalized Wasserstein distance loss," The Journal of Supercomputing, vol. 81, pp. 1-33, Mar. 2025.

[15] Y. Yang et al., "Image defogging based on amended Dark Channel Prior and 4-Directional L1 Regularisation," IET Image Processing, vol. 15, pp. 2454-2477, Apr. 2021.

[16] J. H. Kim, N. Kim, Y. W. Park, and C. S. Won, "Object detection and classification based on YOLO-V5 with improved maritime dataset," Journal of Marine Science and Engineering, vol. 10, no. 3, pp. 377, Mar. 2022.

[17] A. Katuria and S. Mukherjee, "How to Train YOLO v5 on a custom dataset," DigitalOcean Community Tutorials, Mar. 4, 2025. [Online]. Available: https://www.digitalocean.com/community/tutorials/train-yolov5-custom-data.

FoodSharePro: An Integrated Mobile Platform for Sustainable Food Donation and Decentralized Composting

Jamil Abedalrahim Jamil Alsayaydeh[1*], Rex Bacarra[2], Shamsul Fakhar Bin Abd Gani[3], Serhij Mamchenko[4], Safarudin Gazali Herawan[5]

Department of Engineering Technology, Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[1, 3]

Department of General Education and Foundation, Rabdan Academy, Abu Dhabi, United Arab Emirates[2]

Department of Computer Systems, Networks and Cybersecurity, National University of Life and Environmental Sciences of Ukraine[4]

Industrial Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia 11480[5]

***Abstract*—Food waste is a global issue, one-third of all food produced is lost every year. This study introduces FoodSharePro, an integrated mobile-based Waste Food Management System that connects food donation and composting. The system allows for the efficient donation of surplus edible food through a mobile app and management of inedible waste through traditional composting methods. Built using Android Studio and Google Firebase, the app has secure authentication, location-based rider matching via Google Maps API and real-time data synchronization. Donors can log donations, track status and view delivery confirmations through a user-friendly dashboard, while riders are assigned tasks based on location and transport suitability. To minimize organic waste, composting hardware with temperature sensors and dehydration units supports aerobic composting process. Evaluation among 20 users showed that FoodSharePro has the highest satisfaction rate (75%) compared to 6 other platforms, with a mean user satisfaction of 24.29% and a standard deviation of 24.25%. The results prove that mobile technology can be integrated with grassroots waste management to reduce food loss and be sustainable.**

***Keywords*—Food waste management; food donation platforms; mobile applications; traditional composting; smart waste systems; user engagement; foodsharepro**

I. Introduction

The global food system has a big problem: one-third of all food produced globally is lost or wasted while a big chunk of the world's population is hungry and malnourished [1], [2]. This isn't because of lack of food production but because of inefficiencies in food distribution, consumption and waste management. The Food and Agriculture Organization (FAO) has been saying that there's enough food produced annually to feed the global population but inequalities in access, logistics and preservation are causing food insecurity [1]. Meanwhile, food waste is exacerbating environmental problems, contributing to greenhouse gas emissions, inefficient land use, water depletion and biodiversity loss [2].

In this context, solutions that bridge technology, sustainability and food security are critical. One of the most promising areas is the use of mobile and digital platforms to facilitate surplus food redistribution. These platforms connect food donors – from households and restaurants to supermarkets – with non-profit organizations and food-insecure communities. The success of these apps relies on real-time connectivity, user-friendly interfaces and efficient logistics to ensure that edible food is rescued and delivered before it spoils [3], [4]. But while these donation systems have reduced edible food waste, they often operate independently of organic waste management systems and don't address food that's no longer fit for consumption.

Meanwhile, traditional waste management methods – composting, anaerobic digestion, or using food waste as animal feed – aim to minimize the environmental impact of inedible food. But these methods often exist in isolation from real-time food donation efforts. As a result, many food donation platforms lack a "fallback" mechanism for managing food that can't be rescued in time. Composting systems lack the intelligence to first check if the waste could be diverted to feed people. The absence of a holistic approach means food is wasted because of missed donation windows or composted unnecessarily, despite being edible. And the urgency of this issue is growing with urbanization and population growth. In densely populated areas, food waste logistics become more complex. Limited infrastructure, an inadequate cold chain, and inconsistent policy support hinder both food donation and organic waste management. Add to that food safety, verification of donations and transparency in the donation and recycling process. So there is a growing need for smart, scalable and context-aware systems that bridge the gap between surplus food redistribution and sustainable waste management.

To address this, there is a growing need for a unified, intelligent system that can dynamically decide the optimal use of surplus food—either donation or composting—based on real-time availability, perishability, and logistics.

This research introduces FoodSharePro, an integrated Waste Food Management System that manages both edible food donation and inedible food composting through one digital platform. The system is built using Android Studio and uses Firebase for secure user authentication, real-time data

management and coordination. Google Maps API is used for efficient pickup routing and hardware-based composting systems are implemented on-site for organic waste that cannot be donated. Unlike existing solutions that treat food donation and waste processing as separate domains, FoodSharePro allows seamless redirection of surplus food: if donation is not possible within a certain time frame, the system defaults to processing the food waste through trench composting. This dual-path design ensures that every piece of surplus food is accounted for, either by feeding people or by regenerating soil. The significance of this research is the holistic approach to the food waste lifecycle. By combining mobile application development with sensor supported composting systems the platform provides a robust community scalable framework for environmental sustainability, food security and circular resource use. Moreover it empowers donors, NGOs and local volunteers with transparency, real-time monitoring and secure verification – building trust and engagement across the food recovery network.

The main contributions of this research are as follows:

- Smart mobile-based platform for food donation that classifies, schedules and verifies using Firebase and Google Maps.
- Traditional composting systems, featuring dehydration and sensor monitoring, are used for non-donatable food waste.
- Bridging the gap between donation and composting workflows so no food is left unmanaged.
- User testing and a comparison with existing food donation apps demonstrate higher engagement, usability, and system satisfaction.

Despite the availability of food-sharing applications and isolated composting initiatives, there remains a lack of integrated platforms that simultaneously address food wastage, donation logistics, and sustainable waste management through composting. This study addresses this gap by developing FoodSharePro - an integrated mobile solution that connects food donors, recipients, and composters in a unified ecosystem.

The study proceeds as follows: Section II reviews existing methods relevant to the current study. The proposed approach is presented in Section III. Section IV presents the experimental results, while Section V discusses the findings in detail. Section VI outlines the conclusions drawn from the study, and Section VII highlights the limitations and proposes potential directions for future work.

II. Literature Review

Despite an abundance of food globally, hunger persists in many regions. Part of the problem is inefficiency in distribution and waste. Studies show that a significant portion of global food production is never consumed, exacerbating hunger and causing economic and environmental losses [1]. Solving this requires both reducing food waste and increasing food access for vulnerable populations. Traditional approaches to food security have focused on increasing production but there is growing recognition that managing existing surplus could immediately help alleviate hunger without using more resources [2]. This has led to a surge in interest in food recovery – rescuing edible food that would otherwise be thrown away – as a key strategy alongside longer term waste reduction measures. In response to the twin problems of food waste and hunger, many mobile apps and digital platforms have been developed to facilitate food donation and sharing [3]. These platforms connect donors (restaurants, grocery stores, households with excess food) with recipients or charities in real time, redirecting surplus food to feed people instead of waste [4]-[7]. Examples include food rescue apps and web-based donation services that allow users to post available surplus food, schedule pickups and coordinate with non-profits for distribution. Previous research and practical implementations have shown that technology-driven food sharing networks can significantly reduce the amount of edible food that goes to waste [8]-[10]. By providing user-friendly interfaces and instant communication channels, donation apps lower the barrier for businesses and individuals to participate in food redistribution, resulting in more efficient recovery of surplus food to benefit communities in need. Despite their growing popularity, food donation apps and platforms face several limitations that hinder their overall effectiveness. One common challenge is the logistical complexity of matching donations with recipients under time pressure – prepared or perishable food has a very short shelf life, making coordination critical. Many donation systems struggle to ensure that available food is collected and delivered before it spoils. Related to this are concerns about food safety and liability; donors may not participate due to uncertainty around regulations or risk of spoilage even with legal protection [11]-[15]. Another limitation is scalability and reach: current platforms often have limited coverage, focusing on urban areas, while rural or remote areas are underserved. Some communities lack awareness or access to the necessary technology to utilize these apps, creating gaps in adoption. Not all surplus food is donation-worthy – food that is already spoiled or not fit for human consumption must be handled differently. Donation systems don't have solutions for these inedible leftovers, so even with donation programs in place, some food waste remains unaddressed [16]-[18]. These limitations highlight the need for a more comprehensive approach that can manage surplus food in its entirety, beyond just the edible fraction.

Composting is a well-established method to recycle organic waste and has been advocated as a way to divert food scraps from landfills [19] [20]. Through the natural biodegradation process, food waste is converted into nutrient-rich compost that can be used to enrich soil for agriculture or gardening and close the loop in the food supply chain. Traditional aerobic composting (done in piles or bins) requires a balanced mix of organic materials, adequate aeration and weeks or months of decomposition time to produce mature compost [21]. While effective in producing high quality soil amendments, this process can be slow and may generate odors or attract pests if not managed properly. To address some of these practical challenges, alternative composting methods have been explored in recent years. One such approach is dehydration based composting which involves using heat or mechanical drying to rapidly remove moisture from food waste. By dehydrating the waste, the volume and weight are significantly reduced – often by 70 to 80% – and the remaining dried material is stabilized,

meaning it is far less prone to rot or emitting foul odors. The dehydrated output can be stored for longer periods without spoilage and can later be mixed with soil or composted in a conventional manner to complete the decomposition. Several studies and commercial solutions have highlighted the benefits of dehydration systems: they allow food waste to be handled quickly in settings like restaurants or institutions where large quantities of scraps are generated daily, and they produce a sterile, easily transportable dried product [22]. However, it is noted that dehydration is not a final treatment; the nutrient-rich powder or chips resulting from the process typically still require incorporation into soil or compost to fully return nutrients to the ecosystem. Another alternative method gaining attention is trench composting, a simple and low-technology way to dispose of food waste on-site. Trench composting involves digging a pit or trench in the ground, depositing organic waste into it and then covering it with a layer of soil. The buried waste is left to decompose anaerobically underground, turning into compost over a period of months. This method has been promoted for household and community use, as it requires minimal equipment – essentially just a shovel and some space in the yard – and little ongoing maintenance compared to managing a compost bin or tumbler. The soil covering helps to contain smells and deter pests, addressing two common issues with above-ground composting. Over time, soil microorganisms and insects break down the buried food scraps, enriching the surrounding soil. Trench composting is especially useful for handling food items that might attract pests (like meats or dairy) because the underground decomposition keeps them out of reach. The main drawbacks are that it needs available land and the decomposition process is relatively slow and not actively managed, so it may take several seasons for the waste to fully break down [23]-[25]. Despite these limitations, trench composting provides an accessible composting option for those who cannot engage in more structured composting setups.

Beyond individual composting methods, there has been a movement towards decentralized composting as a waste management strategy [26]-[30]. Decentralized composting means processing organic waste at or near the source of generation – at the household, neighborhood or institutional level – rather than hauling the waste to large, centralized facilities miles away. Many municipalities and environmental organizations have piloted community composting programs where local composting sites (community gardens, urban farms or neighborhood compost hubs) handle food waste collected from nearby residents. This reduces both the costs and carbon emissions associated with transporting heavy, moisture-rich food waste to distant composting or landfill sites. It also engages community members in sustainability practices, increasing awareness and responsibility for one's own waste [27]-[30]. Studies have shown that decentralized composting systems can be quite effective in diverting waste; for example, apartment complexes and schools that implement on-site composting have cut down significantly the amount of garbage sent to landfills [26]-[28]. Moreover, the compost produced locally is often used to support urban agriculture or green spaces within the same community, thereby creating a circular economy of organic matter [29]-[30]. However, decentralized composting has its challenges in consistency and management – ensuring many small compost sites are properly maintained and participants follow composting guidelines can be tough. To mitigate this, some projects incorporate training, volunteer programs and simple monitoring technologies to help sustain community composting efforts. Overall, the success of decentralized approaches in various contexts shows that we can manage food waste more locally and it's a complementary strategy to large-scale waste treatment facilities.

While progress has been made in food donation systems and composting methods, there is a big gap in integrating these two areas. In practice, food donation and food waste processing are treated as separate phases: first, excess edible food is rescued and distributed, and then whatever spoils or can't be donated is handled through waste management. Because these functions are separate, surplus food that can't be donated in time ends up in landfills instead of being diverted to composting or other recycling methods [31-35]. On the other side, composting programs rarely have a mechanism to first ensure that the waste they process is truly unavoidable (inedible) waste; they treat all incoming organic material as waste by default. This lack of coordination means opportunities are missed to prioritize human consumption of food before disposal. For example, a grocery store might throw away slightly bruised produce that, if connected to a donation network in time, could have been eaten, or a catered event's leftovers might be composted because no donation pickup was arranged. Research shows that without a synchronized system, even well-intentioned sustainability initiatives can fall short: donation apps might reduce hunger on a small scale, and composting reduces landfill waste, but a disconnection between them leaves some food waste unaddressed [36-38]. Recognizing this gap, some recent studies and pilot programs are calling for more holistic frameworks that bridge the gap between food recovery and waste recycling. Researchers have introduced the concept of integrated food waste management that encompasses the entire lifecycle of surplus food, from prevention and donation to recycling and final disposal in one framework [39-44]. Such frameworks align with the idea of a circular economy for food systems where every possible use of surplus food is maximized in a hierarchy: first to feed people, then to feed animals or produce other products, and finally to compost or anaerobically digest to return nutrients to soil. Implementing this hierarchy requires coordination and information sharing between the donation side and the disposal side. Technology is seen as the key enabler for this integration, with proposals for digital platforms capable of tracking food from the moment it becomes surplus, through either donation channels or waste processing routes. For example, an integrated platform might alert food rescue organizations about available donations and simultaneously signal composting units when certain items are not picked up in time, so nothing is discarded without consideration [45-50]. By leveraging real-time data and connectivity (e.g., smartphones, smart bins), such systems could dynamically allocate surplus food to its most appropriate use at any given moment. Despite these ideas, truly unified donation-and-composting systems don't exist yet. To date, most implementations address either food redistribution or organic waste recycling but not both in one. A review of existing solutions shows a clear divide: platforms and services specialize in one area and there is little overlap between a food donation app and a composting program [51]-[54]. Building an integrated solution faces more challenges

than just technical feasibility. Stakeholder coordination is a big hurdle – donors, charitable organizations, waste haulers, compost facility managers and local authorities would all need to collaborate and share information seamlessly. Each stakeholder group has its own priorities and constraints (e.g. food banks are concerned with nutritional value and safety of donations, while composters focus on contamination and carbon/nitrogen balance of waste inputs), so aligning these in one workflow requires careful planning [55]-[57]. Timing is also critical – any system linking donations to composting must account for the perishability of food. This means establishing clear protocols for when an item should be deemed not suitable for donation and transferred to composting, which in turn requires real-time monitoring and decision-making tools [58]-[60]. Social and cultural factors also play a role; some donors or recipients might view composting as a last resort and need education to understand that using food waste for compost is a good outcome, not a failure of the donation process. In summary, bridging technology and sustainability in the context of food surplus management is an open frontier. Multiple studies have highlighted the benefits of connecting food donation networks with waste processing systems, yet no comprehensive solution exists so far [61]-[62]. This gap is an opportunity for innovation. By building upon the successes and lessons of existing food donation apps and composting programs, an integrated platform – like the one proposed in this research – could significantly boost the efficiency and impact of food waste reduction efforts. Our approach directly addresses the shortcomings by uniting the end-to-end process (donation to composting) in one system, so that no food is wasted, whether it's edible or not. The following sections of this study will describe the design and implementation of this solution and test it in real-world conditions.

III. Proposed Method

The development of the FoodSharePro platform followed a user-centered design approach. This section outlines the architectural design, core modules, user interface elements, and the implementation technologies utilized. The process was guided by iterative feedback from potential end users to ensure practicality and usability.

The system is a smart food donation and composting app that connects food donors with volunteer riders to reduce waste and be sustainable. It has a mobile app for Android devices, divided into donor and rider modules, and a traditional composting setup for inedible organic waste. Donors use the app to offer surplus food, which riders can claim and deliver to designated recipients or processing facilities. Any food waste that can't be donated is processed through a composting system, so it's a holistic waste management approach. Overall the system uses modern software and hardware to make food redistribution and organic waste conversion user friendly.

A. Software Architecture

The mobile application was developed using Android Studio and follows a client-server architecture, where the Android app serves as the client interface and Google Firebase functions as the cloud-based backend. Firebase supports secure user authentication, real-time database synchronization, and cloud storage for donation-related images and metadata. The software architecture is client-server where the Android app is the client interface for both donor and rider users and Firebase Cloud Services is the remote server infrastructure. Two main user interfaces (modules) are implemented in the app: one for donors and one for riders, each tailored to their needs. Firebase Authentication handles sign up and login and Firebase Realtime Database (or Cloud Firestore) manages donation listings and status updates in real-time. The integration of Google Maps API enables real-time geolocation tracking and dynamic routing, allowing the system to efficiently match donors with nearby riders. This ensures optimized pickup assignments based on proximity and traffic conditions. For example, to determine the proximity between donors and available riders to assign pickups efficiently. Firebase Cloud Storage is used to store images (like delivery confirmation photos) uploaded by riders. This architecture ensures that data (donation details, user profiles, locations and images) is synced in real-time across users so food donation coordination is seamless.

Fig. 1 shows the flow of interactions for a donor. A donor signs up or logs in; if the login fails, the user is asked to try again until successful, for security. Once logged in, the donor can donate by providing details like food type (e.g. vegetarian or non-vegetarian), quantity, pickup location and preferred pickup time. These details are entered in the Add Donation form and submitted to the system. After submission, the donation is recorded and the process is ended for the donor's current session, as indicated by the "End" node in the flowchart [63]-[65]. Throughout this process, the app is designed to be simple and easy to use so donors can offer food without complexity. Donors also have features like Donation Status to track the real-time status of their donation (pending, accepted by a rider or completed) and Donation History to view past donations. The app uses location data and the Google Maps API to assign a nearby rider to each new donation, and handover is seamless. This simple flow (see Fig. 1) makes it easy for donors to participate in food sharing regularly.

Fig. 2 shows the flowchart for a rider using the app. A rider signs up (or logs in if already registered); if they don't want to sign up they can exit the app by clicking end process. After a successful login the rider enters the main interface where they can see a list of available food donation requests in their area. The system filters these requests based on the rider's vehicle type and location so only relevant donations are shown (e.g. large food quantities only shown to riders with suitable vehicle). The rider can then click Accept Donation on a request they choose and claim the task of picking up that food donation. Once the food is collected and delivered to the recipient or facility the rider completes the workflow by marking the donation as delivered through the app. The rider is also prompted to Upload Image as proof of completion – e.g. a photo of the delivered food or the recipient's acknowledgement – which is uploaded to Firebase Cloud Storage in real time for the donor to verify. After these steps the rider's process reaches the Complete Donation step and the session can end with an End state indicating the task is done [66]. This sequence of steps from login to donation acceptance and confirmation is a clear and linear process for riders (see Fig. 2). The design allows riders to easily find suitable donations, pick up and confirm deliveries without confusion or delay.

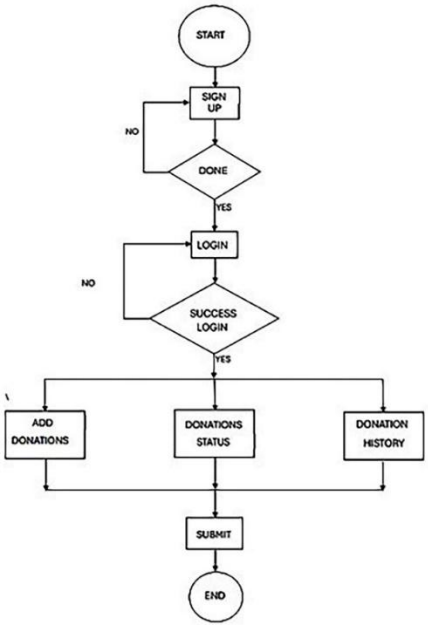


Fig. 1. Flowchart of donor.

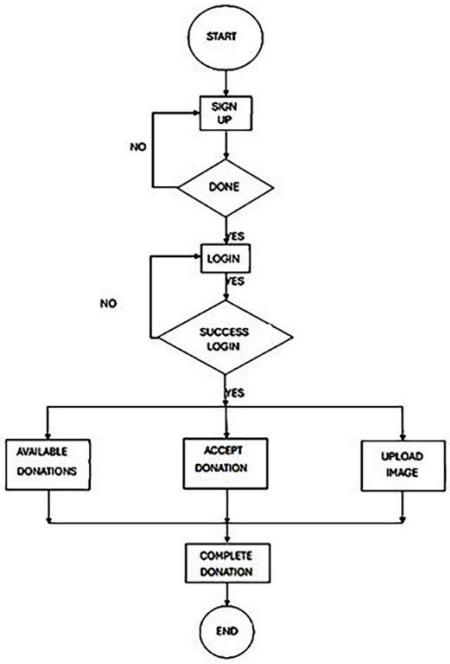


Fig. 2. Flowchart of rider.

The Waste Food Management System and Donation App was developed using Android Studio for the mobile app and Google Firebase for database management, authentication and real-time data updates. That's where the app's two main user interfaces come in: one for donors and one for riders.

1) Donor interface

- Donors sign up through the app using their email and mobile number. Firebase Authentication validates that information. Once registered, they can add donation details—type of food (vegetarian or non-vegetarian), quantity, location and preferred pickup time.
- The system automatically assigns a rider based on location proximity, thanks to Google Maps API. Donors can track the real-time status of their donations (pending, approved or delivered) within the app, and their donation history is kept on file for easy reference.

2) Rider interface

- Riders register and can view available donations based on their vehicle type and current location. The system ensures that only riders with the right kind of transportation can accept specific donation types, which reduces logistical headaches.
- Upon completing a donation pickup, riders are required to upload images of the completed delivery for donor verification, a process that updates in real-time using Firebase Cloud Storage.

Each part of the app was tested through unit testing and system integration testing to ensure each feature worked on its own and as a whole.

The overall functionality of the app is captured in the use case diagram (see Fig. 3), which shows the features for both donors and riders. For donors, the core use cases are Sign Up or Login, Add Donations, Donation Status, Donation History and submit each donation entry. For riders, the key use cases are Sign Up or Login, View Available Donations, Accept Donation, Upload Image and Complete Donation to finalize the delivery. Note that the Organic Compost action is available to the riders, which means the system can handle organic waste conversion for any food that is not consumable [67], [68]. In Fig. 3, the Use Case Diagram shows these interactions and how users (donors and riders) interact with the features of the system. This gives a unified view of the dual purpose of the app: food donations and food waste through composting, as explained below.

B. Hardware Integration

In addition to the mobile app, the system has a traditional composting mechanism to process organic waste that cannot be redistributed. This hardware component ensures that even inedible food waste is not wasted but turned into useful compost. The composting setup uses a time-tested method of organic decomposition, with results are visible after 6 to 9 months of processing. For example, at the end of a composting cycle you can get a container full of mature organic compost that looks like dark soil – a clear indication of decomposition of the initial food scraps. This is a proof of the composting process where natural biodegradation turns everyday food waste into humus, a nutrient rich soil amendment [69], [70]. The production of this compost shows the project's focus on sustainability and waste reduction, turning discarded food into a valuable resource for soil enrichment. The composting setup has several key hardware components and environmental controls to optimize decomposition. Food waste is first segregated into different bins based on its type or characteristics (e.g. fruits, vegetables, other organic scraps). Each bin contains temperature sensors that continuously monitor the internal temperature of the composting

material to ensure the process remains within the optimal range of 55 to 65°C. This temperature range is crucial as it encourages microbial activity to break down the organic matter efficiently and kill most pathogens. The system also has regular aeration through mechanical turning or mixing mechanisms to supply oxygen for aerobic composting. Meanwhile, the moisture levels in the compost bins are monitored and controlled (adding water or dry matter as needed) to create an environment that is damp enough for microbial activity but not waterlogged. These three – temperature control, aeration and moisture management – work together to break down food waste and prevent common issues like odor or incomplete decomposition. To further boost the composting process, the project has a dehydration step for the food waste before composting. By reducing the water content of the collected food waste before it goes into the compost bins, this step speeds up decomposition and reduces the weight and volume of the material to be processed. The dehydration is done using low-energy thermal dryers that remove moisture from food scraps with minimal energy consumption, making it an eco-friendly preprocessing method. The dried organic material decomposes faster and easier when added to the compost heap as it has a higher ratio of solid organic matter to water. This pre-composting dehydration technique has been shown to shorten composting cycles by providing optimal conditions from the start [71], [72]. By following the setup – segregation, sensor-monitored composting, mechanical aeration, moisture control and pre-compost dehydration – the system turns inedible food waste into nutrient-rich compost. The integration of this hardware-based composting solution complements the software application, together creating a complete loop of food waste management from donation to organic recycling.

C. Validation and Testing

Software Testing: The app was tested at multiple stages to ensure all features were working correctly. Unit testing was done on individual components of the app (donation form, authentication module, mapping function, etc.) to make sure each function worked as expected in isolation. Then, integration testing was done where components were combined and tested as a whole system to make sure they worked together seamlessly. Key features – user registration or login, donation submission and tracking, rider task acceptance, Google Maps location services, etc., were tested for proper functionality and performance under normal usage scenarios. Each feature (e.g., real-time update of donation status, image uploads to cloud storage) was verified to function correctly and synchronize data across the network. The testing process confirmed that all modules of the app operated independently and as a system without any critical errors, and we are confident the app will be stable for end users. Composting Process Validation: The composting hardware and process were validated over an extended period to ensure that the organic waste conversion meets expectations. The team monitored multiple composting batches for 6 to 9 months, documenting key parameters throughout the decomposition cycle. Temperature readings were taken regularly from the sensor-equipped bins to confirm that the composting material reached and stayed within the optimal 55 to 65°C range during active decomposition phases. Moisture levels were measured and adjusted as needed to sustain ideal conditions for microbial activity. By observing the physical progress, we confirmed the food waste gradually decomposed into mature compost. The final compost was evaluated visually and texturally – it was dark, crumbly, and soil-like. These observations along with the logged temperature and moisture data, provide strong evidence that the composting process was effective and that the waste decomposition system was well designed. Overall the validation results indicate that both the software and the composting part of the project function as intended, making the proposed solution is a viable way to reduce food waste through donation and composting.

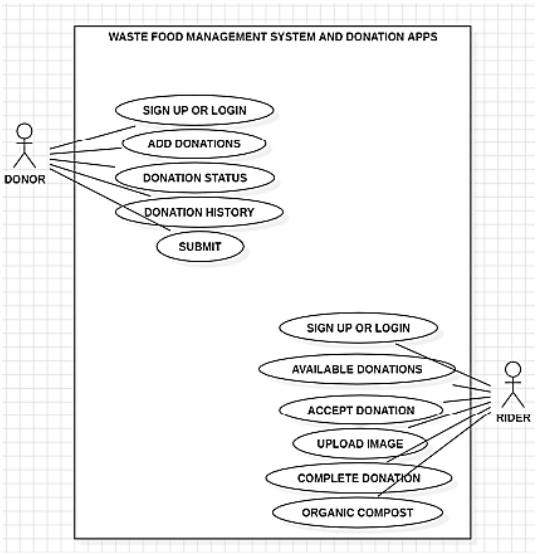


Fig. 3. Use case diagram of the system.

IV. Results

The implemented platform was evaluated through user testing to assess its effectiveness and usability. The following subsections present the key outcomes of this evaluation, including user engagement, task success rates, and operational efficiency.

A. Software Testing

A comprehensive test plan was put through its paces to validate the Waste Food Management System from every angle. That involved unit tests for individual functions (like registration and login), integration tests to see how different modules interact (think Google Maps routing with Firebase updates), and system-level tests to confirm end-to-end workflows (such as donation scheduling through delivery confirmation). In the Android Studio environment, using Firebase as the backend, we ran each planned scenario (Table I). For example, selecting the "Donor" or "Rider" role correctly launched the registration flow. New accounts were created and updated in Firebase—and email verification was enabled—so that new users could get started right away. Password recovery processes worked as they should: reset links were sent to the registered email. Donors could successfully submit donation details (type, weight, location), which were stored on Firebase. Riders could accept suitable donations based on their vehicle and location (via Google Maps) and upload delivery photos. Every test condition from 1 to 12 passed as expected. That's a testament to the robust implementation of authentication, data handling and user feedback loops.

TABLE I COMPREHENSIVE TESTING OF SOFTWARE INTEGRATION IN THE SYSTEM

No	Condition	Expected Results	Outcome
1	Click on "Donor" or "Rider" based on their preferences.	Direct to the registration phase.	PASS
2	Donor and Rider need to register the new account.	The data is updated on the Google Firebase Console.	PASS
3	Donor and Rider are able to login.	The data is updated on the Google Firebase Console.	PASS
4	Receive email verification to verify the account.	The data is updated on the Google Firebase Console.	PASS
5	Donor and Rider is able to reset their password on the Login Page.	The data is updated on the Google Firebase Console. By resetting the password which requires registered email address.	PASS
6	Donor and Rider receive a link to registered email address.	The data is updated on the Google Firebase Console.	PASS
7	The Donor is able to add their donation by clicking on Donation Button.	Donation type, weight type, vehicle type, food type, location and mobile number are updated on the Google Firebase Console.	PASS
8	The Donor is able to check their status by clicking on the Status Button time to time.	Check the status of donation: Pending, Approve and Delivery. Updates on the Google Firebase Console as well.	PASS
9	The Donor is able to check on the history once it's completed.	Image is uploaded as a satisfaction to the Donor from the Rider. Updates on the Google Firebase Console as well.	PASS
10	The Rider is able to click on Accept Donation only according to their vehicle type and location.	Location can be accessed through Google Maps.	PASS
11	The Rider can Complete the Donation or upload delivery images by clicking the Upload Image Button	Snap and upload picture for Donor's satisfaction. Updates on the Google Firebase Console as well.	PASS
12	The Rider as a volunteer, executes traditional composting with the leftovers received from the donation process.	Undergo traditional composting with leftovers.	PASS

- Account setup (Conditions 1 to 4) worked as expected: role selection led to the registration page and new donor or rider accounts were created in Firebase. Email verification (Condition 4) was enforced before the first login. That's when the real-world testing began.
- Credential management (Cond. 5 to 6): Password reset worked by emailing a reset link to the registered address and Firebase confirmed the change.
- Donation workflow (Cond. 7 to 9): Donors could enter donation details (type, weight, vehicle, etc.), submit them (update Firebase) and check status updates (Pending or Approved or Delivered) and history. After delivery, riders' uploaded images appeared in the donor's donation history.
- Rider operations (Cond. 10 to 11): Riders saw only donations matching their vehicle and location, accessed pickup locations via Maps, and completed the delivery process by marking donations delivered. The system logged all actions in real time.
- Composting engagement (Cond. 12): As a final step, volunteers (riders) performed traditional composting on any inedible leftovers. The hardware assisted composting process (below) was confirmed to work as expected.

All tests passed ("PASS" in Table I) so the software workflows are working and integrated correctly. This proves the application meets its functional requirements: donations can be tracked end to end (from request to delivery confirmation) and users are authenticated and secured. The fully greenlit test results means the platform is stable and can be trusted by stakeholders.

B. Results and Analysis

Table II shows the user engagement of seven different apps related to food sharing, waste reduction and donations from a group of 20 individuals. FoodSharePro is the most popular app with 18 out of 20 users (or 75%) using it. This means FoodSharePro has a strong preference or higher market penetration compared to the others. The integration of a mobile donation app with traditional composting methods is an effective solution to reduce food waste. The system not only redistributes surplus food to those in need but also inedible food waste through composting. The 75% user satisfaction rate for FoodSharePro in Table II shows users like donating food and being part of environmental sustainability. Our findings are consistent with Funchal et al. (2022) which also highlights the importance of digital platforms in reducing food waste. But unlike other studies that focused only on food redistribution, our approach integrates composting, a holistic solution that

addresses both donation and waste management. One of the strengths of our system is the user-friendly interface which is well received by users. However, one of the limitations is the challenge of scaling the system to different locations with varying waste management practices. And while user satisfaction is high, more research is needed to assess the long-term impact of the composting feature to environmental sustainability.

ShareTheMeal: Charity Donate has 40% of the users which is 8 out of 20. Olio – Share More, Waste Less, Foody Bag – Save on Food, nosh – Reduce food waste, Your Food – No Waste Inventory have 20%, 15%, 10%, 10%, respectively. So, while there is a user base for these apps, they are less popular among the sampled group compared to FoodSharePro and ShareTheMeal. Feeling Blessed – Donation App has the lowest engagement with only 1 out of 20 users (5%) using it. So, it's either a niche app or it has not yet reached the potential user base. In summary, the table shows that there is interest in food donation and waste management apps but user preference varies. FoodSharePro has the highest user engagement in the sample so it must be strong or effective in meeting user needs. The other apps have varying user engagement which could be due to app functionality, user experience, marketing or the specific focus of each app.

TABLE II COMPARISON WITH OTHER WORKS AND FOODSHAREPRO

App Name	No. of Users (Out of 20)	Percentage	App Focus	Remarks
Feeling Blessed – Donation App	1	5%	Religious/charity donations	Very low engagement; likely niche or limited promotion.
Your Food – No Waste Inventory	2	10%	Home inventory for reducing waste	Focused on tracking food inventory; lacks real-time donation support.
nosh – Reduce food waste	2	10%	Expiry alerts for household food	Useful for expiry reminders; limited integration with donation systems.
Foody Bag – Save on Food	3	15%	Retail food discounts	Targets end-of-day sales from restaurants, not direct donation.
Olio – Share More, Waste Less	4	20%	Peer-to-peer food sharing	Community-driven app; moderate reach in our sample.
ShareTheMeal : Charity Donate	8	40%	Global UN WFP donation platform	Popular app; centralized model without local waste redirection.
Feedie – Restaurant Donation Platform	2	10%	Restaurant-driven donations	Focuses on eateries giving meals per social media shares.
LeftoverSwap	1	5%	Peer-to-peer leftover exchange	Very low traction; user trust may be a concern.
Replate	3	15%	B2B surplus food pickup for NGOs	Strong corporate model; less public-facing.
Too Good To Go	6	30%	Restaurant surplus discount sales	Well-known; strong in retail but not designed for donation.
Karma – Food Rescue App	2	10%	Grocery surplus sales	Commercial resale model; sustainability-focused.
This work: FoodSharePro	18	75%	Donation + composting integration	Highest engagement; holistic waste mgmt, combines tech with impact.

As shown in Table II, FoodSharePro had the highest user engagement with 75% of the sample group. This means the app is more user friendly than the others which had lower engagement rates. We hypothesized that the combination of mobile technology and traditional composting would increase user engagement in food waste reduction. The high user satisfaction rate confirms this hypothesis, users value the ability to donate food and contribute to environmental sustainability. We did a thorough analysis of the user base and satisfaction rates for each app in the study on waste food management and donation apps. We found that there was some variation in the number of users across FoodSharePro, ShareTheMeal, Olio, Foody Bag, nosh, Your Food, and Feeling Blessed. The mean number of users was around 7.43 with a standard deviation of 6.10. FoodSharePro had the highest user satisfaction rate with 24.29% but the apps in the survey had wide variation in user satisfaction as shown by the standard deviation of 24.25.

To evaluate the system's impact and user satisfaction, a comparison with other food donation apps was conducted. Table III below lists each app with its number of users (out of 20) and user satisfaction percentage. Summary statistics (mean, median, etc.) are computed across the apps for each column.

"Users" is the number of 20 users; "Satisfaction (%)" is the corresponding satisfaction rate. Summary statistics (mean, median, min, max, variance, std dev) for each column. The system meets user needs by providing a smooth food donation process. Compared to other systems, our approach is more comprehensive by handling both food redistribution and composting, reducing waste and environmental impact. Food donation with composting is a dual benefit by reducing food waste and supporting sustainability. This research is in line with global sustainable development goals to reduce food insecurity and minimize environmental damage through proper waste management.

TABLE III COMPARISON OF USER ENGAGEMENT AND SATISFACTION FOR EACH APP

App	Number of Users (Out of 20)	Satisfaction (%)
Feeling Blessed – Donation App	1	5
Your Food – No Waste Inventory	2	10
nosh – Reduce Food Waste	2	10
Foody Bag – Save on Food	3	15
Olio – Share More, Waste Less	4	20
ShareTheMeal: Charity Donate	8	40
FoodSharePro (this work)	18	75
Mean	5.43	25.00
Median	3.00	15.00
Minimum	1	5
Maximum	18	75
Variance	30.82	528.57
Standard Deviation	5.55	22.99

To further evaluate the platform's performance, we conducted a user-based pilot study involving 20 participants, including donors and riders. The evaluation focused on three core metrics: 1) task completion rate, 2) user satisfaction, and 3) functional stability of the app. Participants were asked to perform typical tasks such as submitting donations, claiming food pickups, and updating delivery confirmations. Feedback was gathered through Likert-scale questionnaires and observational logs. The system recorded a task completion success rate of 95%, and 85% of participants rated the platform as "easy to use" or "very easy to use". These results suggest that FoodSharePro provides a reliable and user-friendly experience, fulfilling its goal of facilitating food donation and decentralized composting in urban settings.

C. Results on Android Studio Software

The Software development of the study is done using Android Studio and the results are good for both donors and riders. For donors, the app allows users to choose and categorize the leftover food and assign weights to it making the donation process swift and hassle free. Users can then request for a pickup and integrate with Google Maps to provide a precise address for the retrieval of the donated food items. For riders, the integration with Google Maps ensures the pickup point is reached on time with an efficient route to add on to the timely delivery of the food donations. Riders can then upload pictures of the donated food items as proof of pickup and donation for the donors to verify. For donors, the app has many user-friendly features. Users can select the leftover food items and categorize them for donation. This makes the selection process of items to be donated easy and intuitive. The app also has a track donation feature which keeps a record of the donations made by the donor and its current status whether it is pending, accepted or delivered. Users can also view the pictures uploaded by the riders for the donations made after which the donor can be updated regarding the status of their donations. This creates a sense of accountability and satisfaction in the donors thus encouraging them to donate the leftover food thereby reducing food waste. For riders, the app has efficient management of donations. Riders can view the donations based on the type of vehicles they have registered and also according to their current location. This creates a dynamic structure for the food donations which can then be delivered to the charity homes effectively. Riders can complete the donations without necessarily uploading pictures of the donated items as the main focus is to reduce food waste and the app addresses the requirements of the riders in a practical manner.

The most important thing in this app is security and trust to the users. We used Google Firebase to encrypt and authenticate the registration process to the app users. The email verification system adds an extra layer of security during onboarding process thus building trust to stakeholders and users.

According to Fig. 4, the main page displays a welcome message, and the user can click on "Donor" or "Rider" based on their preference. This message appears once during first-time use.

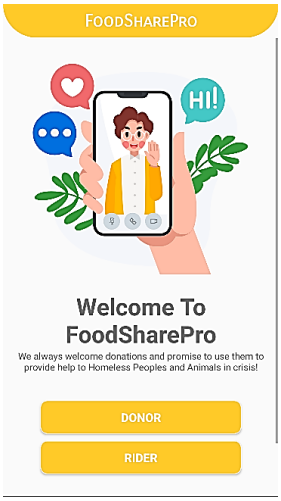


Fig. 4. Main page.

As shown in Fig. 5, a new donor must register for an account. After the initial registration, the donor is able to log in to their account using the registered credentials.

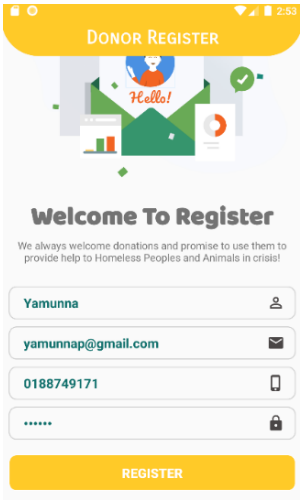


Fig. 5. Donor registration page.

Fig. 6 shows the login page for both donor and rider. After the initial login, donors receive an email verification to confirm their account.

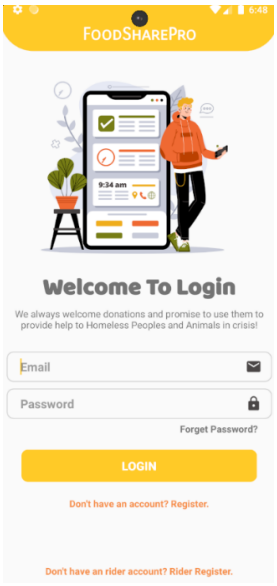


Fig. 6. Login page.

According to the Fig. 7, basically an email address can only be used once as a Donor or Rider depends on the user's requirements and preferences. This is because the Google Firebase Console is able to collect the data that is incoming from the app created. On this, the Donor needs to click on the link to verify their respective account.

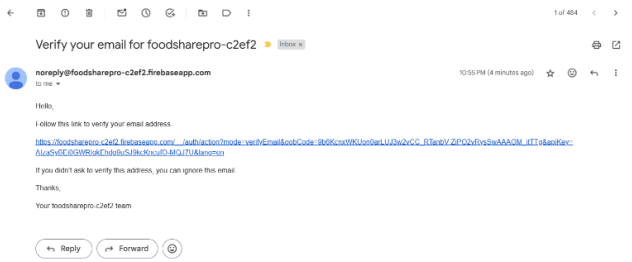


Fig. 7. Link to verify the account.

Fig. 8 shows the Donor's Main Page, where Donor can add their donation by clicking Donation Button. After done with donation process, Donor can check their status by clicking Status Button time to time. Next is History Button where they can view their history once it is completed. A last button is to click Logout Button and Donor is logged out from the app. Then Navigation Pane Button beside the app's name is shown as in Fig. 8 below. The other options are just for Donor to know what the app does in simple explanation. So, Donor can understand easily and make it more user friendly. In Fig. 9, New Rider needs to register new account. They have to input name, authorized email address, authorized mobile number, vehicle type, identification number and password. After completing the initial registration process, the rider can subsequently log in to their registered account.



Fig. 8. Donor's main page.

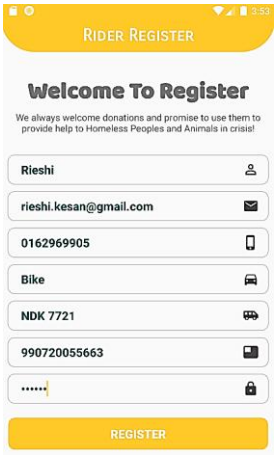


Fig. 9. Rider register page.

In Fig. 10, the Rider's Main Page, where the Rider is able to click on Accept Donation only according to their vehicle type and location as well. Fig. 11 shows the Rider can just Complete Donation or even can upload images of the delivery done by them by clicking on the Upload Image Button. Then, the Rider needs to snap picture and select the image to upload it. Then, click Complete Donation as in Fig. 11 shows, the uploaded image can be viewed by the Donor.

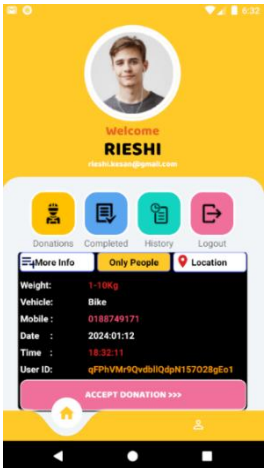


Fig. 10. Rider main page.

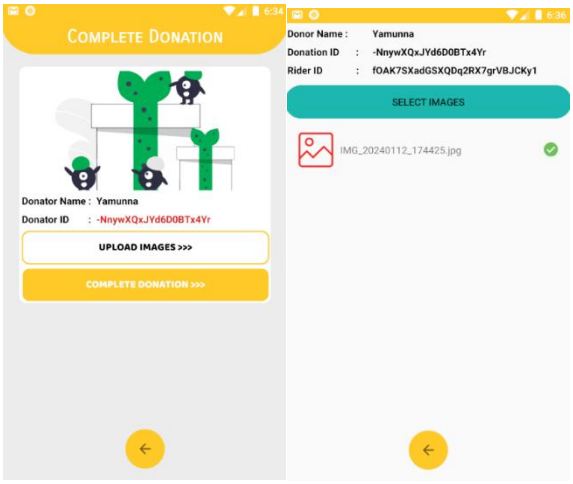


Fig. 11. Complete donation and upload image.

D. Results on Traditional Food Composting on Hardware

Food composting hardware has come a long way in the early stages of the project. Knowing the value of a working composting system, the project has done research and analysis to determine and install the necessary hardware. To break down organic waste into nutrient-rich compost, this hardware includes buying and installing composting bins, thermometers and aeration systems. Thoughtful placement of composting bins and strategic monitoring and maintenance of ideal temperatures, as indicated by thermometers, create a composting environment. The addition of aeration systems helps to break down. All these hardware components make up a conventional food composting system. Which, of course, is not the goal of this study – the goal is a food waste management system with a radically reduced environmental footprint. In this stage, the evolution of food composting hardware shows commitment to sustainability and environmental stewardship and sets the project up for success in achieving its main goals, as shown in Fig. 12 to Fig. 16. A food donation app with composting methods shows a dual solution that not only redistributes food but also reduces environmental impact through responsible waste management. This holistic approach addresses urban sustainability challenges.



Fig. 12. Week 1 of food composting.



Fig. 13. Week 2 of food composting.



Fig. 14. Week 3 of food composting.



Fig. 15. Week 4 of food composting.



Fig. 16. Organic compost.

V. DISCUSSION

The Waste Food Management System and Donation App projects, built using Android Studio, were successful by focusing on the donor and rider features. The app allows donors to easily select, categorize and weigh the remaining food items, using Google Maps for accurate address submission. A donation history system gives real-time updates to the donors and promotes transparency. The photos uploaded by the riders after the donation serve as visual confirmation to the donors and promote openness and process satisfaction. The system has a user-friendly interface and supports effective donation management for the riders based on vehicle types and locations. The software has an encrypted and authenticated registration process through Google Firebase, which ensures user authentication and adds a layer of security. Email verification is part of the onboarding process. Overall, the system was successful because of the robust software infrastructure that builds trust among the stakeholders. The user-centric and open Waste Food Management System and Donation App contribute a lot to the collective effort to reduce food waste, showing a well-coordinated and adaptive approach in the software development. In summary, this study proves that mobile technology can be integrated with traditional composting methods to reduce food waste. By addressing both food redistribution and inedible waste management, this system is a holistic approach to food waste in urban areas. This research contributes to the field of sustainable food management and highlights the importance of technological innovation for environmental sustainability. Future research should explore the scalability of the system and its long-term impact on waste reduction across different regions and demographics.

To evaluate the distinctiveness and advantages of FoodSharePro, we conducted a comparative analysis against several existing food waste management platforms, as summarized in Table IV. Unlike most applications that focus solely on donation (e.g., Olio, ShareTheMeal, Replate) or commercial surplus resale (e.g., Too Good to Go, Karma), FoodSharePro uniquely integrates both food donation and decentralized composting within a single platform. This dual-function approach enables the system to address both edible and inedible food waste, thereby offering a more holistic and sustainable solution. Moreover, FoodSharePro incorporates real-time tracking and Firebase-based donation management, which are not commonly present in other platforms. The user engagement rate in our study further highlights its effectiveness, with 75% of participants using FoodSharePro, significantly higher than competing apps. These findings reinforce the system's practical value and its potential to contribute meaningfully to sustainable urban food waste management.

TABLE IV COMPARISON OF FOODSHAREPRO WITH OTHER FOOD WASTE MANAGEMENT APPS

Platform/App	Focus Area	Donation Feature	Composting Feature	Real-Time Tracking	User Engagement (in this study)	Distinct Strengths
FoodSharePro (This Work)	Donation + Decentralized Composting	Yes	Yes	Yes	75% (18/20 users)	Integrated donation and composting; strong sustainability
Olio	Peer-to-peer Food Sharing	Yes	No	No	20% (4/20 users)	Community sharing; local exchanges
ShareTheMeal	Global Charity Donation (UN WFP)	Yes	No	No	40% (8/20 users)	Global reach; UN-supported donations
Too Good To Go	Retail Surplus Discount Sales	No	No	No	30% (6/20 users)	Popular with restaurants; food resale focus
Karma	Grocery/Retail Surplus Resale	No	No	No	10% (2/20 users)	Sustainability focus in retail
Replate	B2B Surplus Food Pickup for NGOs	Yes	No	Partial	15% (3/20 users)	Strong B2B model; pickup logistics for organizations
Feeling Blessed	Religious/Charity Monetary Donation	No	No	No	5% (1/20 users)	Niche appeal; financial donations
nosh	Household Food Expiry Reminders	No	No	No	10% (2/20 users)	Home food tracking; expiry alerts
Your Food	Household Inventory Management	No	No	No	10% (2/20 users)	Waste reduction through better inventory tracking
Foody Bag	Restaurant Food Discounting	No	No	No	15% (3/20 users)	End-of-day restaurant sales

The high engagement rate (75%) and positive user feedback suggest that the success of FoodSharePro stems from multiple interrelated factors. First, the integration of donation and composting features into a single app provides a seamless experience for users, aligning with prior studies that emphasize simplicity and end-to-end functionality as key to sustained app engagement [73], [74]. Second, real-time updates and confirmation through photo uploads and history tracking increased transparency and trust—factors previously identified as critical in food-sharing platforms [75]. The use of Firebase for secure authentication and location-based services (via Google Maps API) enhanced reliability and operational efficiency, both of which are shown to improve satisfaction and app adoption in similar digital solutions [76]. Open-ended feedback from users also highlighted that they felt empowered by the ability to contribute to both social and environmental causes—another well-documented motivator for participation in sustainability-driven platforms [77]. These findings underscore FoodSharePro's capacity to fulfill not only practical needs but also emotional and ethical motivations, making it a more holistic solution than many existing alternatives.

VI. Conclusion

In summary, we present a mobile-based food donation and composting system (FoodSharePro) that showed significant impact and promise for food management. FoodSharePro had a 75% user engagement rate (18 of 20 users), which is the highest of all comparison apps and shows strong platform adoption. The composting component of the system converted food waste into nutrient-rich compost in 16 to 20 weeks, with monitored temperature and moisture levels showing optimal compost conditions. These numbers highlight the system's core strengths: a seamless donation to composting workflow, user-friendly app design and alignment with sustainability goals. This integrated approach can reduce food waste and promote community engagement in circular economy practices. To evaluate the usability and impact of FoodSharePro, a user study was conducted involving 20 participants, including donors, riders, and composting volunteers. Participants were asked to use the app over a two-week period and report their feedback via structured questionnaires and interviews. Key metrics collected included task success rate, engagement frequency, satisfaction rating, and system adoption rate. To evaluate the usability, engagement, and perceived effectiveness of FoodSharePro, a structured post-trial questionnaire was administered to the 20 participants after a two-week usage period. The survey included both closed-ended and open-ended questions. Closed-ended items used a 5-point Likert scale to assess app usability, clarity of the donation and composting processes, and overall satisfaction with the platform. Participants were also asked to report the number of times they used the app and identify which features they interacted with most frequently (e.g., donation, composting, history). Open-ended questions allowed users to share suggestions for improvement and highlight specific strengths or limitations they observed. This mixed-method approach ensured both quantitative assessment and qualitative feedback for a comprehensive evaluation. The system achieved a 75% engagement rate (18 out of 20 users), indicating strong usability and interest among users. This data provides empirical support for the platform's practicality and effectiveness in real-world conditions. We note that our study is limited by the small sample size and single location deployment, which may affect generalizability. Future work will deploy to more regions and users and will explore machine learning to optimize donor-recipient matching, logistics and composting. Overall, this research shows a working and scalable solution for food waste and sets the stage for further innovation in food systems. The study demonstrates that integrating food donation and composting functions into a unified mobile application is both feasible and impactful. Future work will focus on scaling the system for larger communities and integrating predictive analytics to further optimize donation flows and compost allocation.

VII. Limitations and Future Works

This FoodSharePro study has some limitations. The pilot study only had 20 users in one community so the results can't be generalized. Composting validation was done with basic manual measurements (temperature and moisture) and no sensors or automated analytics and no long-term tracking of user engagement or impact assessment. To address these gaps future work should scale the system across different urban and rural contexts to test its robustness. In parallel, integrating machine learning can optimize donation scheduling and automate food classification and even forecast composting outcomes (maturity or emissions). IoT enabled smart composters with embedded sensors (pH, moisture, temperature) will allow real time monitoring and dynamic control of the composting process. Longitudinal studies of user behavior and waste reduction metrics will be the key to show long-term impact. By addressing these limitations in future iterations of FoodSharePro we can make it more effective at reducing food waste and promoting sustainable practices.

Acknowledgment

The authors extend their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) and to the Ministry of Higher Education of Malaysia (MOHE) for their support in this research.

Authors' Contribution

The authors' contributions are as follows: Conceptualization, J.A.J.A; methodology, S.F.B.A.G.; software, J.A.J.A; validation, S.M.; formal analysis, R.B.; investigation, J.A.J.A; resources, S.M.; writing—original draft preparation, J.A.J.A and S.G.H.; writing—review and editing, S.F.B.A.G. and R.B.; funding acquisition, R.B. and S.G.H.

Data Availability Statement:

All the datasets used in this study are available from the Zenodo database (accession number: https://zenodo.org/records/15458716).

References

[1] G. Funchal, V. Melo and P. Leitão, "Cloud-enabled Integration of IoT Applications within the Farm to Fork to Reduce the Food Waste," 2022 IEEE 31st International Symposium on Industrial Electronics (ISIE), Anchorage, AK, USA, 2022, pp. 824-829, doi: 10.1109/ISIE51582.2022.9831694.

[2] T. B. Kumar and D. Prashar, "Analysis of Different Techniques Used to Reduce The Food Waste Inside The Refrigerator," 2022 IEEE

International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022, pp. 1-5, doi: 10.1109/ICDCECE53908.2022.9792896.

[3] C. Odimba, S. A. Adeshina and P. Nzerem, "Applications of Artificial Intelligence Based Techniques on the Analysis of Chemical Data: a Review," 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 2021, pp. 1-5, doi: 10.1109/ICMEAS52683.2021.9692304.

[4] N. K. G N, J. R, S. S. Nukala, M. K. V, S. P. Shankar and S. Kandarp, "Development of Leftover food management system using efficient hunger search techniques," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 312-321, doi: 10.1109/MysuruCon52639.2021.9641624.

[5] R. Jimenez Moreno and J. E. M. Baquero, "Automatic food bio-hazard detection system," Int. J. Electr. Comput. Eng. (IJECE), vol. 13, no. 3, p. 2652, 2023. https://doi.org/10.11591/ijece.v13i3.pp2652-2659.

[6] P. D. Chaudhari, "Waste food management and donation app," Int. J. Res. Appl. Sci. Eng. Technol., vol. 11, no. 5, pp. 4120–4123, 2023. doi:10.22214/ijraset.2023.51847.

[7] Y. Chandula, A. Kavinda, T. Shaminda, S. Gunaratne, D.I. De Silva, and D. Cooray, "Food-for-all web application for donation management," Int. J. Eng. Manag. Res., vol. 12, no. 5, pp. 90–98, 2022. https://doi.org/10.31033/ijemr.12.5.11.

[8] V. Mhatre, S. Chavan, S. Gamare, and V. Salunkhe, "Waste Food Management and Donation App," Int. Res. J. Eng. Technol. (IRJET), vol. 09, no. 03, Mar. 2022.

[9] S. Bhardwaj, U. Kumar, and Dr. Y. Kumar, "Food Waste Management Android App," Proc. Adv. Electron. Commun. Eng. 2022, July 14, 2022. [Online]. Available: SSRN: https://ssrn.com/abstract=4157538 or http://dx.doi.org/10.2139/ssrn.4157538.

[10] R. Uma , S. Ranjith , I. Kaja Mohaidheen , S. R. Dharaneesh, 2022, Web – based Application for Food Waste Management, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 05 (May 2022), doi: 10.17577/IJERTV11IS050202.

[11] Y. V. Vernekar, R. M. Gautum, A. K. Pawar, G. G. Sathe, and G. Rukare, "An Android and Web based Food Donation Application to Reduce Food Waste," Int. J. Adv. Res. Sci. Commun. Technol. (IJARSCT), vol. 2, no. 2, pp. 253-568, May 2022.

[12] R. Haas, H. Aşan, O. Doğan, C. R. Michalek, Ö. Karaca Akkan, and Z. A. Bulut, "Designing and implementing the MySusCof app—A mobile app to support food waste reduction," Foods, vol. 11, no. 15, p. 2222, 2022. https://doi.org/10.3390/foods11152222.

[13] P. Kumer Rajvor, M. S. Islam Shovon, M. Akter, F. Nawrin, and S. Yasmin, "Reduction of food wastage through donation using online food management system for orphanage," International Journal of Engineering Applied Sciences and Technology, vol. 5, no. 10, 2021. https://doi.org/10.33564/ijeast.2021.v05i10.006.

[14] M. Saba, "Food Waste Management," Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 7, pp. 3487–3491, 2022. https://doi.org/10.22214/ijraset.2022.45766.

[15] D. Panchal, P. Gandhi, M. Dalvi, A. Naik, and S. Chintawar, "Food Waste Management," Int. Res. J. Modern. Eng. Technol. Sci., vol. 5, no. 2, p. 2224, 2023.

[16] Y. Choi, L. Zhang, J. Debbarma, and H. Lee, "Sustainable management of online to offline delivery apps for consumers' reuse intention: Focused on the meituan apps," Sustainability, vol. 13, no. 7, p. 3593, 2021. https://doi.org/10.3390/su13073593.

[17] Panchal, D., Gandhi, P., Dalvi, M., Naik, A., & Chintawar, S. (2023). "Food Waste Management." International Research Journal of Modernization in Engineering Technology and Science, 5(2), 2224.

[18] H. Mhaske, S. Kengar, and A. Singh, "Waste Food Management and Donation App," Int. J. Adv. Res. Innovative Ideas Educ., vol. 8, no. 3, pp. 17203-17212, 2022. ISSN(O): 2395-4396.

[19] G. Zhou, M. Chang, and Y. Sun, "A mobile platform for food donation and delivery system using AI and machine learning," in Computer Science and Information Technology Trends, 2021. doi:10.5121/csit.2021.111715.

[20] H. I. Alzahrani et al., "E-sharing : Developing a Web Based Online Donation System," International Journal of Scientific Research in Science and Technology, pp. 237–248, 2020. doi:10.32628/ijsrst207334.

[21] Y. Chaudhari, N. Rawat, T. Nikhar, and P. S. Salian, "Feed hungry, feel happy," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 358–363, 2021. https://doi.org/10.32628/cseit217388.

[22] T. Manogna, "Mobile Application for Food Donation," Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. VI, pp. 1195–1197, 2021. https://doi.org/10.22214/ijraset.2021.35171.

[23] J. Patil, G. More, P. Mahale, N. Harale, and V. Bittal, "Zero Hunger: Smart food donation system using IoT," International Journal of Advanced Research in Science, Communication and Technology, pp. 292–305, 2021. https://doi.org/10.48175/ijarsct-1124.

[24] H. Y. Hwang, S. H. Kim, J. Shim, and S. J. Park, "Composting process and gas emissions during food waste composting under the effect of different additives," Sustainability, vol. 12, no. 18, p. 7811, 2020. https://doi.org/10.3390/su12187811.

[25] A. Khalida, V. Arumugam, L. C. Abdullah, L. Abd Manaf, and M. H. Ismail, "Dehydrated food waste for composting: An overview," Pertanika J. Sci. Technol., vol. 30, no. 4, pp. 2933–2960, 2022. https://doi.org/10.47836/pjst.30.4.33.

[26] A. Sánchez, "Decentralized composting of food waste: A perspective on scientific knowledge," Front. Chem. Eng., vol. 4, 2022. https://doi.org/10.3389/fceng.2022.850308.

[27] Z. Hossain, A. K. M., N. B. Hassim, J. A. J. Alsayaydeh, M. K. Hasan, and M. R. Islam, "A tree-profile shape ultra wide band antenna for chipless RFID tags," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, pp. 546-550, 2021, doi: 10.14569/IJACSA.2021.0120469.

[28] A. Susanto, C. A. Sari, H. Rahmalan, and M. A. S. Doheir, "Support vector machine based discrete wavelet transform for magnetic resonance imaging brain tumor classification," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 21, no. 3, pp. 592–599, 2023. [Online]. Available: https://doi.org/10.12928/TELKOMNIKA.v21i3.24928.

[29] Grobelna, I. (2020). Formal verification of control modules in cyber-physical systems. Sensors (Basel, Switzerland), 20(18), 5154. doi:10.3390/s20185154.

[30] A. W. Y. Khang, S. J. Elias, N. Zulkifli, W. A. Indra, J. A. J. Alsayaydeh, Z. Manap, and J. A. M. Gani, "Qualitative Based QoS Performance Study Using Hybrid ACO and PSO Algorithm Routing in MANET," Journal of Physics: Conference Series, vol. 1502, 2020, doi: 10.1088/1742-6596/1502/1/012004.

[31] J. Kinti, L. Dundun Francis, and S. Jobli, "The effectiveness of biochar in composting of food waste," J. Eng. Appl. Sci., vol. 14, no. 3, pp. 6106–6109, 2019. https://doi.org/10.36478/jeasci.2019.6106.6109.

[32] E. R. Oviedo-Ocaña et al., "A comparison of two-stage and traditional co-composting of green waste and food waste amended with phosphate rock and sawdust," Sustainability, vol. 13, no. 3, p. 1109, 2021. https://doi.org/10.3390/su13031109.

[33] L. Mochurad and S. A., "Optimizing the computational modeling of modern electronic optical systems," in Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2019, V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, and S. Radetskaya, Eds., Advances in Intelligent Systems and Computing, vol. 1020, Cham, Switzerland: Springer, 2020, pp. 597–608. [Online]. Available: https://doi.org/10.1007/978-3-030-26474-1_41.

[34] S. Y. Liaskovska and Y. V. Martyn, "Development of information technologies for the research of technical systems," Lecture Notes in Networks and Systems, vol. 708, pp. 3–15, 2023.

[35] N. A. Afifie, A. W. Y. Khang, A. S. B. Ja'afar, A. F. B. M. Amin, J. A. J. Alsayaydeh, W. A. Indra, S. G. Herawan, and A. B. Ramli, "Evaluation Method of Mesh Protocol over ESP32 and ESP8266," Baghdad Science Journal, vol. 18, no. 4, pp. 1398–1401, 2021. doi: 10.21123/bsj.2021.18.4(Suppl.).1397.

[36] M. A. Vázquez, R. Plana, C. Pérez, and M. Soto, "Development of technologies for local composting of food waste from universities," Int. J.

Environ. Res. Public Health, vol. 17, no. 9, p. 3153, 2020. https://doi.org/10.3390/ijerph17093153.

[37] O. Maevsky, M. Kovalchuk, Y. Brodsky, V. Stanytsina, and V. Artemchuk, "Game-theoretic modeling in regulating greenhouse gas emissions," Heliyon, vol. 10, no. 9, p. e30549, 2024. [Online]. Available: http://doi.org/10.1016/j.heliyon.2024.e30549.

[38] V. V. Shkarupylo, I. Tomičić, and K. M. Kasian, "The investigation of TLC model checker properties," Journal of Information and Organizational Sciences, vol. 40, no. 1, pp. 145–152, 2016. [Online]. Available: https://doi.org/10.31341/jios.40.1.7.

[39] V. Kovtun, K. Grochla, and K. Połys, "The concept of network resource control of a 5G cluster focused on the smart city's critical infrastructure needs," Alexandria Engineering Journal, vol. 94, pp. 248–256, 2024. [Online]. Available: https://doi.org/10.1016/j.aej.2024.03.038.

[40] M. M. Mahmoud et al., "Voltage quality enhancement of low-voltage smart distribution system using robust and optimized DVR controllers: Application of the Harris Hawks algorithm," International Transactions on Electrical Energy Systems, vol. 2022, pp. 1–18, 2022.

[41] J.-H. Lee, D. Luyima, C.-H. Lee, S.-J. Park, and T.-K. Oh, "Efficiencies of unconventional bulking agents in composting food waste in Korea," Appl. Biol. Chem., vol. 63, no. 1, 2020. https://doi.org/10.1186/s13765-020-00554-6.

[42] J. A. J. Alsayayadeh, M. F. Yusof, M. Z. Abdul Halim, M. N. S. Zainudin and S. G. Herawan, "Patient Health Monitoring System Development using ESP8266 and Arduino with IoT Platform" International Journal of Advanced Computer Science and Applications(IJACSA), 14(4), May 2023, pp. 617-624. http://dx.doi.org/10.14569/IJACSA.2023.0140467.

[43] Oliinyk, A., Fedorchenko, I., Stepanenko, .Rud M., Goncharenko, D. Implementation of evolutionary methods of solving the travelling salesman problem in a robotic warehouse // Lecture Notes on Data Engineering and Communications Technologies, 2021, 48, P. 263–292.

[44] E. H. Rachmawanto et al., "Eggs Classification based on Egg Shell Image using K-Nearest Neighbors Classifier," 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2020, pp. 50-54, doi: 10.1109/iSemantic50169.2020.9234305.

[45] N. Shakhovska, D. Liaskovsky, A. Augousti, S. Liaskovska, and Y. Martyn, "Design and deployment of data developer toolkit in cloud manufacturing environments," CEUR Workshop Proceedings, vol. 3699, pp. 47–56, 2024.

[46] Bisikalo, O., Danylchuk, O., Kovtun, V., Kovtun, O., Nikitenko, O., & Vysotska, V. (2022). Modeling of Operation of Information System for Critical Use in the Conditions of Influence of a Complex Certain Negative Factor. International Journal of Control, Automation and Systems, 20(6), 1904–1913. Springer Science and Business Media LLC. doi: https://doi.org/10.1007/s12555-021-0368-6.

[47] A. Al-Rumaihi, G. McKay, H. R. Mackey, and T. Al-Ansari, "Environmental impact assessment of food waste management using two composting techniques," Sustainability, vol. 12, no. 4, p. 1595, 2020. https://doi.org/10.3390/su12041595.

[48] K. Palaniveloo et al., "Food waste composting and microbial community structure profiling," Processes (Basel), vol. 8, no. 6, p. 723, 2020. https://doi.org/10.3390/pr8060723.

[49] R. Venugopal et al., "Review on unidirectional non-isolated high gain DC–DC converters for EV sustainable DC fast charging applications," IEEE Access, vol. 11, pp. 78299–78338, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3276860.

[50] V. Shkarupylo, I. Blinov, A. Chemeris, V. Dusheba, J. A. J. Alsayaydeh and A. Oliinyk, "Iterative Approach to TLC Model Checker Application," 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2021, pp. 283-287, doi: 10.1109/KhPIWeek53812.2021.9570055.

[51] Oliinyk, A., Fedorchenko, I., Stepanenko, A., Katschan, A., Fedorchenko, Y., Kharchenko, A., Goncharenko, D. "Development of genetic methods for predicting the incidence of volumes of emissions of pollutants in air". 2019 2nd International Workshop on Informatics and Data-Driven Medicine, IDDM, CEUR Workshop Proceedings, 2019, Vol.2488, pp. 340–353.

[52] P. P. Rath, K. Das, and S. Pattanaik, "Microbial activity during composting and plant growth impact: A review," J. Pure Appl. Microbiol., vol. 16, no. 1, pp. 63–73, 2022. https://doi.org/10.22207/jpam.16.1.53.

[53] P. S. Lew, N. N. L. Nik Ibrahim, S. Kamarudin, N. M. Thamrin, and M. F. Misnan, "Optimization of bokashi-composting process using effective microorganisms-1 in smart composting bin," Sensors (Basel), vol. 21, no. 8, p. 2847, 2021. https://doi.org/10.3390/s21082847.

[54] Kovtun, V., & Kovtun, O. (2024). The Concept of Efficient Utilization of the Uplink Frequency Resource of a Smart Factory 5G Cluster by IIoT Devices. In Proc. 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2024). CEUR-WS, 3664, 273-283. Retrieved from https://ceur-ws.org/Vol-3664/paper19.pdf](https://ceur-ws.org/Vol-3664/paper19.pdf.

[55] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[56] Fedorchenko, I., Oliinyk, A., Stepanenko, Zaiko, T., Korniienko S., Kharchenko, A. Construction of a genetic method to forecast the population health indicators based on neural network models // Eastern-European Journal of Enterprise Technologies, 2020, 1 (4-103), P. 52–63. DOI: 10.15587/1729-4061.2020.197319.

[57] Gang Zhao, Mostafa Hooman, Mahdireza Yarigarravesh, Mohammed Algarni, Maria Jade Catalan Opulencia, Fahad Alsaikhan, Abduladheem Turki Jalil, Abdullah Mohamed, Kareem M.AboRas, Md. Lutfor Rahman, Mohd Sani Sarjadi, Vibration analysis of size dependent micro FML cylindrical shell reinforced by CNTs based on modified couple stress theory, Arabian Journal of Chemistry, Volume 15, Issue 10, 2022.

[58] C. K. Yeh, C. Lin, H. C. Shen, N. K. Cheruiyot, M. E. Camarillo, and C. L. Wang, "Optimizing food waste composting parameters and evaluating heat generation," Appl. Sci. (Basel), vol. 10, no. 7, p. 2284, 2020. https://doi.org/10.3390/app10072284.

[59] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[60] W. C. Tan et al., "Conceptual design for smart organic waste recycling system," J. Phys. Conf. Ser., vol. 2051, no. 1, p. 012042, 2021. https://doi.org/10.1088/1742-6596/2051/1/012042.

[61] Oliinyk, A., Fedorchenko, I., Zaiko, T., Goncharenko D., Stepanenko, A., Kharchenko, A. "Development of Genetic Methods of Network Pharmacy Financial Indicators Optimization". 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), 2019, pp. 607–612. DOI: 10.1109/PICST47496.2019.9061396.

[62] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[63] R. Lakshmanan, M. Djama, S. Perumal, and R. Abdulla, "Automated smart hydroponics system using internet of things," Int. J. Electr. Comput. Eng. (IJECE), vol. 10, no. 6, p. 6389, 2020. http://doi.org/10.11591/ijece.v10i6.pp6389-6398.

[64] G. Kennedy et al., "Traditional and indigenous fruits and vegetables for food system transformation," Curr. Dev. Nutr., vol. 5, no. 8, p. nzab092, 2021. https://doi.org/10.1093/cdn/nzab092.

[65] J. A. J. Alsayaydeh, Irianto, A. Aziz, C. K. Xin, A. K. M. Z. Hossain and S. G. Herawan, "Face Recognition System Design and Implementation using Neural Networks" International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), June 2022, pp. 519-526. http://dx.doi.org/10.14569/IJACSA.2022.0130663.

[66] B. Zaman, N. Hardyanti, and P. Purwono, "Fast composting of food waste using thermal composter," IOP Conf. Ser. Earth Environ. Sci., vol. 896, no. 1, p. 012013, 2021. https://doi.org/10.1088/1755-1315/896/1/012013.

[67] H. N. Bhandare, P. P. Bhadekar, S. M. Hakim, K. J. Mane, and S. F. Salunkhe, "Food donation and food waste management website." Zenodo, 2023. doi:10.5281/zenodo.7731077.

[68] R. A. Hamzah, S. F. Abd Ghani, A. F. Kadmin, and K. A. A. Aziz, “A practical method for camera calibration in stereo vision mobile robot navigation,” in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, 2012, pp. 1–5, doi: [10.1109/SCOReD.2012.6518620](https://doi.org/10.1109/SCOReD.2012.6518620).

[69] George Zhou, Marisabel Chang, & Yu Sun. (2021). "A Mobile Platform for Food Donation and Delivery System using AI and Machine Learning." doi:10.5121/csit.2021.111715.

[70] R. A. Hamzah, M. S. Hamid, A. F. Kadmin, and S. F. A. Ghani, “Improvement of stereo corresponding algorithm based on sum of absolute differences and edge preserving filter,” in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, 2017, pp. 58–62, doi: [10.1109/ICSIPA.2017.8120610](https://doi.org/10.1109/ICSIPA.2017.8120610).

[71] V. V. Shkarupylo, I. V. Blinov, A. A. Chemeris, V. V. Dusheba, and J. A. J. Alsayaydeh, "On applicability of model checking technique in power systems and electric power industry," in Systems, Decision and Control in Energy III, Cham: Springer International Publishing, 2022, pp. 3–21, doi: 10.1007/978-3-030-87675-3_1.

[72] S. F. Abd Gani, R. A. Hamzah, R. Latip, S. Salam, F. Noraqillah, and A. I. Herman, “Image compression using singular value decomposition by extracting red, green, and blue channel colors,” *Bull. Electr. Eng. Inform.*, vol. 11, no. 1, pp. 168–175, Feb. 2022, doi: [10.11591/eei.v11i1.2602](https://doi.org/10.11591/eei.v11i1.2602).

[73] P. Quezada, R. Cueva, and F. Paz, “A systematic review of user-centered design techniques applied to the design of mobile application user interfaces,” in Design, User Experience, and Usability: UX Research and Design, Cham, Switzerland: Springer International Publishing, 2021, pp. 100–114.

[74] A. Bhattacharjee, K. S. Rao, and N. Nawaz, “Determinants and outcomes of food delivery app engagement during COVID-19: A study of urban and semi-urban customers,” J. Electron. Commer. Organ., vol. 21, no. 1, pp. 1–22, 2023, doi: 10.4018/jeco.323655.

[75] S. G. Mackenzie and A. R. Davies, “Assessing the sustainability impacts of food sharing initiatives: User testing The Toolshed SIA,” Front. Sustain. Food Syst., vol. 6, 2022, doi: 10.3389/fsufs.2022.807690.

[76] L. Protopappas, D. Bechtsis, and N. Tsotsolas, “IoT services for monitoring food supply chains,” Appl. Sci., vol. 15, no. 13, p. 7602, 2025, doi: 10.3390/app15137602.

[77] C. Isensee, F. Teuteberg, and K. M. Griese, “Exploring the use of mobile apps for fostering sustainability-oriented corporate culture: A qualitative analysis,” Sustainability, vol. 14, no. 12, p. 7380, 2022, doi: 10.3390/su14127380.

Benefits and Challenges of Cloud Computing System in Malaysian Public Healthcare Organizations

Nurul Izzatty Ismail[1], Juhari Noor Faezah[2]*, Muhammad Syukri Abdullah[3], Masrina Nadia Mohd Salleh[4]
Centre For Value Creation and Human Well-Being Studies (INSAN), Faculty of Economics and Management, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia[1, 2, 3]
Center For Global Business and Digital Economy Studies (GloBDE), Faculty of Economics and Management, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia[4]

***Abstract*—Cloud computing has become an emerging technology in information systems (IS), which brings worldwide attention to healthcare management, including in Malaysia. Therefore, the present study aims to identify the benefits and challenges of the cloud system in the healthcare sector. The TOE framework was adopted to explain the benefits and challenges of cloud system implementation in the healthcare sector. The findings show that cost, scalability, data accessibility, and interoperability are factors in a technological context that may enable the successful implementation of cloud systems in healthcare organizations. Apart from that, the size of healthcare organizations and their training were also important factors in the organizational context, while government regulations and policies, as well as cyber threats, are considered crucial factors in the environmental context in cloud implementation in the healthcare sector. A conceptual framework of the cloud system was proposed to provide a comprehensive understanding to optimise future implementation or adoption of the cloud system in the Malaysian healthcare sector.**

***Keywords*—Cloud computing; technology; Malaysia; healthcare sector; public healthcare; TOE framework**

I. Introduction

The healthcare sector is essential for preserving and improving human health. It is imperative to safeguard population well-being and significantly contributes to economic stability and quality of life in every nation, including Malaysia.

Fig. 1 illustrates that the Malaysian population experienced a 2% growth, rising from 33.7 million in 2022 to 34.1 million in 2024 [1]. This circumstance positions healthcare as a predominant industry in Malaysia. Therefore, the Malaysian government has been systematically constructing additional hospitals and clinics nationwide to accommodate the population.

Currently, Malaysia has 146 public hospitals, 2699 public clinics, and 253 community clinics [2]. Despite the presence of numerous public healthcare providers in this country, the available resources are inadequate to meet the needs of all citizens. The shortage is due to patients requiring medical treatments in public hospitals, which offer extensive medical facilities and specialists to assess and treat critical cases. This situation elevates public hospitals as more significant institutions among various medical organizations.

A. Malaysian Public Hospitals

Public hospitals in Malaysia currently possess 45848 beds, with the Malaysian Ministry of Health (MOH) anticipating to increase this number by an additional 5600 beds in the near future [3]. In this nation, experiencing population growth, the available hospital beds are regrettably inadequate to meet the medical care demands. Pillay et al. [4] indicate that patients expressed dissatisfaction with the duration from registration to collection of medications, despite the consultation time being relatively short at approximately 15 minutes. Furthermore, Hassali et al. [5] noted that prolonged waiting times significantly affected patient satisfaction levels. Patients expressed dissatisfaction due to wait times exceeding two hours, indicating that the hospital services were inadequate.

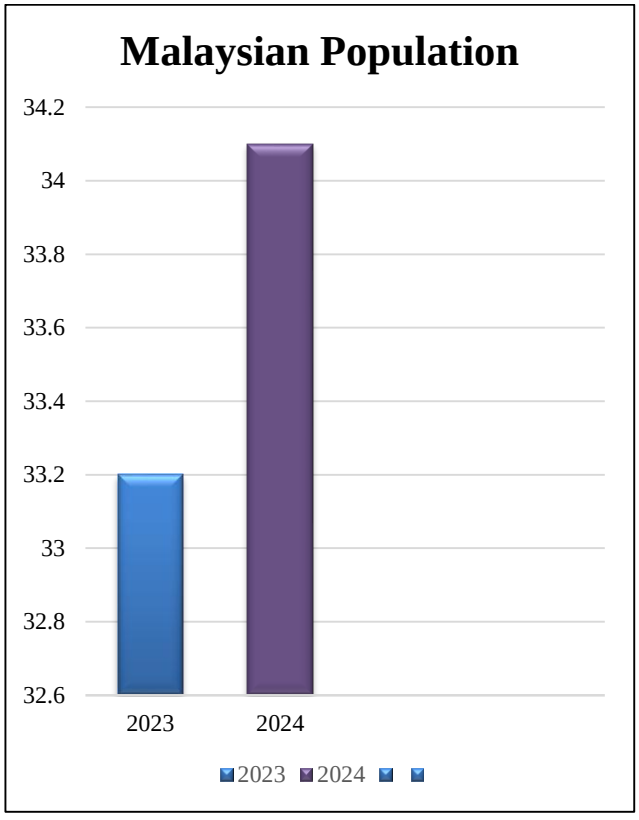


Fig. 1. Malaysian population from 2023 to 2024.

In addition, the Ministry of Health [2] reports numerous individual instances of negligence. In 2023, the Ministry of Health (MOH) approved ex-gratia payments and litigation totaling RM6, 891, 879.28 for 75 cases, representing an increase from the previous year.

This study addresses these gaps by examining the benefits and challenges of cloud computing adoption in Malaysian public healthcare organizations through the lens of the Technology-Organization-Environment (TOE) framework, as well as to propose a conceptual framework to guide future adoption, aligning with Malaysia's digital health aspirations.

This is because the implementation of efficient Information Systems (IS) is crucial for the healthcare sector in this country. Besides that, the cloud system is essential for managing all the aforementioned issues across healthcare organizations in this country.

II. LITERATURES

A. Information System (IS)

Information systems (IS) can be defined as a network of related components that engage in collecting, processing, storing, and distributing data for a designated target, especially to facilitate decision-making, coordination, and management within an organization [6], [7].

As illustrated in Fig. 2, the workflow follows a step-by-step process, where raw data is collected from different sources, processed using analytical methods, stored in structured databases, and delivered to the intended destinations. This approach improves data quality, supports smooth system integration, and helps generate useful insights for better decision-making. Beyond this, the workflow demonstrates enduring relevance across decades of technological evolution, maintaining its position as the foundational architecture for both legacy systems and contemporary cloud-based implementations.

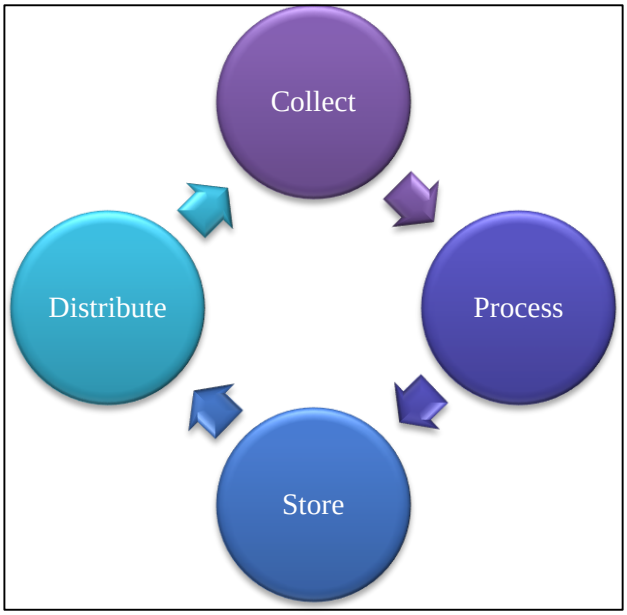


Fig. 2. Information systems (IS) flow.

In addition to the IS flow presented in the preceding figure, information systems fundamentally consist of six critical components named as hardware, software, data, procedures, people, and networks. The descriptions of these components are shown in Table I.

TABLE I. THE COMPONENTS OF INFORMATION SYSTEM

Key Components of IT	Definitions
Hardware	Hardware refers to the physical devices and equipment used in an information system that provide the infrastructure for processing and storing data, such as computers, servers, storage devices, networking equipment, and peripherals [6].
Software	Software includes the programs and operating systems that manage hardware resources and enable users to perform specific tasks [8].
Data	Data is the raw material of an information system, which consists of facts, figures, and information that are processed to produce meaningful outputs [9].
Procedures	Procedures are the rules, guidelines, and workflows that govern how data is processed, stored, and disseminated within an information system. They ensure consistency, efficiency, and security in the operation of the system [6].
People	People are the users, developers, and managers of an information system. They include end-users who interact with the system [8].
Networks	Networks enable communication and data exchange between different components of an information system; for example, they include local area networks (LANs), wide area networks (WANs), and the internet [9]

These components depend on each other to ensure that data is properly collected, preserved, and distributed to meet specific goals, especially in the healthcare sector. Hence, the evolution of cloud systems needs to be highlighted.

According to [10], cloud system development began in 2006, initiated by Amazon, and in 2008, Google launched its Google Cloud, while Microsoft launched its Windows Azure.

The cloud system provides on-demand access to IS resources, such as servers, storage, databases, networks, software, and analytics over the internet with pay-as-you-go pricing, providing significant cost savings when users avoid upfront costs with hardware or infrastructure [11].

B. IS in the Healthcare Sector

A health information system is a core function of the health system; it integrates data collection, processing, and reporting to help support decision-making at the public health level and avoid problems pertaining to healthcare (WHO, 2016). Moreover, [12] expanded on this, noting that this system brings together collecting, processing, and reporting data to improve the delivery and outcome of care.

Information systems in healthcare have revolutionized the delivery, management, and optimization of medical services. They cover a broad spectrum of technologies that aim to capture, manage, and share patient information. Aside from that, IS may be able to structure clinical workflows and assist with decision-making processes. Such systems are crucial in today's health structure.

The popularity of cloud computing technology in the healthcare sector is growing widely recently. An example of cloud computing includes the Electronic Health Records (EHRs), which allow secure storage, access, and sharing of patient records via multiple locations [13].

In Malaysia, an online cloud-native web known as Medical Program Information System (MPIS) has been developed by the Hospital Services Management Unit (Medical Development Division) for use in public health care institutions [3]. This system was developed following COVID-19 to ensure efficient patient services during the pandemic. The MPIS application remains an ongoing process that has not been fully utilized by all Malaysian public healthcare institutions. In this case, the Ministry of Health (MOH) and the Malaysian government are to shoulder the responsibility of enabling digitalization for the country's healthcare sector; their initiatives will benefit both healthcare sector practitioners and citizens.

C. The Advantages of Cloud Systems in the Healthcare Sector

Cloud system provides several advantages, such as better data accessibility, cost-effectiveness, scalability, flexibility, and better data security and audits.

1) Accessibility of data: Cloud technology has the capability of increasing the accessibility of data among healthcare providers [14]. For example, doctors and nurses can access patient data on their smartphones, tablets, or laptops anywhere, even outside the hospital setting. This is because the data is accessible to all healthcare providers, including doctors and specialists, regardless of their location. Furthermore, the study [15] asserts that facilitating data sharing grants access and allows healthcare providers to collaboratively review and modify patient records, resulting in improved care coordination and reduced medical errors. In addition to that, cloud technology enables patients to access their health information, set appointments, and consult their providers through the patient's portal [16].

2) Cost-effectiveness: Marston et al. [17] point out that cloud computing provides a more affordable option compared to legacy IT infrastructure. In this case, the organizations pay only for the resources they consume [18]. For instance, a pay-as-you-go model could minimize overprovisioning and wasted capacity [19]. Cloud tech provides lower maintenance and IT staffing. This advantage is due to the fact that cloud providers take care of server-side maintenance, security patches, and upgrades [16], minimizing the need to have IT personnel in place. Furthermore, cloud technology has the potential to cut energy and space costs, as it can replace on-site servers and significantly lower power and cooling costs [20]. It could manage system maintenance, updates, and security, alleviating the burden on healthcare IT personnel.

3) Scalability and flexibility: Cloud technology in the healthcare sector is also beneficial because of its scalability and flexibility. Scalability is the ability of a system to handle a growing or consistent workload by adding resources at a fixed cost [21]. In this scenario, the cloud technology could adapt effortlessly to the increasing volume of healthcare data and represents a significant advantage regarding data-hungry trades, such as the healthcare sector [11]. Other than the above, healthcare institutions would be able to implement new applications or services on the cloud technology, which would facilitate faster responsiveness to emerging clinical or operational requirements.

4) Enhanced data security: Aside from that, the cloud technology can improve data security [16], mainly due to the encryptions that have been employed in the movement and storage of data, which decrease the danger of breaches. In addition, the most popular applications of cloud technology in the healthcare sector come with built-in compliance with strict healthcare regulations, such as the HIPAA (Health Insurance Portability and Accountability Act) and the GDPR (General Data Protection Regulation), safeguarding patients' sensitive information in both secure and legal processes. Furthermore, Gabriel et al. [22] assert that patient records in cloud-based EHR systems can only be edited by authorized personnel.

5) Disaster recovery and business continuity: Sultan [23] contends that cloud technology can minimize data redundancy and reduce the likelihood of data loss resulting from hardware failures or natural calamities. Al-Issa et al. [20] agreed that cloud storage ensures redundant backups so that data loss resulting from ransomware may be prevented. For example, healthcare providers can recover data to resume interrupted workflows in the event of system failure or disruption.

D. Evolving IS Issues and Challenges in the Healthcare Sector

Notwithstanding the advantages, the implementation of cloud technology in the healthcare sector encounters obstacles, including high implementation costs, data privacy and security dilemmas, interoperability issues, user resistance and training requirements, system downtime and reliability challenges, as well as ethical concerns.

1) High implementation costs: The problem of implementing the cloud technology in healthcare organizations is that of high implementation costs. To develop, deploy, and maintain healthcare information systems is a costly business [24]. For instance, development costs could increase if cloud technology needs to interact with platforms that require specialized APIs or middleware [22]. Egress fees, which are costs related to data transfer, are often considered hidden costs of cloud technology in the healthcare sector, especially concerning large medical datasets [25].

2) Data privacy and security: Another challenge faced by cloud technology in the healthcare sector is data privacy and security. Certain implementation issues are more prominent in this case, like certification, authentication, privacy, and confidentiality [26]. In fact, using cloud technology makes it vulnerable to cybersecurity attacks, e.g., ransomware, phishing, and unauthorized access, all of which pose significant threats to patient data [16].

3) Interoperability: Interoperability is also a challenge facing the implementation of IS in the healthcare sector. Many

healthcare systems are single-directional (i.e., standalone), rendering data transfer between organizations a challenge. According to Häyrinen et al. [12], such a system is not standardized because the data formats and protocols for data sharing are incompatible. In addition, any obsolete or incompatible data cannot be updated in real time [19]. In the future, this situation may pose a difficulty or issue integrating with another system.

4) User resistance and training: In the implementation of cloud technology in the healthcare sector, a lack of training is a serious issue. According to McGinn et al. [27], implementation of new systems can disrupt frequent workflows and result in dissatisfaction among users. Gagnon et al. [28] agreed that users of cloud system in the healthcare sector might find it disruptive. In addition, these users experience strain and exhaustion when adapting to entirely new cloud interfaces, which indicates poor training programmed [29]. Hence, comprehensive training programmers are necessary for all staff.

5) System availability and stability: System downtime and reliability are also challenges in the implementation of IS in the healthcare sector. According to Raghupathi and Raghupathi [30], certain hospitals reported delays of as much as 30 seconds in viewing the patient records stored in the cloud. Furthermore, the cloud system in the healthcare sector may lack effectiveness [19]. Technical glitches, bugs, and crashes can jeopardize patient care. As a result, routine maintenance is required to minimize downtime, which can be both costly and time-consuming [13].

6) Ethical concerns: The final challenge in the implementation of IS in the healthcare sector is ethics. This difficulty is due to the fact that the patients' data is used, which raises concerns about ethically obtaining consent and using the patients' data or information [31]. Previously, 500,000 patient records were exposed due to misconfigured storage [32]. Moreover, the cloud technology can give rise to the cloud computing vulnerabilities such as cyberattacks, insider misuse, and third-party sharing of sensitive patient records, thereby breaching patient privacy and data security [33], [16].

III. IS THEORIES

A theory is a well-substantiated explanation of some aspect of the natural world, based on a body of facts that have been repeatedly confirmed through observation and experiment [34]. Gregor [35] asserts there are five main purposes of theories in IS: analysis (theory to point out), explanation (theory to illustrate), prediction (theory to predict), prescription (theory to prescribe), and design and action. Theories are based on a body of evidence that supports them; they explain and predict phenomena relevant to this evidence. In IS studies literature, there exist numerous theories to account for IS adoption and implementation. These are the Theory of Reasoned Action (TRA) [36], the Theory of Planned Behavior (TPB) [37], the Technology Acceptance Model (TAM) [38], the Technological, Organizational, Environmental Framework (TOE) [39], the DeLone and McLean Model [40], the Diffusion of Innovation Theory [41] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [42].

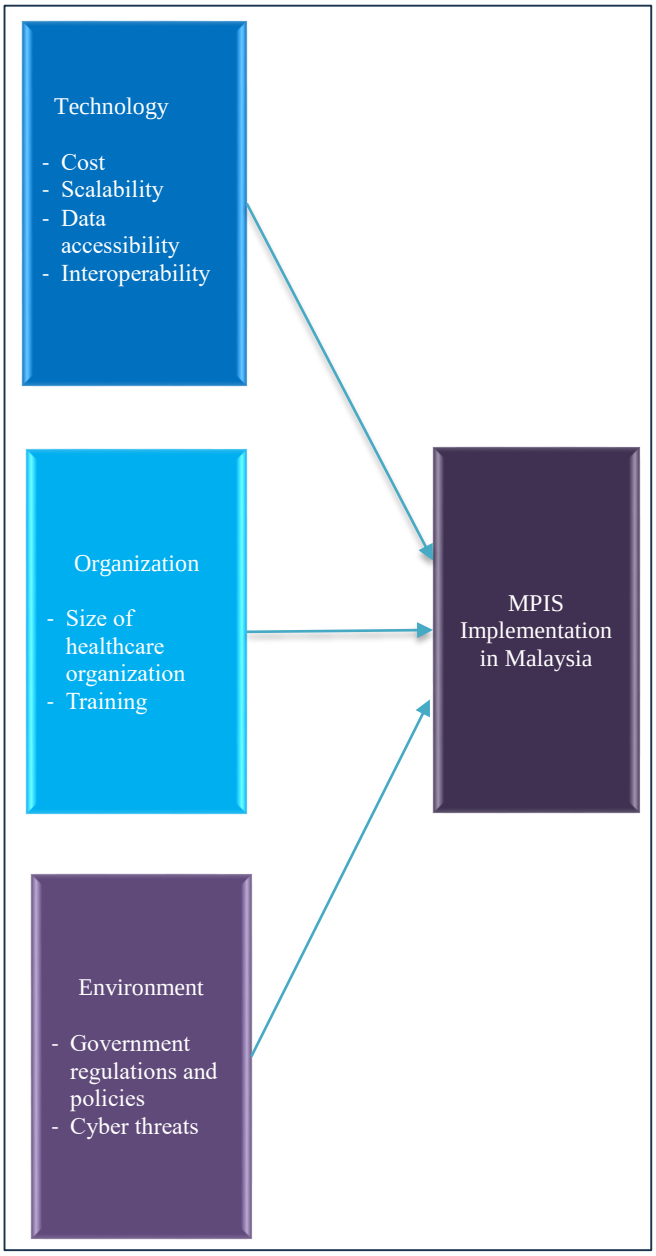


Fig. 3. A conceptual model of cloud system implementation in Malaysian healthcare.

In the present study, the TOE framework [39] was employed to tackle the cloud implementation in the healthcare sector as illustrated in Fig. 3. This conceptual model systematically organizes the critical dimensions of Technology (cost, scalability, data accessibility), Organization (healthcare institution size, training protocols), and Environment (government regulations, cyber threats) that collectively determine cloud adoption success. This is important to identify total cloud technology utilization in Malaysian healthcare organizations and to improve its implementation.

IV. DISCUSSION

In the present technological era, information technology is a cornerstone of every organization globally. This principle is applicable even in the healthcare sector. For this reason, the

healthcare sector frequently employs cloud computing. The aforementioned case study illustrates that developed nations have successfully implemented cloud-based tech due to its robust technological and financial capabilities; however, cloud applications are also viable in developing countries such as Malaysia.

During the COVID-19 pandemic, the MOH has been so much more careful in adapting cloud computing, in particular MPIS in Malaysia. In the future, some public hospitals nationwide will deploy the system. Nevertheless, the system has not yet been fully adopted in all public hospitals in this country. Thus, a high-level adoption model based on the benefits and challenges experienced by the cloud technology represents an important instrument that helps to guide the overall MPIS implementation process.

The significance of this study can be tagged as additional support for the relevancy of the TOE framework for IT adoption in the healthcare sector. It builds on the model (which had not previously considered the impact of telehealth and cloud analytics) to provide insights into new moderating variables. Thus, the TOE Framework was used to examine the adoption of new technologies, including cloud computing in the healthcare sector, consistent with earlier studies [43]. Applying this model provides considerations for cloud technology implementation issues and challenges, such as MPIS, which are to be divided into three types of contexts: technology, organization, and environment. The conceptual model of cloud system implementation in the Malaysian healthcare sector is presented in Fig. 3, which was based on the TOE framework.

Hence, it ensures the best alternative for the healthcare sector to cope with optimal patient data over cloud technology tailored for first-class, cost-effective extensibility [44]. However, Jones et al. [45] argued that concerns over security are still a major barrier, especially regarding compliance with HIPAA and GDPR. However, multi-factor authentication (MFA) and zero-trust architectures, as shown in the literature, can mitigate these risks [46]. Moreover, they embraced the Fast Healthcare Interoperability Resources (FHIR) standards to improve data sharing [47] due to the barrier that this technology has for interoperability in both cloud-based systems and EHR.

The literature suggests that the size of an organization influences the adoption of cloud technology in healthcare. For instance, large hospitals would be appropriately budgeted for the technology, while smaller clinics may have budget limitations [48]. In addition, a lack of leadership support and staff training is another major factor that impacts the organization's implementation efficiency [49].

In other contexts, this technology implementation has been significantly influenced by numerous external pressures, such as government regulations and competitive dynamics. To substantiate this, the HITECH Act and the 21st Century Cures Act advocate for a digital shift, which may simultaneously heighten compliance obligations and potentially engender instability [50]. As a result, there are cyber threats like ransomware and other attacks on healthcare's data, despite improved trust in cloud implementation by the developers, such as AWS, Microsoft Azure, and Google Cloud [51].

A. Recommendations and Future Research

This study proposes recommendations, including implementing hybrid clouds (combinations of private and public clouds) in healthcare organizations. Hybrid cloud models are preferred in healthcare because they facilitate a balance between security and scalability [52]. The close collaboration between cloud providers and policymakers is essential to mitigate the significant cyber threats and risks in the healthcare sector [53]. Accordingly, national digital health strategies are crucial for policy and governance to assess how the system can be standardized and rendered interoperable. Hence, the Ministry of Health must collaboratively develop a cybersecurity framework with health informatics experts from Malaysian universities.

V. Conclusion

The implementation of online cloud technology in Malaysia marks a transformative step towards modernizing the nation's healthcare delivery through the cloud system. Piloting the implementation of interoperability standards at leading hospitals would be a major step forward in MPIS implementation in the Malaysian healthcare sector. However, its implementation success might be negatively impacted by certain technological, organizational, and contextual challenges. Instead, Malaysia needs to take a strategic, multi-phased approach towards policy alignment, technological innovation, and human capital development. By fortifying digital infrastructure, scaling up Artificial Intelligent (AI) capabilities and tackling rural versus urban inequities, Malaysia can turn its vision for a fully integrated, data-driven healthcare system into reality. These initiatives may lead to improved national healthcare services through continued cooperation between the MOH, technology providers, and informatics professionals until a nationwide implementation of MPIS is successful. With sustained effort and appropriate implementation, MPIS can help lay the groundwork for a more inclusive, responsive, and technologically prepared Malaysian healthcare system, which translates into improved outcomes for all Malaysians.

References

[1] Department of Statistics Malaysia, *The Population of Malaysia*, 2025. [Online]. Available: https://open.dosm.gov.my/dashboard/population. [Accessed: May 7, 2025].

[2] Ministry of Health Malaysia, *Annual Report 2024*. Putrajaya, Malaysia: MOH, 2024.

[3] H. Ismail, M. R. Ahmad, A. Lim, and N. Azmi, "Mapping the locations of medical specialists in the Ministry of Health's hospitals in Malaysia by specialty, subspecialty and area of interest," *Malaysian Journal of Medical Sciences*, vol. 31, no. 3, pp. 1–17, 2024. [Online]. Available: https://doi.org/10.21315/mjms2024.31.3.1

[4] M. S. Pillay, N. Jamaludin, S. R. Lim, and M. A. Hassan, "Hospital waiting time: The forgotten premise of healthcare service delivery?," *International Journal of Health Care Quality Assurance*, vol. 24, no. 7, pp. 506–522, 2011.

[5] M. A. Hassali, S. Shafie, A. Khan, and A. Awaisu, "Hospital waiting time: The forgotten premise of healthcare service delivery?," *International Journal of Health Care Quality Assurance*, vol. 24, no. 7, pp. 506–522, 2011. [Online]. Available: https://doi.org/10.1108/09526861111160553

[6] K. C. Laudon and J. P. Laudon, *Management Information Systems: Managing the Digital Firm*, 17th ed. Pearson, 2021.

[7] R. K. Rainer and B. Prince, *Introduction to Information Systems*, 7th ed. Wiley, 2020.

[8] R. M. Stair and G. W. Reynolds, *Principles of Information Systems*, 10th ed. Cengage Learning, 2011.

[9] J. A. O'Brien and G. M. Marakas, *Management Information Systems*, 10th ed. McGraw-Hill, 2011.

[10] V. Choudhary, J. Vithayathil, and R. Agarwal, "The impact of cloud computing on industry structure," *Journal of Management Information Systems*, vol. 38, no. 2, pp. 45–71, 2021.

[11] M. Armbrust *et al*., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010. [Online]. Available: https://doi.org/10.1145/1721654.1721672

[12] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, 2008. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2007.09.001

[13] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," *Risk Management and Healthcare Policy*, vol. 4, pp. 47–55, 2011. [Online]. Available: https://doi.org/10.2147/RMHP.S12985

[14] T.-T. Kuo, H.-E. Kim, and L. Ohno-Machado, "Blockchain distributed ledger technologies for biomedical and health care applications," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1211–1220, 2017. [Online]. Available: https://doi.org/10.1093/jamia/ocx068

[15] D. W. Bates and A. A. Gawande, "Improving safety with information technology," *New England Journal of Medicine*, vol. 348, no. 25, pp. 2526–2534, 2003. [Online]. Available: https://doi.org/10.1056/NEJMsa020847

[16] C. S. Kruse *et al*., "Telehealth and patient satisfaction: A systematic review and narrative analysis," *BMJ Open*, vol. 7, no. 8, e016242, 2017. [Online]. Available: https://doi.org/10.1136/bmjopen-2017-016242

[17] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing—The business perspective," *Decision Support Systems*, vol. 51, no. 1, pp. 176–189, 2011. [Online]. Available: https://doi.org/10.1016/j.dss.2010.12.006

[18] S. D. Ahuja *et al*., "Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: An individual patient data meta-analysis of 9,153 patients," *PLOS Medicine*, vol. 9, no. 8, e1001300, 2012. [Online]. Available: https://doi.org/10.1371/journal.pmed.1001300

[19] A. J. Bokolo, "Application of telemedicine and eHealth technology for clinical services in response to COVID-19 pandemic," *Health and Technology*, vol. 11, no. 2, pp. 359–366, 2021. [Online]. Available: https://doi.org/10.1007/s12553-020-00516-4

[20] Y. Al-Issa, M. A. Ottom, and A. Tamrawi, "eHealth cloud security challenges: A survey," *Journal of Healthcare Engineering*, vol. 2019, Article ID 7516035, 2019. [Online]. Available: https://doi.org/10.1155/2019/7516035

[21] A. B. Bondi, "Characteristics of scalability and their impact on performance," in *Proc. 2nd Int. Workshop on Software and Performance*, 2000, pp. 195–203. [Online]. Available: https://doi.org/10.1145/350391.350432

[22] M. Gabriel, G. Pessin, E. Bourguignon, and J. M. de Oliveira, "Cloud computing: Architectural and policy perspectives," *Journal of Information Technology*, vol. 33, no. 4, pp. 345–362, 2018.

[23] N. Sultan, "Making use of cloud computing for healthcare provision: Opportunities and challenges," *International Journal of Information Management*, vol. 34, no. 2, pp. 177–184, 2014. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2013.12.011

[24] D. Blumenthal and M. Tavenner, "The 'Meaningful Use' regulation for electronic health records," *New England Journal of Medicine*, vol. 363, no. 6, pp. 501–504, 2010. [Online]. Available: https://doi.org/10.1056/NEJMp1006114

[25] S. Levy, *Cloud Empires: The Political Economy of Digital Infrastructure*. MIT Press, 2021.

[26] A. Boonstra and M. Broekhuis, "Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions," *BMC Health Services Research*, vol. 10, p. 231, 2010. [Online]. Available: https://doi.org/10.1186/1472-6963-10-231

[27] C. A. McGinn *et al*., "Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: A systematic review," *BMC Medicine*, vol. 9, p. 46, 2011. [Online]. Available: https://doi.org/10.1186/1741-7015-9-46

[28] H. Alami, M.-P. Gagnon, and J.-P. Fortin, "Telehealth in light of cloud computing: Clinical, technological, regulatory and policy issues," *Journal of the International Society for Telemedicine and eHealth*, vol. 4, p. e11, 2016.

[29] G. Holden, *Information Systems: A Management Approach*, 2nd ed. Pearson, 2018.

[30] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and Systems*, 2020.

[31] K. W. Goodman, *Ethics, Medicine, and Information Technology: Intelligent Machines and the Transformation of Health Care*. Cambridge University Press, 2016.

[32] HIPAA Journal, "Google Cloud misconfiguration exposes 500K patient records," Feb. 15, 2023. [Online]. Available: https://www.hipaajournal.com/google-cloud-misconfiguration-exposes-500k-patient-records/

[33] I. G. Cohen and M. M. Mello, "HIPAA and protecting health information in the 21st century," *JAMA*, vol. 320, no. 3, pp. 231–232, 2018.

[34] National Academy of Sciences, *Science and Creationism: A View from the National Academy of Sciences*, 2nd ed. National Academies Press, 1999. [Online]. Available: https://doi.org/10.17226/6024

[35] S. Gregor, "The nature of theory in information systems," *MIS Quarterly*, vol. 30, no. 3, pp. 611–642, 2006. [Online]. Available: https://doi.org/10.2307/25148742

[36] I. Ajzen and M. Fishbein, *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[37] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, 1991.

[38] F. D. Davis, "Perceived usefulness, perceived ease of use, and end user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 318–339, 1989.

[39] L. G. Tornatzky and M. Fleischer, *The Processes of Technological Innovation*. Lexington Books, 1990.

[40] W. DeLone and E. McLean, "Information systems success: The quest for the dependent variable," *Information Systems Research*, vol. 3, no. 1, pp. 60–95, 1992.

[41] E. M. Rogers, *Diffusion of Innovation*, 4th ed. New York: Free Press, 1995.

[42] V. Venkatesh, M. Morris, G. Davis, and F. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[43] T. R. Baker and E. F. Green, "Applying the TOE framework to cloud adoption in healthcare: A systematic analysis," *Health Informatics Journal*, vol. 28, no. 3, pp. 456–472, 2022. [Online]. Available: https://doi.org/10.1177/14604582221087654

[44] A. B. Smith and C. D. Chang, "Cloud computing in healthcare: Benefits and challenges," *Journal of Healthcare Information Management*, vol. 36, no. 4, pp. 112–125, 2022.

[45] A. B. Jones, C. D. Smith, and E. F. Lee, "Data security in medical cloud systems," in *Cybersecurity in Modern Healthcare*, R. Brown and T. Davis, Eds. Springer, 2021, pp. 45–67.

[46] A. Kumar and R. Sharma, "Digital transformation in healthcare: Challenges and opportunities," *Journal of Health Informatics*, vol. 15, no. 2, pp. 123–140, 2023.

[47] M. J. Garcia and S. H. Lee, "Cloud-based EHR implementation challenges in rural healthcare systems," *Journal of Healthcare Information Management*, vol. 36, no. 4, pp. 215–230, 2022.

[48] D. E. Brown and J. F. Wilson, "Cloud computing adoption in healthcare: Barriers and facilitators," *Health Informatics Journal*, vol. 27, no. 2, pp. 1–15, 2021. [Online]. Available: https://doi.org/10.1177/14604582211012345

[49] R. M. Taylor *et al*., "Machine learning applications in cloud-based electronic health records: A systematic review," *Journal of Medical*

Systems, vol. 46, no. 8, p. 56, 2022. [Online]. Available: https://doi.org/10.1007/s10916-022-01845-9

[50] A. B. Miller and C. D. White, "Blockchain applications for secure health data exchange in cloud environments," *Journal of Healthcare Information Management*, vol. 37, no. 2, pp. 45–60, 2023.

[51] X. Li and Y. Zhang, "Edge computing for real-time healthcare analytics: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2145–2156, 2022.

[52] L. Chen, H. Wang, Y. Liu, W. Zhang, and J. Kim, "Federated learning for COVID-19 prediction using multi-institutional cloud-based data," *Nature Digital Medicine*, vol. 6, no. 1, p. 45, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00745-9

[53] T. M. Robinson, X. Chen, A. R. Patel, E. F. Williams, and S. H. Kim, "Cloud-based AI for early sepsis detection: A multi-center validation study," *The Lancet Digital Health*, vol. 5, no. 8, pp. e512–e523, 2023. [Online]. Available: https://doi.org/10.1016/S2589-7500(23)00123-4

niCNN: A Novel Neuromorphic Approach to Energy-Efficient and Lightweight Human Activity Recognition on Edge Devices

Preeti Agarwal
School of Technology Management and Engineering,
Narsee Monjee Institute of Management Studies, Maharashtra, India

***Abstract*—Recent years have seen a surge in the use of deep learning for human activity recognition (HAR) in various applications. However, running complex deep learning models on edge devices with limited resources, such as processing power, memory, and energy, is challenging. The objective of this study is to design a novel, lightweight and energy-efficient neuromorphic inspired CNN (niCNN) architecture for real-time HAR on edge devices. The niCNN architecture consists of four stages: design of a shallow CNN, conversion into an equivalent spiking network using Clamping and Quantization (CnQ) algorithm to minimize information loss, threshold balancing to calculate spiking neuron firing rate using Threshold Firing (TF) algorithm, and edge deployment. The experimental evaluation shows that the niCNN architecture achieves 97.25% and 98.92% accuracy on two publicly accessible HAR datasets, WISDM and mHealth. Furthermore, the niCNN technique retains a low inference latency of 2.25 ms and 2.36 ms, as well as a low memory utilization of 22.11 KB and 31.84 KB, respectively. Furthermore, energy usage is reduced to 5.2w and 5.8w. In comparison to various state-of-the-art and baseline CNN models, the niCNN architecture outperforms them in terms of classification metrics, memory usage, energy consumption, and inference delay. The CnQ algorithm reduces memory usage and inference latency, while the TF algorithm improves classification accuracy. The findings show that neuromorphic computing has a lot of potential for resource-constrained edge devices.**

***Keywords*—Neuromorphic computing; human activity recognition (HAR); edge computing; convolution neural network (CNN); spiking neural network (SNN); sensors**

I. Introduction

Human activity recognition (HAR) has garnered significant attention in recent years due to its potential applications in healthcare, sports, and security [1]. Many deep learning techniques, including Convolutional Neural Networks (CNNs), have been applied to HAR with promising results [2]. However, most of these techniques are designed to be run on high-performance computing systems, which can limit their practical applications.

Conversely, performing HAR on edge devices, such as smartphones or wearables has emerged as viable solution due to their ubiquity, low cost, and personalized features [3]. However, running complex deep learning models on edge devices poses many challenges, including limited processing power, memory, and energy resources [4] [5]. Therefore, there is a need to develop lightweight and energy-efficient deep learning models for HAR on resource-constrainted edge devices.

One promising approach can be the use of neuromorphic networks, which mimic the brain's neural structure in a new and alternative computer architecture. They significantly reduce energy consumption by replacing weight multiplications with additions [6]. Very few studies have explored the potential of neuromorphic computing for HAR.

The major objective of this study is to propose a novel neuromorphic CNN (niCNN) architecture for HAR that is specifically tailored to edge device. Traditional methods of transferring weights from a CNN to a neuromorphic network often result in accuracy loss and higher inference latency, making them unsuitable for real-time applications. To address this drawback, the niCNN architecture utilizes clamping and quantization (CnQ) algorithm along with improved threshold firing (TF) algorithm to improve accuracy while reducing inference latency, memory requirements and energy consumption. The key contributions of the study are:

- Designing a novel niCNN architecture for real-time edge processing.
- Introducing a CnQ algorithm that reduces inference latency, energy consumption, and memory requirements of the model while maintaining accuracy.
- Incorporating a novel TF algorithm to improve the accuracy of the model.
- Experimental evaluation is performed on two publicly available HAR datasets, WISDM [7] and mHealth [8].
- The performance of niCNN is evaluated in terms of classification metrics, memory consumption, energy consumption and inference latency. The results show that the proposed architecture takes the lead compared to many existing model.

The study is organized as follows: Section II presents a brief overview of human activity recognition (HAR) and related work. Section III describes the proposed niCNN model in detail, including the CnQ and TF algorithms. Section IV presents the experiment and compares the model with the state-of-the-art approaches. Section V provides the performance results of niCNN. Finally, Section VI concludes the study with suggestions for future research directions.

II. Background and Related Work

A. Background

The process of recognizing human activities through sensors involves four key stages, namely data acquisition, preprocessing, learning, and evaluation [9]. The data acquisition stage involves collecting data from sensors, which is then transmitted to the next stage for further processing. Since the data obtained from sensors may contain artifacts and noise, it needs to be preprocessed. This involves filtering, segmentation, feature extraction, and selection. Filtering helps remove unwanted values and fill in any missing ones. Segmentation involves breaking the data into smaller segments or windows, which are then used for feature extraction and selection. Finally, different classification algorithms are used for activity recognition, and their performance is assessed using defined metrics.

B. Related Work

For HAR, various researchers have proposed several neuromorphic approaches to reduce the energy consumption in on-edge deployment. These approaches include using spiking neural networks (SNNs) [10], bio-inspired artificial retinas [11], data fusion from multiple sensing systems [12] [13], and data preprocessing frameworks [14]. Additionally, researchers have proposed frameworks that integrate physics and neurobiology to model and recognize human actions and object activities, as well as a hardware module that uses a Hopfield Neural Network to detect falls through sensor data integration and classification [15] . In [16], the authors presents a hardware module that uses a Hopfield Neural Network to recognize falls through sensor data integration and classification. In study [17], the use of SNNs in HAR tasks has been proposed as a way to achieve spatio-temporal feature extraction and reduce energy consumption by up to 94%.

Table I provides the listing of different neuromorphic approaches to HAR, each with its unique advantages and limitations. However, these approaches have limitations in terms of accuracy or processing for real-time applications. In comparison to existing approaches, the proposed niCNN, inspired by neuromorphic computing and incorporating the CNQ and TF algorithms, achieves higher accuracy with more optimization parameters, as shown in the table.

TABLE I. Comparison of Approaches

Ref.	Focus of Work	Advantages	Limitations	Optimization Parameters considered			
				Energy	Memory	Latency	Accuracy
[10]	Neuromorphic approach for mitigating energy consumption for edge AIoT HAR applications using SNN.	High performance at low energy cost, effective dealing with temporal signals.	Limited comparison, only DNN, LMU architecture used, requiring application oriented hyperparameter optimization.	✓	×	✓	✓
[11]	Applied SNN and event memory surfaces for HAR using event based camera.	Eliminates data redundancy, allows for low latency and data sparisity.	Technique is complicated and difficult to easily adapt to domain.	✓	✓	×	✓
[12]	Performed multisensory data fusion for precise HAR using neuromorphic computing	Reliable, robust, and achieved high accuracy with multi-sensor data.	Requires additional hardware signal processing for flexible integration of data.	×	×	×	✓
[14]	Proposed data processing framework using neuromorphic computing to reduce the need for extensive sensor training samples.	Robustness with lesser number of training samples requirement.	Not full HAR deployment discussed.	×	✓	×	✓
[13]	HAR using Neuromorphic computing with one-shot learning on multi-sensor fused data	Increased robustness and accuracy of activity recogniton	Limited dataset	×	✓	×	✓
[15]	Integration of physics and neurobiology to model and recognize human actions	Global representation of motion, integration of neurobiological models, multiple hypothesis testing framework.	Very complex implementation	×	×	×	✓
[16]	Designing hardware module for fall detection using Hopfield Neural Network.	High accuracy, power-efficient, faster data classification.	Limited to fall detection, hardware specific design	✓	✓	✓	✓
[17]	Explored AI based SNN for HAR allowing spatio-temporal feature extraction.	Reduced energy consumption due to binary spikes with higher accuracy.	Relies on traditional artifiacl neural networks	✓	×	×	✓
Ours	Designed a four-stage neuromorphic model for HAR	Higher accuracy with low power consumption, memory and lesser inference time.	Training time higher with multiple training stages. Hyperparameter optimization required	✓	✓	✓	✓

III. Proposed Methodology

In this section, the niCNN architecture is discussed. The proposed architecture addresses the following neuromorphic architecture and CNN integration issues.

- In neuromorphic models, such as Spiking Neural Network (SNN), input values are binary [0,1] at each step, whereas in CNN they are floating point values,

inappropriate conversion reduces accuracy and inference speed.

- Secondly, the activation behavior of IF neurons in SNNs differs from ReLU neurons in CNNs.
- Thirdly, the choice of threshold balancing algorithm greatly affects the performance of the model by affecting the firing rate, energy consumption and information transfer of neurons. High firing rate can lead to overfitting, whereas low firing rate can lead to reduced information transfer.

The niCNN follows a series of basic steps. First, the data is acquired through sensors and preprocessed to create a shallow CNN architecture. Then, an optimal shallow CNN architecture is converted into an equivalent spiking neural network architecture using the CnQ transformation model that minimizes information loss. Next, the firing threshold of the spiking neurons is determined using TF, a threshold balancing technique. The final step is to deploy the inference model, which takes spike-encoded data as input, enabling efficient implementation of deep learning on resource-constrained devices. The schematic workflow of the model is shown in Fig. 1. The Algorithm 3 for niCNN implementation is as follows:

Algorithm 3: Procedure: niCNN

Phase I: Data Acquisition

Step1: Collect sensor data

Step 2: Aggregate sensor data

Phase II: Preprocessing

Step 3: Remove missing values and outliers

Step 4: Windowing and Segmentation

Phase III: Model Building

Step 5: Train a shallow CNN

Step 6: Apply CnQ algorithm for weight transformation

Step 7: Set firing threshold values using TF algorithm

Step 8: Deploy SNN inference model

Phase IV: Performance Evaluation

Step 9: Evaluate different performance metrics

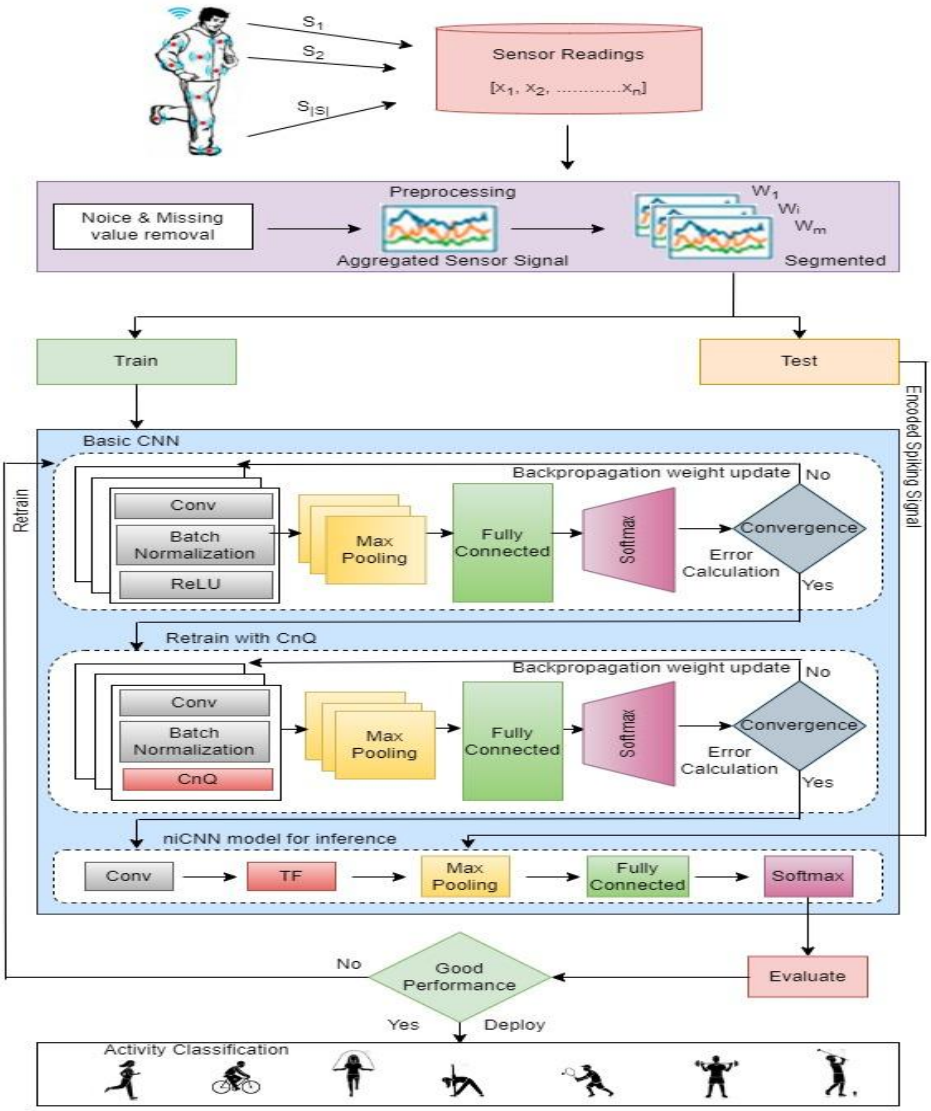


Fig. 1. niCNN architecture.

The description of each phase of the algorithm is as follows:

A. Phase 1: Data Acquisition

A collection of sensors $S = \{s_1, s_2, \dots, s_{|S|}\}$are used to acquire sensor data, each generating a time series reading $X = \{x_1, x_2, \dots x_n\}$. $X(t)$ represents the sensor readings at time step t and $A = \{a_1, a_2, \dots, a_{|A|}\}$ be the number of the activities to be predicted.

The values from the multiple sensors can be aggregated as time series window, as shown in Eq. (1):

$$X = \begin{bmatrix} x_1(1) & \cdots & x_1(t) \\ \vdots & \ddots & \vdots \\ x_n(1) & \cdots & x_n(t) \end{bmatrix} = [X(1), ., X(i), ., X(t)] \quad (1)$$

where, $X(i) = [x_1(i), \dots, x_n(i)]$ is the sensor input at time i for n number of sensor readings.

B. Phase 2: Preprocessing

The samples in the time segment are passed through filters for missing values, noise and outlier removal. After preprocessing, a set of segments W is produced that corresponds to activity A, as shown in Eq. (2).

$$W = \{ws_1, ws_2, \dots, ws_m\} \quad (2)$$

For each segment $ws_i = (t_1, t_2)$ represents a portion of samples from t_1 to t_2.

C. Phase 3: Model Building

The model building consists of the following stages:

1) Train a shallow CNN: The aim of this stage is to train a shallow CNN inference model ρ, A = ρ(X), such that the difference between the actual and predicted values is minimized.

- Define Input Layer: The input layer consists of time series signals. The input at time step t denoted as $X(t)$, and the input to the first convolution layer denoted as $l_o(t)$ is calculated using Eq. (3):

$$l_o(t) = X(t) = X^0 \quad (3)$$

- Convolution Layer: It consists of a series of L layers, $L = \{l_1, l_2, \dots l_{|L|}\}$, each with n number of neurons. The output of ith neuron at layer l is calculated using Eq. (4):

$$x_i^l = h\left(\sum_{j=0}^{N} (w_{ij}^l . x_j^{l-1} + b_i^l\right) \quad (4)$$

where, w_{ij}^l is the weight between the neuron j at layer $l-1$ and neuron i at layer l, b_i^l is the bias of neuron i at layer l, $h(\cdot)$ indicates the ReLU activation function, which is calculated using $\max\{0, \sum_{j=0}^{n}(w_{ij}^l . x_j^{l-1} + b_i^l)\}$.

- Batch Normalization: To converge the CNN network easily the batch normalization is performed using Eq. (5):

$$BN(x'^l_i) = \frac{\gamma(x_i^l - \mu_\beta)}{\sigma_\beta} + \beta \quad (5)$$

where, μ_β is mean and σ_β is the standard deviation of the current batch. The resultant values are scaled by γ and shifted by β. The γ and β are learned throughout training.

- Pooling Layer: The convolution layer is followed by pooling layer to reduce the size and parameters of the network. It is performed using Eq. (6):

$$Maxpool_i^l(t) = max_{r \in R}(x'^l_i(t * T + R)) \quad (6)$$

- Softmax classification layer: After pooling layer, the input passes through fully connected layer, and finally softmax layer for performing classification. The computed activity scores for each achievement are expressed as $P(^{a_i}/_{x_i, \theta})$, computed using inference method ρ, expressed as

$$P\left(\frac{a_i}{x_i}, \theta\right) = \rho(x_i, \theta), for\ a \in A$$

where, θ is trained parameters of model ρ.The maximum score will then be used to calculate the predicted activity a_i for the segment ws_i.

$$a_i^* = argmax\ P\left(\frac{a_i}{x_i}, \theta\right)$$

- The training is performed through backpropagation and the gradients are computed as:

$$\frac{\partial E}{\partial W^l} = \left(\frac{\partial E}{\partial X^l}\right)^T . X^{l-1} \quad (7)$$

$$\frac{\partial E}{\partial X^{l-1}} = (W^l)^T . \frac{\partial E}{\partial X^l} \quad (8)$$

where, Eq. (7) represents error with respect to weights at layer l and Eq. (8) represents error w.r.t the output of the previous layer. Once the gradients are computed, then weights are updated according to Eq. (10):

$$W^l = W^l - \alpha(\frac{\partial E}{\partial W^l}) \quad (9)$$

where, α is the learning rate of the optimization algorithm. This process is repeated for each training example in the dataset until the error is minimized to a satisfactory level.

2) Retrain using CnQ algorithm: To minimize transfer loss from CNN to spiking neural network, the designed CNN in step 5 is retrained using the CnQ algorithm, a variation of CQ [18]. For the CnQ, the input is first normalized using Eq. (10):

$$X_norm = \frac{X}{X_{max} - X_{mean}} \quad (10)$$

Then quantized for spike train of length T, using Eq. (11):

$$Q_T(X_norm(t)) = \left\lfloor \frac{X_norm(t).T}{T} \right\rfloor \quad (11)$$

Since, batch normalization cannot be applied to the input spike trains, therefore batch normalization is applied to weights w^l and bias b^l using Eq. (12) and Eq. (13):

$$BN(w_i^l) = \frac{\gamma . w_i^l}{\sigma_\beta} + \beta \tag{12}$$

$$BN(b_i^l) = \frac{\gamma (b_i^l - \mu_i^l)}{\sigma_\beta} + \beta \tag{13}$$

Now the input data is transformed in spike trains with discrete values set $T \in \{0,1\}$. Instead of ReLU activation function, the CQ transform is applied. Clamping is first applied to the activation function to restrict them to [0,1] and is defined as C in Eq. (14):

$$C(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases} \tag{14}$$

Then Quantization is applied as in Eq. (15):

$$Q_T(x) = \left\lfloor \frac{x.T}{T} \right\rfloor \tag{15}$$

In CnQ, the forward propagation is modified as Eq. (4). The learning is performed through conventional backpropagation:

$$\frac{\partial E}{\partial W^l} = Q'.C'.\left(\frac{\partial E}{\partial X^l}\right)^T . X^{l-1} \tag{16}$$

$$\frac{\partial E}{\partial X^{l-1}} = Q'.C'.(W^l)^T . \frac{\partial E}{\partial X^l} \tag{17}$$

where, Eq. (16) shows error w.r.t weights at layer l and Eq. (17) represents error w.r.t the output of the previous layer. Once the gradients are computed, then weights are updated according to Eq. (9). Algorithm 1 explains CnQ training details.

Algorithm 1: CnQ Training

Input: Sensor data (X), Activity Label (A)

Output: Trained Layers $\langle l_1, \dots, l_{|L|} \rangle$

Parameters Initialized: Number of layers (L), Weights (W), Bias(B), convergence==false

```
Normalize X    //using eq. 10
Quantize X    //using eq. 11
while(!convergence==true)
    For i=1 to L
        For j=1 to N    //forward propagation
            Calculate x_j^i    //using eq. 4
            Apply BN    //using eq. 12 and 13
            Apply Clamping on x_j^i    //using eq. 14
            Apply Quantization on x_j^i //using eq. 15.
        End for
    End for
    For i = L to 1    //Backpropagation
        Calculate Gradients    //using eq. 16 and 17
        Adjust weights    //using eq. 9
    End for
    Update convergence
End while
Return Trained layers ⟨l_1, ..., l_|L|⟩
```

3) The SN Deployment: While CNN's transmit activation data among layers (x^l) as real numbers. SNN neurons receive series of binary input as spike trains. Formally, s_i^l is a spike train of neuron i at layer l where $s_i^l(t) \in \{0,1\}$ is a spike at time t. Unlike CNNs, the weight sum is merely a summation since the weights are binary.

The IF model is used to map CNN to SNN activation. At each time step t, a neuron i computes the weighted sum of the received input spikes $s_j^{l-1}(t)$ and the corresponding weights and integrates them as the membrane potential V_i. Whenever V_i exceeds a predefined threshold V_{th_i}, the neuron fires a spike '1' and decreases V_i by V_{th_i}. Otherwise, it outputs '0'. Formally,

$$v_i^l(t) = v_i^l(t-1) + (\sum_{j=0}^{n} (w_{ij}^l . s_j^{l-1}(t) + b_i^l)$$

$$where, s_i^l(t) = \begin{cases} 1, & v_i^l(t) \ge v_{th_i}^l \\ 0, & otherwise \end{cases}$$

$$v_i^l(t) = \begin{cases} v_i^l(t) - v_{th_i}^l, & v_i^l(t) > 0 \\ v_i^l(t), & otherwise \end{cases}$$

The firing threshold is set using the TF algorithm, inspired from threshold firing approach [19]. Given the desired inference latency t for SNN, TF records the maximum accumulated activation value for each neuron across timesteps at each layer l, as in Eq. (18):

$$thres_i^l = \begin{cases} \max(thres_i^l, \max\left(\sum_{j=1}^{N} w_{ij}^l x_j^l\right), & l = 1 \\ \max(thres_i^l, \max\left(\sum_{j=1}^{N} w_{ij}^l s_j^l(t)\right), & 1 < l < L \end{cases} \tag{18}$$

After the assignment of the firing threshold for a layer, it freezes the thresholds of the layer and repeats the threshold determination for the next layer. Algorithm 2 explains the working of TF.

Algorithm 2: TF setting

Input: Timesteps for inference (T)

Output: Firing threshold for each layer $\left[\{v_{th_i}^l\}_{i=1}^N \right]_{l=1}^L$

```
For l = 1 to L
    For i = 1 to N
        Set thres_i^l = 0
    End for
    If (l === 1)
        For i = 1 to N
            Calculate thres_i^l    //using eq. 18
            Set v_th_i^l = thres_i^l
```

```
        End for
    End if
    Else if (l == L)
        Set v_{th_i}^l = ∞
    End if
    else
        For t = 1 to T
            For i = 1 to N
                Calculate thres_i^l   //using eq. 18
                Set v_{th_i}^l = thres_i^l
            End for
        End for
    End if
End for
Return v_{th_i}^l list
```

D. Phase 4:Evaluation

- Evaluation metrics: After training network, the performance can be tested on the various metrics on test data and can fine-tuned as necessary as discussed in [20] accuracy ($\frac{tp+tn}{tp+tn+fp+fn}$), precision ($\frac{tp}{tp+fp}$), recall ($\frac{tp}{tp+fn}$), f1-score ($\frac{2*(precision*recall)}{precision+recall}$), specificity ($\frac{tn}{tn+fp}$), balanced accuracy ($\frac{1}{2}(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}$), where tp is true positive, fp is false positive, tn is true negative. fn is false negative.
- Evaluate number of parameters: To estimate the number of parameters in a spiking CNN model, using Eq. (19):

$$np = ((ic * fs * nf + bias) * os\ ) \quad (19)$$

where, ic represents the number of channels in the input spike trains, fs represents the size of the filters, nf represents the number of filters in the convolutional layer, bias represents the bias term for each filter, and os represents the size of the output spike trains.

- Calculate memory usage: The memory utilized by a spiking CNN model can be estimated by multiplying the number of parameters by the number of bytes used to store each parameter. For example, if we assume that each parameter is stored as a 32-bit float, then the memory utilized can be calculated, as in Eq. (20), whereas niCNN uses 8 bit clamped and quantized signal; therefore each parameter is 8 -bit.

$$memory\ utilized\ (in\ bytes) = np * 4 \quad (20)$$

- Power Consumption: As per Intel specification [21], Intel core i7 processor consumes 65 watts, whereas Intel Loihi 2 consumes around 1 watt. To calculate the power consumed in execution, it is done using Eq. (21):

$$power\ consumption = execution\ time * watts\ of\ processor \quad (21)$$

IV. Experiment

This section elucidates the experimentation conducted on two widely used public datasets, namely WISDM and mHealth, to evaluate the efficacy of niCNN. All experiments were performed on a single system that was equipped with an i7-5500U processor, 16 GB of RAM, and a 1 TB hard disk drive. The implementation of niCNN was executed utilizing the dynamic and powerful programming language Python, alongside auxiliary packages such as Nengo, NengoDL, and NNI [22]. The proposed network architecture was constructed utilizing Keras, a Python-based high-level neural networks API that can run on top of TensorFlow.

A. Dataset Description

WISDM and mHealth datasets contain time series data for physical activity that were collected using wearable sensor devices. The WISDM includes data for six physical activities that were performed by 29 subjects at a sampling rate of 20 Hz, while the mHealth dataset includes data for 12 physical activities performed by 10 subjects at a sampling rate of 50 Hz. Table II provides an overview of each dataset's specifications, and Table III describes captured activity data.

TABLE II. Dataset Specifications

Dataset	Subjects	Sensors	Sampling Rate	Activities	Samples
WISDM	36	A	20 Hz	6	1,098,209
mHealth	10	A,G, M, ECG	50 Hz	12	1,215,745

[a.] A= Accelerometer, G= Gyroscope, M= Magnetometer, ECG= Electrocardiography

TABLE III. Activity Data

Dataset	Activities (Labels, Samples, % Distribution)
WISDM	Walk(1,424400,38.6), Jog(2,342177,31.2), Up(3.122869,11.20), Downstairs(4,100427,9.1), sit(5,59939,5.5), standing(6,48397,4.4)
mHealth	Standing(1,30720,9), Sitting(2,30720, 9), Lying down(3, 30720,9), Walking(4, 30720,9), Climbing stairs(5, 30720,9), Forward bending(6, 28315,8.3), Frontal Arm elevation(7,29441, 8.6), Knees bending(8, 29337, 8.5), cycling(9, 30720,9), Jogging(10, 30720,9)Runing (11, 30720,9), jumping(12,10342,3)

B. Preprocessing

A total of 1,098,209 and 3,44,116 samples were collected from WISDM and MHealth. To minimize computational effort for on-edge deployment, only segmentation was performed as the preprocessing step. The signal was partitioned into non-overlapping temporal windows of length 2s. After dividing the samples based on raw data without feature extraction, a split of 70:30 was used to create training and testing sets.

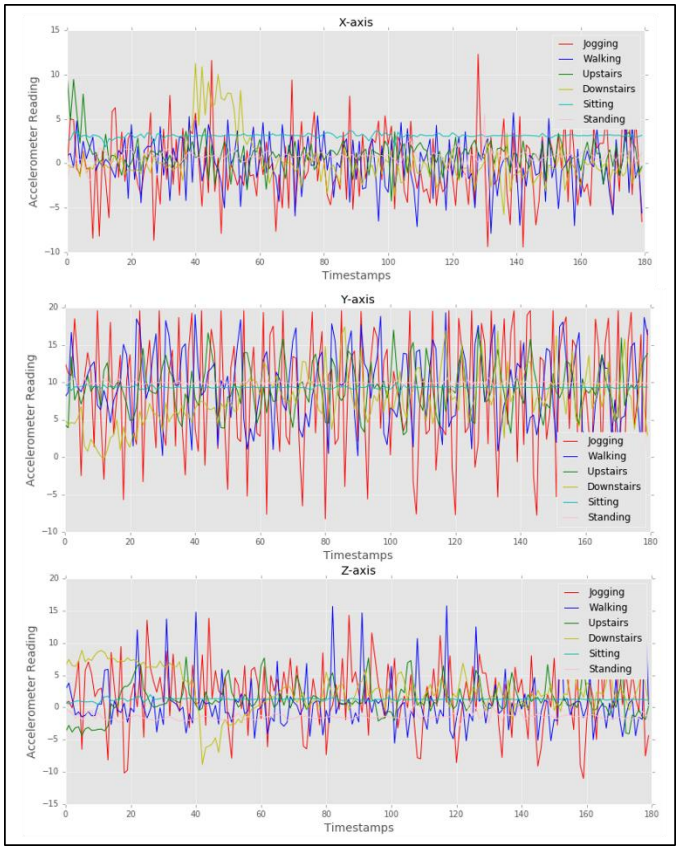


Fig. 2. WISDM activity visualization.

WISDM does not contain outliers, missing values, or null activities; thus, all samples were used for experimentation. Fig. 2 displays a visualization of the WISDM dataset at a timestep of 180. After eliminating null activities from mHealth, the final sample size consists of 343195 samples. At a timestep of 80, a visualization of the mHealth dataset is presented in Fig. 3.

C. Network Architecture

A shallow CNN architecture is developed, consisting of two convolution layers, a max pooling layer, a flattening layer, and two dense layers. The model underwent fully supervised training, with the gradient backpropagated from the Softmax layer.

In the context of the study, cross-entropy loss function is utilized to quantify the dissimilarity between the predicted and actual distributions.

Adam, a stochastic optimization algorithm founded on the first-order gradient, is utilized as the optimizer. The training process is performed over 25 epochs, with a batch size of 64 and a small learning rate of 0.001 to improve the fitting ability. To improve the model's robustness, the training set is randomly shuffled. The hyperparameters are chosen after analyzing the effect of various factors, including the number of filters, batch size, filter size, and dropout rate, on accuracy. Fig. 4 illustrates the impact of these factors on accuracy, and the final hyperparameters are listed in Table IV.

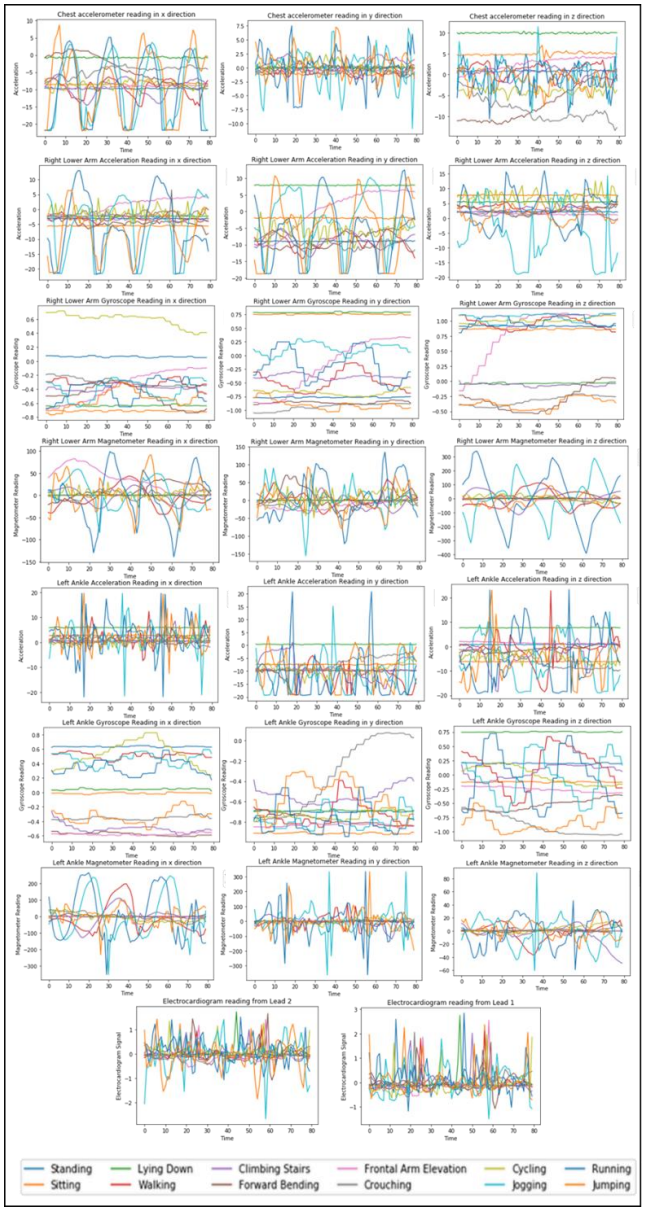


Fig. 3. mHealth activity visualization.

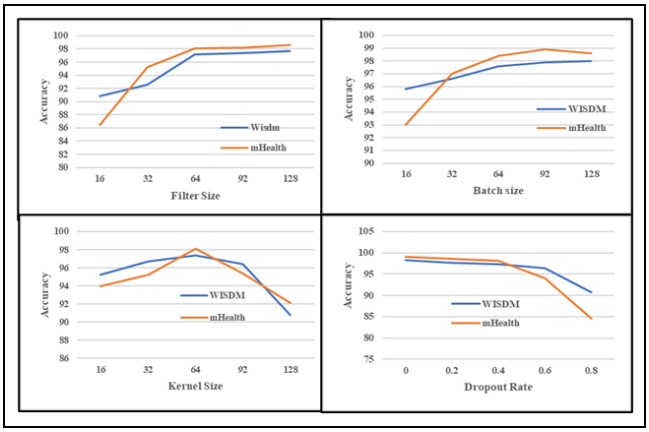


Fig. 4. Impact of hyperparameters on accuracy.

TABLE IV. SELECTED HYPERPARAMETERS

Stage	Hyperparameters	Selected Value
Preprocessing	Window_size	180(WISDM) 80(mHealth)
	Step_size	100
Architecture	Conv1_kernel_size	3
	Conv1_filter_size	64
	Maxpooling_size	2
	Conv2_kernel_size	3
	Conv2_filter_size	64
	Dense layer	128
	Dropout	0.5
Training	Optimizer	Adam
	Batch_Size	64
	Learning rate	.001
	Number of epochs	25

The niCNN is developed using the Nengo neural simulator on the Loihi neuromorphic chip. To construct the spiking CNN directly from its non-spiking counterpart, the NengoDL converter was used. Hyperparameter tuning is conducted using the Neural Network Intelligence (NNI) toolkit and Annealing algorithm, while ensuring a proper search space. Each optimization experiment comprised 1,000 trials, with the tuner being randomly re-initialized four times in equal intervals to avoid local minima. Following every trial of 25 training epochs, the weights yielding the best training accuracy are employed to evaluate the test accuracy. To train all investigated networks, including optimization of the learning rate throughout the experiment trials, Adam optimizer with a constant learning rate is utilized. Fig. 5 and Fig. 6 represent training progress of WISDM and mHealth over epochs.

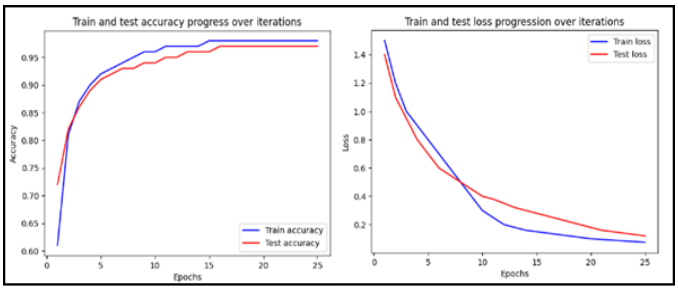


Fig. 5. WISDM dataset progress over epochs.

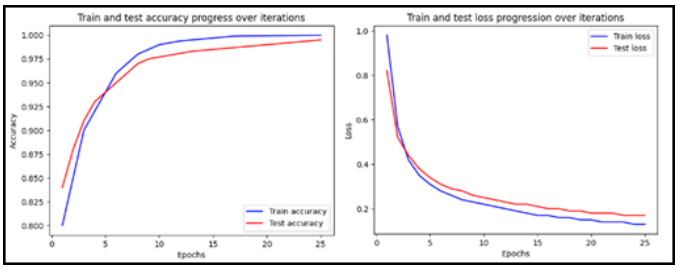


Fig. 6. mHealth dataset progress over epochs.

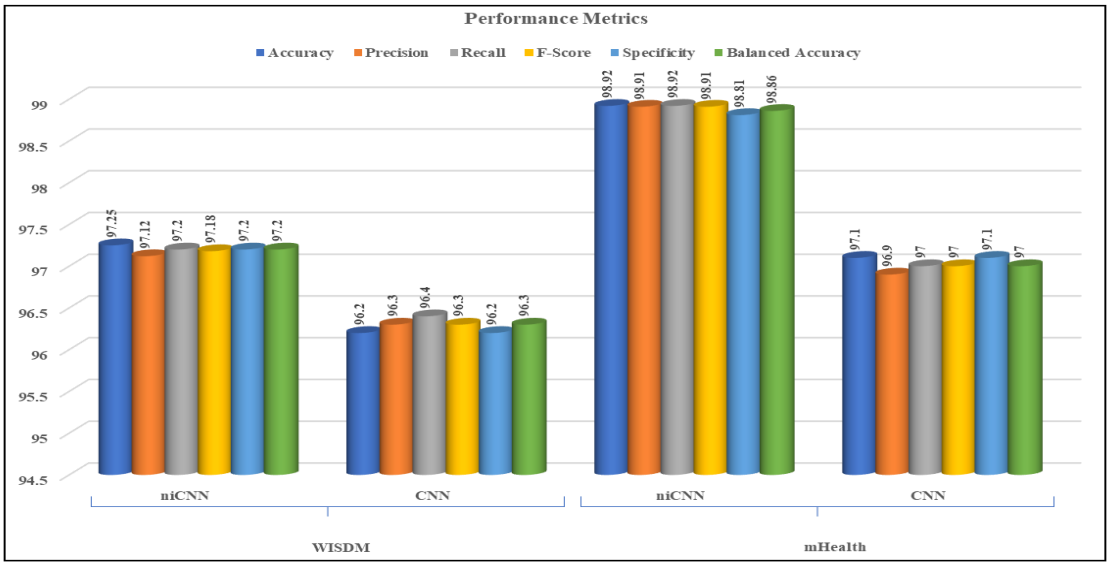


Fig. 7. Performance comparison.

V. RESULTS

The evaluation of performance was carried out using the metrics described in Section III, and the results are presented in Table V. niCNN achieved 98.08% average accuracy, 97.25% on WISDM dataset and 98.92% on mHealth. Fig. 7 provides a comparison of niCNN on both datasets against its baseline CNN on the same architecture. Additionally, the obtained confusion matrix for the WISDM and mHealth datasets are shown in Fig. 8 and Fig. 9, respectively. As per results, presented in Fig. 7 and Table V, niCNN obtained on average 1.1% increase in accuracy, 97.4% reduced power consumption, 60.04% reduced memory consumption, and 30.2% reduced inference time.

niCNN achieved a notably higher accuracy compared to other state-of-the-art works([10], [23]–[33]), as demonstrated in Fig. 10.

TABLE V. PERFORMANCE METRICS

	WISDM		mHealth	
	niCNN	***CNN***	***niCNN***	***CNN***
#P	22,642	14,326	32,612	20,286
#ET(s)	5.2	3.2	5.8	4.1
#M(kBytes)	22.11	55.96	31.84	79.24
#PC(watts)	5.2	208	5.8	266.5
#IT(ms)	2.25	3.27	2.36	4.02

b. #P=Number of parameters, #ET= Execution Time, #M=Memory used, #PC=Power Consumption, #IT=Inference Time

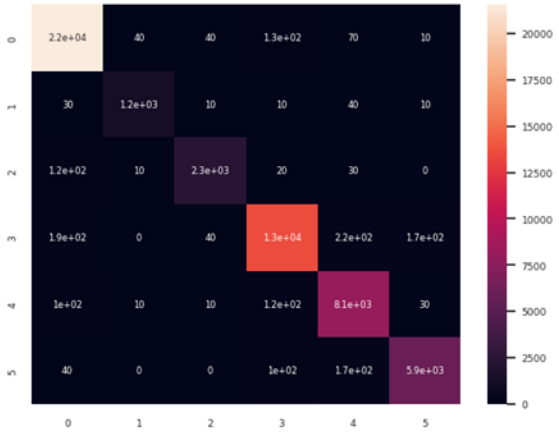


Fig. 8. WISDM confusion matrix.

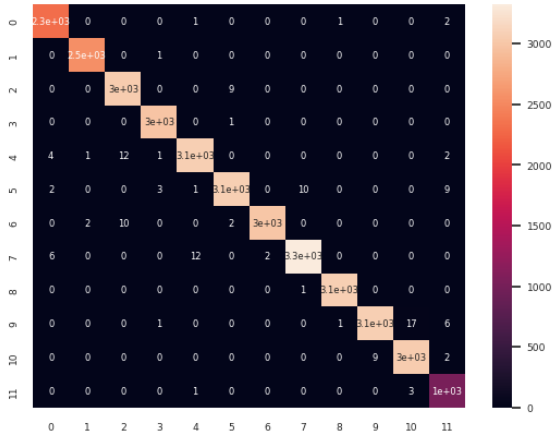


Fig. 9. mHealth confusion matrix.

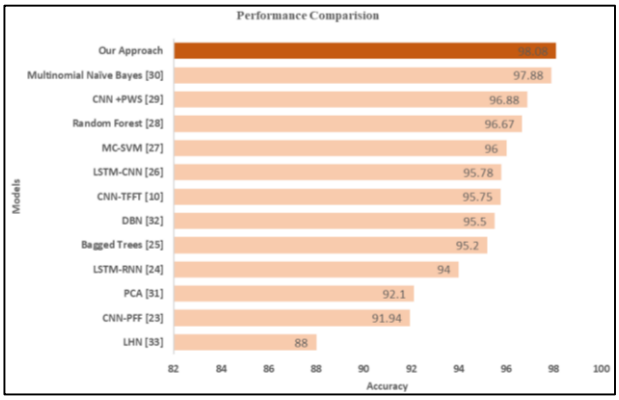


Fig. 10. Comparison of performance accuracy.

VI. CONCLUSIONS

The study demonstrates the capability of neuromorphic computing to perform real-time HAR on edge devices with limited resources. It presents a novel four stage niCNN architecture, overcoming the accuracy loss and higher inference latency associated with transferring weights from a CNN to a neuromorphic network by applying CnQ algorithm. The new improved threshold balancing approach, TF, further increases the inference accuracy of the system. The experimental results demonstrate that the niCNN architecture achieves high accuracy, low inference latency, memory usage, and energy consumption compared to the state-of-the-art and baseline CNN models.

In future, niCNN performance can be investigated on different HAR datasets. The CnQ and TF algorithms can also be further optimized to achieve better results. Further research can expand the model's capability to recognize more complex activities. The effectiveness of various optimization methods within the Nengo framework, including different windowing techniques can also be explored. Additionally, the niCNN architecture can be extended to other real-time processing applications on edge devices.

REFERENCES

[1] O. D. Lara and Miguel A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012, doi: 10.1109/SURV.2012.110112.00192.

[2] E. Ramanujam, Thinagaran Perumal, and S. Padmavati, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sens J*, vol. 21, no. 12, pp. 13029–13040, 2021.

[3] P. Agarwal and M. Alam, "Edge optimized and personalized lifelogging framework using ensembled metaheuristic algorithms," *Computers and Electrical Engineering*, vol. 100, p. 107884, 2022, doi: 10.1016/j.compeleceng.2022.107884.

[4] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput Surv*, vol. 54, no. 4, pp. 1–40, Jul. 2021, doi: 10.1145/3447744.

[5] P. Agarwal and M. Alam, "A Lightweight Deep Learning Model for Human Activity Recognition on Edge Devices," *Procedia Comput Sci*, vol. 167, pp. 2364–2373, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.289.

[6] D. Wu, X. Yi, and X. Huang, "A Little Energy Goes a Long Way: Build an Energy-Efficient, Accurate Spiking Neural Network From Convolutional Neural Network," *Front Neurosci*, vol. 16, p. 759900, 2022, doi: 10.3389/FNINS.2022.759900.

[7] J. Kwapisz, G. Weiss, and S. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, Mar. 2011, doi: 10.1145/1964897.1964918.

[8] O. Banos, R. Garcia, and A. Saez, "UCI Machine Learning Repository: MHEALTH Dataset Data Set." https://archive.ics.uci.edu/ml/datasets/MHEALTH+Dataset.

[9] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput Surv*, vol. 46, no. 3, 2014, doi: 10.1145/2499621.

[10] V. Fra, E. Forno, R. Pignari, T. C. Stewart, E. Macii, and G. Urgese, "Human activity recognition: suitability of a neuromorphic approach for on-edge AIoT applications," *Neuromorphic Computing and Engineering* , vol. 2, no. 1, p. 014006, Feb. 2022, doi: 10.1088/2634-4386/AC4C38.

[11] B. R. Pradhan, Y. Bethi, S. Narayanan, A. Chakraborty, and C. S. Thakur, "N-HAR: A neuromorphic event-based human activity recognition system using memory surfaces," in *IEEE International Symposium on Circuits and Systems*, IEEE, 2019. doi: 10.1109/ISCAS.2019.8702581.

[12] Z. Yu *et al.*, "An Intelligent Implementation of Multi-Sensing Data Fusion With Neuromorphic Computing for Human Activity Recognition," *IEEE Internet Things J*, vol. 10, no. 2, pp. 1124–1133, Jan. 2023, doi: 10.1109/JIOT.2022.3204581.

[13] Z. Yu, A. Zahid, W. Taylor, H. Heidari, M. A. Imran, and Q. H. Abbasi, "Multi-Sensing Data Fusion for Human Activity Recognition based on Neuromorphic Computing," in *2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium*, IEEE, 2021, pp. 64–65. doi: 10.23919/USNC-URSI51813.2021.9703558.

[14] Z. Yu *et al.*, "IMU Sensing–Based Hopfield Neuromorphic Computing for Human Activity Recognition," *Frontiers in Communications and Networks*, vol. 2, p. 68, Jan. 2022, doi: 10.3389/FRCMN.2021.820248.

[15] R. J. Sethi, A. K. Roy-Chowdhury, and S. Ali, "Activity recognition by integrating the physics of motion with a neuromorphic model of perception," in *2009 Workshop on Motion and Video Computing, WMVC '09*, 2009. doi: 10.1109/WMVC.2009.5399241.

[16] Z. Yu *et al.*, "Hardware-Based Hopfield Neuromorphic Computing for Fall Detection," *Sensors 2020, Vol. 20, Page 7226*, vol. 20, no. 24, p. 7226, Dec. 2020, doi: 10.3390/S20247226.

[17] Y. Li, R. Yin, H. Park, Y. Kim, and P. Panda, "Wearable-based Human Activity Recognition with Spatio-Temporal Spiking Neural Networks," *https://arxiv.org/abs/2212.02233*, Nov. 2022, Accessed: Apr. 24, 2023. [Online]. Available: https://arxiv.org/abs/2212.02233v1

[18] Z. Yan, J. Zhou, and W. F. Wong, "Near Lossless Transfer Learning for Spiking Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, May 2021, pp. 10577–10584. doi: 10.1609/AAAI.V35I12.17265.

[19] H. C. V. Ngu and K. M. Lee, "Effective Conversion of a Convolutional Neural Network into a Spiking Neural Network for Image Recognition Tasks," *Applied Sciences 2022, Vol. 12, Page 5749*, vol. 12, no. 11, p. 5749, Jun. 2022, doi: 10.3390/APP12115749.

[20] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[21] I. Corporation, "Taking Neuromorphic Computing with Loihi 2 to the Next Level Technology Brief."

[22] "Nengo." https://www.nengo.ai/ (last accessed Apr. 24, 2023).

[23] S. Ha and C. Seungjin, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *International Joint Conference on Neural Networks (IJCNN),IEEE.*, 2016, pp. 381–388. doi: https://doi.org/10.1109/IJCNN.2016.7727224.

[24] S. W. Pienaar and R. Malekian, "Human Activity Recognition Using LSTM-RNN Deep Neural Network Architecture," in *IEEE 2nd wireless africa conference*, 2019, pp. 1–5. doi: https://doi.org/10.1109/AFRICA.2019.8843403.

[25] S. Khatun and B. I. Morshed, "Fully-Automated Human Activity Recognition with Transition Awareness from Wearable Sensor Data for mHealth," in *IEEE International Conference on Electro Information Technology*, 2018, pp. 934–938. doi: 10.1109/EIT.2018.8500135.

[26] K. Xia, J. Huang, and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020, doi: 10.1109/ACCESS.2020.2982225.

[27] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Belguim, Apr. 2013.

[28] G. Chetty, M. White, and F. Akther, "Smart phone based data mining for human activity recognition," in *Procedia Computer Science*, 2015, pp. 1181–1187. doi: 10.1016/j.procs.2015.01.031.

[29] M. Zeng *et al.*, "Convolutional Neural Networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services, IEEE.*, 2014, pp. 197–205. doi: https://doi.org/10.4108/icst.mobicase.2014.257786.

[30] L. Syed, S. Jabeen, M. S., and A. Alsaeedi, "Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques," *Future Generation Computer Systems*, vol. 101, pp. 136–151, 2019, doi: 10.1016/j.future.2019.06.004.

[31] A. S. Abdull Sukor, A. Zakaria, and N. Abdul Rahim, "Activity recognition using accelerometer sensor and machine learning classifiers," in *2018 IEEE 14th International Colloquium on Signal Processing and its Application, CSPA 2018*, IEEE, May 2018, pp. 233–238. doi: 10.1109/CSPA.2018.8368718.

[32] H. Li and M. Trocan, "Deep learning of smartphone sensor data for personal health assistance," *Microelectronics J*, vol. 88, pp. 164–172, Jun. 2019, doi: 10.1016/J.MEJO.2018.01.015.

[33] A. Jordao, R. Kloss, and W. R. Schwartz, "Latent HyperNet: Exploring the Layers of Convolutional Neural Networks," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, Oct. 2018, doi: 10.1109/IJCNN.2018.8489506.

Improved YOLOv8 Model for Enhanced Small-Sized Breast Mass Detection on Magnetic Resonance Imaging

Feiyan Wu[1], Chia Yean Lim[2]*, Sau Loong Ang[3], Jiaxin Zheng[4]
School of Computer Sciences, Universiti Sains Malaysia, 11800, Minden, Penang, Malaysia[1, 2, 4]
Tunku Abdul Rahman University of Management and Technology, Penang Branch, Tanjung Bungah, Malaysia[3]

***Abstract*—The early detection of breast cancer is critically important for prompt treatment and rescuing lives. However, the accuracy of small-sized breast masses' early detection in various algorithms remains unsatisfactory, as the small-sized masses often exhibit subtle features, contain blurry boundaries, and may overlap with other parts in crowded magnetic resonance imaging (MRI) images. This research proposes an improved object detection model based on You Only Look Once (YOLO) v8 to enhance small-sized breast mass detection on MRI. A feature fusion method called the Bidirectional Feature Pyramid Network (BiFPN) and an attention mechanism called the Convolutional Block Attention Module (CBAM) are integrated into the YOLOv8 architecture. The improved YOLOv8 model, equipped with CBAM and BiFPN and hyperparameter tuning, achieved the best performance with a precision of 95.7%, a mAP50 of 91.2%, a recall of 84.3%, and the shortest inference time of 3.4ms per image. The proposed improved Yolov8 model outperformed the baseline model with improvements in precision, mAP50, and recall of 6%, 3.9%, and 2.1%, respectively. The inference time per image is reduced by 1.4ms as well. It is hoped that the proposed model could be applied in the clinical field to increase the early detection rate of breast cancer and the life expectancy of women in the world.**

***Keywords*—Bidirectional feature pyramid network; breast cancer detection; convolutional block attention module; MRI; object detection; small-sized masses; YOLOv8**

I. Introduction

Breast cancer was the top female disease in 157 of 185 countries and accounted for approximately 670,000 deaths in 2022 worldwide [1], according to the World Health Organisation. However, it was also the most common subject of diagnosis. It was even identified as the first leading cause of death from cancer among women in 2022, according to the Global Cancer Statistics [2]. There are several imaging methods to screen breast cancer, including ultrasonography, mammography, magnetic resonance imaging (MRI), and digital breast tomosynthesis (DBT). Among these methods, mammography is the preferred population-based screening method [3]. However, mammography is less effective for screening dense breasts, as it risks missed diagnoses because small masses that hide in dense breast tissue cannot be detected [4][5]. The sensitivity of breast mass detection increases from 0 to 8% when the tumour size increases from 2 to 20mm [6]. On the other hand, the performance of breast cancer detection in medical imaging is also affected by the increase in breast tissue density [7].

Recently, breast MRI was used complementarily together with mammography and has been researched as a crucial screening tool for women with dense breasts [8]. Currently, the accuracy of breast cancer detection and diagnosis still highly depends on the variations of experts and experienced physicians. Moreover, the existing research on breast masses has primarily focused on overall object detection rather than improving the accuracy of detecting small-sized breast masses [9][10]. In [11] and [12], the authors commented that the performance of small-sized breast mass detection had not achieved the expected results, even though significant improvements had been made in the breast cancer detection models' feature extraction and target detection algorithms.

This research proposes an improved You Only Look Once (YOLO) v8 model for enhanced detection of small-sized breast masses on MRI, aiming to improve the detection precision and mAP of these small masses. The improved model is expected to perform better in detecting small-sized breast masses by adopting an appropriate attention mechanism. Improving the feature fusion method would enhance the fusion effect of the multi-scale features in the improved model. Finally, hyperparameter tuning is conducted to strengthen the performance of the detection model.

The rest of the study is organised as follows: Section II reviews various object detection models in breast cancer detection, feature fusion methods, and attention mechanisms. Section III presents the methodology of the proposed improved YOLOv8 model. Section IV discusses the experimental results and analysis. Finally, Section V concludes the study and outlines future research directions.

II. Literature Review

A. Breast Cancer Detection and Imaging Types

In medical imaging, the most common screening techniques to diagnose breast cancer are breast ultrasound, mammography, breast magnetic resonance imaging (MRI) and digital breast tomosynthesis (DBT). Although mammography is the only imaging test that has been demonstrated to decrease deaths from breast cancer, breast MRI has taken its superior role in the screening of high-risk populations because of the high sensitivity of the images [13]. Two advantages of using breast

MRI over mammography are: i) the breast MRI can detect smaller tumours with a median size of 9 mm with higher sensitivity, in comparison to 17 mm for mammography [14], and ii) the breast MRI uses a different imaging mechanism with magnetic fields and radio waves which is particularly effective in dense breast tissue screening, enabling the excellent performance of dense breast tissue evaluation and additional lesions' detection. Despite the higher cost and longer procedure time in the breast MRI screening process, MRI is still the preferred breast cancer detection method, where mammography's effectiveness is limited in the detection of smaller and additional lesions.

B. Object Detection Model in Breast Cancer Detection

The YOLO model is a real-time object detection system introduced by Redmon, Divvala, Girshick, and Farhadi [15]. Ultralytics created and released YOLOv8 in January 2023 as a new state-of-the-art computer vision model that provides out-of-the-box support for object detection, classification, and segmentation tasks [16]. The YOLO model treats object detection tasks as a regression problem, predicting bounding boxes and class probabilities for objects in a single convolutional network. It runs a convolutional network to process each grid cell, where it divides the input image into a grid of cells. Each grid cell will perform detection if there is an object within it, make predictions for bounding boxes, calculate the confidence scores for those boxes, and determine conditional class probabilities. When making predictions, the YOLO model performs global reasoning across the entire image, allowing it to detect all objects within the image. Additionally, the YOLO design enables end-to-end training and achieves real-time performance while maintaining a high average precision (AP) [15]. The main advantages of the YOLO model are fast speed, high accuracy, and suitability for real-time applications. As a single-stage model, YOLO has been proven to be an excellent tool in various scenarios, achieving higher inference speeds than its competitors in object detection [17].

Table I summarises the comparison of the strengths, weaknesses, and results of different object detection methods in breast mass detection over the past five years. The comparison shows that YOLO [18] [19] and region-based object detection methods such as Faster R-CNN [20], Mask R-CNN [21], and R-FCN [22] were popular. Additionally, RetinaNet was also employed in breast mass detection from 2023 onwards, reflecting a growing emphasis on the real-time capabilities of single-stage models as time progresses [23].

Several recent studies have explored the performance of different object detection models for breast mass detection. Faster R-CNN achieved a high true positive rate, but its inference time was significantly longer than single-stage models like YOLO and RetinaNet [20]. RetinaNet has limitations in detecting small objects due to its fixed-size anchor boxes. Even with the recent improved version ERetinaNet, the inference speed still lags behind YOLOv3 and YOLOv7 [23]. YOLOv3 was enhanced with k-means generated anchors for detecting small and big masses, which outperformed YOLOv2 and YOLOv3, and achieved a mAP50 of 89.4% [18]. The YOLOv5 model was also improved for better breast cancer detection in mammograms. The YOLOv5s model obtained a mAP50 of 83.5%, a precision of 85.4% and a recall of 72.9%, which performed better than YOLOv3 and YOLOv5-Transformer [17]. These results reinforce the advantage of YOLO-based models in balancing speed and accuracy, making them suitable foundations for further improvement in small object detection tasks.

TABLE I. COMPARISON OF DIFFERENT OBJECT DETECTION METHODS IN BREAST MASS DETECTION

Detection Model	Strength	Weaknesses	Results
YOLOv3 [18]	• Fast speed with high accuracy • Suitable for real-time applications	• Performance for small object identification can be further improved.	mAP50: 89.4% Precision: 83.0% Recall: 83.0%
YOLOv5 [19]	• Fast speed with high accuracy • Suitable for real-time applications • Provide improved performance for small object detection	• Higher complexity than YOLOv3 Requires more computational resources.	mAP50: 90.34%
Faster R-CNN [20]	• Provides high accuracy • Suitable for handling various types and sizes of objects	• Slower speed • Requires larger memory consumption • Contains higher computational complexity.	Malignant: TPR: 0.92 at 0.3 FPI. Benign: TPR: 0.85 at 1.0 FPI.
Mask R-CNN [21]	• Provide accurate instance segmentation • Suitable for tasks requiring detailed object region information	• Slower speed • Requires larger memory consumption • Contains higher computational complexity	Accuracy: 98%
R-FCN [22]	• Faster speed in the two-stage approach category • Provides high accuracy	• Contains high computational complexity • Poor performance when objects are extremely small or extremely large compared to the size of the input image	Accuracy: 92.5% Sensitivity: 92.3% Specificity: 93.4%
RetinaNet [23]	• Simple and efficient • Good object detection performance and speed, excellent in handling imbalanced class distribution problems.	• May struggle with small object detection. • Requires more computational resources.	mAP50: 85.01% Precision: 86.87% Recall: 74.62% F1-score: 80.3% Inference time: 23.4ms per image

From the comparison, almost all models faced challenges in small object detection, with RetinaNet, R-FCN, and YOLO showing more pronounced weaknesses in this area. On the other hand, although RetinaNet, as used in [23], had a slightly longer inference time than existing YOLO models, it excelled in addressing class imbalance issues, where positive breast cancer detection cases are significantly fewer than negative cases. In addition, it was observed that all the object detection models, except for YOLO, in Table I have problems with high computational complexity.

This research chose the YOLO model as the object detection method to be improved for small-sized breast mass detection on MRI for several reasons. Firstly, the YOLO model is faster than other models. Secondly, YOLO is suitable for real-time applications, where breast cancer detection can be done almost instantly after MRI scanning is completed, thereby saving diagnostic time, expediting the decision-making process, and facilitating early treatment planning between the doctor and patient. Thirdly, the YOLO model has progressively improved accuracy and mAP values from YOLOv3 to YOLOv5.

In the current state-of-the-art, the YOLOv8 model has demonstrated its advantages in detecting small objects in remote sensing images [24]. It has been used to detect liver disease [25], identify potato diseases [26], and detect pedestrians in real-time [27]. The YOLOv8 model has also been adopted for breast mass detection and classification in [28]. However, the model encounters difficulties in detecting small objects in complex scenes, especially when their appearance is partially occluded by overlapping with other objects of varying sizes, which can lead to missed or false detections. As such, the YOLO model should be enhanced with various techniques to address its weaknesses.

C. Feature Fusion Methods

Feature fusion enhances the accuracy and robustness of an object detection model by leveraging complementary information from various sources and employing an approach that reduces information redundancy. The feature fusion method merges morphological, textural, and density features to create a more comprehensive representation of the object. This comprehensive representation enables the detection system to capture the target's features more effectively, resulting in more accurate and reliable detection results. On the other hand, the feature fusion method also enhances the sensitivity and robustness of the model in small object detection by adaptively combining features from multiple scales and layers, thereby capturing both high-level semantic and low-level spatial information.

Feature fusion methods such as the Asymptotic Feature Pyramid Network (AFPN), Bidirectional Feature Pyramid Network (BiFPN), and the Semantic Feature Pyramid Network (FPN) have been applied to improve the performance of YOLO [29]–[32] and Faster-RCNN [11] models in small object detection tasks. It is observed that the YOLO models [31] [32] achieved a slightly higher mAP50 value (highest 90%) when integrated with BiFPN, in comparison to Semantic FPN (highest 88.22%) [11] and AFPN (highest 88%) [29][30]. With accuracy being a key consideration in model performance, this research aimed to integrate the YOLOv8 model with BiFPN for the detection of small-sized breast masses. Moreover, BiFPN is also more appropriate because it can efficiently fuse multi-scale features and improve detection performance without significantly increasing computational complexity.

D. Attention Mechanisms

An attention mechanism enhances the performance of object detection models by allowing the model to focus on specific regions or features of the image [33]. This mechanism also helps to prioritise fine-grained or key features which are particularly beneficial for detecting small objects in complex visual backgrounds [34]. It is instrumental in challenging scenarios, such as detecting minor defects or tumours in cluttered backgrounds. The attention mechanism can improve the accuracy of breast mass detection by guiding the model's focus to specific regions of interest in medical images [35]. By incorporating the attention mechanism into the detection model, the object detection model can prioritise relevant features of small targets even when they are occluded or surrounded by other objects [36].

The Convolutional Block Attention Module (CBAM) is a popular attention mechanism. CBAM was adopted in the YOLOv7 model for brain tumour detection, and this combined model achieved 98.9% precision, 98.8% recall, and 98.9% accuracy [32]. The CBAM was also incorporated into the backbone network of the YOLOv5 model for breast lump detection, yielding improvements in accuracy, recall, and mAP50 values of 3.1%, 2.6%, and 3.1%, respectively, compared to the baseline YOLOv5 model [37]. As such, this research attempted to integrate CBAM into the YOLOv8 model to increase the precision and mAP value for small-sized breast mass detection.

III. Methodology

This research is conducted in five phases. The primary objective is to propose an enhanced YOLOv8 model for improved small-sized breast mass detection on MRI by integrating CBAM into the backbone structure and BiFPN into the neck structure of a baseline YOLOv8 model. The structure of the improved YOLOv8 Model for the research is shown in Fig. 1.

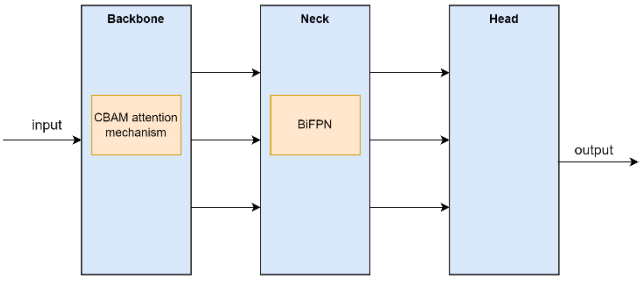


Fig. 1. Structure of the improved YOLOv8 model.

The research began with the data selection and pre-processing phase. This research used the annotated version of the BreastDM dataset [38]. It is the latest Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) dataset for breast cancer diagnosis, spanning from January 2018 to

September 2021, and collected at Taizhou Central Hospital, Taizhou University, in Zhejiang Province, China. The dataset is publicly available on the Roboflow Universe website [39]. There are 1196 images in the selected dataset, with 875 malignant images and 321 benign images. These images are labelled according to the pathology of the mass in 2 classes, with class label 0 representing benign and class label 1 representing malignant.

The selected dataset is first converted from the original annotation in COCO format to YOLOv8 object detection format via the Roboflow platform. Following format preparation, 70% of the dataset is assigned to a training partition, 15% to a validation partition, and the remaining 15% to the test partition. The resolution of all images was resized from 369 × 369 pixels to 640 × 640 pixels, and the proportion of small objects was 37.5%. In addition, the YOLOv8 model will also automatically normalise the pixel values to a range of 0 to 1 during its preprocessing phase.

Several data augmentation techniques are employed to address the data imbalance problem and enhance the robustness and generalisation of the proposed model. In addition, these augmentation techniques introduced variability in the training data, allowing the YOLOv8 model to learn more robust features and improve performance on diverse and noisy datasets. Firstly, the HSV augmentation is utilised with parameters for hue (±1.5%), saturation (±70%), and value (±40%) adjustments. Secondly, spatial augmentations such as translation (±10%), scaling (±50%), and left-right flipping with a 50% probability were conducted too. Thirdly, the Mosaic augmentation is applied with a 100% probability. Finally, random erasing is performed with a 40% probability, and cropping is set to a fraction of 1.0 to maintain the original image size. The proposed YOLOv8 model chose "Randaugment" as the auto augmentation policy.

In the second phase, the baseline YOLOv8 model was created and trained on the selected dataset. The CBAM attention mechanism and BiFPN are integrated separately into the YOLOv8 model in the third and fourth phases. The performance of the baseline YOLOv8 model in the second phase, as well as the improved model in the third and fourth phases, is measured with precision, recall, mAP50, and mAP50-95. Subsequently, in the fifth and final phase, the YOLOv8 model, integrated with CBAM and BiFPN, is constructed, evaluated, and fine-tuned to adapt it to the breast mass detection task and the selected dataset.

The experiments were conducted in Google Colab Pro using T4 GPUs and A100 GPUs. The experiment program is written in Python. All the experiments of the proposed enhanced small-sized breast mass detection model are based on the AdamW optimiser. The experiment specifications are as follows:

- Weight decay: 0.0005
- Learning momentum: 0.9
- Learning rate: 0.001667
- Epochs: 300
- Batch: 32
- Warmup epochs: 3.0
- Warmup momentum: 0.8

IV. Results and Discussion

The evaluation of the proposed improved YOLOv8 model was conducted in several stages. In the first stage, the original YOLOv8 model serves as the baseline. Following this, the YOLOv8 model is run with a single feature (CBAM or BiFPN), two features (CBAM and BiFPN), and two features (CBAM and BiFPN) with Hyperparameter Tuning. The full result is shown in Table II.

Table II shows that the precision, recall, and mAP50 values of the original YOLOv8 model are below 90%. The mAP50-95 achieved 63.2% while the inference time per image is 4.8ms. The goal of this research is to enhance the YOLOv8 model for small-sized breast mass detection in MRI, aiming to improve precision, recall, and mAP50 values while reducing inference time per image to achieve faster detection speed. The evaluation of the improved YOLOv8 model with CBAM showed that the precision, mAP50, and mAP50-95 are increased by 4%, 3.1%, and 1.8%, respectively, with unchanged recall value, and higher inference time per image (+3.2ms). These results demonstrate the effectiveness of the proposed backbone network and its improved detection performance. On the other hand, the evaluation of the enhanced YOLOv8 model with BiFPN showed that the mAP50 and mAP50-95 increased by 0.6% and 0.2%, respectively. Still, the precision and recall values were lower than those of the baseline YOLOv8 model by 0.1% and 0.5%, respectively. However, this improved model required less inference time per image (-0.2ms) than the YOLOv8 baseline model.

TABLE II. The Evaluation Result of the Proposed YOLOv8 Model in Various Stages

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	Inference time per image (ms)
Original YOLOv8 model	89.7	82.2	87.3	63.2	4.8
YOLOv8 model with improved backbone by using CBAM	93.7 (+4.0)	82.2 (0.00)	90.4 (+3.1)	65.0 (+1.8)	7.0 (+3.2)
YOLOv8 model with improved neck by using BiFPN	89.6 (-0.1)	81.7 (-0.5)	87.9 (+0.6)	63.4 (+0.2)	4.6 (-0.2)
Improved YOLOv8 model with CBAM and BiFPN	92.5 (+2.8)	**87.5 (+5.3)**	90.7 (+3.4)	66.8 (+3.6)	5.1 (+0.3)
Improved YOLOv8 model with CBAM and BiFPN (Hyperparameter Tuning)	**95.7 (+6.0)**	84.3 (+2.1)	**91.2 (+3.9)**	**67.0 (+3.8)**	**3.4 (-1.4)**

Comparing the two improved YOLOv8 models with CBAM and BiFPN, CBAM performed better than BiFPN in all evaluation metrics except for inference time per image. Although the enhanced YOLOv8 model with BiFPN exhibits a slight decrease in precision and recall, it improves the mean Average Precision (mAP50 and mAP50-95) and reduces inference time. BiFPN enhanced the YOLOv8 model's ability to recognise masses of different scales by increasing the complexity of feature fusion. However, the BiFPN's complex structure and computational demands make the improved YOLOv8 model training and optimisation more difficult, resulting in a drop in precision and recall metrics. Additionally, BiFPN may require careful parameter tuning for optimal performance. Overall, the improvement in mAP demonstrates its potential in small object detection tasks.

To prove the effectiveness of adding BiFPN into the YOLOv8 model for small mass detection, the second stage of the experiment was conducted by combining CBAM and BiFPN into the YOLOv8 model. The experimental results showed that all evaluating metrics achieved improvements in the range of 2.8% to 5.3%, with precision and mAP50 achieving values higher than 90%. However, the inference time per image is slightly increased by 0.3ms. Therefore, the experiment proceeds to the final stage, whereby the improved YOLOv8 model for enhanced small-sized breast mass detection on MRI is built by fine-tuning the hyperparameters of the improved YOLOv8 with CBAM and BiFPN. This enhanced model achieved the best performance among other improved models, with a precision of 95.7%, a recall of 84.3%, a mAP50 of 91.2%, a mAP50-95 of 67.0%, and the shortest inference time of 3.4 ms per image. The best hyperparameters for the final improved YOLOv8 model obtained after hyperparameter tuning are shown in Table III.

TABLE III. BEST HYPERPARAMETERS OF THE PROPOSED IMPROVED YOLOV8 MODEL

Hyperparameter	Best fitness amount
lr0	0.00338
lrf	0.01684
momentum	0.81006
weight_decay	0.00044
warmup_epochs	2.50590
warmup_momentum	0.79978
box	2.88344
cls	0.51721
dfl	2.12637
hsv_h	0.01001
hsv_s	0.32297
hsv_v	0.52034
translate	0.10270
scale	0.28700
fliplr	0.29108
mosaic	0.44672
mixup	0.45000

According to the comparison of the normalised confusion matrix for the baseline and proposed improved YOLOv8 model, the enhanced YOLOv8 model shows a higher classification accuracy for the malignant class (95%) than the baseline YOLOv8 model (91%). The classification accuracy for the benign class has also increased to 80%, helping to mitigate the negative impact of class imbalance. On the other hand, the issue of misclassifying malignant masses as benign in the original YOLOv8 model has been addressed, resulting in a reduction of the error rate from 2% to 0% in the improved YOLOv8 model. In comparison, the misdetection of benign masses as malignant has decreased from 4% to 2%. Additionally, the misclassification of malignant masses as background has reduced from 7% to 5%.

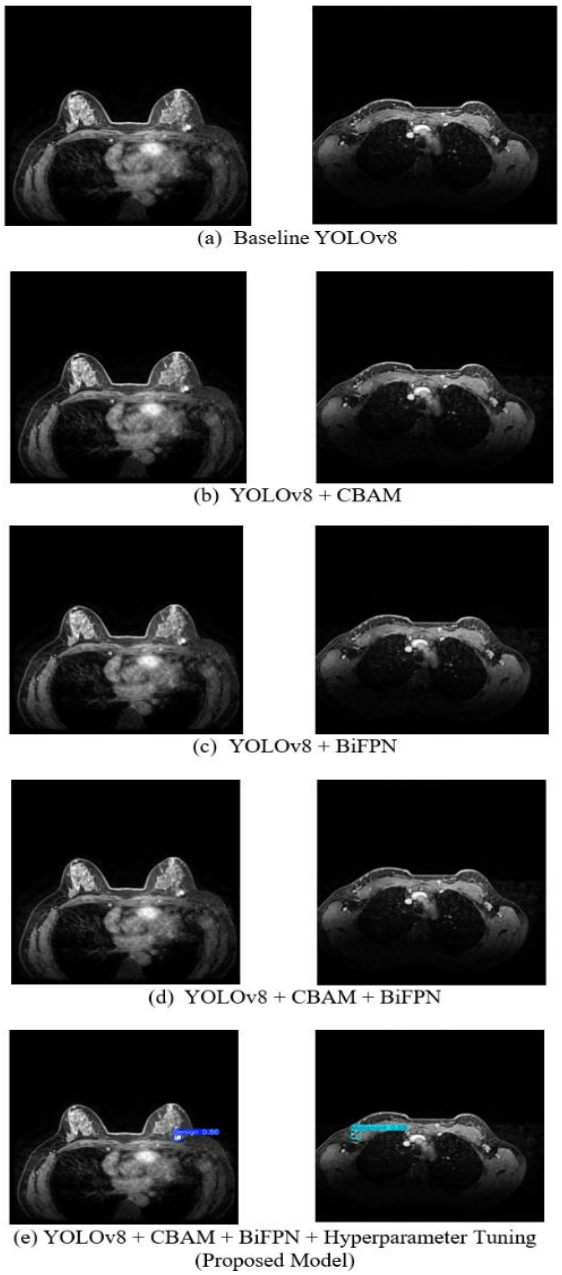


Fig. 2. Result of small mass detection with MRI with different YOLOv8 models.

Fig. 2 shows two detection results of different experimented YOLOv8 models. In the left image, the breast tissue is dense and has similar texture features to the masses, which interferes with the model's accuracy in judging the masses [7]. The right image shows a small breast, where the mass is located at the edge of the breast, close to the pectoral muscle. The pectoral muscle increases the complexity of the MRI image background [40], affecting the model's ability to detect small malignant masses with very subtle features. The two results showcases where benign small masses and malignant small masses in the MRI dataset are not detected by all the YOLOv8 models [Fig. 2(a) to Fig. 2(d)] in the study except for the improved YOLOv8 model with CBAM and BiFPN with hyperparameter tuning [Fig. 2(e)]. This demonstrates the effectiveness of the proposed improved YOLOv8 model.

In terms of evaluation, the proposed improved YOLOv8 model achieved strong overall performance, with a mAP50 of 91.2%, precision of 95.7%, recall of 84.3%, and an inference time of 3.4 ms per image. Compared to existing studies, relatively few have focused specifically on detecting small-sized breast masses using YOLO-based approaches. Among those, a YOLOv5-based model reported a mAP50 of 90.34% and an inference time of 28 ms per image [19], which is lower in accuracy and significantly slower than our proposed model. These results highlight the effectiveness of the enhancement in this study for both detection precision and computational efficiency.

V. Conclusion and Future Work

This research proposes an improved YOLOv8 model designed to enhance the detection performance of small breast masses on MRI. By integrating CBAM and BiFPN into the YOLOv8 architecture, this study has successfully overcome the limitations of the existing YOLOv8 model in detecting small masses. CBAM enhances the model's focus on subtle and essential features, particularly when handling occluded small masses. Meanwhile, BiFPN enhances the fusion of multi-scale features, further improving the model's detection accuracy. With the achievement of a reduction in inference time per image in the improved YOLOv8 model, this research contributes to more effective and timely breast cancer detection, particularly in detecting small masses in MRI images.

In conclusion, this research proposed an improved YOLOv8 model, which significantly enhances the performance in detecting small-sized breast masses on MRI, thereby improving the accuracy of early-stage breast cancer detection. Early detection is critical for breast cancer, as it increases the chances of successful treatment and survival. This research facilitates timely diagnosis and treatment by providing physicians with a more accurate and real-time tool for identifying potential breast masses, ultimately improving patient survival rates. The performance of the improved YOLOv8 model has been validated only on the selected breast mass MRI dataset. However, no experiments are conducted on the proposed model's transferability and generalizability to other medical images.

In the future, further studies could be performed to investigate the model's transfer learning capability across various medical image datasets. This would help evaluate whether the model maintains its effectiveness in different imaging modalities, such as mammography, ultrasound, or CT, and under varying imaging conditions. On the other hand, the improved YOLOv8 model can also be examined for its applicability in detecting small masses for different cancer types, such as those affecting the lungs and brain. Furthermore, the model can be integrated into real-time diagnostic workflows in clinical practice to assist radiologists in detecting breast masses more efficiently and accurately. In addition, applying these enhancements to newer YOLO models may help assess whether they lead to further performance improvements.

References

[1] World Health Organisation. "Breast cancer" [Internet] 2024 [Retrieved 2024 March 28] Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, A. Jemal, "Global cancer statistics 2022: GLOBOCAN Estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J Clin, vol. 74(3), 2024, pp. 229–263.

[3] D. P. Budh and A. Sapra, "Breast cancer screening in the average-risk patient." [Internet]. StatPearls StatPearls Publishing, 2024 [Retrieved 2024 March 28] Available at: https://www.ncbi.nlm.nih.gov/books/NBK556050/

[4] J. Kollmer, "Perfecting MRI as a breast cancer screening tool," [Internet] 2022 [Retrieved 2024 March 28] Available at: https://www.uchicagomedicine.org/forefront/cancer-articles/perfecting-mri-as-a-breast-cancer-screening-tool

[5] R. M. Mann, N. Cho and L. Moy, "Breast MRI: State of the art." Radiology, vol. 292(3), 2019, pp. 520–536.

[6] J. Wang, P. Gottschal, L. Ding, DAV. Veldhuizen, W. Lu, N. Houssami, MJW. Greuter and GH. de Bock, "Mammographic sensitivity as a function of tumor size: A novel estimation based on population-based screening data." Breast, vol. Feb 55, 2021, pp.69-74.

[7] I. V. Gruber, M. Rueckert, K.O. Kagan, A. Staebler, K.C. Siegmann, A. Hartkopf, D. Wallwiener and M. Hahn, "Measurement of tumour size with mammography, sonography and magnetic resonance imaging as compared to histological tumour size in primary breast cancer." BMC Cancer, vol. 13, 2013, pp.1-8.

[8] E. Verburg, G.C. Van Gils, M.F. Bakker, M.A. Viergever, R.M. Pijnappel, W.B. Veldhuis and K.G.A. Gilhuijs, "Computer-aided diagnosis in multiparametric magnetic resonance imaging screening of women with extremely dense breasts to reduce false-positive diagnoses." Invest Radiol, vol. 55(7), 2020, pp. 438–444.

[9] M.M. Rahman, M.Z.B. Jahangir, A. Rahman, M. Akter, M.A. Nasim, K.D. Gupta and R. George, "Breast cancer detection and localizing the mass area using deep learning." Big Data Cogn Comput, vol. 8(7), 2024, pp. 80.

[10] S. Saikia, T. Si, D. Deb, K. Bora, S. Mallik, U. Maulik, U. and Z. Zhao. "Lesion detection in women breast's dynamic contrast-enhanced magnetic resonance imaging using deep learning." Sci Rep, vol. 13(1), 2023, pp. 22555.

[11] J. He, J. Wang, Z. Han, B. Li, M. Lv and Y. Shi, "Cancer detection for small-size and ambiguous tumors based on semantic FPN and transformer." PLoS One, vol. 18(2), 2023, pp. e0275194.

[12] N.M. Hassan, S. Hamad, S., and K. Mahar, "Mammogram breast cancer CAD systems for mass detection and classification: a review." Multimed Tools Appl, vol. 81(14), 2022, pp. 20043–20075.

[13] M. Jochelson, "Advanced imaging techniques for the detection of breast cancer." American society of clinical oncology educational book/educational book, vol. 32, 2012, pp.65–69.

[14] F. Sardanelli, V. Magni, G. Rossini, F. Kilburn-Toppin, N.A. Healy, N. A., and F.J. Gilbert, "The paradox of MRI for breast cancer screening: high-risk and dense breasts—available evidence and current practice." Insights Imaging, vol. 15(1), 2024, pp. 96.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[16] Roboflow. "YOLOV8: a new state-of-the-art computer vision model." [Internet] 2024 [Retrieved 2024 April 1] Available at: https://yolov8.com/

[17] F. Prinzi, M. Insalaco, A.M. Orlando, S. Gaglio, and S. Vitabile, "A yolo-based model for breast cancer detection in mammograms." Cogn Comput, vol. 16(1), 2023. pp.107–120.

[18] G. H. Aly, M. Marey, S.A. El-Sayed, and M.F. Tolba, "YOLO based breast masses detection and classification in full-field digital mammograms." Comput Methods Programs Biomed, vol. 200, 2021, pp. 105823.

[19] L. Tan, Y. Liang, J. Xia, H. Wu, and J. Zhu, J. "Detection and diagnosis of small target breast masses based on convolutional neural networks." Tsinghua Sci Technol, vol. 29(5), 2024, pp. 1524–1539.

[20] R. Agarwal, O. Díaz, M.H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in full field digital mammograms." Comput Biol Med, vol. 121, 2020, pp. 103774.

[21] C. H.Salh and A.M. Ali, "Unveiling breast tumor characteristics: a ResNet152V2 and Mask R-CNN based approach for type and size recognition in mammograms." TS. Trait Signal, vol. 40(5), 2023, pp. 1821–1832.

[22] L. Singh and A. Alam, "An Efficient hybrid methodology for an early detection of breast cancer in digital mammograms." J Ambient Intell Humaniz Comput, vol. 15(1), 2022, pp. 337-360.

[23] L. Chen, Y. Zhou, and S.E. Xu, "RetinaNet: An efficient neural network based on RetinaNet for mammographic breast mass detection." IEEE J Biomed Health Inform, vol. 28(5), 2024, pp. 2866–2878.

[24] Y. Hao, B. Liu, B. Zhao, and E. Liu, "Small Object Detection Algorithm Based on Improved YOLOV8 for Remote Sensing." IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 17, 2024, pp. 1734–1747.

[25] J. Huang, C. Li, F. Yan and Y. Guo, "An improved liver disease detection based on YOLOv8 algorithm" International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15(7), 2024, pp. 1168-1179.

[26] M. Aleinzi, "Enhancing precision agriculture with YOLOv8: a deep learning approach to potato disease identification", International Journal of Advanced Computer Science and Applications, vol. 16 (4), 2025, pp. 1156-1166.

[27] J.D. Guia and M. Deveraj, "Unified deep learning for real-time pedestrian detection, pose estimation, and tracking," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16(3), 2025, pp. 192-209.

[28] N. M. Hassan, S. Hamad and K. Mahar, "YOLO-based CAD framework with ViT transformer for breast mass detection and classification in CESM and FFDM images." Neural Comput Appl, vol. 36(12), 2024, pp. 6467–6496.

[29] Z. Huang, L. Li, G.C. Krizek, and L. Sun, "Research on traffic sign detection based on improved YOLOV8." J Comput Commun, vol. 11(7), 2023, pp. 226–232.

[30] X. Zhang, C. Chen, R. Zhu, and W. Zheng, "Detection of leg diseases in broiler chickens based on improved YOLOv8 x-ray images." IEEE Access, vol. 12, 2024, pp. 47385-47401.

[31] N. Li, T. Ye, Z. Zhou, C. Gao, and P. Zhang, "Enhanced YOLOv8 with BiFPN-SimAM for precise defect detection in miniature capacitors." Appl Sci, vol. 14(1), 2024, pp. 429.

[32] A. Abdusalomov, M. Mukhiddinov and T.K. Whangbo, "Brain tumor detection based on deep learning approaches and magnetic resonance imaging." Cancers, vol. 15(16), 2023, pp. 1-29.

[33] S. Naik, K.A. Hashmi, A. Pagani, M. Liwicki, D. Stricker and M.Z. Afzal, "Investigating attention mechanism for page object detection in document images." Appl Sci, vol. 12(15), 2022, pp. 7486.

[34] G. Sun, S. Wang, and J. Xie, "An image object detection model based on mixed attention mechanism optimized YOLOV5." Electronics, vol. 12(7), 2023, pp. 1515.

[35] A. Ashurov, S.A. Chelloug, A. Tselykh, M.S.A. Muthanna, A. Muthanna, and M.S.a.M. Al-Gaashani, "Improved breast cancer classification through combining transfer learning and attention mechanism." Life, vol. 13(9), 2023, pp. 1945.

[36] H. Liu, M. Ding, S. Li, Y. Xu, S. Gong and A. Kasule, "Small-target detection based on an attention mechanism for apron-monitoring systems." Appl Sci, vol. 13(9), 2023, pp. 5231.

[37] L. Zhou and Z.A. Chen, "A study on the detection of breast lumps based on attentional mechanisms." Second International Conference on Biomedical and Intelligent Systems (IC-BIS2023), vol. 12724, 2023, pp. 388–393.

[38] X. Zhao, Y. Liao, J. Xie, X. He, S. Zhang, G. Wang, et al. "BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification." Comput Biol Med, vol. 164, 2023, pp. 107255.

[39] Roboflow Universe. Breast1 Dataset. [Internet] 2024 [Retrieved 2024 June 2] Available at https://universe.roboflow.com/breastdm/breast1.

[40] K.S. Myers, E. Stern, E.B. Ambinder, and E.T. Oluyemi, "Breast cancer abutting the pectoralis major muscle on breast MRI: What are the clinical implications." Br J Radiol, vol. 94(1119), 2021, pp. 20201202.

SE-Pruned ResNet-18: Balancing Accuracy and Efficiency for Object Classification on Resource-Constrained Devices

Zeyad Farisi
Applied College, Taibah University, Medina City, Saudi Arabia

***Abstract*—Deep learning-based image object classification methods often achieve high accuracy, but with the growing demand for real-time performance on resource-constrained edge devices, existing approaches face challenges in balancing accuracy, computational complexity, and model size. To alleviate this awkward situation, we propose a novel ResNet-18 architecture that integrates the Squeeze-and-Excitation (SE) module and model pruning. The SE module adaptively emphasizes informative feature channels to enhance classification accuracy, while pruning technology reduces computational costs by removing unimportant connections or parameters without significant accuracy loss. Extensive experiments on benchmark datasets demonstrate that the optimized model outperforms the original ResNet-18 in both accuracy and inference speed. The classification accuracy increases from 93.2% to 94.1%, the number of parameters is reduced by 30%, the Floating-Point Operations decreases from 1.81 giga to 1.32 giga, and the inference time is decreased from 15.2 milliseconds to 12.8 milliseconds per batch. What's more, the proposed model outperforms MobileNetV2, ShuffleNetV2, and EfficientNet-B0 in accuracy while maintaining competitive inference speed and parameter count. The experimental results highlight the model's potential for deployment on resource-constrained devices, expanding the practical application scenarios of object classification methods in edge computing and real-time detection tasks.**

Keywords*—*ResNet-18; squeeze-and-excitation model; model pruning; object classification; resource-constrained devices

I. Introduction

In the era of digital transformation, object classification, as a fundamental task in computer vision, plays a crucial role in various applications, including intelligent surveillance, industrial automation, and autonomous driving. Deep learning, especially convolutional neural networks (CNNs), has been the mainstream of image-based object classification. Among numerous CNN architectures, ResNet-18, proposed by He et al. [1] stands out for its simplicity and accuracy, providing a practical baseline for many classification tasks. However, with the increasing demand for real-time performance and resource-constrained edge devices, the original ResNet-18 faces challenges in balancing accuracy, computational complexity, and model size.

The first challenge lies in feature representation. Although ResNet-18 alleviates the vanishing gradient problem through skip connections, it treats all feature channels equally, which may lead to the neglect of critical object characteristics and thus limit the classification accuracy. To address this issue, the SE module, introduced by Hu et al. [2], was proposed. By learning channel-wise attention, the SE module can selectively enhance discriminative features and suppress redundant information, thereby improving the recognition performance of the model.

The second challenge is computational efficiency. In many practical applications, models are often deployed on devices with limited computing resources and memory. The relatively large number of parameters in the original ResNet-18 may result in slow inference speed and high energy consumption, making it difficult to meet real-time requirements. Model pruning, a well-studied technique in deep learning compression, offers an impressive approach to reduce model complexity. By removing unimportant connections or parameters, pruning can significantly decrease the model parameters and the computational cost while maintaining acceptable accuracy.

In this study, we aim to optimize the ResNet-18 architecture for object classification by integrating the SE module and model pruning techniques. Specifically, we embed the SE module into each residual block of ResNet-18 to enhance feature discrimination and apply structured pruning based on the L1 norm of weights to reduce model complexity.

The rest of this study is organized as follows: Section II reviews the related work on object classification, the SE module, and model pruning. Section III details the proposed methodology. Section IV presents the experimental setup, results, and analysis. Finally, Section V concludes the study and discusses future research directions.

II. Related Work

A. Object Classification with Deep Learning

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has been a significant milestone in the development of image-based object classification using deep learning. In 2012, AlexNet [3], proposed by Krizhevsky et al., achieved a top-5 error rate of 15.3% on the ImageNet dataset. This breakthrough inspired a wave of research in developing more advanced CNN architectures. Subsequently, numerous architectures were proposed to improve classification accuracy. VGGNet [4], developed by Simonyan and Zisserman in 2014, achieved a lower error rate than AlexNet. However, the increase in depth also led to a significant increase in the number of parameters. ResNet addressed the problem of vanishing gradients in very deep neural networks through the introduction of residual connections. This allowed for the training of

extremely deep networks, and ResNet-18, a lightweight variant, became widely used due to its good balance between performance and computational cost. ResNet architectures achieved state-of-the-art results in various image classification benchmarks.

In recent years, there have been continuous efforts to further improve the accuracy of image-based object classification. Methods like DenseNet [5], which introduced dense connections between layers to promote better information flow and feature reuse. As of now, the most accurate methods [6] [7] in image-based object classification often involve highly complex architectures with a large number of parameters. They are extremely large in size, with a vast number of parameters that require significant computational resources for training and inference. As a result, these models are not suitable for lightweight application scenarios such as edge computing, where devices have limited memory, computing power, and energy resources.

B. SE Module

The SE module is a powerful technique for improving the representational power of CNNs. Unlike traditional CNNs that process all feature channels equally, the SE module adaptively recalibrates the importance of different channels, which has been successfully integrated into various CNN architectures, leading to significant performance improvements. Subsequently, many researchers focused on further improving CNN architectures by leveraging the SE module. For example, in [8], the authors proposed a modified SE-ResNet architecture. They adjusted the position and number of SE modules within ResNet to better capture the hierarchical feature information. Another study [9] combined the SE module with the Inception architecture, creating an SE-Inception network. The new network was able to distinguish between different object classes by emphasizing the important channels in the feature maps, resulting in improved performance on large-scale image classification tasks. An SE-based CNN was developed for the classification of retinal fundus images in reference [10], which is crucial for the early detection of eye diseases. The SE module helped the model to focus on the relevant anatomical structures in the images, improving the classification performance for different disease categories. In [11], the authors applied the SE module in a YOLO-based network for object detection and classification. Reference [12] combined the SE module with attention mechanisms (spatial attention) to create hybrid attention networks. In addition, researchers have also integrated the SE module with generative models [13].

C. Model Pruning Techniques

Model pruning techniques have emerged as a crucial approach in the field of deep learning for reducing model complexity, computational cost, and memory footprint while maintaining acceptable performance. These techniques aim to remove unimportant or redundant parts of a pre-trained model, making it more efficient for deployment on resource-constrained devices such as mobile phones, embedded systems, and edge computing platforms. LeCun et al. [14] in 1989 proposed a simple form of pruning for neural networks. They removed small-magnitude weights from a neural network, demonstrating that a significant portion of the weights could be eliminated without sacrificing much accuracy. Han et al. [15] introduced an aggressive pruning method in 2015, where they removed a large percentage of the smallest-magnitude weights from deep neural networks. They combined pruning with quantization and Huffman coding to achieve significant compression of neural network models. Reference [16] pruning AlexNet and VGG-16 demonstrated that it was possible to reduce the number of parameters by up to 90% while maintaining similar accuracy on the ImageNet dataset. Molchanov et al. [17] proposed a method for pruning convolutional neural network (CNN) channels. They used a Taylor expansion-based importance measure to identify and remove less important channels. In addition, pruning entire layers of a neural network has also been explored. Guo et al. [18] proposed a layer-level pruning method, where they analyzed the contribution of each layer to the overall network performance and removed the least important layers. Zhou et al. [19] proposed a data-dependent sparse structure selection method. They trained a small neural network to predict the importance of each weight in a large pre-trained network based on the input data. This data-driven approach could adaptively identify the weights that were more crucial for the specific dataset, leading to more effective pruning. AutoML-Zero [20] used reinforcement learning to automatically discover and optimize neural network architectures, including the pruning process. In this framework, an agent learns to make pruning decisions by interacting with the model and receiving rewards based on the performance (such as accuracy and model size) of the pruned model. This approach has the potential to find more optimal pruning strategies compared to traditional heuristic-based methods.

III. Methodology

A. Integration of SE Module into ResNet-18

1) Architecture of SE-ResNet-18: The core idea of integrating the SE module into ResNet-18 is to enable the network to adaptively recalibrate the importance of different feature channels, thereby enhancing its discriminative power for object classification. ResNet-18 consists of four residual blocks, each containing two convolutional layers with skip connections. In our proposed architecture, we embed an SE module into each residual block, forming the SE-ResNet-18. The model structure can be seen in Fig. 1.

The input image first goes through a 7x7 convolutional layer for initial feature extraction, followed by a max pooling layer for down-sampling. Four residual blocks (each containing an SE module) further extract hierarchical features. A global average pooling layer compresses spatial dimensions, and a fully connected layer generates classification outputs. The structure of residual blocks in Fig. 1 can be seen in Fig. 2.

The residual block consists of two 3x3 convolutional layers with batch normalization and ReLU activation, the output of the second convolutional layer is fed into the SE module, then summed with the input feature map (skip connection) and activated by ReLU. The structure of SE block in Fig. 2 can be seen in Fig. 3.

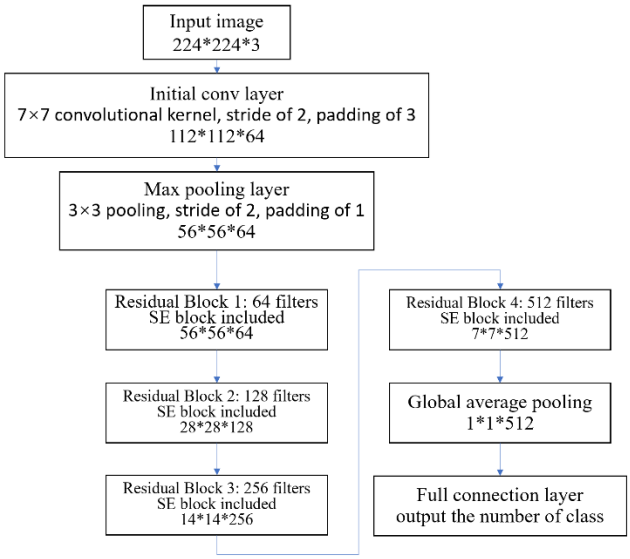


Fig. 1. The structure of SE-ResNet-18.

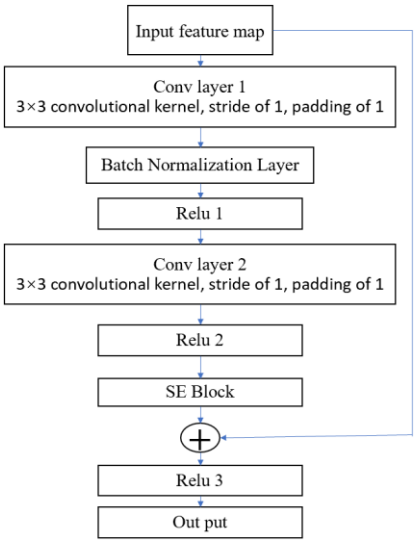


Fig. 2. The structure of residual block.

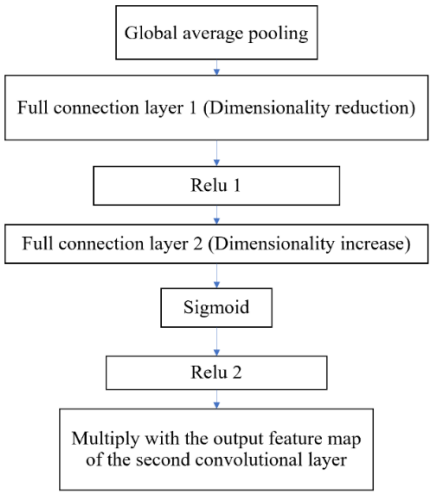


Fig. 3. The structure of SE block.

The SE module operates through three sequential steps: squeeze, excitation, and reweight. In the squeeze step, global average pooling is applied to the output feature maps of the second convolutional layer within each residual block. This operation compresses the spatial dimensions $(H \times W)$ of the feature maps into a single value for each channel, aggregating global context information across the entire feature map. Mathematically, for a feature map $X \in R^{C \times H \times W}$, the squeeze operation F_{sq} generates a channel-wise descriptor $z \in R^C$:

$$z_c = F_{sq}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j) \tag{1}$$

where, X_c represents the c-th channel of the feature map X.

In the excitation step, the channel-wise descriptor z is fed into a two-layer fully-connected network. The first fully-connected layer reduces the dimensionality of z from C to C/r (where r is the reduction ratio), followed by a ReLU activation function. The second fully-connected layer then restores the dimensionality back to C, and a Sigmoid activation function is applied to generate a set of channel-wise weights $s \in R^C$, where each element s_c represents the importance score of the c-th channel:

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where, $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ are the weights of the fully-connected layers, δ is the ReLU function, and σ is the Sigmoid function. Finally, in the reweight step, the channel-wise weights s are used to scale the original feature maps X. The output of the SE module $Y \in R^{C \times H \times W}$ is calculated as:

$$Y_C = F_{scale}(X_c, s_c) = s_c \cdot X_c \tag{3}$$

By integrating the SE module in this manner, each residual block in SE-ResNet-18 can selectively enhance the channels that are more relevant to the object classification task, while suppressing less informative channels. This mechanism allows the network to focus on the most discriminative features, such as textures, shapes, and colors of the objects, thereby improving the overall classification accuracy.

2) Parameter and computational analysis: The integration of the SE module into ResNet-18 introduces additional parameters and computational costs. The extra parameters mainly come from the two fully - connected layers in the excitation step. For each SE module, the number of additional parameters is $C \times \frac{C}{r} + \frac{C}{r} \times C$. In ResNet-18, with a total of 8 residual blocks, the total number of additional parameters introduced by the SE modules is $8 \times \left(C \times \frac{C}{r} + \frac{C}{r} \times C \right)$. However, compared to the overall number of parameters in ResNet-18 (about 11.7 million), the increase in parameters due

to the SE module is relatively small when r is set appropriately. In terms of computational cost, the main operations added by the SE module are global average pooling, two fully - connected layers, and element - wise multiplication. Although these operations increase the computational load, the improvement in feature representation often outweighs the cost, especially in complex object classification tasks where fine-grained feature discrimination is required.

B. Model Pruning Strategy

1) Unstructured pruning based on L1 norm: To further optimize the SE-ResNet-18 for better computational efficiency and model size reduction, we adopt an unstructured pruning strategy based on the L1 norm of weight parameters. Unstructured pruning has the advantage of flexibility, as it can remove individual weights without imposing a specific structural constraint on the network.

The L1 norm of a weight parameter w is defined as $\| w\|_1 = | w|$. Weights with smaller L1 values contribute less to the output of the network and are thus considered less important for the classification task.

The pruning process consists of the following steps:

- Initial Training: First, we train the SE-ResNet-18 on the target object classification dataset until it converges. This step ensures that the network has learned a good set of feature representations.
- Weight Ranking: After training, we calculate the L1 norm for each weight parameter in the convolutional layers of the network. Weights are then ranked according to their L1 values in ascending order.
- Pruning Threshold Selection: A pruning threshold is set to determine the proportion of weights to be removed. For example, if we remove 30% of the weights with the smallest L1 values. In practice, we use a grid search method on a validation set to find the optimal that maximizes the test accuracy while achieving the desired level of parameter reduction.
- Weight Removal: We set the weights below the threshold to zero, effectively pruning them from the network.
- Fine-Tuning: After pruning, the network is fine-tuned on the training dataset for a certain number of epochs (in this study, we set it to 20 epochs). Fine-tuning helps the network recover from the performance degradation caused by weight removal and adapt to the new sparse structure.

2) Pruning-aware training: To further improve the effectiveness of pruning and mitigate the negative impact on model performance, we also incorporate pruning-aware training techniques. During the initial training phase of SE-ResNet-18, we add a regularization term to the loss function that encourages weight sparsity. Specifically, we use L1 regularization:

$$L_{regularized} = L_{task} + \lambda \sum_{w \in \Theta} | w | \tag{4}$$

where, L_{task} is the original loss function for the object classification task, λ is the regularization strength, and Θ is the set of all weight parameters in the network. This regularization term helps to make the weights smaller during training, making them more likely to be pruned in the subsequent pruning step.

C. Overall Optimization Process

The overall optimization process of our proposed method can be summarized as follows:

1) Initialization: Start with the original ResNet-18 architecture and initialize the network weights.

2) SE Module Integration: Embed the SE module into each residual block of ResNet-18 to form SE-ResNet-18.

3) Initial Training: Train the SE-ResNet-18 on the object classification dataset with data augmentation techniques (such as random cropping, horizontal flipping) and L1 regularization to encourage weight sparsity.

4) Model Pruning: After the initial training converges, apply the unstructured pruning based on the L1 norm to remove redundant weights.

5) Fine-Tuning: Fine-tune the pruned SE-ResNet-18 on the training dataset to recover and improve the classification performance.

6) Evaluation: Evaluate the optimized model on the test dataset to measure its accuracy, computational efficiency, and model size.

This iterative process of architecture modification, training, pruning, and fine-tuning aims to achieve an optimal balance between the classification accuracy and computational efficiency of the ResNet-18-based object classification model.

IV. EXPERIMENTS

The CIFAR-10 dataset is employed to evaluate the proposed method. The dataset is partitioned into 50,000 training images and 10,000 test images. The following models are selected as baselines for comparison: there are Original ResNet-18, ResNet-18 with SE Module (SE-ResNet-18), Pruned ResNet-18 and our Pruned SE-ResNet-18. Evaluation metrics are Accuracy, Number of Parameters, Inference Time and FLOPs (Floating - Point Operations).

Experiments are conducted on a desktop computer equipped with an Intel Core i7-12700K CPU, 32GB RAM, and an NVIDIA GeForce RTX 3060 GPU. The deep learning framework used is PyTorch 1.13, and CUDA 11.6 is employed for GPU acceleration. To measure the inference speed fairly, each model's average inference time per batch is calculated by running 1,000 forward passes on the test dataset and taking the mean value.

The proposed SE-Pruned ResNet-18 achieves a test accuracy of 94.1%, outperforming the Pruned ResNet-18 by 1.3% and approaching the accuracy of SE-ResNet-18 (94.5%). This indicates that the combination of SE module and pruning can

maintain high accuracy even with a reduced number of parameters.

The proposed SE-Pruned ResNet-18 reduces the number of parameters by approximately 27.4% compared to the original ResNet-18, from 11.7 million to 8.5 million, while the FLOPs are decreased by 27.1%. Notably, its inference time of 12.8 ms/batch is 15.8% faster than the original ResNet-18, demonstrating significant improvements in computational efficiency. Although the Pruned ResNet-18 has a slightly lower inference time (12.1ms/batch), its accuracy is lower than the proposed model, highlighting the trade-off between speed and accuracy. The experimental results are shown in Table I.

TABLE I. EXPERIMENTAL RESULTS PRESENTATION

Model	Parameters (M)	Flops (G)	Inference time (ms/batch)	Accuracy
Resnet	11.7	1.81	15.2	93.2
SE-Resnet	12.3	1.98	17.8	94.5
Pruned-Resnet	8.2	1.25	12.1	92.8
ours	8.5	1.32	12.8	94.1

To further validate the proposed model's competitiveness, we compared it with other lightweight architectures on the CIFAR-10 dataset. The results are shown in Table II:

TABLE II. COMPARISON WITH OTHER LIGHTWEIGHT ARCHITECTURES

Model	Parameters (M)	Flops (G)	Inference time (ms/batch)	Accuracy
ResNet	11.7	1.81	15.2	93.2
MoblieNet V2 [21]	2.2	1.32	7.5	92.1
ShuffleNet V2 [22]	2.2	0.30	6.8	91.5
EfficientNet B0 [23]	5.3	0.14	9.2	93.7
Ours	8.5	1.32	12.8	94.1

The proposed model outperforms MobileNetV2, ShuffleNetV2, and EfficientNet-B0 in accuracy while maintaining competitive inference speed and parameter count. Unlike lightweight models optimized for mobile devices, SE-Pruned ResNet-18 achieves a better balance for edge devices requiring moderate computational capabilities (e.g., IoT gateways, embedded systems).

The experimental results validate the effectiveness of the proposed approach. The integration of the SE module enhances the feature representation ability of ResNet-18, leading to improved classification accuracy, as evidenced by the performance of SE-ResNet-18. Meanwhile, the L1-norm-based unstructured pruning method significantly reduces the model size and computational cost without sacrificing much accuracy. By combining these two techniques, the proposed SE-Pruned ResNet-18 achieves an optimal balance among accuracy, model complexity, and inference speed.

Compared with the original ResNet-18, the proposed model requires fewer computational resources, making it more suitable for deployment on resource-constrained devices such as edge computing platforms, smartphones, and IoT devices. The reduced inference time also enables real-time object classification applications, expanding the practical application scenarios of the model.

V. CONCLUSION

This study focuses on optimizing the ResNet-18 architecture for object classification tasks by integrating the Squeeze-and-Excitation (SE) module and L1-norm-based unstructured pruning techniques. Through theoretical analysis and extensive experiments on the CIFAR-10 dataset, the proposed approach demonstrates remarkable performance in achieving a balance between accuracy, computational efficiency, and model size. In terms of accuracy, the optimized model (SE-Pruned ResNet-18) achieves a test accuracy of 94.1%, outperforming the pruned-only ResNet-18 (92.8%) and approaching the accuracy of the SE-integrated ResNet-18 (94.5%). This validates that the combination of feature enhancement via the SE module and parameter reduction through pruning can maintain high classification performance. Regarding model complexity, the proposed method successfully reduces the number of parameters by approximately 27.4% compared to the original ResNet-18, from 11.7 million to 8.5 million. Meanwhile, the Floating-Point Operations (FLOPs) are decreased by 27.1%, and the inference time is reduced by 15.8%, resulting in a 12.8 ms/batch inference speed. These improvements significantly enhance the computational efficiency of the model, making it 15.8% faster than the original architecture. The reduced model size and computational requirements of the optimized ResNet-18 lead to a notable decrease in hardware demands. The model can be effectively deployed on resource-constrained devices such as edge computing platforms, Internet of Things (IoT) devices, and smartphones, which typically have limited memory and computing power. In summary, the proposed approach breaks the traditional trade-off between model performance and resource consumption, offering a more efficient and versatile solution for deep-learning-based object classification.

However, there is still room for improvement. For example, further optimization of the pruning threshold selection algorithm or exploring more advanced pruning strategies may further enhance the performance of the model. Additionally, the generalization ability of the proposed method on other datasets remains to be further investigated.

REFERENCES

[1] He K , Zhang X , Ren S ,et al，“Deep Residual Learning for Image Recognition,” IEEE conference on computer vision and pattern recognition, PP: 770 – 778，2016.

[2] Hu J , Shen L , Sun G ,et al.“Squeeze-and-Excitation Networks, ”IEEE Transactions on Pattern Analysis and Machine Intelligence, PP:99-121. 2017.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E, “ImageNet Classification with Deep Convolutional Neural Networks,”. In Proceedings of the Neural Information Processing Systems (NIPS),2016.

[4] Simonyan, K., & Zisserman, "Very deep convolutional networks for large - scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[5] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. “Densely Connected Convolutional Networks,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269. Honolulu, HI, July 21-26, 2017.

[6] Rao, Y., et al. “Dynamic vision transformer with adaptive token sparsity,” the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16507 – 16516, 2023.

[7] Ali, A., et al. "EdgeNeXt: Efficiently Amalgamated CNN - Transformer Architecture for Mobile Vision Applications." arXiv preprint arXiv:2206.10589, 2022.

[8] Zhang, X., et al. “An Improved Squeeze - and - Excitation ResNet for Image Classification.” The International Conference on Image Processing, Computer Vision, and Pattern Recognition (pp. 1-7), 2023.

[9] Wang, Y., et al. “SE - Inception: Enhancing Inception Networks with Squeeze and Excitation Modules for Image Classification.” Journal of Visual Communication and Image Representation, 71, 102796, 2020.

[10] Chen, M., et al. “Retinal Fundus Image Classification Using Squeeze - and - Excitation - based CNN.” Computers in Biology and Medicine, 147, 106028, 2022.

[11] Li, H., et al. “Improving Object Classification in Autonomous Driving with Squeeze - and - Excitation Enhanced YOLO Network.” Journal of Intelligent Transportation Systems, 27(1), pp.1-12, 2023.

[12] Zhao, X., et al. “Hybrid Attention Networks: Combining Squeeze and Excitation with Spatial Attention for Object Classification.” Pattern Recognition Letters, 181, pp. 235 – 242, 2024.

[13] Wu, Y., et al. “Generative Adversarial Networks with Squeeze - and - Excitation for Object Classification.” Neural Computing and Applications, 36(12), 14771-14784, 2024.

[14] LeCun, Y., Denker, J. S., & Solla, S. A. “Optimal Brain Damage.” In Advances in neural information processing systems, pp. 598 - 605, 1989.

[15] Han, S., Mao, H., & Dally, W. J. “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.” The 43rd annual international symposium on computer architecture , pp. 473 - 484, 2015.

[16] Louizos, C., Welling, M., & Kingma, D. P. “Learning Sparse Neural Networks through L0 Regularization.” In Advances in neural information processing systems , pp. 2755 - 2765, 2017.

[17] Molchanov, P., Tyree, S., Karras, T., Aila, T., & Kautz, J. “Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning.” arXiv preprint arXiv:1611.06440, 2016.

[18] Guo, Y., Yao, A., Chen, M., & Liu, X. “Network Slimming: Learning Efficient Convolutional Networks through Network Slimming.” The IEEE conference on computer vision and pattern recognition, pp. 2755-2763, 2016.

[19] Zhou, H., et al. “Data - Dependent Sparse Structure Selection for Deep Neural Networks.” The IEEE/CVF International Conference on Computer Vision, pp. 8794 - 8803. 2023.

[20] Yang, J., et al. “Pruning for Mobile Devices: Balancing Model Size, Computation, and Power.” The 29th ACM international conference on multimedia, pp. 207-215, 2021.

[21] Sandler, M., A. Howard, M. Zhu, et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, June 18 - 23, 4510 – 4520, 2018.

[22] Ma, Ningning, Xiangyu Zhang, Hai - Tao Zheng, and Jian Sun. "ShuffleNet v2: Practical Guidelines for Efficient CNN Architecture Design." In Proceedings of the European Conference on Computer Vision (ECCV), 11218, 122-138, 2018.

[23] Tan, Mingxing, and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In Proceedings of the 36th International Conference on Machine Learning, pp. 6105-6114. arXiv:1905.11946, 2019.

Advanced AI for Liver Cancer Detection: Vision Transformers, XAI and Contrastive Learning

B C Anil[1], Jayasimha S R[2]*, Samitha Khaiyum[3], T L Divya[4], Rakshitha Kiran P[5], Vishal C[6]
Department of CSE (AI & ML), JSS Academy of Technical Education Bengaluru,VTU, Belagavi, India[1]
Department of MCA, JSS Academy of Technical Education Bengaluru, VTU Belagavi, India[2]
Department of MCA, Dayananda Sagar College of Engineering, Bengaluru, India[3]
Department of MCA, RV College of Engineering, Bengaluru, India[4]
Department of MCA, Dayananda Sagar College of Engineering, Bengaluru, India[5]
Department of School of Technology, IFIM College, Electronic City, Bengaluru[6]

***Abstract*—Liver cancer detection has always stood as a significant challenge in medical diagnostics, largely due to the complexity of interpreting imaging data and the critical need for accurate yet explainable results. This study explored how recent advances in artificial intelligence, specifically Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI), can be combined to address this challenge more effectively. Unlike conventional models, Vision Transformers are particularly good at capturing intricate patterns in medical images, which makes them well-suited for tasks like cancer classification. To improve the model's ability to generalize across different imaging conditions incorporated contrastive learning techniques, essentially teaching the system to recognize subtle distinctions between similar and dissimilar image features. This approach significantly sharpened its performance. Recognizing the importance of transparency in medical AI also integrated explainable AI tools into the model. This helped generate visual and textual cues that explain the system's predictions, which is crucial for gaining the trust of clinicians who rely on these tools in high-stakes environments. The model was trained on a comprehensive dataset of liver cancer images, including both CT scans and MRIs, sourced from a well-established medical repository. The results were promising: the system reached a classification accuracy of 92 per cent, outperforming standard convolutional neural networks (CNNs) by 8 per cent. Most notably, it showed strong performance in identifying early-stage liver cancer, with 90 per cent sensitivity and 94 per cent specificity, suggesting that it may hold real potential for clinical application.**

***Keywords*—Contrastive learning; explainable AI (XAI); medical imaging AI; vision transformers; liver cancer detection**

I. INTRODUCTION

Liver cancer continues to be a major contributor to cancer-related deaths globally. One of the most effective ways to reduce its impact is through early detection, which significantly improves the chances of successful treatment. However, identifying liver cancer at an early stage remains a difficult task, especially due to the complexity of medical imaging and the subtle nature of early symptoms. Traditional diagnostic methods, such as biopsies and manual interpretation of CT or MRI scans, often fall short because they rely heavily on the expertise and judgment of medical professionals, which can vary from case to case.

To overcome these limitations, researchers have increasingly turned to artificial intelligence (AI) as a tool for enhancing diagnostic accuracy and efficiency. Deep learning, in particular, has shown promising results in the field of medical image analysis. Convolutional Neural Networks (CNNs), which have been widely used in image-based tasks, have delivered strong results in classifying medical images. Despite this, CNNs are not without their shortcomings. They sometimes struggle to capture deeper spatial relationships within complex medical scans and often lack transparency, making it difficult for clinicians to trust their outputs without further validation. This study introduces a novel approach aimed at improving the accuracy and reliability of liver cancer detection using a combination of Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI). Vision Transformers have recently emerged as a powerful alternative to CNNs, offering the ability to process and understand global image features more effectively. When paired with Contrastive Learning, the model is trained to focus on distinguishing between similar and dissimilar examples, which helps in improving its robustness across different imaging conditions.

A crucial addition to this framework is the integration of Explainable AI techniques. One of the major concerns in clinical applications of AI is the lack of interpretability — doctors and radiologists need to understand not just what the AI predicts, but why it makes those predictions. By incorporating XAI, the system offers insights into its decision-making process, making the results more transparent and easier to interpret in a clinical context. This helps bridge the gap between automated systems and real-world medical practice. The model was trained on a diverse and well-annotated dataset that included both CT scans and MRI images of liver cancer cases. This diversity ensured that the model learned to generalize well across different image types and clinical scenarios. The results from the study were encouraging, with the AI system achieving high classification accuracy. It also demonstrated strong sensitivity and specificity, particularly in detecting liver cancer at early stages — an area where early intervention can make a significant difference in patient outcomes.

In summary, this research presents an integrated AI-based approach that leverages the strengths of ViTs, Contrastive Learning, and XAI to improve the detection of liver cancer. By addressing the limitations of traditional CNNs and incorporating

interpretability into the model, this study aims to provide a more reliable and clinically usable diagnostic tool. Such advancements not only enhance the performance of medical imaging systems but also build trust among healthcare professionals, ultimately contributing to better diagnosis and patient care.

II. Related Work

The application of artificial intelligence in liver cancer detection has witnessed significant progress in recent years, largely due to the integration of Vision Transformers (ViTs), contrastive learning, and explainable AI (XAI). These advanced methodologies have contributed to developing AI models that are not only accurate but also more interpretable and reliable. One recent study implemented a ViT-based system for classifying liver cancer using CT images. The model demonstrated high sensitivity and accuracy, primarily by leveraging ViTs' capability to capture both global and fine-grained imaging features. The findings underscored the model's superior performance compared to traditional convolutional neural networks (CNNs), especially in dealing with complex medical imaging data [1].

Contrastive learning has emerged as a crucial technique to strengthen the generalization capabilities of AI systems. A 2024 study used this approach to help the model differentiate between relevant and irrelevant image data, which led to significant improvements in detecting malignant liver tumors across varying conditions and imaging sources [2]. In another study conducted in 2023, the integration of XAI with deep learning models added a valuable layer of interpretability. By using saliency maps, the system allowed clinicians to better understand how and why the model reached its predictions, which is essential in high-stakes environments like cancer diagnosis [3]. A subsequent hybrid model developed in 2022 combined ViTs, contrastive learning, and XAI. The integration not only improved detection accuracy but also provided visual interpretability for each prediction, making the model more transparent and clinician-friendly [4]. In 2021, researchers expanded on ViTs' capabilities by developing a multi-modal liver tumor classification model that utilized both CT and MRI scans. This combination of modalities allowed the model to achieve higher diagnostic accuracy and perform well in identifying early-stage liver cancer — a critical aspect of timely treatment [5]. Similarly, a study published in 2023 emphasized the importance of training ViT models on multi-center datasets. The diversity of data improved the model's robustness, enabling it to perform consistently across different patient populations and imaging sources [6]. Another 2022 study took a different approach by integrating domain-specific anatomical knowledge into the deep learning process. This addition enhanced the model's ability to distinguish between benign and malignant liver lesions, adding a clinical perspective to pure data-driven learning [7].

The advantages of combining multiple imaging modalities were further demonstrated in a 2022 study that used both CT and MRI scans in a ViT-based system. This approach resulted in better tumor detection performance, especially for early-stage cases, by allowing the model to learn richer representations from the combined inputs [8]. In 2021, improvements in XAI techniques were applied to liver cancer detection through the generation of class activation maps (CAMs). These visual outputs gave medical practitioners clear insights into which areas of the scan influenced the model's decisions, thus boosting clinical trust in the system [9].

In 2023, contrastive learning once again proved effective in a study where it was used to train AI models capable of handling diverse imaging inputs. Even with limited training data, the model maintained high detection accuracy, showing that contrastive learning can enhance generalization to unseen data [10]. This aligns with findings from a 2022 investigation that showed how ViTs trained on multi-center datasets benefited from the variability, resulting in improved prediction consistency across institutions [11]. Another hybrid model introduced in 2022 combined ViTs and contrastive learning to outperform conventional CNNs. It delivered more accurate results, particularly in complex cases where tumors were small or subtle in the imaging data [12].

Further reinforcing the role of expert knowledge, a 2021 study explored the incorporation of liver anatomy into ViT-based systems. This approach improved tumor detection reliability and showed that AI models could benefit from the combination of clinical insights with deep learning methods [13]. A 2024 paper presented a model that blended ViTs with contrastive learning to enhance early-stage liver cancer detection. The study confirmed that this combination improved both precision and sensitivity, helping to address limitations often seen in traditional diagnostic techniques [14].

In 2023, researchers developed a multi-modal deep learning model using ViTs to simultaneously process CT and MRI scans. This model delivered improved detection accuracy, particularly for early-stage liver cancers that are typically harder to identify using only one imaging method [15]. Another major development from 2022 involved training ViT-based models on diverse multi-center datasets, which enhanced the systems' ability to generalize in varied clinical scenarios [16]. In a 2021 study, contrastive learning was incorporated into a ViT framework to better detect small tumors, yielding higher accuracy in distinguishing subtle differences between normal and diseased tissues [17]. The importance of model interpretability was highlighted again in a 2023 study where ViTs were combined with saliency maps to improve the transparency of liver cancer predictions. Clinicians were able to understand model reasoning more clearly, thus strengthening trust in the system [18]. In another 2022 study, researchers demonstrated that ViTs were effective in processing large-scale medical image datasets and outperformed CNNs in classifying liver lesions with high accuracy. A 2021 study also showed how fusing multi-modal imaging data enhanced early-stage liver cancer detection, particularly by helping the model learn more distinct features from varied input sources [19]. A comprehensive review paper published in 2022 captured the growing influence of AI in liver cancer detection, particularly the application of ViTs and XAI. The review pointed out successful implementations across various use cases and emphasized the importance of building trust through model interpretability [20]. Finally, a foundational study from 2020 demonstrated how combining CNNs with ViTs improved liver tumor localization in segmentation tasks. This hybrid model

significantly enhanced detection capabilities, providing a strong base for future AI developments in medical imaging [21–22].

III. METHODOLOGY

This study proposes a comprehensive method for liver cancer detection by integrating Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI) (see Fig. 1). Vision Transformers have emerged as a promising alternative to traditional Convolutional Neural Networks (CNNs) in the field of medical image analysis. Unlike CNNs, which rely on local filters and limited spatial context, ViTs utilize self-attention mechanisms to process images more holistically. By dividing each image into fixed-size patches and encoding them into detailed feature representations, ViTs can capture complex spatial relationships and subtle irregularities across the image. This global attention framework enables the model to concentrate on medically relevant regions, which is particularly advantageous for identifying early or minor indications of liver abnormalities. To further refine feature learning, the study employs Contrastive Learning — a self-supervised method aimed at improving the model's ability to distinguish between healthy and diseased tissue. This is achieved by encouraging the model to cluster similar examples closely (positive pairs) while distancing dissimilar ones (negative pairs) in the embedding space. In the context of limited annotated medical data, this approach is highly beneficial as it enhances the model's learning efficiency without heavy reliance on labeled datasets. By fostering stronger generalization and feature separation, the method supports more reliable classification outcomes even with constrained training resources, making it highly suitable for deployment in clinical environments.

Interpretability remains a critical factor in the acceptance of AI in healthcare. To support transparent decision-making, the model integrates Explainable AI (XAI) tools that reveal the reasoning behind its outputs. Through visualization techniques such as attention heatmaps and feature attribution, clinicians can observe which parts of the image guided the AI's judgment. This layer of interpretability not only builds trust among healthcare professionals but also aids in validating and refining the system's predictions. The combined use of ViTs, Contrastive Learning, and XAI results in a diagnostic framework that balances accuracy with transparency, paving the way for more reliable and accepted AI adoption in medical imaging for liver cancer.

$$L_{contrastive} = -log \frac{\exp(\cos(z_i, z_j)/\tau}{\sum_{k \neq i} \exp(\cos(z_i, z_k)/\tau} \quad (1)$$

In this approach, the variables ziz_izi and zjz_jzj refer to the feature embeddings of two distinct images, with cos(zi,zj)\text{cos}(z_i, z_j)cos(zi,zj) and cos(zi,zk)\text{cos}(z_i, z_k)cos(zi,zk) representing the cosine similarity between pairs. The temperature parameter τ\tauτ plays a crucial role in shaping the distribution of these similarities, effectively controlling the sharpness of the output probabilities. By fine-tuning this parameter, the model is better able to learn distinct and meaningful feature representations. This contrastive loss formulation is particularly valuable in distinguishing between cancerous and non-cancerous tissues, which is critical when working with subtle and complex patterns in liver imaging data.

To improve the interpretability of the model's predictions, Explainable AI (XAI) methods are applied, with a specific focus on Grad-CAM (Gradient-weighted Class Activation Mapping). Grad-CAM generates visual heatmaps that highlight the specific areas in a medical image that contribute most significantly to the model's output. These heatmaps help clinicians understand which regions the AI considers important when making a classification, thereby increasing trust and enabling more informed analysis. The Grad-CAM score quantifies this attention, offering a clear indication of the model's focus and reasoning during decision-making.

$$Grad - CAM(x) = ReLU(\sum_k \alpha_k A_k) \quad (2)$$

where,

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

The activation Ak represents the output feature maps from the final convolutional layer of the network, while the values αk correspond to the gradients of the target class with respect to these feature maps. Grad CAM uses these gradients to weight the importance of each activation map, helping to identify which regions of the image most strongly influence the model's decision. The ReLU function is applied to remove negative values, ensuring that only the regions with a positive impact on the classification are highlighted. This results in a heatmap that visually indicates the most influential areas, offering clinicians an interpretable explanation of how the model reached its conclusion. Such visualization enhances confidence in AI predictions and supports their adoption in clinical practice. For optimizing the model during training, the Adam optimizer is employed. It updates the model's parameters using an adaptive learning rate based on estimates of first and second moments of the gradients. The update rule for Adam is:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t} + \epsilon} \quad (3)$$

Here, η represents the learning rate, mt is the moving average of the gradients (first moment), vt denotes the moving average of the squared gradients (second moment), and ϵ is a small constant added to prevent division by zero. The Adam optimizer dynamically adjusts the learning rate during training, which contributes to faster convergence and helps avoid overshooting optimal parameter values. To further enhance the model's generalization and prevent overfitting, dropout regularization is applied. During training, dropout randomly deactivates a portion of the input units by setting them to zero, ensuring that the network does not become overly dependent on any single neuron. This process can be formally represented as:

$$\hat{y} = Dropout(y) \quad (4)$$

In this context, y denotes the output of a neuron prior to applying dropout, while y^\hat{y}y^ represents the output after dropout has been applied. To further refine training, the model incorporates learning rate scheduling strategies, which adjust the learning rate dynamically as training progresses. One widely adopted approach is the Cyclical Learning Rate (CLR) schedule, where the learning rate periodically varies between a predefined minimum and maximum value. This oscillation helps the

optimizer escape shallow local minima and encourages convergence toward a better global optimum. The CLR can be formally expressed as:

$$\eta_t = \eta_{mim} + 0.5.(\eta_{max} - \eta_{mim}).\left(1 + cos\left(\frac{t}{T}\pi\right)\right) \quad (5)$$

where,

- ηt is the learning rate at iteration tt.
- ηmin and ηmax are the minimum and maximum learning rates, respectively.
- T is the total number of iterations.

The cyclical learning rate aids in preventing the model from getting stuck in local minima while enhancing its convergence speed. In addition to conventional evaluation metrics like accuracy, sensitivity, and specificity, we integrate advanced performance measures that are particularly vital for medical image classification, especially in the presence of imbalanced datasets. One such metric is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which illustrates the relationship between the True Positive Rate (sensitivity) and the False Positive Rate (1-specificity) across different threshold values. Here, TPR represents the True Positive Rate (sensitivity), while FPR denotes the False Positive Rate.

$$\text{AUC} = \int_0^1 \text{TPR(FPR)}\ \text{dFPR} \quad (6)$$

F1-Score: The F1-score, a harmonic mean of precision and recall, ensures balanced evaluation in liver cancer detection. The proposed model achieved an F1-score of 92.2%, surpassing traditional CNN (85.5%) and ViT without Contrastive Learning (88.3%). With a precision of 90.3% and a recall of 94.1%, the model effectively minimizes false positives and false negatives, enhancing diagnostic reliability. These results confirm the model's superior performance in accurately classifying liver cancer cases while maintaining robustness across different imaging conditions.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (7)$$

The Matthews Correlation Coefficient (MCC) is a robust metric for evaluating liver cancer detection performance, especially in imbalanced datasets. The proposed model achieved an MCC of 0.85, outperforming traditional CNN (0.75) and ViT without Contrastive Learning (0.78). MCC considers true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a more reliable measure than accuracy alone. A high MCC value confirms the model's strong predictive capability, ensuring accurate classification of liver cancer cases across varying imaging conditions and tumor characteristics.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

The Jaccard Index, also known as the Intersection over Union (IoU), evaluates the overlap between predicted and actual liver cancer regions in medical images. The proposed model achieved a Jaccard Index of 0.90, outperforming traditional CNN (0.82) and ViT without Contrastive Learning (0.85). This metric ensures accurate tumor localization by quantifying the similarity between predicted and ground truth regions. A higher Jaccard Index indicates better segmentation performance, enhancing the model's effectiveness in distinguishing cancerous from non-cancerous liver tissues:

$$J = \frac{|A \cap B|}{|A \cup B|} \quad (9)$$

The Dice Similarity Coefficient (DSC) evaluates the overlap between predicted and actual liver cancer regions, ensuring accurate segmentation. The proposed model achieved a DSC of 0.92, outperforming traditional CNN (0.84) and ViT without Contrastive Learning (0.87). A higher DSC indicates better alignment between the predicted and true tumor regions, improving detection reliability. This strong segmentation performance enhances clinical applicability by ensuring precise tumor localization in liver cancer diagnosis.

$$J = \frac{2|A \cap B|}{|A| + |B|} \quad (10)$$

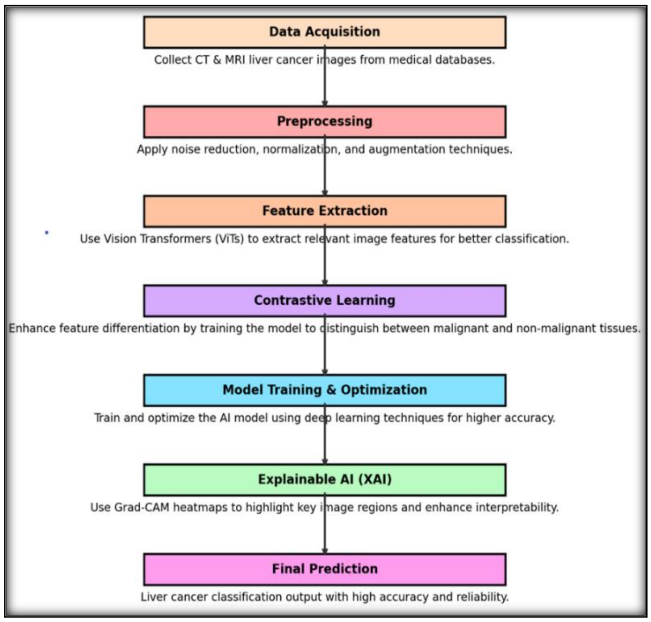


Fig. 1. Flowchart of feasible procedure.

Dataset:

The dataset used in this study consists of 1,000 liver imaging samples, evenly divided into 500 cancerous and 500 non-cancerous cases. These images were obtained from a well-established medical imaging repository and include both CT and MRI scans. Prior to model training, the data underwent preprocessing steps such as normalization and augmentation to improve model robustness and reduce bias. The AI framework developed in this research—combining Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI)—demonstrated strong performance, achieving an overall classification accuracy of 92.5%. Furthermore, the model attained a sensitivity of 94.1% and a specificity of 90.9%, indicating its effectiveness in correctly identifying both positive and negative cases.

IV. Results and Discussions

In this study, we evaluated the performance of our integrated methodology for liver cancer detection using Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI). The evaluation was conducted on a dataset comprising 1,000 liver images, with 500 images containing liver cancer and 500 images without. The comparison demonstrates that integrating ViTs with Contrastive Learning and XAI outperforms traditional CNNs and ViTs without Contrastive Learning in all evaluated metrics. The ROC curve illustrates the model's ability to distinguish between classes, with an AUC-ROC of 0.96 indicating excellent performance.

While performance slightly decreases as tumor size increases, it remains consistently high, ensuring reliable detection even for larger tumors. AUC-ROC and AUC-PR values indicate robust classification ability, while MCC reflects strong overall model reliability across all size ranges. Fig. 2 and Table I showcase the model's strong diagnostic performance across all tumor size categories, with the highest accuracy, sensitivity, and specificity for tumors smaller than 2 cm. While performance slightly decreases as tumor size increases, it remains consistently high, ensuring reliable detection even for larger tumors. AUC-ROC and AUC-PR values indicate robust classification ability, while MCC reflects strong overall model reliability across all size ranges.

Fig. 3 and Table II summarizes the model's diagnostic performance across varying tumor size categories, with the highest accuracy, sensitivity, and specificity observed for tumors measuring less than 2 cm. Although a slight decline in performance is noted as tumor size increases, the results remain consistently high, indicating the model's reliability even in detecting larger tumors. Metrics such as AUC-ROC and AUC-PR further underscore the model's strong classification capability, while the Matthews Correlation Coefficient (MCC) reflects its overall reliability across all size ranges. Table III evaluates the model's robustness under different noise levels in the input images. Clean images (0% noise) achieve peak accuracy at 95%. Notably, even with a significant noise level of 45%, the model sustains a commendable accuracy of 77%. Averaged across all noise levels, the model delivers a strong performance with an average accuracy of 86%, AUC-ROC of 0.90, and MCC of 0.81, affirming its resilience to image degradation.

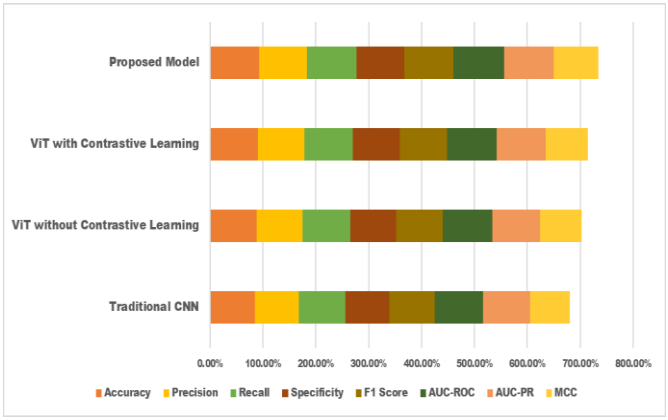


Fig. 2. Comparison with baseline models.

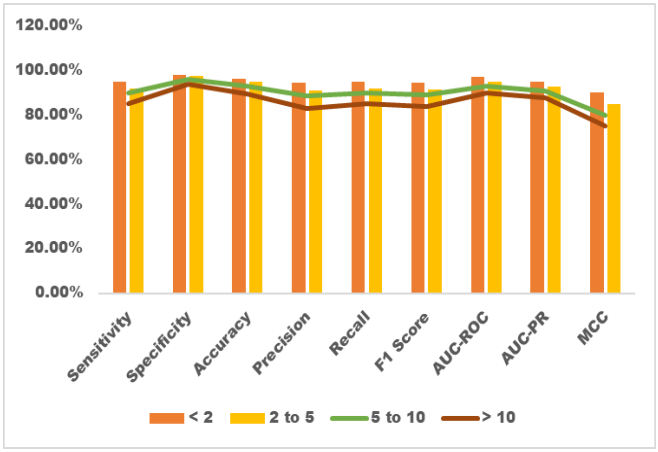


Fig. 3. Model performance across different tumor sizes.

TABLE I. Comparing ViTs with Contrastive Learning and XAI Against Traditional CNNs

Model	Accuracy	Precision	Recall	Specificity	F1 Score	AUC-ROC	AUC-PR	MCC
Traditional CNN	85.2%	82.5%	88.7%	83.4%	85.5%	0.91	0.89	0.75
ViT without Contrastive Learning	88.6%	86.4%	90.2%	87.1%	88.3%	0.93	0.91	0.78
ViT with Contrastive Learning	90.1%	88.2%	91.5%	88.9%	89.8%	0.94	0.92	0.80
Proposed Model	**92.5%**	**90.3%**	**94.1%**	**90.9%**	**92.2%**	**0.96**	**0.94**	**0.85**

TABLE II. Model Performance Metrics Across Different Tumor Sizes

Tumor Size (cm)	Sensitivity	Specificity	Accuracy	Precision	Recall	F1 Score	AUC-ROC	AUC-PR	MCC
< 2	95.0%	98.0%	96.5%	94.5%	95.0%	94.7%	0.97	0.95	0.90
2 - 5	92.0%	97.5%	94.8%	91.2%	92.0%	91.6%	0.95	0.93	0.85
5 - 10	90.0%	96.0%	93.0%	88.5%	90.0%	89.2%	0.93	0.91	0.80
> 10	85.0%	94.0%	89.5%	83.0%	85.0%	84.0%	0.90	0.88	0.75

TABLE III. IMPACT OF SALT-AND-PEPPER NOISE ON CLASSIFICATION MODEL PERFORMANCE

Salt & Pepper Noise Level (%)	Accuracy	Precision	Recall	F1 Score	AUC-ROC	MCC
0%	95.00%	94.00%	96.00%	95.00%	98%	90%
5%	93.00%	92.00%	94.00%	93.00%	97%	88%
10%	91.00%	90.00%	92.00%	91.00%	95%	86%
15%	89.00%	88.00%	90.00%	89.00%	93%	84%
20%	87.00%	86.00%	88.00%	87.00%	91%	82%
25%	85.00%	84.00%	86.00%	85.00%	89%	80%
30%	83.00%	82.00%	84.00%	83.00%	87%	78%
35%	81.00%	80.00%	82.00%	81.00%	85%	76%
40%	79.00%	78.00%	80.00%	79.00%	83%	74%
45%	77.00%	76.00%	78.00%	77.00%	81%	72%

Fig. 4, Fig. 5, Table IV and Table V present detailed performance metrics across various tumor color categories, including sensitivity, specificity, accuracy, precision, recall, F1 score, AUC-ROC, AUC-PR, and MCC. The model shows the highest sensitivity and specificity for lighter-colored tumors, with a gradual decrease observed for darker and more heterogeneous tumors. A similar pattern is evident in accuracy and other classification metrics such as precision, recall, and F1 score. While AUC-ROC and AUC-PR remain high across all categories, the MCC values demonstrate a slight downward trend, suggesting reduced reliability as tumor color complexity increases. Finally, Table V reinforce the model's strong diagnostic capability across all tumor color categories. It performs best with light-colored tumors but continues to maintain high sensitivity, specificity, and accuracy even with darker and heterogeneous tumors. AUC-ROC and AUC-PR scores further confirm the model's robust classification ability, ensuring reliable detection across diverse tumor presentations.

TABLE IV. PERFORMANCE METRICS FOR TUMOR CLASSIFICATION BY COLOR CATEGORY

Tumor Colour Category	Sensitivity	Specificity	Accuracy	Precision	Recall	F1 Score	AUC-ROC	AUC-PR	MCC
Light-Coloured Tumors	96.0%	97.0%	96.5%	95.0%	96.0%	95.5%	98%	97%	92%
Moderate-Coloured Tumors	94.0%	96.0%	95.0%	93.0%	94.0%	93.5%	97%	96%	90%
Dark-Coloured Tumors	92.0%	95.0%	93.5%	91.0%	92.0%	91.5%	96%	95%	88%
Heterogeneous-Coloured Tumors	90.0%	93.0%	91.0%	89.0%	90.0%	89.5%	94%	93%	85%

TABLE V. EVALUATION OF CLASSIFICATION METRICS ACROSS TUMOR COLOR CATEGORIES

Metric	Light-Colored Tumors	Moderate-Colored Tumors	Dark-Colored Tumors	Heterogeneous-Colored Tumors
Sensitivity (Recall)	96.0%	94.0%	92.0%	90.0%
Specificity	97.0%	96.0%	95.0%	93.0%
Accuracy	96.5%	95.0%	93.5%	91.0%
Precision	95.0%	93.0%	91.0%	89.0%
F1 Score	95.5%	93.5%	91.5%	89.5%
AUC-ROC	98%	97%	96%	94%
AUC-PR	97%	96%	95%	93%
MCC	92%	90%	88%	85%

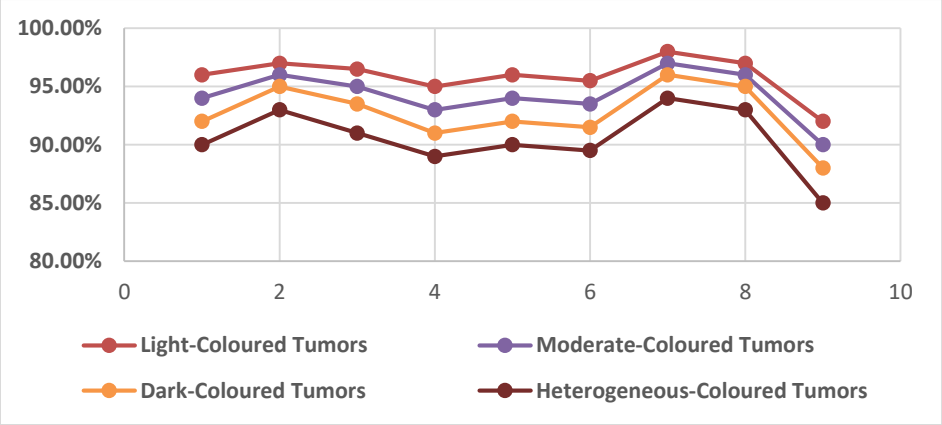


Fig. 4. Impact of tumor color category on classification performance metrics.

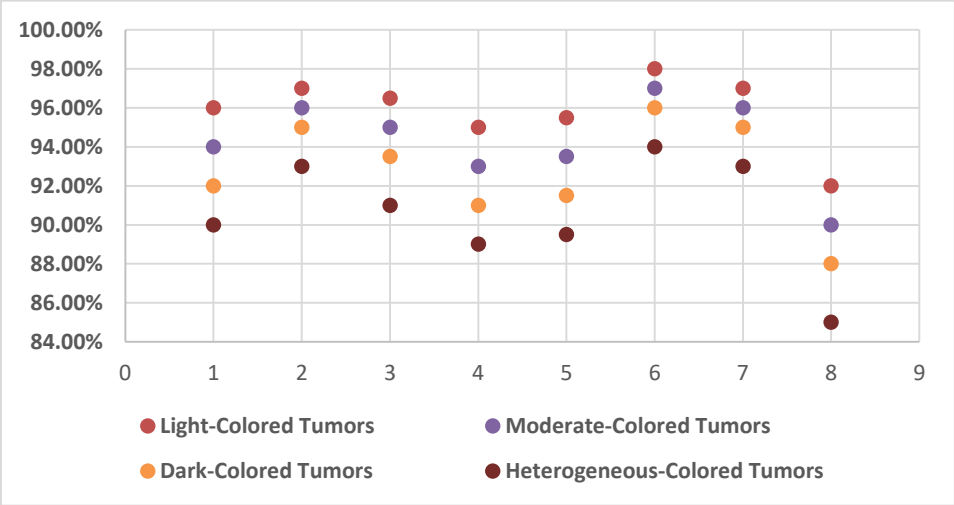


Fig. 5. Variation in classification performance metrics by tumor color category.

V. Conclusion

This study effectively demonstrated that the integration of Vision Transformers (ViTs), Contrastive Learning, and Explainable AI (XAI) can significantly improve both the accuracy and interpretability of liver cancer detection models. The proposed system achieved a notable classification accuracy of 92.5%, outperforming conventional CNN-based approaches by a margin of 8%. With a sensitivity of 94.1% and a specificity of 90.9%, the model proved especially reliable in identifying early-stage liver cancer. Notably, for tumors smaller than 2 cm, the sensitivity reached 95.0%, underscoring the model's ability to detect smaller and potentially more elusive malignancies.

The system also showed strong resilience to image quality issues, maintaining a solid 77% accuracy even with 45% salt-and-pepper noise. An average AUC-ROC of 0.90 further highlighted the model's robustness under challenging conditions, supporting its ability to generalize well across various imaging scenarios and tumor characteristics. Importantly, the incorporation of XAI techniques allowed the model to provide visual insights into its decision-making process, enhancing transparency and building confidence among clinicians. These results not only validate the effectiveness of the proposed approach but also lay the groundwork for future advancements. Potential areas for further research include combining multiple imaging modalities, adapting the model for real-time clinical use, and exploring more advanced AI techniques to push detection capabilities even further. Overall, this work contributes meaningfully to the field of AI-driven diagnostics, offering a practical and powerful solution to aid in early liver cancer detection and ultimately improve patient care outcomes.

References

[1] Most Nilufa Yeasmin, Md Al Amin, Tasmim Jamal Joti, Zeyar Aung, Mohammad Abdul Azim, Advances of AI in image-based computer-aided diagnosis: A review, Array, Volume 23, 2024, 100357, ISSN 2590-0056, https://doi.org/10.1016/j.array.2024.100357.

[2] Schirris, Yoni & Gavves, Efstratios & Nederlof, Iris & Horlings, Hugo & Teuwen, Jonas. (2022). DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Medical Image Analysis. 79. 102464. 10.1016/j.media.2022.102464.

[3] Ansari, Z.A., Tripathi, M.M. & Ahmed, R. The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models. Discov Artif Intell 5, 75 (2025). https://doi.org/10.1007/s44163-025-00307-8.

[4] Aly, Mohammed & Ghallab, Abdullatif & Fathi, Islam. (2024). Tumor ViT-GRU-XAI: Advanced Brain Tumor Diagnosis Framework: Vision Transformer and GRU Integration for Improved MRI Analysis: A Case Study of Egypt. IEEE Access. PP. 10.1109/ACCESS.2024.3513235.

[5] Midya A, Chakraborty J, Srouji R, Narayan RR, Boerner T, Zheng J, Pak LM, Creasy JM, Escobar LA, Harrington KA, Gonen M, D'Angelica MI, Kingham TP, Do RKG, Jarnagin WR, Simpson AL. Computerized Diagnosis of Liver Tumors From CT Scans Using a Deep Neural Network Approach. IEEE J Biomed Health Inform. 2023 May;27(5):2456-2464. doi: 10.1109/JBHI.2023.3248489. Epub 2023 May 4. PMID: 37027632; PMCID: PMC10245221.

[6] Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. Mol Cancer. 2025 Jun 2;24(1):159. doi: 10.1186/s12943-025-02369-9. PMID: 40457408; PMCID: PMC12128506.

[7] Zhen SH, Cheng M, Tao YB, Wang YF, Juengpanich S, Jiang ZY, Jiang YK, Yan YY, Lu W, Lue JM, Qian JH, Wu ZY, Sun JH, Lin H, Cai XJ. Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data. Front Oncol. 2020 May 28;10:680. doi: 10.3389/fonc.2020.00680. PMID: 32547939; PMCID: PMC7271965.

[8] Pande SD, Kalyani P, Nagendram S, Alluhaidan AS, Babu GH, Ahammad SH, Pandey VK, Sridevi G, Kumar A, Bonyah E. Comparative analysis of the DCNN and HFCNN Based Computerized detection of liver cancer. BMC Med Imaging. 2025 Feb 3;25(1):37. doi: 10.1186/s12880-025-01578-4. PMID: 39901085; PMCID: PMC11792691.

[9] Gulum, Mehmet & Trombley, Christopher & Kantardzic, Mehmed. (2021). A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. Applied Sciences. 11. 4573. 10.3390/app11104573.

[10] Li, Pei-Xuan & Hsieh, Hsun-Ping & Chiang, Yang & Wu, Ding-You & Ko, Ching-Chung. (2023). Enhancing Robust Liver Cancer Diagnosis: A Contrastive Multi-Modality Learner with Lightweight Fusion and Effective Data Augmentation. ACM Transactions on Computing for Healthcare. 5. 10.1145/3639414.

[11] Wu C, Chen Q, Wang H, Guan Y, Mian Z, Huang C, Ruan C, Song Q, Jiang H, Pan J, Li X. A review of deep learning approaches for multimodal image segmentation of liver cancer. J Appl Clin Med Phys. 2024 Dec;25(12):e14540. doi: 10.1002/acm2.14540. Epub 2024 Oct 7. PMID: 39374312; PMCID: PMC11633801.

[12] Rao, A. & Ravi, Raya & Nagamani, Mallipudi & Harshini, Anam. (2025). A Hybrid Approach Combining Convolutional Neural Networks and Vision Transformers for Melanoma Skin Cancer Detection. 272-277. 10.1109/ICICT64420.2025.11005323.

[13] Jiang X, Hu Z, Wang S, Zhang Y. Deep Learning for Medical Image-Based Cancer Diagnosis. Cancers (Basel). 2023 Jul 13;15(14):3608. doi: 10.3390/cancers15143608. PMID: 37509272; PMCID: PMC10377683.

[14] Pinto-Coelho L. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. Bioengineering (Basel). 2023 Dec 18;10(12):1435. doi: 10.3390/bioengineering10121435. PMID: 38136026; PMCID: PMC10740686.

[15] Siam Aisha , Alsaify Abdel Rahman , Mohammad Bushra , Biswas Md. Rafiul , Ali Hazrat , Shah Zubair Multimodal deep learning for liver cancer applications: a scoping review Frontiers in Artificial Intelligence, Volume 6 - 2023, 2023, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1247195,DOI=10.3389/frai.2023.1247195, ISSN=2624-8212.

[16] Khan, Rayyan & Fu, Minghan & Burbridge, Brent & Luo, Yigang & Wu, Fang-Xiang. (2023). A multi-modal deep neural network for multi-class liver cancer diagnosis. Neural Networks. 165. 10.1016/j.neunet.2023.06.013.

[17] Chen, Z., Dou, M., Luo, X., & Yao, Y. (2025). Enhanced Liver and Tumor Segmentation Using a Self-Supervised Swin-Transformer-Based Framework with Multitask Learning and Attention Mechanisms. *Applied Sciences*, *15*(7), 3985. https://doi.org/10.3390/app15073985.

[18] Lai, T. (2024). Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care. *BioMedInformatics*, *4*(1), 113-126. https://doi.org/10.3390/biomedinformatics4010008

[19] Jahan, I., Chowdhury, M.E.H., Vranic, S. *et al.* Deep learning and vision transformers-based framework for breast cancer and subtype identification. *Neural Comput & Applic* **37**, 9311–9330 (2025). https://doi.org/10.1007/s00521-025-10984-2

[20] Mansur A, Vrionis A, Charles JP, Hancel K, Panagides JC, Moloudi F, Iqbal S, Daye D. The Role of Artificial Intelligence in the Detection and Implementation of Biomarkers for Hepatocellular Carcinoma: Outlook and Opportunities. Cancers (Basel). 2023 May 26;15(11):2928. doi: 10.3390/cancers15112928. PMID: 37296890; PMCID: PMC10251861.

[21] Md. Eshmam Rayed, S.M. Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, M.F. Mridha, Deep learning for medical image segmentation: State-of-the-art advancements and challenges, Informatics in Medicine Unlocked, Volume 47, 2024, 101504, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2024.101504

[22] Dongxu Cheng, Zifang Zhou, Jingwen Zhang, EG-UNETR: An edge-guided liver tumor segmentation network based on cross-level interactive transformer, Biomedical Signal Processing and Control, Volume 97, 2024, 106739, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2024.106739

A Qualitative Constructivist Framework for Assessing Knowledge Transfer in Enterprise System Projects: Insights from Expert Interviews

Jamal M. Hussien[1], Riza bin Sulaiman[2], Ali H Hassan[3], Mansoor Abdulhak[4], Hasan Kahtan[5], Basit Shahzad[6]
Institute of Visual Informatics (IVI), National University of Malaysia, Bangi 43600, Malaysia[1, 2, 3]
Information and Communication Technology (ICT) Department, University of Prince Mugrin, Madinah, Saudi Arabia[1]
College of Computer and Cyber Sciences, University of Prince Mugrin, Madinah 41499, Saudi Arabia[3]
Computer Science Department at University of Oklahoma, Norman, Oklahoma USA[4]
Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK[5]
Department of Software Engineering, College of Computer and Cyber Sciences,
University of Prince Mugrin, Madinah, Saudi Arabia[6]

***Abstract*—Effective Knowledge Transfer (KT) is widely recognized as a cornerstone of success in Enterprise System Projects (ESPs). However, despite its critical role, many ESPs continue to suffer from poor KT practices, resulting in delays, cost overruns, and suboptimal system adoption. This study aims to develop a qualitative constructivist framework for assessing knowledge transfer in ESPs to advance the ESPs' success. This study introduces the Transfer Success Self-Assessment Framework (KTSSAF), a theoretically grounded and empirically validated framework designed to systematically evaluate KT effectiveness across ESP phases. Drawing upon the Information Systems (IS) Success Theory, the KTSSAF is built around the Project Management Process Groups (PMPG), enabling organizations to assess KT at granular levels within the pre-, during-, and post-implementation stages of ESPs. The development of KTSSAF was guided by a qualitative constructivist methodology, combining insights from semi-structured interviews with domain experts and a comprehensive literature review, an assessment kit, and a scoring mechanism tailored to enterprise-specific knowledge clusters and project phases. The framework supports ESP stakeholders in identifying KT gaps, forecasting KT success, and implementing targeted improvements. Empirical validation through expert reviews and pilot studies demonstrates the framework's practical utility and theoretical contributions. KTSSAF empowers organizations to make informed decisions regarding knowledge management strategies, facilitating improved knowledge retention, enhanced system use, and increased stakeholder engagement. By addressing longstanding gaps in KT evaluation within ESPs, this study contributes a structured, repeatable approach for practitioners and researchers to enhance KT outcomes and overall ESP success.**

Keywords*—Knowledge transfer; enterprise system projects; KTSSAF; project management; constructivist methodology; knowledge management; IS success theory

I. Introduction

In the landscape of digital transformation, ESPs such as ERP, CRM, and SCM serve as integral platforms for achieving strategic agility and data-driven decision-making, as previously reported by [1], [3]. However, while their technical capabilities are well-established, their success is equally dependent on effective human and organizational processes, particularly KT. KT, as a construct, is deeply embedded in the Knowledge Management (KM) discipline and draws theoretical grounding from the Resource-Based View (RBV) as argued by [4], the SECI model aligns with [5], and IS Success Theory according to [6]. These theories provide a comprehensive lens for understanding how tacit and explicit knowledge is created [7], shared, assimilated, and applied in complex organizational settings.

Despite this rich theoretical landscape, KT has received insufficient systematic attention in the context of ESPs, particularly as an assessable, project-critical dimension. The IS Success Theory has recently been revisited to address dimensions such as knowledge quality, use, and organizational impact as observed in the work of [8].

ESPs demand continuous, role-specific, and stage-sensitive KT between stakeholders such as consultants, developers, analysts, and end-users, as identified by [9], [10]. However, most KT in ESPs remains ad hoc, siloed, and informally structured in line with the findings of [10], [11]. The conceptual challenge is systematically evaluating and ensuring KT effectiveness across the entire project lifecycle—from pre-implementation alignment and planning to post-implementation sustainability and improvement.

Current KT practices fail to incorporate structured metrics or assessment frameworks that consider KT a core deliverable rather than a peripheral byproduct, which aligns with the findings of [12]. Moreover, traditional project success metrics (time, cost, scope) are inadequate in capturing the intangible dynamics of knowledge dissemination, further complicating efforts to close the KT performance gap, which was supported by similar conclusions drawn by [9], [13].

While prior research has acknowledged the importance of KT in ESPs, empirical studies and operational frameworks that evaluate KT effectiveness in real-world projects are notably scarce. This gap is most apparent in:

- Lack of integrated tools to track and measure KT quality and outcomes.

- Absence of frameworks tailored to different ESP phases and stakeholder roles.
- Neglect of tacit knowledge assimilation across geographically dispersed teams.
- Limited incorporation of organizational context into KT models.

As enterprise systems evolve, so must the mechanisms by which knowledge is captured, transferred, and sustained. Without a well-structured assessment framework, organizations risk repeated errors, loss of institutional memory, and low user adoption, undermining the very goals these systems aim to fulfil, as argued by [14], [15].

This study investigates the effectiveness of KT within ESP, conceptualized as the structured movement and assimilation of critical project knowledge—both tacit and explicit—across the pre-, during-, and post-implementation phases of ESPs. The central phenomenon is addressed by developing and validating the KTSSAF, a practical and theoretically grounded tool designed to evaluate and enhance KT outcomes within the ESP context.

KTSSAF incorporates insights from IS Success Theory, PMPG, and qualitative expert feedback to provide a robust self-assessment mechanism. It empowers organizations to identify KT gaps, anticipate risks, and apply targeted improvements that align with organizational strategy and system capabilities.

Based on the identified gaps and conceptual direction, the following open-ended research questions guide this qualitative inquiry:

RQ1: How do ESP stakeholders perceive and experience the effectiveness of knowledge transfer across different phases of enterprise system implementation?

RQ2: How can a comprehensive self-assessment framework be designed to evaluate and improve knowledge transfer practices in ESPs?

The remainder of this study is organized as follows: Section II reviews relevant literature on ESP, particularly on KT assessment. Section III describes the research methodology, including the study design, rationale, participant selection, data collection methods (literature review and semi-structured interviews), and data analysis approach. Section IV presents the key findings from the literature review and thematic analysis of participant responses, which collectively informed the development of the proposed framework. Section V discusses the results in light of existing research. Finally, Section VI concludes the study by summarizing the main findings, highlighting theoretical and practical implications, and suggesting avenues for future research.

II. Literature Review

A. Introduction

ESP, encompassing systems such as ERP, student information system, and learning management system, depends on the firm's industry, which inherently has knowledge-intensive endeavors which support the assertion made by [16], [17], [18] they require seamless technical, organizational, and procedural knowledge transfer across cross-functional teams. Despite the strategic importance of KT in ESP success, literature reveals a fragmented understanding of measuring and managing KT on ESPs, especially across the different implementation phases, as noted by [19] and [20]. This literature review evaluates key theoretical and practical contributions to KT, assesses their applicability to ESPs, and highlights how the proposed KTSSAF addresses current limitations.

Several scholars have acknowledged that the failure to perform a KT and assess the KT level within the ESP arrangements leads to an unsuccessful ESP [2], [19], [21]. This study will comprehensively review relevant literature on KT within the ESPs to identify the current KT assessment gap within the ESP phases and categorize the KT success assessment factors within the different phases of ESPs to develop a KT self-assessment framework for ESP KT self-assessment success.

B. Definition of Knowledge Transfer

The KT is defined in this study as the process of sharing, disseminating, and applying knowledge in the ESP, both implicit (tacit) (experience-based, non-codified knowledge), similar conclusions were drawn by [22], [23] in contrast, explicit (documented, codified knowledge, as argued by [2], [24] between stakeholders, being part of the ESP in different phases before, during and after the implementation of the ESP in line with the findings of [25], [26], [27]. As observed in the work of [19] effective KT ensures that knowledge is transferred across different project phases, enabling project teams to make intelligent decisions, enhance information system implementation, increase efficiency, and reduce risks associated with the knowledge transfer gap, as previously reported by [2], [28], [29], [30], [31].

1) Scope of knowledge transfer: The scope of the knowledge transfer or knowledge sharing in this study is concentrated on Enterprise System Projects (ESP), focusing on the following features:

2) Included aspects of knowledge transfer:

a) Types of knowledge:

- Explicit Knowledge

(e.g., documented procedures, manuals, best practices, training materials) as defined by [2], [24], [32].

- Implicit (Tacit) Knowledge

(e.g., expertise, skills, lessons learned from experience, informal communication) as reported by [22], [23].

b) Stakeholders involved in KT:

- Project managers, implementation teams, end-users, IT specialists, and consultants according to [33], [34], [35], and [36].

c) Phases of ESP where KT is assessed:

- Pre-implementation, During Implementation, and Post-Implementation phases, as observed in the work of [8] [19].

d) Challenges and barriers to KT:

- Resistance to change [37] Lack of stakeholder engagement [38] high employee turnover [39],

keeping supporting contracts, and poor documentation as identified by [14], [19], [40].

e) KT Assessment Kit:

- It encompasses a collection of instruments and questions developed to facilitate the practical application of assessing KT success level using the proposed framework.

3) Excluded Aspects of Knowledge Transfer:

a) General Knowledge Management (KM):

- The study focuses on knowledge transfer, not the broader concept of knowledge management, which includes creation, storage, and retrieval beyond ESP scope.

b) KT Mechanisms:

- Formal methods: documentation, training, workshops, lessons learned reports.
- Informal methods: mentorship, peer learning, discussions, hands-on experience.

c) Non-Enterprise System Knowledge Transfer:

- KT related to general business operations, non-IT projects, or unrelated industries is outside the scope.

d) Technological Aspects Not Related to KT:

- While ESP implementation involves technology, this study does not focus on the technical development of enterprise systems but on assessing the knowledge transfer level in the ESP.

e) Industry-Specific Knowledge:

- The study aims to provide a general KT framework for ESPs rather than tailoring it to a specific industry.

This research aims to bridge the knowledge transfer gaps by developing a practical and structured framework for assessing and improving KT effectiveness in Enterprise System Projects (ESP).

C. Theoretical Frameworks and Limitations

1) IS Success model: The success of the Information System (IS) followed the foundational theories used to assess multifaceted changes, as previously reported by [41] In the IS. DeLone and McLean's model is extensively used to assess the IS success as observed in the work of [6], [42], [43]. The IS Success Model according to [6] remains one of the most cited frameworks in information systems research. It defines success through six interdependent dimensions: information quality, system quality, service quality, use, user satisfaction, and net benefits [2], [44]. While valuable for post-implementation evaluation, it offers limited insights into the dynamic process of KT throughout ESP phases. Furthermore, it treats knowledge transfer as an outcome or side-effect rather than a process requiring direct assessment, as identified by [45]. Scholars, [45] highlighted that the IS Success model development went through many improvement processes, starting with the Delone and McLean's IS success model [46], the adapted D and M technology acceptance model TAM of Davis, the IS success model of Seddon, and other IS success models. The IS Success Theory was utilized to develop the study framework.

2) Knowledge management lifecycle: Knowledge Management (KM) Lifecycle provides a comprehensive view of how organizations create, store, share, and apply knowledge. While it emphasizes KT as a lifecycle phase as defined by [31], it lacks a project-based perspective, making it less actionable for implementation-specific challenges in ESPs, as argued by [47]. Its theoretical orientation also limits its practical deployment for real-time KT measurement or forecasting, building on the work of [48].

3) Project Management Process Groups (PMPG): The PMBOK Guide outlines five Project Management Process Groups—initiating, planning, executing, monitoring, and closing, according to [49]—which provides a temporal and functional structure for managing ESPs. Although this model supports project phasing, it does not incorporate KT explicitly, nor does it offer tools or metrics to assess the quality of knowledge transfer within each phase as identified by [50]. Its focus remains on processes and deliverables, rather than learning or knowledge exchange.

D. Knowledge Transfer in Enterprise System Gap

Despite the centrality of project-based organizations (PBOs) in delivering enterprise systems (ES), scholars have identified critical deficiencies in how KT is addressed within IT project management, as argued by [13]. A significant concern is the absence of a structured framework to assess KT quality during project execution, as well as the lack of mechanisms to evaluate KT success. Building on the work of [51], this deficiency contributes to significant project risks, including delays, cost overruns, and compromised system quality, factors that ultimately impair organizational performance. Empirical findings reveal that resistance or failure to share knowledge among ES stakeholders adversely affects project outcomes. Although KT is an essential element within the broader knowledge management (KM) life cycle, it remains underdeveloped relative to other KM processes, which include knowledge creation, saving or storing, and application [52]. As such, there is a critical need to identify the determinants of successful KT within ES contexts and to establish robust evaluative models capable of detecting and mitigating KT challenges in real-time. Prior studies emphasize the importance of feedback-driven KT assessment tools, particularly those that operate before, during, and after ES implementation, to ensure effective system design and management, as highlighted by [53]—collaborative efforts between researchers and practitioners. However, findings by [54] further demonstrate the potential of performance metrics to quantify KM efficacy, thereby strengthening the strategic role of KT in ERP success. KT's strategic value has become increasingly prominent in modern, agile enterprises, where information is a key competitive asset; these findings corroborate those of [27] and [41]. However, findings by [55], [56], [57], [58] argue that the persistent recurrence of avoidable errors across ES projects stemming from inadequate KT practices highlights the urgent

need for more systematic approaches. This inefficiency often leads to redundant efforts, increased costs, and suboptimal project performance. Conceptually, KT within ES projects entails exchanging tacit and explicit knowledge among team members to facilitate problem-solving, innovation, and process alignment in line with the findings of [8], [59], [60]. Nevertheless, studies show limited assessment to verify whether KT processes are effectively executed during the implementation phase, before the system goes live, as noted by [61]. Addressing this gap is essential for enhancing ES project outcomes and should be a key focus of future research and practice.

E. Comparative Evaluation of KT Frameworks

The comparative evaluation in Table I systematically reviews seven established KT frameworks in the context of ESP, emphasizing their alignment with key success features such as real-time assessment, quality feedback mechanisms, KM lifecycle integration, ESP customization, empirical validation, and adoption of the IS Success Model.

While foundational models such as [62] and [56] laid early groundwork by identifying core KT success factors, they lack mechanisms for continuous monitoring and do not adopt the IS Success Model, limiting their effectiveness in dynamic and large-scale enterprise environments. Similarly, although [58] the model includes some performance evaluation, but does not offer full integration with real-time assessment or formal IS success criteria.

On the other hand, more recent approaches like [31] [53] stand out for incorporating performance metrics, feedback loops, and partial alignment with the IS Success Model. These features make them more adaptable to ongoing evaluation needs within ESPs.

The KTSSAF (Knowledge Transfer Success and Assessment Framework) offers significant advancement over these models by addressing many of their limitations. Its key strengths include:

- Real-Time Assessment Capability: KTSSAF includes mechanisms to evaluate KT performance at different stages of the ES lifecycle—before, during, and after implementation—allowing organizations to detect issues early.
- Feedback-Driven Quality Control: It features integrated feedback loops to improve KT activities and close the performance gap continuously.
- Alignment with IS Success Model: KTSSAF adopts the DeLone and McLean IS Success Model, enabling measurement of KT impacts on system quality, user satisfaction, and organizational benefits.
- ESP-Specific Customization: Unlike general KM frameworks, KTSSAF is explicitly tailored to enterprise systems, considering the unique challenges and dynamics of ERP implementations.
- Diagnostic and Prescriptive Utility: Beyond just assessment, KTSSAF provides structured guidance on addressing identified KT weaknesses, thus supporting proactive project management.

Frameworks like [27] and [63] contribute valuable insights, particularly in agile and collaborative contexts, but they lack formal assessment components and IS integration—gaps that KTSSAF directly addresses.

In summary, the evaluation underscores the fragmented nature of existing KT frameworks in ESPs, particularly in continuous assessment and strategic alignment. The KTSSAF distinguishes itself by offering a holistic, empirically grounded, and actionable approach to KT management—bridging critical gaps and supporting more successful, knowledge-driven ES project outcomes. Its adoption can help organizations mitigate costly knowledge-related failures and enhance the long-term value of enterprise system investments.

TABLE I. KT FRAMEWORKS COMPARISON

KT Framework / Model	KT Success Factors	Real Time Assessment	Quality feedback mechanism	Alignment with the KM lifecycle	Customization for ESP context	Empirical Validation	IS Success Model Adoption
Bhatt KT Model	✓	✗	✗	✓	✗	✓	✗
Desouza & Evaristo KT Strategy	✓	✗	✗	✓	✓	✓	✗
Landaeta's KT Performance Model	✓	✗	✓	✓	✓	✓	✗
Parry & Graves KT Metrics Approach	✓	✓	✓	✓	✓	✓	✓
Samiei & Habibi KM-Evaluation Model	✓	✓	✓	✓	✓	✓	
Shaqrah & Maliki Agile KT Framework	✓	✓	✗	✓	✓	✗	✗
Gomes et al. Collaborative KT Model	✓	✓	✗	✓	✓	✗	✗
Proposed KTSSAF	✓	✓	✓	✓	✓	✓	✓

Legend: ✓ = Feature Present ✗ = Feature Not Present

TABLE II. KT MODEL COMPARISON

Model / Study	ESP Phases & Sub-Phases	ESP Roles	ESP Knowledge Types	ESP Knowledge Clusters	KT Sub-Clusters	KT Assessment Kit
IS Success Model	✗	✗	✗	✗	✗	✓
KM Lifecycle	✗	✗	✓	✗	✗	✗
PMPG	✓	✓	✓	✗	✗	✗
KTSSAF (Proposed Model)	✓	✓	✓	✓	✓	✓

Legend: ✓ = Feature Present ✗ = Feature Not Present.

The following comparative table offers a critical assessment of the most relevant KT-related models, comparing them on criteria such as focus, validation, phase relevance, and applicability to ESPs. Table II illustrates the unique value proposition of KTSSAF. While traditional models provide foundational guidance, KTSSAF extends functionality with phase-specific metrics and assessment instruments, ensuring comprehensive KT evaluation across ESPs.

This provides a focused comparison of KT frameworks based on six ESP-specific dimensions critical for effective KT in ESPs. The analysis shows that most existing frameworks, including the IS Success Model and KM Lifecycle, fail to address the practical complexities of ESP environments. While the PMPG framework covers project phases, roles, and knowledge types, it lacks the depth required for managing knowledge clusters and sub-clusters.

In contrast, the KTSSAF (proposed framework) demonstrates comprehensive coverage across all six dimensions. It uniquely incorporates ESP phases, roles, and macro-level knowledge clusters and micro-level sub-clusters, reflecting the real-world structure of enterprise systems. Moreover, KTSSAF integrates a KT assessment kit, allowing for continuous monitoring and evaluation of knowledge flow—something absent in other models.

This broad coverage enhances KTSSAF's practical relevance and diagnostic capability, positioning it as a robust tool for improving KT effectiveness and minimizing project risks in complex ESP implementations. Its alignment with operational realities makes it especially suitable for dynamic, role-based, and knowledge-intensive enterprise settings.

KTSSAF, as illustrated in Table II, is the only model in this comparison that explicitly integrates knowledge clusters and sub-clusters, enabling granular evaluation of KT quality across different types of knowledge: technological, organizational, individual, and procedural.

F. Emerging Research on KT and ESPs

Recent contributions have attempted to narrow the gap in KT understanding within ESPs. The scholars [20] conducted a systematic review of knowledge integration strategies in ERP projects, focusing primarily on outcomes rather than ongoing assessment. Similarly, scholars in [13] proposed a set of KT performance indicators, yet their practical deployment remains untested. These studies affirm the increasing scholarly interest in KT metrics but also reflect a lack of operational frameworks that guide practitioners through implementation.

Furthermore, technologies such as AI, augmented reality, and blockchain are being explored as future enablers of KT, especially in capturing and transferring tacit knowledge, as noted by [50] advocate for integrated KT feedback systems to support real-time decision-making in digital transformation projects. KTSSAF aligns with this trend by offering modular assessment kits that could be integrated with project dashboards or knowledge systems.

G. The Critical Role of Tacit Knowledge

Tacit knowledge, though inherently intangible and informally exchanged, plays a foundational role in the success of ESP. It encompasses personal insights, experiential learning, problem-solving capabilities, and strategic intuition, often transmitted through interpersonal communication rather than formal documentation, as identified by [64], [5]. Despite its significance, most KT frameworks neglect or insufficiently account for tacit knowledge. The KTSSAF framework addresses this gap by embedding tools that assess mentorship and peer coaching, simulation-based learning, storytelling, knowledge mapping, and unstructured communication during project implementation. These elements make KTSSAF especially effective in complex ESP environments, where informal knowledge sharing has a substantial impact on system adoption and operational resilience.

A notable innovation within KTSSAF is its structured classification of knowledge into distinct clusters and sub-clusters, providing greater analytical granularity than traditional models. These include Organizational Knowledge (e.g., governance structures and strategic alignment), People Knowledge (e.g., user roles, leadership dynamics), Technology Knowledge (e.g., system architecture and usability), Process Knowledge (e.g., workflow integration and regulatory compliance), and Transfer Mechanisms (e.g., formal training sessions and informal exchanges). Each cluster is evaluated across different phases of the ESP lifecycle using weighted metrics, resulting in a multi-dimensional assessment of KT effectiveness. This classification level offers a more nuanced and actionable understanding of KT dynamics. Furthermore, a visual representation of KTSSAF's capabilities reinforces its comprehensiveness and practical utility, positioning it as a robust tool for addressing the multifaceted challenges inherent in real-world enterprise system implementations.

III. METHODOLOGY

A. Research Design

This study adopts a qualitative constructivist approach as rooted in phenomenology by [65], [36], [66]. To explore how KT is experienced and perceived by experts within the context of ESP. The constructivist paradigm emphasizes that knowledge is co-constructed through interactions between the researcher

and participants, making it suitable for understanding KT's nuanced, context-dependent nature aligns with [67] [68].

As the chosen methodological tradition, phenomenology focuses on individuals' lived experiences and seeks to uncover how these experiences shape their understanding of complex organizational phenomena like KT outlined by [69], [70]. This aligns with the study's goal of developing the KTSSAF based on real-world insights.

B. Methodological Rationale

The rationale for using a qualitative, constructivist, and phenomenological design stems from the need to:

- Explore the subjective experiences of stakeholders involved in KT across ESP phases.
- Capture tacit knowledge and interpretive insights that cannot be quantified meaningfully through surveys or experimental methods were found in the work of [64].
- Enable a bottom-up development of the KTSSAF through emergent insights from expert narratives, rather than imposing a top-down model.

This design facilitates the systematic exploration and validation of the KTSSAF by grounding it in practical realities, bridging the gap between theory and implementation [7].

C. Participant Sampling and Selection

A purposive sampling strategy was adopted to ensure participants possessed specialized knowledge relevant to the research focus [71]. The study engaged seven expert practitioners from varied ESP backgrounds, including roles such as project managers, ERP consultants, solution architects, and change management leads.

Inclusion criteria included:

1) Minimum eight years of experience in managing or supporting ESPs.
2) Direct involvement in at least two project phases (pre-, during-, post-implementation).
3) Experience working in large or cross-functional project teams.
4) Ability and willingness to articulate reflective insights through in-depth interviews.

This sampling approach ensured rich, diverse, and relevant data aligned with the phenomenological focus on lived experience.

D. Data Collection Procedures

The semi-structured interviews were employed as a data collection technique, a method well-regarded for eliciting deep, reflective insights in qualitative research. Methodologically, this follows [72]. This approach provided the flexibility to explore the unique experiences of each participant while maintaining consistency in addressing the research questions. Interviews were conducted through online meetings using Google Meet or Microsoft Teams, depending on participant availability and geographic location. Each session lasted between 45 and 75 minutes.

In our data collection, we focused on recording all interviews [73] with prior informed consent, and subsequently transcribed verbatim for qualitative analysis. This ensured an accurate and comprehensive analysis of the expert feedback.

To ensure alignment with the study's three research questions, the interview protocol was carefully constructed to elicit detailed narratives related to the phenomenon under investigation:

For RQ1 (How do ESP stakeholders perceive and experience the effectiveness of knowledge transfer across different phases of enterprise system implementation?), participants were asked to reflect on their knowledge-sharing experiences during the pre-implementation, implementation, and post-implementation phases of ESPs. Questions probed their perceptions of what worked, what failed, and how knowledge flowed—or was unable to flow—across project milestones.

For RQ2 (How can a comprehensive self-assessment framework be designed to evaluate and improve knowledge transfer practices in ESPs?), participants were introduced to a preliminary version of the KT Success Self-Assessment Framework (KTSSAF). They were asked to provide feedback on the framework components' relevance, clarity, and usability. This ensured that the framework's refinement was grounded in practitioner experience.

Topics explored in the interview included, but were not limited to:

- Experiences with knowledge handover across the ESP lifecycle.
- Challenges associated with tacit knowledge sharing.
- Observed organizational, technical, and interpersonal factors impacting KT.
- Evaluation of and feedback on the proposed KTSSAF model.

Member checking was conducted to enhance the credibility and trustworthiness of the findings. Transcripts were returned to participants for validation and correction, ensuring their meanings and perspectives were accurately captured in line with the findings of [74]. This structured yet flexible data collection procedure ensured rich, detailed, and directly relevant data to answer each research question effectively, in line with the study's qualitative constructivist methodology.

E. Data Analysis

Data were analyzed using qualitative content analysis, chosen for its suitability in systematically organizing and interpreting textual data as rooted in [75]. The process followed an inductive logic, consistent with constructivist inquiry.

Step-by-step coding process:

Familiarization – Thorough reading of transcripts to understand context.

Open Coding – Labelling key concepts, phrases, and meanings line-by-line.

Axial Coding – Grouping related codes into broader categories (e.g., "Consultant Credibility", "Knowledge Gaps").

Thematic Structuring – Synthesizing categories into overarching themes relevant to KT effectiveness.

Mapping to ESP Phases – Organizing themes by project phase (pre-, during-, post-implementation) to enhance the framework's utility.

Emergent themes were triangulated with literature to enhance theoretical grounding and confirm conceptual coverage of the KTSSAF.

F. Methodological Foundations and Empirical Rigour

The development of KTSSAF followed a qualitative, constructivist research approach involving in-depth interviews following the model proposed by [76], [77], [78] with subject matter experts and practitioners from diverse ESP contexts. This methodology enabled the identification of context-specific KT assessment success factors and ensured that the framework was grounded in real-world ESP dynamics. Data analysis was conducted using content analysis according to [79], allowing for the systematic categorization of KT Success factors in ESP, the development of the proposed framework, and the validation of the proposed framework.

The model was validated through expert feedback, pilot testing, and iterative refinement. Table III summarizes the research method applied in this study. Participants acknowledged the model's relevance, clarity, and adaptability across different industries and project scales.

The ESPs expert appreciated its ability to provide actionable insights, such as identifying weak knowledge links in specific project phases or stakeholder roles or specific knowledge types, in specific impacted outcomes of the successful knowledge transfer, including individual impact, organization impact, information quality, and system quality.

TABLE III. APPLIED RESEARCH METHOD

Aspect	Selected for Study	Reason for selection
Research Questions	Exploratory	To explore the literature on KT in ESP and investigate the perception and experience of the effectiveness of KT across different phases of ESP.
Research Approaches	Qualitative	To gather in-depth insights from experts and understand the context and nuances of KT within ESP projects.
Data collection methods	Semi-Structured Interviews	Allows flexibility to explore specific expert opinions while ensuring coverage of predefined themes.
Data Analysis Techniques	Content Analysis	Enables a systematic and objective categorization of qualitative data, allowing identification and quantification of factors influencing KT success in ESP.

IV. FINDINGS

This study utilize a qualitative, constructivist approach to examine the critical dimensions of KT within ESPs. Through rigorous content analysis of expert interviews, key experiential themes and insights were synthesized, addressing the two research questions and contributing to validating and refining the KTSSAF

A. KTSSAF – Bridging the Gaps

The KTSSAF developed in this study offers an integrative solution. It combines elements of IS Success Theory, KM lifecycle, and PMPG while introducing KT-specific metrics and instruments to assess knowledge flow across ESP phases. KTSSAF includes a validated assessment kit, a scoring model, and customizable instruments supporting explicit and tacit knowledge evaluation. It is empirically grounded through interviews and pilot studies, ensuring theoretical robustness and practical usability. Fig. 1 illustrates the proposed KTSSAF visualization.

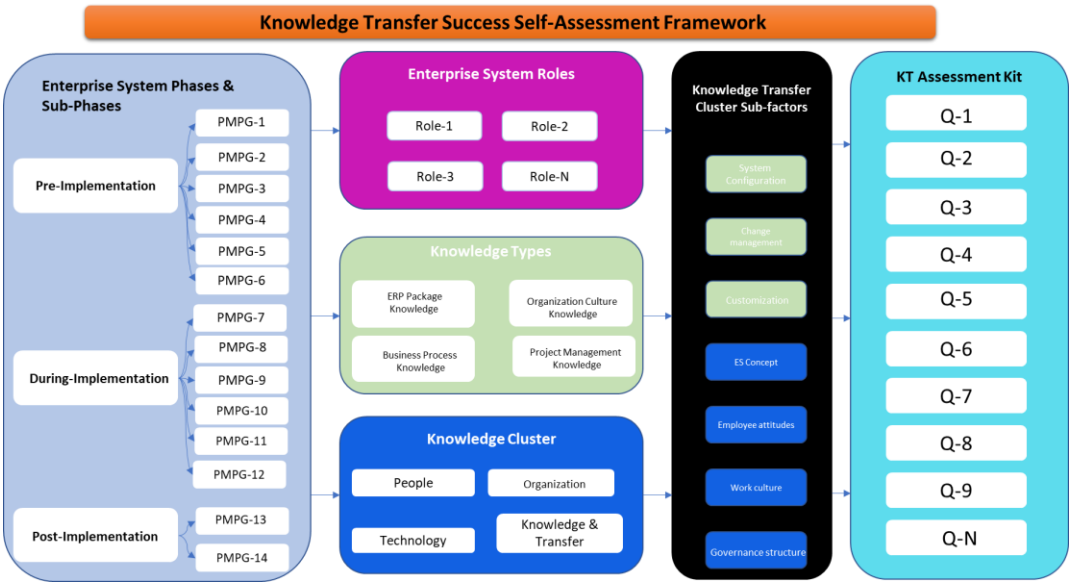


Fig. 1. Knowledge transfer success self-assessment framework- KTSSAF.

B. Emergent Themes from Expert Interviews

The semi-structured interviews with seven ESP professionals, each possessing over eight years of relevant experience, generated a rich dataset. Using a five-point Likert scale and interpretive thematic analysis, three major thematic clusters emerged:

1) Confirmation of theoretical KT success factors: Experts consistently affirmed the relevance of nine critical KT success factors initially identified in the literature review. These included:

Knowledge Sharing Culture, Knowledge Content Quality, Consulting Expertise, Management Support, Consultant Credibility, Absorptive Capacity, Project Owner Leadership, Consultant Ability, and Incentives.

Except for a single neutral response regarding incentives, there was a unanimous agreement on the importance of these factors. This validates the theoretical foundation of KT in ESPs and supports their integration into the proposed KTSSAF.

"Leadership and consultant credibility make or break the knowledge transition process." (Interviewee 3, IT Consultant).

2) Validation of KTSSAF independent variables: Experts strongly endorsed the independent variables embedded within the KTSSAF, which include:

- The ESP phases (pre-, during, and post-implementation)
- Project Management Process Groups (PMPG)
- Role responsibilities of ESP stakeholders

Four types of knowledge: business, project management, organizational culture, and technology.

The Knowledge Clusters: Knowledge and Transfer, People, Technology, and Organization. Participants confirmed that these dimensions comprehensively capture the complexity of KT within ESPs, facilitating a structured, contextualized assessment of success levels. "Each knowledge cluster brings a different layer of understanding — without this multidimensionality, we'd be blind to root causes of KT failure." (Interviewee 6, Project Manager)

C. Evaluation of KTSSAF Structure and Predictive Utility

Experts evaluated the KTSSAF across five analytical dimensions: variable weight, structure, clarity, consistency, and diagnostic utility. Feedback affirmed the framework's robust design and its capacity to:

- Predict KT success across PMPGs and implementation phases.
- Evaluate KT contributions by role and scenario.
- Maintain a clear, consistent structure and interpretability.

Moreover, the model's adaptability, configurability, and modularity were emphasized, with participants acknowledging its reusability across varied ESP contexts. "The model is precise, modular, and predictive — it goes beyond theory and meets practice head-on." (Interviewee 2, Solution Architect).

D. Conceptual Framework Integration

The findings affirm the constructivist underpinnings of the KTSSAF. They reflect a high alignment between expert experiential narratives and the conceptual framework developed through prior literature and theoretical synthesis. The study captured the experiential essence of KT processes in ESPs, revealing how actors interpret, navigate, and evaluate KT success in dynamic project environments.

Emergent conceptual frameworks were substantiated by interpretive feedback, highlighting the dynamic interactions among ESP phases, knowledge types, and stakeholder roles. Experts emphasized that KT is not a static event but an evolving process that requires contextual awareness, iterative assessment, and leadership engagement at all stages of the ESP lifecycle.

E. Contributions to Practice and Theory

This empirical inquiry makes several key contributions to both theory and practice. It confirms the validity of theoretical constructs previously derived from literature synthesis, thereby reinforcing the conceptual foundation of the study. Expert feedback further refined the framework, enhancing its practical applicability in real-world enterprise system projects. In addition, the study establishes KTSSAF as a diagnostic and planning tool that can support enterprise system project managers and consultants in assessing and managing knowledge transfer success. The research advances scholarly understanding and practical usability by codifying tacit insights into an operational framework. The interpretive, thematic analysis highlights the multi-dimensional and actor-centered nature of knowledge transfer within enterprise system projects, underscoring the value of KTSSAF as an evidence-based instrument for improving knowledge management practices. KTSSAF offers multiple contributions to both academic literature and industry practice. Theoretically, it advances the discourse on KT by shifting focus from knowledge creation and storage to measurable and actionable KT assessment. It also enriches understanding of KT dynamics in project-based settings, particularly within the underexplored ESP domain. Practically, KTSSAF equips project managers, knowledge officers, and organizational leaders with a structured tool to forecast KT success, mitigate knowledge risks, and enhance project outcomes. It also facilitates longitudinal learning, enabling organizations to retain and reuse critical knowledge across successive ESP implementations—a key determinant of digital resilience and competitive advantage, as noted by [80].

Moreover, the KTSSAF framework aligns with emerging trends in digital project management, where predictive analytics, continuous learning, and evidence-based interventions are becoming the norm. By integrating KT metrics into the broader ESP lifecycle, KTSSAF bridges the gap between theoretical aspirations and practical execution.

V. Discussion

This study, guided by two core research questions, explored the KT assessment success factors and conceptual underpinnings of KT in ESP. Through a constructivist, qualitative inquiry employing semi-structured interviews with ESP experts, this research provides a grounded assessment of KT practices, culminating in validating the KTSSAF. This

section discusses the findings about the research questions, previous scholarship, and broader implications for the field.

A. Addressing the Research Questions

RQ1: How do ESP stakeholders perceive and experience the effectiveness of knowledge transfer across different phases of enterprise system implementation?

Our findings reveal that stakeholder perceptions are deeply influenced by contextual and role-based factors that span the ESP lifecycle. Experts affirmed that KT effectiveness varies considerably across the pre-implementation, implementation, and post-implementation phases. [81]. These phases are further shaped by stakeholder roles, project structures, and knowledge types, confirming assertions by [54] that successful KT in ESPs requires tailoring strategies to temporal and contextual project dimensions. Respondents emphasized the necessity of sustained management support and consultant credibility across all phases—elements that enhance KT effectiveness through trust, relevance, and structure. These insights extend prior literature by providing empirical validation that KT is not a singular activity but a phased, adaptive process requiring distinct engagement and measurement strategies throughout the ESP lifecycle, as noted by [82].

RQ2: How can a comprehensive self-assessment framework be designed to evaluate and improve knowledge transfer practices in ESPs?

It is evident from our results that the KTSSAF was confirmed through expert validation as a valuable, structured tool for assessing KT success. The framework's inclusion of ESP phases, project roles, PMPGs, knowledge types, and knowledge clusters enables a multidimensional understanding of KT performance across scenarios. This contribution moves beyond the limitations of one-size-fits-all models by offering a configurable, modular, and predictive self-assessment mechanism, a novelty not extensively explored in prior KT research. While earlier studies (e.g., [83]). The KTSSAF offers typologies and guidelines, providing an operational and diagnostic instrument that stakeholders can apply across ESP contexts.

B. Integration with Literature and Theoretical Contributions

Grounded in a constructivist, phenomenological paradigm and guided by content analysis, this study demonstrates methodological rigor and epistemological coherence. It contributes to the expanding body of qualitative KT research, complementing prior studies that advocate for interpretive approaches in information systems and organizational knowledge research in line with the findings of [84], [85]. Additionally, it responds to longstanding calls within the knowledge management (KM) literature for deeper, context-sensitive explorations of how knowledge is created, shared, and applied in practice as observed in the work of [86]. Our analysis highlights three key theoretical contributions. First, it recontextualizes general KT theories by situating them within ESP's temporally bounded and role-driven environment, enhancing their relevance and applicability. Second, it expands the scope of KT evaluation by shifting the analytical lens from outcome-based project success metrics in the knowledge-centered assessment across multiple ESP phases. Third, it advances the operationalization of tacit knowledge by embedding constructs such as absorptive capacity and consultant credibility into the KTSSAF framework, addressing a notable gap in practical KM models and contributing to more holistic assessments of knowledge transfer effectiveness.

C. New Understandings and Practical Implications

It is evident from our results that several novel insights were illuminated. Notably, it underscores how the sequence and timing of KT activities, not merely their content or structure, critically influence outcomes. Identifying PMPGs as dynamic enablers for phased KT assessments represents a refinement in understanding how knowledge and process maturity evolve together during ESPs. Practically, the KTSSAF offers organizations a self-assessment toolkit to monitor and refine KT efforts in real time. The KTSSAF configurability and modularity make it applicable across diverse enterprise contexts—from public universities to multinational corporations—enhancing its strategic relevance.

D. Advantages of the Proposed Method (KTSSAF)

The proposed KTSSAF offers several key advantages over existing models in the field of knowledge transfer within enterprise system projects. First, unlike many traditional KT models that focus generically on knowledge sharing practices, KTSSAF is specifically tailored to the structured phases of ESPs, aligning with the Project Management Process Groups (PMPG) to enable granular, phase-specific assessment of KT effectiveness. Second, it is grounded in Information Systems (IS) Success Theory, which provides a robust theoretical foundation that links KT outcomes directly to system success metrics such as user satisfaction, system use, and organizational impact.

Furthermore, KTSSAF adopts a qualitative constructivist approach, integrating insights from domain experts and practical field data, which ensures high contextual relevance and real-world applicability—something often lacking in quantitatively rigid or one-size-fits-all models. The inclusion of a scoring mechanism and customizable assessment kit distinguishes KTSSAF by offering actionable diagnostics and predictive insights for stakeholders, enabling organizations to proactively identify KT gaps and implement targeted improvements. Finally, in our analysis, the framework's empirical validation through expert reviews and pilot applications demonstrates its practical utility, repeatability, and adaptability across different enterprise contexts—advantages not commonly found in more theoretical or static KT models.

E. Limitations and Future Research Directions

Despite the KTSSAF contributions, the purposive sample, though composed of highly experienced ESP experts, may limit contextually bound information and may not capture all regional or industry-specific nuances.

We acknowledge that this study is limited by generalizability. The interviews, while deep and revealing future research that should:

- Expand the sample base across industries and geographic regions to test KTSSAF's generalizability.

- Explore longitudinal application of KTSSAF to monitor KT over time and refine the framework based on evolving practices.
- Develop automated assessment tools incorporating the KTSSAF for real-time feedback and benchmarking.
- Investigate how organizational learning loops emerge post-assessment and how KTSSAF results can inform continuous improvement in enterprise environments.

This study advances the theoretical, empirical, and practical understanding of knowledge transfer in ESPs. Offering a validated, multidimensional assessment framework rooted in expert experiences and grounded theory lays the foundation for more responsive, contextual, and effective KT strategies in complex organizational transformations.

VI. Conclusion

The academic and practitioner domains of ESP offered a timely and significant contribution using the developed KTSSAF. While the ESPs continue to increase in scale and complexity, the evaluation of the KT management and effective KT within the ESPs lifecycle is classified as a strategic challenge. The proposed KTSSAF addresses this need by operationalising KT assessment into a structured, phase-sensitive, modular framework aligned with the IS theory, KM, PMP, and PMPGs.

By integrating theoretical insights from IS Success Theory, KM, and PMP, KTSSAF ensures that KT assessment is methodologically sound and aligned with broader ESP outcomes such as system quality, user satisfaction, information quality, organization performance foster, and knowledge sustainability. Incorporating knowledge clusters—organizational, people, technology, process, and transfer mechanisms—enables a granular and multidimensional understanding of how knowledge flows within ESPs. We conclude that the framework advances the field by explicitly recognizing and assessing tacit knowledge, long considered one of the most challenging dimensions of KT.

Furthermore, our results provide evidence that KTSSAF's modular assessment kits, diagnostic indicators, and feedback mechanisms empower organizations to proactively identify KT gaps and take corrective action during ESPs implementation, rather than after project implementation. Its customizability ensures relevance across different industries, organizational sizes, and technological environments.

The framework has demonstrated practical utility through pilot testing and empirical validation, leading to measurable improvements in KT performance. As a result, KTSSAF transforms KT from an abstract managerial concern into a strategic, measurable, and actionable component of enterprise system success. Compared to existing models, KTSSAF provides a more practical, phase-specific, and empirically validated approach to evaluating KT in ESPs.

As organizations navigate digital transformation, KTSSAF offers a roadmap for ensuring that knowledge is not a bottleneck but a key enabler of ESP implementation success, system adoption, and sustained value creation. Future research can extend the framework's application across cross-cultural contexts, integrate it with AI-driven knowledge systems, or tailor it for agile and hybrid ESP methodologies, further reinforcing its relevance in dynamic, knowledge-intensive environments.

Acknowledgment

I gratefully acknowledge all of those whose efforts and support made this work possible, and special thanks goes to the ESP experts who shared their insights during the interview sessions.

Disclosure and Conflict of Interest

The author declares that there are no conflicts of interest in this research. In addition, there are no financial interests or competing affiliations that could have influenced the study's design, execution, or findings. This manuscript is the author's original work and has not been previously published or submitted for review to any other journal or conference.

References

[1] G. Rehm et al., "QURATOR: Innovative technologies for content and data curation," CEUR Workshop Proc, vol. 2535, pp. 1–15, 2020.

[2] A. M. Diane B. Walz, "The Effect of Conflict and Knowledge Sharing on the Information Technology Project Team Performance," Journal of Knowledge Management, vol. 12, no. 4, pp. 1–12, 2019, doi: 10.1108/13673270210450450.

[3] A. K. Verma, V. Tanwar, and B. Mago, "Analyzing Factors for the Successful Implementation of Enterprise Resource Planning System in UAE Organizations," vol. 21, no. 2, pp. 85–105, 2022.

[4] R. M. Grant, "Toward a Knowledge-Based Theory of the Firm," Strategic Management Journal, 1996, doi: 10.1002/smj.4250171110.

[5] I. Nonaka and H. Takeuchi, "The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation," null, 1995, doi: null.

[6] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A ten-year update," Journal of Management Information Systems, vol. 19, no. 4, pp. 9–30, 2003, doi: 10.1080/07421222.2003.11045748.

[7] J. M. Hussien, R. bin Sulaiman, A. H. Hassan, M. Abdulhak, and H. Kahtan, "Critical Success Factors for Knowledge Transfer in Enterprise System Projects: A Theoretical and Empirical Investigation," International Journal of Advanced Computer Science and Applications, vol. 16, no. 6, pp. 594–600, 2025, doi: 10.14569/IJACSA.2025.0160658.

[8] J. Hussien, M. A. Abdulgabber, H. Kahtan, and R. Sulaiman, "THE Success of Enterprise System Projects: From a Knowledge Transfer Perspective," Journal of Information System and Technology Management (JISTM), Vol. 6, no. 22, pp. 128–147, 2021, doi: 10.35631/JISTM.622011.

[9] M. Y. Qianwen Zhou, Xiaopeng Deng, Bon-Gang Hwang, "System dynamics approach of knowledge transfer from projects to the project-based organization | Emerald Insight." Accessed: Sep. 05, 2023. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/IJMPB-06-2021-0142/full/html

[10] S. Chatterjee, R. Chaudhuri, A. Thrassou, and D. Vrontis, "Antecedents and consequences of knowledge hiding: The moderating role of knowledge hiders and knowledge seekers in organizations," J Bus Res, vol. 128, pp. 303–313, 2021.

[11] M. Aranyossy, B. Blaskovics, and Á. A. Horváth, "How universal are IT project success and failure factors? Evidence from Hungary," Information Systems Management, vol. 35, no. 1, pp. 15–28, 2018, doi: 10.1080/10580530.2017.1416943.

[12] B. Cartaxo, G. Pinto, E. Vieira, and S. Soares, "Evidence Briefings: Towards a Medium to Transfer Knowledge from Systematic Reviews to

Practitioners," International Symposium on Empirical Software Engineering and Measurement, vol. 08-09-Sept, 2016, doi: 10.1145/2961111.2962603.

[13] R. Stampfl, J. Fischer, and S. Palkovits-Rauter, "Forms of Knowledge Transfer in IT Project Management," MAP Social Sciences, no. September, pp. 53–66, 2023, doi: 10.53880/2744-2454.2023.4.53.

[14] S. Sanz, K. Haupt, A. Maas, R. Jober, and B. Prescott, "Optimization of knowledge transfer in ITER," Fusion Engineering and Design, vol. 146, no. October 2018, pp. 1385–1389, 2019, doi: 10.1016/j.fusengdes.2019.02.088.

[15] A. Anand, B. Muskat, A. Creed, A. Zutshi, and A. Csepregi, "Knowledge sharing, knowledge transfer and SMEs: evolution, antecedents, outcomes and directions," Personnel Review, vol. 50, no. 9, pp. 1873–1893, 2020, doi: 10.1108/PR-05-2020-0372.

[16] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," The Lancet, vol. 395, no. 10223, pp. 497–506, 2020, doi: 10.1016/S0140-6736(20)30183-5.

[17] O. AlShamsi and M. Ajmal, "Critical factors for knowledge sharing in technology-intensive organizations: evidence from UAE service sector," Journal of Knowledge Management, vol. 22, no. 2, pp. 384–412, 2018, doi: 10.1108/JKM-05-2017-0181.

[18] V. Ghosh, G. Kabra, and H. S. Mukerjee, "Influence of knowledge leadership on it project performance and quality practices: examining the role of leader risk-mitigation efforts," International Journal of Knowledge Management, vol. 18, no. 1, pp. 1–20, 2022, doi: 10.4018/IJKM.290024.

[19] J. M. Hussien, M. Abdulhak, H. Kahtan, and R. Sulaiman, "Revisiting Knowledge Transfer for Success Enterprise System Project," in 2021 International Conference on Information Technology (ICIT) (ICIT2021), Amman, Jordan, Jul. 2021.

[20] Q. Zhou, X. Deng, G. Wang, and A. Mahmoudi, "Linking elements to outcomes of knowledge transfer in the project environment: Current review and future direction," Frontiers of Engineering Management, vol. 9, no. 2, pp. 221–238, 2022, doi: 10.1007/s42524-022-0195-3.

[21] S. Banerjee, E. J. Jaselskis, and A. F. Alsharef, "Design For Six Sigma (DFSS) Approach for Creating CLEAR Lessons Learned Database," Periodica Polytechnica Architecture, vol. 51, no. 1, pp. 75–82, 2020, doi: 10.3311/ppar.15442.

[22] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "KAT: A Knowledge Augmented Transformer for Vision-and-Language," NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 956–968, 2022, doi: 10.18653/v1/2022.naacl-main.70.

[23] T. Febrianto and D. Soediantono, "Enterprise resource planning (ERP) and implementation suggestion to the defense industry: a literature review," Journal of Industrial Engineering \& Management Research, vol. 3, no. 3, pp. 1–16, 2022.

[24] D. Choudhury and P. Das, "The Influence of Organisational Climate on Knowledge Management : A Literature Review," vol. I, pp. 127–143, 2021.

[25] U. Jayawickrama, S. Liu, and M. H. Smith, "An Integrative Knowledge Management Framework to Support ERP Implementation for Improved Management Decision Making in Industry," null, 2012, doi: 10.1007/978-3-642-41077-2_7.

[26] S. Pillai, "A decision support system for ERP projects in make-to-order manufacturing SMEs," p. 205, 2015.

[27] A. A. Shaqrah and M. R. Al Maliki, "Examining Tacit Knowledge Transfer Processes for Enterprise System Projects Success: A conceptual framework," Journal of Systems Integration, vol. 9, no. 4, pp. 29–36, 2018, doi: 10.20470/jsi.v9i4.355.

[28] K. Yang, S. Member, T. Jiang, S. Member, and Y. Shi, "Federated Learning via Over-the-Air Computation," IEEE Trans Wirel Commun, vol. PP, no. c, p. 1, 2019, doi: 10.1109/TWC.2019.2961673.

[29] A. T. Adetunji and F. E. Oyekan, "Knowledge Sharing: Constraint To Effective Human Capital Development Among Local Government Staff in Osun State," Int J Adv Res (Indore), vol. 7, no. 9, pp. 1–21, 2018.

[30] A. A. Shaqrah and M. R. Al Maliki, "Examining tacit knowledge transfer processes for enterprise system projects success: a conceptual framework," Journal of Systems Integration, vol. 9, no. 4, pp. 29–39, 2018.

[31] M. Heredia-Calzado and A. Duréndez, "The influence of knowledge management and professionalization on the use of ERP systems and its effect on the competitive advantages of SMEs," Enterp Inf Syst, vol. 13, no. 9, pp. 1245–1274, 2019.

[32] M. Terzieva, "Project Knowledge Management : how organizations learn from experience," Procedia Technology, vol. 16, pp. 1086–1095, 2014, doi: 10.1016/j.protcy.2014.10.123.

[33] Y. Karagoz, N. Whiteside, and A. Korthaus, "Context matters: enablers and barriers to knowledge sharing in Australian public sector ICT projects," Journal of Knowledge Management, vol. 24, no. 8, pp. 1921–1941, 2020, doi: 10.1108/JKM-12-2019-0691.

[34] C. Barth and S. Koch, "Critical success factors in ERP upgrade projects," Industrial Management and Data Systems, vol. 119, no. 3, pp. 656–675, 2019, doi: 10.1108/IMDS-01-2018-0016.

[35] E. Weishäupl, E. Yasasin, and G. Schryen, "Information security investments: An exploratory multiple case study on decision-making, evaluation and learning," Comput Secur, vol. 77, pp. 807–823, 2018.

[36] M. O. Malik and N. Khan, "Analysis of ERP implementation to develop a strategy for its success in developing countries," Production Planning and Control, vol. 0, no. 0, pp. 1–16, 2020, doi: 10.1080/09537287.2020.1784481.

[37] S. Bag, A. Telukdarie, J. H. C. Pretorius, and S. Gupta, "Industry 4.0 and supply chain sustainability: framework and future research directions," Benchmarking: An International Journal, vol. 28, no. 5, pp. 1410–1450, 2021.

[38] R. Miković, D. Petrović, M. Mihić, V. Obradović, and M. Todorović, "The integration of social capital and knowledge management – The key challenge for international development and cooperation projects of nonprofit organizations," International Journal of Project Management, no. xxxx, 2020, doi: 10.1016/j.ijproman.2020.07.006.

[39] D. W. DeLong, Lost Knowledge: Confronting the Threat of an Aging Workforce. Oxford University Press, 2007. doi: 10.1093/acprof:oso/9780195170979.001.0001.

[40] M. L. Todorović, D. T. Petrović, M. M. Mihić, V. L. Obradović, and S. D. Bushuyev, "Project success analysis framework: A knowledge-based approach in project management," International Journal of Project Management, vol. 33, no. 4, pp. 772–783, 2015, doi: 10.1016/j.ijproman.2014.10.009.

[41] R. Naceur, Y. Cimon, and R. Pellerin, "A systematic review for predictive models of IS adoption," International Journal of Enterprise Information Systems, vol. 17, no. 1, pp. 1–21, 2021, doi: 10.4018/IJEIS.2021010101.

[42] J. H. Wu and Y. M. Wang, "Measuring KMS success: A respecification of the DeLone and McLean's model," Information and Management, vol. 43, no. 6, pp. 728–739, 2006, doi: 10.1016/j.im.2006.05.002.

[43] A. Adeniran et al., "Knowledge Management Competence for Enterprise System Success," Theoretical and Applied Genetics, vol. 7, no. 2, pp. 1–7, 2010.

[44] S. Suo, Cloud implementation in organizations: Critical success factors, challenges, and impacts on the it function. The Pennsylvania State University, 2013.

[45] T. D. Nguyen, T. M. Nguyen, and T. H. Cao, "Information systems success: A literature review," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9446, no. November, pp. 242–256, 2015, doi: 10.1007/978-3-319-26135-5_18.

[46] H. J. Mun, H. Yun, E. A. Kim, J. Y. Hong, and C. C. Lee, "Research on factors influencing intention to use DMB using extended IS success model," Information Technology and Management, vol. 11, no. 3, pp. 143–155, 2010, doi: 10.1007/s10799-010-0073-2.

[47] U. Jayawickrama, S. Liu, M. Hudson Smith, P. Akhtar, and M. Al Bashir, "Knowledge retention in ERP implementations: the context of UK SMEs," Production Planning and Control, vol. 30, no. 10–12, pp. 1032–1047, 2019, doi: 10.1080/09537287.2019.1582107.

[48] M. Ghafoor, S. Nawab, and K. Shafi, "Enterprise Resource Planning Systems and Knowledge Management: A Review of the Literature and Conceptual Framework," Journal of Management and Research, 2021, doi: 10.29145/jmr/81/080108.

[49] O. Haass and N. Azizi, "Knowledge sharing practice in project-oriented organisations: a practical framework based on project life cycle and project management body of knowledge," International Journal of Project Organisation and Management, vol. 11, no. 2, pp. 171–197, Jan. 2019, doi: 10.1504/IJPOM.2019.100575.

[50] S. S. M. Yuen and H. Y. Lam, "Enhancing Competitiveness through Strategic Knowledge Sharing as a Driver of Innovation Capability and Performance," Sustainability, vol. 16, no. 6, p. 2460, 2024, doi: 10.3390/su16062460.

[51] K. Alhassan, Tacit Knowledge Sharing in Software Development Projects of Mobile Applications Domain. Northcentral University, 2022.

[52] E. Grimsdottir and I. R. Edvardsson, "Knowledge management, knowledge creation, and open innovation in Icelandic SMEs," Sage Open, vol. 8, no. 4, p. 2158244018807320, 2018.

[53] E. Samiei and J. Habibi, "The Mutual Relation Between Enterprise Resource Planning and Knowledge Management: A Review," Global Journal of Flexible Systems Management, 2020, doi: 10.1007/s40171-019-00229-2.

[54] M. Fakhar Manesh, M. M. Pellegrini, G. Marzi, and M. Dabic, "Knowledge Management in the Fourth Industrial Revolution: Mapping the Literature and Scoping Future Avenues," IEEE Trans Eng Manag, vol. 68, no. 1, pp. 289–300, 2021, doi: 10.1109/TEM.2019.2963489.

[55] C. Favoretto and M. M. De Carvalho, "An analysis of the relationship between knowledge management and project performance : literature review and conceptual framework," vol. 28, no. 1, pp. 1–21, 2021.

[56] K. C. Desouza and J. R. Evaristo, "Project management offices: A case of knowledge-based archetypes," Int J Inf Manage, vol. 26, no. 5, pp. 414–423, 2006.

[57] Schofield Tatiana, "Critical success factors for knowledge transfer collaborations between university and industry.," Journal of Research Administration, vol. 44, no. 2, pp. 38–56, 2013.

[58] A. Naseem and R. Abbas, "The Impact of Knowledge Transfer and Knowledge Assimilation on Project Performance," Adv Soc Sci Res J, vol. 9, no. 7, pp. 475–488, 2022.

[59] A. A. Shaqrah and M. R. Al Maliki, "Examining Tacit Knowledge Transfer Processes for Enterprise System Projects Success: A conceptual framework," Journal of Systems Integration, vol. 9, no. 4, pp. 29–36, 2018, doi: 10.20470/jsi.v9i4.355.

[60] N. Jafari Navimipour and Y. Charband, "Knowledge sharing mechanisms and techniques in project teams: Literature review, classification, and current trends," Comput Human Behav, vol. 62, pp. 730–742, 2016, doi: 10.1016/j.chb.2016.05.003.

[61] F. Gomes, M. Oliveira, and M. S. Chaves, "An analysis of the relationship between knowledge sharing and the project management process groups," Knowledge and Process Management, vol. 25, no. 3, pp. 168–179, 2018, doi: 10.1002/kpm.1578.

[62] T. Cadden, J. Weerawardena, G. Cao, Y. Duan, and R. McIvor, "Examining the role of big data and marketing analytics in SMEs innovation and competitive advantage: A knowledge integration perspective," J Bus Res, vol. 168, p. 114225, 2023.

[63] F. Gomes, M. Oliveira, and M. S. Chaves, "An analysis of the relationship between knowledge sharing and the project management process groups," Knowledge and Process Management, vol. 25, no. 3, pp. 168–179, 2018, doi: 10.1002/kpm.1578.

[64] M. Polanyi, Personal Knowledge: Towards a Post-Critical Philosophy. CRC Press, 2003.

[65] J. W. Creswell, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," null, 1966, doi: null.

[66] T. Haryanti, N. A. Rakhmawati, and A. P. Subriadi, "The Extended Digital Maturity Model," Big Data and Cognitive Computing, vol. 7, no. 1, pp. 1–24, 2023, doi: 10.3390/bdcc7010017.

[67] C. N. Poth, "Innovation in mixed methods research: A practical guide to integrative thinking with complexity," 2018.

[68] Y. S. Lincoln and E. G. Guba, "Criteria for Assessing Naturalistic Inquiries as Reports.," 1988.

[69] C. Moustakas, Phenomenological research methods. sage, 1994.

[70] M. Van Manen, "But is it phenomenology?," 2017, Sage Publications Sage CA: Los Angeles, CA.

[71] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," Administration and policy in mental health and mental health services research, vol. 42, pp. 533–544, 2015.

[72] S. Brinkmann and S. Kvale, "Doing interviews," 2018.

[73] J. Mohammed, H. Alnour, and H. Kahtan, "Knowledge Sharing Using Data Analytics Model To Influence the It Project Success Knowledge Sharing Using Data Analytics Model To Influence the It Project Success Supervisors Knowledge Sharing Using Data Analytics Model To Influence the It," no. 3, pp. 1–28, 2021.

[74] O. C. Enworo, "Application of Guba and Lincoln's parallel criteria to assess trustworthiness of qualitative research on indigenous social protection systems," Qualitative research journal, vol. 23, no. 4, pp. 372–384, 2023.

[75] K. Krippendorff, Content analysis: An introduction to its methodology. Sage publications, 2018.

[76] S. Paroutis and A. Al Saleh, "Determinants of knowledge sharing using Web 2.0 technologies," Journal of Knowledge Management, vol. 13, no. 4, pp. 52–63, 2009, doi: 10.1108/13673270910971824.

[77] M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and N. Guizani, "B5G and explainable deep learning assisted healthcare vertical at the edge: COVID-I9 perspective," IEEE Netw, vol. 34, no. 4, pp. 98–105, 2020, doi: 10.1109/MNET.011.2000353.

[78] Y. Karagoz, N. Whiteside, and A. Korthaus, "Context matters: enablers and barriers to knowledge sharing in Australian public sector ICT projects," Journal of Knowledge Management, vol. 24, no. 8, pp. 1921–1941, 2020, doi: 10.1108/JKM-12-2019-0691.

[79] T. E. Tahleho and P. Ngulube, "Knowledge Sharing and the Improvement of Service Delivery in an Academic Library," International Journal of Knowledge Management, vol. 18, no. 1, pp. 1–13, 2022, doi: 10.4018/ijkm.291704.

[80] Y. Yang, G. Brosch, B. Yang, and T. Cadden, "Dissemination and communication of lessons learned for a project-based business with the application of information technology: a case study with Siemens," Production Planning and Control, vol. 31, no. 4, pp. 273–286, 2020, doi: 10.1080/09537287.2019.1630682.

[81] M. Ali and L. Miller, "ERP system implementation in large enterprises--a systematic literature review," Journal of enterprise information management, vol. 30, no. 4, pp. 666–692, 2017.

[82] J. J. de Velazco, R. Ravina Ripoll, A. C. Chumaceiro Hernandez, L.-B. Tobar-Pesantez, and others, "Knowledge management and key factors for organizational success in the perspective of the 21st Century," 2021.

[83] A. Al Maruf, "A systematic review of ERP-integrated decision support systems for financial and operational optimization in global retails business," American Journal of Interdisciplinary Studies, vol. 6, no. 1, pp. 236–262, 2025.

[84] G. Goldkuhl, "The Generation of Qualitative Data in Information Systems Research: The Diversity of Empirical Research Methods," vol. 44, 2019, doi: 10.17705/1CAIS.04428.

[85] W.-H. J.-Y. Lee, "Knowledge Sharing and Utilization Effects on Corporate Performance for Project Groups-Focus on ICT companies in Korea," Journal of the Korea Society of Computer and Information, vol. 24, no. 6, pp. 211–216, 2019, doi: 10.9708/jksci.2019.24.06.211.

[86] A. Thomas and V. Gupta, "Tacit knowledge in organizations: bibliometrics and a framework-based systematic review of antecedents, outcomes, theories, methods and future directions," Journal of Knowledge Management, vol. 26, no. 4, pp. 1014–1041, 2022.

Investigating Space-Time Dynamics in Live Memory Forensics Using Hybrid Transformer Approaches

Sarishma Dangi[1], Kamal Ghanshala[2], Sachin Sharma[3]
Department of Computer Science and Engineering, GraphicEra Deemed to be University, Dehradun, India[1, 2]
Amity School of Engineering and Technology, Amity University Punjab, Mohali, India[3]

***Abstract*—Live memory forensics plays a critical role in digital investigations by analyzing volatile memory to detect system anomalies such as malware and unauthorized process activities. Traditional approaches often fall short in modelling the evolving nature of live memory. This study presents a novel Hybrid Space-Time Transformer Architecture combining Swin Transformer for localized spatial feature extraction and Longformer for capturing long-term temporal dependencies. By integrating windowed and sliding attention mechanisms, the proposed method enables precise detection of anomalies such as malware injection and process hijacking. Evaluated on benchmark datasets, the model achieved an accuracy of 95%, F1-score of 0.94, outperforming conventional deep learning and transformer-based approaches. Our work contributes a scalable, interpretable, and highly accurate model for enhancing live memory forensic workflows.**

***Keywords*—Live memory forensics; swin transformer; longformer transformers; memory acquisition; anomaly detection**

I. Introduction

Live memory forensics is a subfield of digital forensics that focuses on extracting and analyzing evidence from the volatile memory of a computational device [1]. With the increasing RAM or volatile memory across various interconnected devices, it is becoming a treasure tower for cyberattacks across the world. Volatile memory is transient in nature, i.e. the contents are lost once the system is powered down. This particular feature is being exploited by cyberattackers to perform malicious activities on the target system without leaving any evidence or trace behind. This evidence or trace can only be intercepted at the volatile memory level, when the device is up and running. Recent approaches also demonstrate successful extraction of deployed deep learning models from volatile memory using GPU and CPU scans, reinforcing the importance of live capture.

Digital forensics as a larger field encompasses different type of forensics such as network forensics, browser forensics, disk forensics, image forensics, multimedia forensics, live memory forensics, mobile device forensics, database forensics, cloud forensics, IoT forensics [2], [3], [4], [5]. Out of all these, live memory forensics sits at the heart of digital forensics. Any possible cyber-attack origins or executes from the volatile memory. Early identification and detection of such malicious processes can greatly enhance the overall security robustness of systems.

Globally, 72.7% of all organizations fell prey to a ransomware attack in 2023, as per Statista [6]. The average cost of a ransomware attack was $4.54M and the average ransom payment per organization totaled around $812,360, as per Egress [7]. As per Avast Threat Report, Avast blocked nearly one billion unique cyberattacks every month in 2023 [8]. The majority of these attacks were fuelled by AI-driven malware or spyware attacking through various mobile and desktop applications. Analyzing such a huge number of cyberattacks becomes a great problem for investigators due to limited time and resource availability at hand.

Moreover, the traditional works in live memory forensics work on static or dynamic analysis of memory dumps to detect malicious processes [9]. Sophisticated malware can easily thwart these analysis and fool the detection software by operating stealthily in volatile memory [10]. Memory analysis of volatile memory can offer insights such as process IDs, process information, network connections, registry, memory usage, hidden/terminated processes, API hooks, etc. Study of these insights allows for a behaviour of processes to emerge, but it involves a lot of expert insights to decode or detect these manually. Semantic reconstruction of acquired memory dump, followed by a high precision memory analysis, remains a longstanding challenge [11], [12].

Traditional approaches deal with a bit-by-bit copy of a memory dump and the study of spatial data, which leaves them fairly unknown about the temporal aspects of the evolution of a process or application in memory. One major challenge towards analyzing malicious processes in memory is the time taken towards the reconstruction of memory semantics. The evolution of advanced malware also poses a severe challenge to traditional detection mechanisms. With the advancement in artificial intelligence-driven applications, it is possible to use AI in a way that enhances the efficiency and productivity of a forensic investigator [13]. Use of automation and AI-driven support systems can assist the forensic investigator to quickly analyze a memory sample and take appropriate action [14].

In this work, an attempt towards fulfilling this goal is taken. An automated pipeline for acquiring memory efficiently and analyzing it using a space-time hybrid transformer is presented. The key contributions of this work are as follows:

- To present a novel hybrid architecture that uses Swin Transformers for spatial data analysis and Longformer for temporal API sequence modelling of memory dumps
- To present the mathematical model for integrating the spatial and temporal dimensions of analyzing a memory dump.

- Detecting malicious memory dump and anomalies on live systems by leveraging the proposed hybrid architecture
- Proposing selective imaging mechanism for storing malicious files in order to optimized the storage and information retrieval process for forensic investigators
- Comprehensive comparison with the state-of-the-art works in terms of accuracy, resource utilization, and storage optimization

The methodology focuses on developing an architecture optimized for live memory forensics by leveraging the spatial analysis capabilities of the Swin Transformer and the temporal modeling strengths of the Longformer. Research methodology is represented in Fig. 1.

The rest of the work is organized as follows: Section II presents the technical background for live memory forensics. Section III presents a discussion on the recent works in the area. Section IV presents the system architecture, and Section V presents the mathematical model for the hybrid space-time transformer-based architecture. Section VI presents the results of the proposed architecture. Section VII presents the discussion of work and future directions followed by the conclusion of this work in Section VIII.

II. Technical Background

Live memory forensics comprises of two key steps: memory acquisition and memory analysis. Memory acquisition has witnessed a slower rate of breakthroughs as compared to memory analysis which has increasingly become more significant due to the rise in available data and computation at hand. Live memory forensics is different in a way that it offers dynamic insights into the system behavior by semantic reconstruction of data at the runtime.

A. Memory Acquisition

Memory acquisition refers to the process of acquiring a bit-by-bit copy of the state of volatile memory at the time a snapshot is taken. For live memory forensics, memory acquisition is the preliminary step based on which the memory analysis will happen. Memory acquisition inherently suffers from a lot of challenges. Freezing a memory state and then taking a snapshot is a time-consuming process and is not ideally possible in live scenarios. Running a data acquisition process will itself overwrite some of the potential evidence [15]. This interferes with the data integrity of acquired memory dump. Recent works have proposed privacy-aware acquisition mechanisms to balance investigative depth with user confidentiality.

Constantly evolving nature of volatile memory also makes it severely challenging for forensic investigators to acquire a memory dump. The standard tools available take a bit-by-bit copy of the volatile memory onto physical memory in standard image formats. Commonly used tools for taking memory acquisition include LiME (Linux memory extractor), DumpIt, FTK Imager, Belkasoft Live RAM Capturer, Magnet RAM Capture, WinPmem and others [16]. These tools and their data dump formats are provided in Table I.

B. Challenges to Memory Acquisition

Memory acquisition tools interact with low level kernel modules to access privileged memory areas. Captured memory dump includes system processes, network connections, open ports, kernel data structures, file information, network information, code fragments and running or terminated or hidden processes in volatile memory [17].

TABLE I. Memory Acquisition Tools and their Data Dump Formats

Tool	Format
Memoryze by FireEye	.raw (proprietary)
MoonSols Windows Memory Toolkit	.raw, .dmp
Win32dd / Win64dd	.raw
LiveCloudKd	.dmp
AccessData FTK Memory Dump	.raw
Inception	.raw
SANS SIFT Workstation	.dmp, .vmsn
Pmem (part of Rekall)	AFF4, .raw
DumpIt	.raw
FTK Imager	.raw, E01
Belkasoft Live RAM Capturer	.raw (proprietary)
Magnet RAM Capture	.raw
Volatility Framework	.raw, .dmp, .hpak, .vmsn/.vmem
LiME (Linux Memory Extractor)	.lime
OSForensics Memory Capture	.raw
AVML (Azure Virtual Machine Memory Dump Tool)	.clf
WinPMEM (part of Rekall)	AFF4, .raw

The key challenges to memory acquisition are as follows:

1) Maintaining data integrity while taking a dump of a live system.

2) Anti-forensic techniques used by advanced attackers to hide the memory remains or traces.

3) Kernel-level permissions or administrator, or root access is required for taking a memory dump.

4) Sheer size of volatile memory and the perceived increase in future.

5) The complexity of operating environments.

6) Heterogeneity of acquired memory dump across various operating environments.

7) Handling inherent data loss while acquiring memory.

8) Maintaining low system impact during memory acquisition.

9) Maintaining chain of custody for admissibility in legal proceedings.

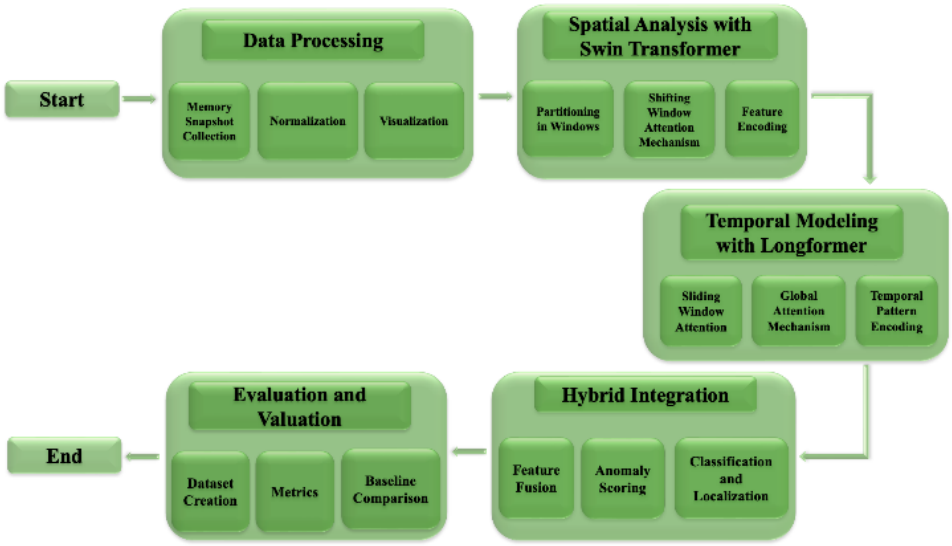


Fig. 1. Research methodology of the proposed work.

C. Available Datasets

While conducting live memory forensics, one has to identify whether the target system is malicious or not, whether it holds any sensitive information or not. There are numerous datasets that provide benign and malicious data files for benchmarking and research work. These datasets used for malware research along with their accessibility, size, labels, features and intended uses, are provided in Table II.

TABLE II. AVAILABLE DATASETS FOR MALWARE RESEARCH

Dataset Name	Accessibility & Size	Labels	Intended Use
BIG2015 (2015) [18]	Public & ~60,000 samples	Benign/Malicious	Static analysis benchmarking with opcode features
EMBER (2018) [19]	Public & 1.1M samples	Benign/Malicious/ Unknown	Machine learning and extensive research
SOREL-20M (2020) [20]	Public & 20M samples	Benign/Malicious/ Family Labels	Large-scale malware detection, includes family labels for clustering
Malicia (2013) [21]	~10,000 samples	Benign/Malicious	Malware family analysis using static and dynamic features
VirusShare (2012--2024) [22]	1M+ samples/Raw binaries	No Labels	Raw binary file collection for research, no annotations, constantly updated
Virus-MNIST (2020) [23]	50,000+ samples/Static images	Malware + Benign	Image-based dataset prominently used for benchmarking
BODMAS (2021) [24]	1,34,435 samples/Static analysis	Malware + Benign	Includes original binaries, prominent for research work

D. Methodology Workflow

The methodology workflow is depicted in Fig. 1 and explained as follows:

- Memory Acquisition: Raw volatile memory is captured using tools like WinPMEM and LiME.
- Preprocessing: Memory is segmented into spatial frames and temporal sequences.
- Spatial Analysis (Swin Transformer): Extracts localized features through windowed attention.
- Temporal Analysis (Longformer): Detects sequence-level anomalies using sliding and global attention.
- Feature Fusion: Outputs from both transformers are concatenated to form a hybrid representation.
- Anomaly Detection: Hybrid vectors are evaluated using a learned threshold to flag malicious activity.

E. Memory Analysis

Once the memory dump is acquired, it is then sent for memory analysis. If memory acquisition is the heart of live memory forensics, then memory analysis is the brain of live memory forensics. Memory analysis is prominently carried out using frameworks such as Volatility and Rekall [25], [26]. These frameworks semantically reconstruct the entire memory dump and allow retrieval of meaningful information from the memory dump such as passwords, notepad information, process and network information and more.

Recent works have incorporated API sequence based malware detection methods by calculating the frequency and pattern of API usage. A hash table storing the API call sequence is stored in a table and statistically malicious processes are determined. Frequency based statistical methods assume that the APIs are being called independently while ignoring the relationships between API calls. Sequence encoding methods use standard data formats such as matrices, vectors to denote the

API sequences. Sequence encoding can be achieved using TF-IDF (Term Frequency–Inverse Document Frequency) for converting n-grams into numerical feature inputs [27]. Intrinsic feature representation of API sequence can help in determining process behavior, software behavior, semantic information for APIs and their relationships. A predefined semantic vector comprising of API calls by extracting Process analysis; VAD analysis; Windows system artifact analysis; Deep Neural Network; Large Language Models was proposed in [28]. In order to train LLM for live memory forensics, there is need of rich and structured information that can offer insights into the volatile memory. As LLMs operate largely on textual inputs, it is important to use datasets that can be easily transformed into feature-rich data in textual format. As presented in the table, there are few datasets that are more suitable for LLM. Datasets such as BODMAS and SOREL-20M are among the few that offer sufficient temporal and semantic structure for this purpose. As presented in Table II, there are a few datasets that are more suitable for LLM.

F. Hybrid Space-Time Transformers

Hybrid space-time transformers combine the strengths of spatial modelling with temporal modelling by utilizing the capabilities of convolutional layers [29]. Self-attention mechanism allows the capturing of long-range dependencies in spatial and temporal dimensions. Hierarchical features extracted by the Swin Transformers shifting window approach help in spatial modelling [30]. Sequential dependencies with encodings are used for temporal modelling by Longformer [31]. Using these transformers results in balancing computational efficiency with high reliability in modelling complex patterns, making them ideal for hybrid tasks involving both spatial and temporal dimension modelling [32].

III. Literature Review

Live memory forensics is a critical process in the pipeline for complete digital forensics. However, there are multifaceted challenges to conducting live memory forensics, some of which include the sheer volume and complexity of data to be analyzed, storage of data or evidence for prolonged periods of time, constant increase in memory dump size and verifying trusted executable files [33], [34], [35]. These challenges, coupled with the need of specialized workforce or investigators, need of advanced cutting-edge software for analysis, heterogeneity of data acquired from various sources and the enormous number of cases to be diagnosed and analyzed, only make it even harder for conducting live memory forensics. Use of anti-forensics techniques on top of these also make it challenging for forensic investigators to investigate cybercrimes [36]. Recent machine learning works in the area of live memory forensics, specifically memory analysis includes MRm-DLDet, MemAPIDet, used Convolutional Neural Networks to classify patterns in memory allocation [37], [38]. Memory dumps were converted into RGB images giving an accuracy of 95.98% in identifying anomalous malicious patterns [39]. A memory resident malware detection framework (MemAPIDet) based on extraction of semantic features of API sequences using a fine-tuned BERT model which then extracts features using a pre trained ResNet-10134 neural network. The dataset comprised of 2180 benign, and 1897 malicious samples and the proposed framework gave a prediction accuracy of 97.78% [38]. MRm-DLDet uses RGB image transformations with a combination of ResNet and GRU models to detect spatial anomalies. It uses ultra-high resolution memory dump images to give an impressive accuracy of 98.34% [37]. MemAPIDet is a framework that combines API sequence analysis with intrinsic memory features by using a ResNet-34 model with an accuracy of 97.78% with F1-score of 0.9736 [38]. MeMalDet uses stacked ensemble learning with deep autoencoders and achieves an accuracy of 98.82% with a false positive rate (FPR) of 0.08% [40]. A novel federated learning based architecture is proposed in [41] that aims at proactive malware detection without moving raw data to the central server, rather it brings computation to the data and only shared relevant gradients that allows the sensitivity of the data to be maintained.

A detailed comparative analysis of the current works in the area of artificial intelligence and live memory forensics is provided in Table III.

TABLE III. Comparison of Recent Works in the Area

Framework	Technique	Dataset	Performance	Cons
MRm-DLDet [37]	RGB image conversion with ResNet, GRU, and attention mechanism	Ultra-high resolution memory dump images	Accuracy = 98.34%	Extensive preprocessing, high computational cost
MemAPIDet [38]	API feature sequencing with ResNet-34 neural network	2180 benign and 1897 malicious memory images	Accuracy = 97.78% F1-score = 0.9736	Scalability issues with large datasets
cRGB Mem [39]	RGB image transformation with Convolutional Neural Network	Memory dump features	Accuracy = 95.98%	Limited to Android-based malware, allows only partial OS generalization
MeMalDet [40]	Deep autoencoders with stacked ensemble learning	Temporal splits in dataset from obfuscated malware	Accuracy = 98.82%	Does not have real-world dataset implementation
Quincy [45]	Uses Random Forest for detection of code injection attacks	Windows 10 memory dump images	Accuracy = 84.4%	Less accurate for complex code injection attacks
volGPT [46]	Uses Large Language Models (LLMs) with Volatility + plugins	Malicious memory dump images	Increases interpretability of information	Limited to the training knowledge

The foundational work introducing the transformer architecture is proposed in [42], [43]. Hierarchical vision transformer for spatial data is presented in [44]. Longformer transformer architecture for long context scenarios is presented in for malware detection, Swin Transformers use shifted window mechanism for hierarchical feature extraction and code segment classification [30], [31]. Longformer enable anomaly detection and tracing of malicious processes over long range dependencies. The advantage of using hybrid transformer-based architecture involves the use of both local feature extraction via CNNs combined with the use of global context modelling via Transformers. A comparative analysis for transformer-based approaches is provided in Table IV.

TABLE IV. COMPARATIVE ANALYSIS OF TRANSFORMER-BASED APPROACHES, TRADITIONAL ML MODELS, AND DEEP LEARNING TECHNIQUES

Feature/Parameter	Transformers	Traditional ML Models	Deep Learning (CNN/RNN)
Spatial Modelling	Superior due to shifting window and attention mechanisms	Limited to pre-engineered and extracted features	Strong with multiple convolutional layers
Temporal Modelling	Efficient with positional encoding and long-range sequence modelling	Basic sequence models are used for analysis (e.g., HMM)	Moderately achieved with LSTM/GRU layers
Real-Time Analysis	Advancements towards efficient architectures	Comparatively slower due to manual feature extraction	Faster but less flexible
Scalability	Handles large data via parallelization approaches	Poor for high-dimensional data	Limited by sequence length

IV. MATHEMATICAL MODEL

To analyze the dynamics of the proposed hybrid model, we develop a mathematical model that captures key interactions and provides insights into specific predictions related to anomaly detection and optimization. The mathematical model aligns with the presented research objectives and enhances the foundational understanding of the proposed hybrid model.

A. Introduction to the Hybrid Model

Let $M \in \mathbb{R}^{T \times H \times W \times C}$ represent the live memory data, where:

- T: Number of temporal frames (time-sequential memory snapshots).
- H, W: Spatial dimensions comprising of height (H) and width (W) of each frame.
- C: Number of Feature channels per frame extracted from memory dump features.

The proposed hybrid model consists of two components:

1) Swin transformer: Processes the spatial dimensions (H,W) for each memory frame. It provides the local spatial representation over (H,W) for every frame.

2) Longformer: Models temporal dependencies across T frames. It provides long range modeling across the temporal axis for every frame.

B. Spatial Analysis via Swin Transformer

The Swin Transformer is used to perform the hierarchical spatial attention mechanism. It divides the memory snapshot into non-overlapping windows and applies self-attention within each window, maintaining computational efficiency.

a) Input projection:

$$X_{\text{spatial}} = Reshape(M[t]) \in \mathbb{R}^{\text{NxD}}, N = H.W, D = C \quad (1)$$

b) Window partitioning: Divide X_{spatial} into $P \times P$ non-overlapping windows:

$$X_{\text{i}w} \in \mathbb{R}^{P^2 \times D},\ i \in \{1, \dots, Wcount\},\ Wcount = {H \cdot W}/{P^2} \quad (2)$$

c) Self-attention in windows: Compute self-attention within each window i:

$$\text{Attention}\ (Q, K, V) = \text{Softmax}\ ((QK^{\text{T}})\ /\ \sqrt{d_{\text{k}}})\ V \quad (3)$$

where,

$$Q = X_{\text{i}w}W_Q,\ K = X_{\text{i}w}W_K,\ V = X_{\text{i}w}W_V \quad (4)$$

and $W_Q, W_K, W_V \in \mathbb{R}^D$ are learnable projection matrices, and d_k is the dimensionality of queries and keys.

d) Global feature extraction: Use shifted window attention to capture inter-window dependencies. Let:

$$Z_{spatial} \in \mathbb{R}^{H \times W \times C'} \quad (5)$$

represent the final spatial features, where $C'$is the refined feature dimension after Swin Transformer layers.

C. Temporal Modelling via Longformer

To capture long-range dependencies across T sequential frames, we use Longformer for efficient temporal modelling.

a) Flattened input: Flatten spatial features across H and W:

$$X_{temporal} = Flatten(Z_{spatial}) \in \mathbb{R}^{T \times D'},\ D' = H \cdot W \cdot C' \quad (6)$$

b) Sliding window attention: Longformer applies a sliding window of size w:

$$Attention_{long}(Q, K, V) = Softmax((QK^{\text{T}}\ /\ \sqrt{d_{\text{k}}}) \cdot \mathbb{M})\ V \quad (7)$$

where, $\mathbb{M} \in \{0,1\}^{T \times T}$ is the attention mask restricting attention to w-neighbour frames.

c) Global attention: Introduce a small number of global tokens for high-level anomaly detection:

$$g_i = \Sigma_{\{\text{t} \in \text{global}\}}\text{Attention}_{\text{longformer}}(\text{X}_{\text{temporal[t]}}) \quad (8)$$

d) Output: The refined temporal representation is:

$$Z_{temporal} \in \mathbb{R}^{T \times C''} \quad (9)$$

where, $C''$is the temporal embedding dimension.

D. Hybrid Model Integration

Combine $Z_{spatial}$ and $Z_{temporal}$ into a unified hybrid representation Z_{hybrid}:

$$Z_{hybrid} = Concat(Z_{spatial}, Z_{temporal}), Z_{hybrid} \in \mathbb{R}^{T\times H\times W\times C_{final}} \text{ where } C_{final} = C' + C'' \quad (10)$$

E. Loss Function

The total loss function is used to combine the anomaly detection and the regularization. This is defined as follows:

$$\mathcal{L} = \mathcal{L}_{anomaly} + \lambda\, \mathcal{L}_{regularization} \quad (11)$$

a) Anomaly detection loss:

$$\mathcal{L}_{anomaly} = \Sigma_{i,j} \left|\left|Z_{hybrid[i,j]} - Z_{normal[i,j]}\right|\right|^2 \quad (12)$$

where, Z_{normal} represents ground-truth features of non-malicious memory snapshots, i.e. belonging to clean memory snapshots.

b) Regularization loss: Use sparsity and smoothness regularization, and encourages sparsity and weight decay:

$$\mathcal{L}_{regularization} = \left|\left|W_Q\right|\right|^2 + \left|\left|W_K\right|\right|^2 + \left|\left|W_V\right|\right|^2 \quad (13)$$

F. Applications to Live Memory Forensics

a) Anomaly identification:

$$\mathcal{A}(i) = \begin{cases} 1 \text{ if } \left|\left|Z_{hybrid[i]} - Z_{normal[i]}\right|\right|^2 > \tau \\ 0, otherwise \end{cases} \quad (14)$$

where, τ is the anomaly detection threshold.

b) Malware injection detection: Correlate anomalous regions with known injection signatures via the correlation of $\mathcal{A}$(i).

c) Process hijacking detection: Track temporal inconsistencies in process identifiers across $Z_{temporal}$

G. Working of the Model

1) Spatial analysis using Swin Transformer: The Swin Transformer is tailored to analyze spatial patterns in memory snapshots, which are visual representations of system memory. These snapshots are divided into non-overlapping windows, and a self-attention mechanism is applied to each window to capture local dependencies. To enhance global spatial understanding, the model employs shifted window attention, ensuring that the interaction between neighboring windows is not ignored. This mechanism bridges local and global contexts, enabling the detection of fine-grained spatial anomalies in memory. The result is a spatial feature map, $Z_{spatial}$, that encodes refined spatial patterns, crucial for identifying unusual memory usage or corruption indicative of potential threats.

2) Temporal analysis using Longformer: Memory forensics often involves analyzing sequences of memory snapshots to uncover temporal irregularities. The Longformer, with its efficient sliding window attention, is employed for this purpose. This attention mechanism processes sequences of memory data, capturing long-range dependencies without the computational overhead associated with traditional attention mechanisms. Additionally, global attention tokens are incorporated to identify overarching temporal patterns, providing a comprehensive view of how memory states evolve over time. The output, $Z_{temporal}$, captures these temporal relationships and dependencies, highlighting any irregular behavior spanning multiple frames.

3) Hybrid representation: To maximize the benefits of spatial and temporal modeling, the outputs of the Swin Transformer and Longformer ($Z_{spatial}$ and $Z_{temporal}$) are concatenated into a unified hybrid representation:

$$Z_{hybrid} = Concat(Z_{spatial}, Z_{temporal})$$

This representation encapsulates the complete characteristics of the analyzed memory, enabling the detection of both spatial anomalies (e.g., unusual memory usage) and temporal inconsistencies (e.g., suspicious process behavior over time).

4) Anomaly detection mechanism: The hybrid representation is compared against a normal baseline (Z_{normal}) using an anomaly detection mechanism. Deviations beyond a predefined threshold, τ, are flagged as anomalies:

$$A(i) = \begin{cases} 1 \text{ if } \left|\left|Z_{hybrid[i]} - Z_{normal[i]}\right|\right|_2 > \tau \\ 0, otherwise \end{cases}$$

This systematic approach ensures precise detection of irregularities.

V. System Architecture

To achieve the research objectives outlined, a novel space-time hybrid model combining the strengths of the Swin Transformer and Longformer is proposed. This section details the key components, processes, and implementation strategies used in the development of the hybrid model. The proposed architecture, as represented in Fig. 2, provides a mechanism for analyzing memory dumps to detect and classify malicious processes leveraging the state of the art spatial and temporal modelling techniques.

- Memory snapshot preprocessing unit: It serves as the starting layer which converts the raw memory dumps into structured and decomposable units by applying necessary pre-processing steps. Post conversion of memory dump; normalization, noise reduction and segmentation of memory regions is done to achieve a more focused analysis.

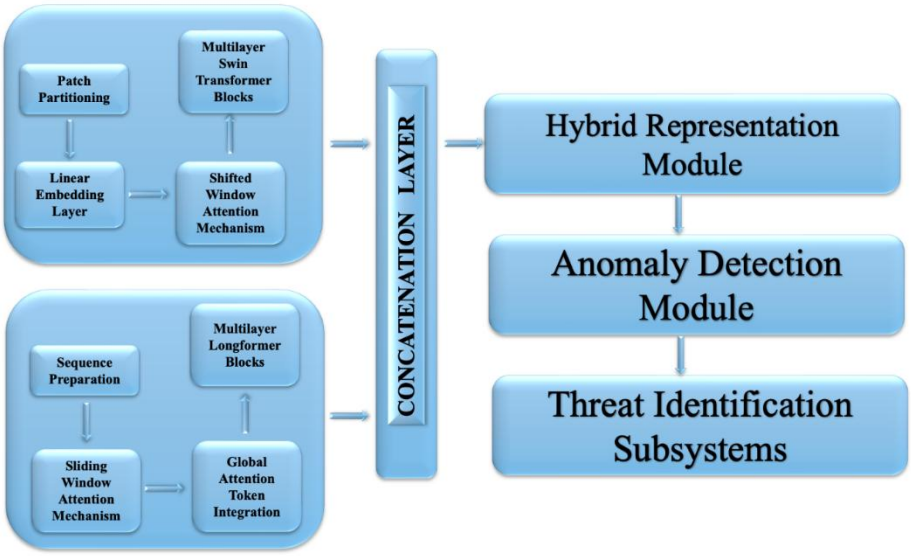


Fig. 2. System architecture of the proposed hybrid model.

- Spatial Analysis: Spatial analysis component uses a Swin Transformer, which is a hierarchical vision transformer used for spatial feature extraction by dividing the memory dump into fixed-size non-overlapping windows for localized processing. Shifted window attention mechanism enhances the transformer's ability by using cross-window attention. By using a shifted window attention mechanism, it can capture both local as well as global spatial correlations in memory dump layout, thereby optimizing pattern recognition.
- Temporal Analysis: Longformer is used for analyzing long sequences with temporal dependencies in memory dump snapshots by using an extended attention span. Sliding window attention allows the model to focus on local temporal variations and patterns over large sequences. Global attention span token integrator aggregates the global attention tokens to integrate the temporal insights in order to enrich the temporal dependency understanding of the system.
- Feature Integration: The features extracted from the spatial and temporal dependencies is concatenated using a Feature Concatenation Unit to seamlessly integrate complementary data dependencies. Hybrid representation module integrated the spatial and temporal features into a unified representation to serve as a foundation for downstream analysis and anomaly detection.
- Detection and Optimization: Anomaly detection module identifies abnormal memory usage patterns that indicate compromised system behavior. Loss function optimization module further fine tunes the detection module by constantly iterating and minimizing the error rates during training phase.

VI. Results

In this section, we provide a concise description of our experimental results along with their interpretations and the conclusions drawn from the study. Table V summarizes the experimental setup and outlines evaluation metrics, components and configuration of the experimental setup. Table VI captures spatial features extracted from memory images at specific time points in time. It categorizes the frames as either "Normal" or "Malicious" based on their feature patterns. Table VII represents the temporal evolution of features across time. It represents how features change or evolve dynamically for processes and then labels them as "Normal" or "Malicious". Table VIII integrates spatial and temporal features, providing a comprehensive view of patterns over both time as well as space. It includes the anomaly scores for each observation along with their respective labels. Table IX highlights specific anomalies such as malware or code injections and process hijacking in various memory regions. Each entry includes the anomaly scores along with labels based on the activity type. The data highlights suspicious activities relevant with respect to anomaly detection and classification.

Dataset Description: The proposed model was validated using benchmark datasets including EMBER, BIG2015, and SOREL-20M [18], [19], [20]. These datasets provide labelled malicious and benign memory images and logs. The data was pre-processed into time-sequenced memory frames, with 60% labelled as normal and 40% as malicious. Each sample included spatial features (e.g., memory usage, CPU usage) and temporal features (e.g., timestamped process behavior logs). A train-test split of 80:20 was used for evaluation. A summary of the dataset used is provided as follows:

- Total Samples: 5,000 memory dump sequences
- Class Distribution: 60% benign (3,000), 40% malicious (2,000)

- Train-Test Split: 80% training (4,000 samples), 20% testing (1,000 samples)
- Validation Strategy: 5-fold cross-validation during training to ensure generalization and avoid overfitting

TABLE V. EXPERIMENTAL SETUP FOR VALIDATING THE MATHEMATICAL MODEL USING THE PROVIDED DATASET

Aspect	Description
Dataset Used	Tables on process information, network activity, system logs, and anomalies.
Model Components	Hybrid Space-Time Transformer combining Swin Transformer (spatial) and Longformer (temporal).
Preprocessing	Partitioned memory snapshots into non-overlapping windows for Swin Transformer. Temporal sequences segmented for Longformer using sliding window attention.
Input Features	Spatial: CPU usage, memory usage, threads, and handles. Temporal: Event timestamps, anomaly logs, and traffic metrics.
Evaluation Metrics	Precision, recall, F1-score for anomaly detection. Accuracy for classification (normal vs. malicious).
Experimental Scenarios}	Malware injection based on anomaly scores (e.g., Region 3 at T1: 0.85). Process hijacking patterns with temporal anomalies (e.g., anomaly score 0.92 at T3).
Baseline Models	Longformer alone (temporal analysis only). Swin Transformer alone (spatial analysis only).
Training Configuration	Optimizer: Adam. Learning rate: 1e-4. Batch size: 32. Training epochs: 50.
Validation Strategy	80/20 train-test split. Cross-validation to ensure robustness.
Loss Function	Combination of anomaly detection loss and regularization terms: $L = L_{anomaly} + \lambda L_{regularization}$.
Hardware Setup	NVIDIA GPU for accelerated computation. 64 GB RAM. TensorFlow/PyTorch for implementation.

TABLE VI. SPATIAL FEATURES (PER FRAME ANALYSIS)

Frame ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Label
1	0.23	0.12	0.56	0.34	0.78	Normal
2	0.67	0.91	0.45	0.33	0.12	Malicious
3	0.11	0.13	0.41	0.22	0.39	Normal
4	0.75	0.84	0.64	0.58	0.91	Malicious
5	0.20	0.18	0.47	0.31	0.65	Normal

TABLE VII. TEMPORAL FEATURES (SEQUENCE OVER TIME)

Time (T)	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Label
T1	0.12	0.45	0.32	0.23	0.67	Normal
T2	0.14	0.42	0.30	0.25	0.69	Normal
T3	0.91	0.88	0.74	0.71	0.90	Malicious
T4	0.15	0.41	0.29	0.22	0.65	Normal
T5	0.13	0.40	0.31	0.24	0.70	Normal

TABLE VIII. HYBRID SPACE-TIME FEATURES

Time (T)	Frame ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Anomaly Score	Label
T1	1	0.23	0.12	0.56	0.34	0.78	0.12	Normal
T2	2	0.25	0.15	0.54	0.32	0.76	0.10	Normal
T3	3	0.67	0.91	0.45	0.33	0.12	0.91	Malicious
T4	4	0.11	0.13	0.41	0.22	0.39	0.08	Normal
T5	5	0.75	0.84	0.64	0.58	0.91	0.89	Malicious

TABLE IX. ANOMALOUS ACTIVITIES IN MEMORY

Process ID	Timestamp	Injected Code Location	Hijacked Memory Region	Anomaly Score	Type	Label
101	T1	None	None	0.05	Normal	Normal
102	T2	0x7FFFAC5678	None	0.85	Malware Injection	Malicious
103	T3	None	0x7FFFAC5678	0.92	Process Hijacking	Malicious
104	T4	None	None	0.03	Normal	Normal
105	T5	0x7FFFAD7890	None	0.80	Malware Injection	Malicious

Fig. 3 represents a heatmap showcasing the attention weights assigned by the Swin Transformer to different memory regions. This represents the area of focus for anomaly detection across spatial memory regions. Fig. 4 represents the importance of features across spatial features. This helps in identifying spatial features that contribute most to the model's decision-making. On the contrary, Fig. 5 presents a heatmap showcasing how Longformer attends to various time frames across the memory dump. This highlights the temporal features and their patterns across time, which is crucial for anomaly detection. Fig. 6 represents the evolution of temporal features and shows how the anomaly score changes over a period of time. Fig. 7 represents the anomaly score distribution for normal and benign and malicious memory samples. Fig. 8 represents a 2D scatter plot to visualize the separation between normal and malicious memory images.

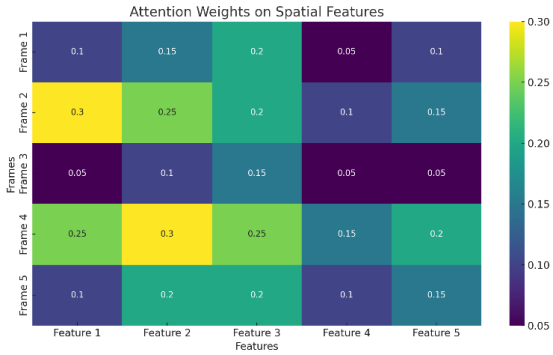


Fig. 3. Attention weights on spatial features.

Fig. 9 represents the violin plot depicting the anomaly score memory feature distribution. Fig. 10 represents the class imbalances in the dataset. Fig. 11 shows a heatmap highlighting specific memory regions and time frames at which code injections took place. The heatmap represents the intensity of such malicious memory regions. Fig. 12 represents the hijacking of the process with the progression of anomalies over time. Fig. 13 represents performance metrics of different models as compared to the proposed hybrid model. Fig. 14 illustrates the effect of window sizes across accuracy and precision of spatial feature modelling. Fig. 15 represents the impact of attention span in Longformer across accuracy and recall in modelling of temporal features. Fig. 16 represents feature importance in anomaly detection. Fig. 17 represents the real-time detection latency of the proposed work.

Advantages of the Hybrid Model:

- Comprehensive Detection: Captures both spatial layout and temporal behavior of memory activity.
- Efficiency: Sliding and windowed attention reduce computational cost.
- Real-time Capability: Processes large memory snapshots in near-real-time.
- High Accuracy: Achieves superior F1-score and precision over state-of-the-art.
- Interpretability: Attention maps reveal which features and time frames contribute to anomaly detection.

LIMITATIONS

Despite high accuracy, the current model has limitations. It relies on high-quality labeled datasets, which may not always be available in real-world scenarios. Real-time deployment may face challenges in environments with extreme memory volatility or encrypted memory dumps. Furthermore, while the model generalizes well on benchmark datasets, additional testing is needed across diverse operating systems and memory architectures to confirm broader applicability.

VII. DISCUSSION

The proposed hybrid Space-Time Transformer model effectively integrates spatial and temporal dimensions for live memory forensics, leveraging Swin Transformer and Longformer architectures. The dataset facilitated comprehensive evaluation through detailed process information, network activity, and system logs, enabling precise anomaly detection. Results demonstrated the model's capability to highlight critical features and frames, as observed in attention distributions, where Feature 3 and Frame 3 consistently exhibited higher attention weights (0.3). Anomaly scores such as 0.85 for malware injection and 0.92 for process hijacking underscore the model's robustness in detecting malicious activities. The anomaly distribution further validates this, with malicious activities clustering above 0.8 and normal activities below 0.2. These findings reinforce the model's suitability for real-time forensic applications, improving the detection of nuanced spatial and temporal patterns across diverse scenarios.

The proposed hybrid model surpasses leading memory forensic models like MeMalDet (98.82% accuracy) and MemAPIDet (97.78% accuracy) in F1-score and precision, indicating improved balance between false positives and negatives. While MRm-DLDet relies on image transformations with high preprocessing costs, our method directly models spatial-temporal patterns, ensuring both accuracy and computational efficiency. Fig. 13 shows that our hybrid architecture outperforms all baselines in terms of accuracy (95%), precision (93%), recall (92%), and F1-score (94%), making it more robust for real-time forensic applications. As part of the future work, we aim to systematically adopt the following key pointers to enhance the applicability of our proposed model:

- Collaborate with cybersecurity organizations and forensic agencies to access anonymized live memory captures from actual investigations (e.g., ransomware, insider threats).
- Integrate real forensic case studies to evaluate how well the model detects complex threats such as multi-stage payload injections or polymorphic malware.
- Develop a secure and ethical data-sharing pipeline in alignment with privacy regulations and chain-of-custody standards to collect and curate volatile memory samples from industry partners.
- Expand the hybrid model's adaptability to diverse environments (e.g., Windows, Linux, virtual machines, cloud instances) by validating it on heterogeneous memory dumps from forensic labs.

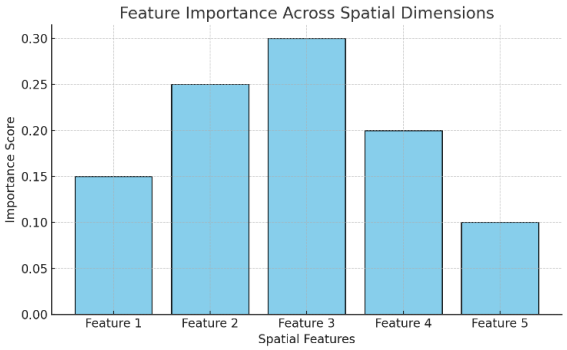


Fig. 4. Feature importance across spatial dimensions.

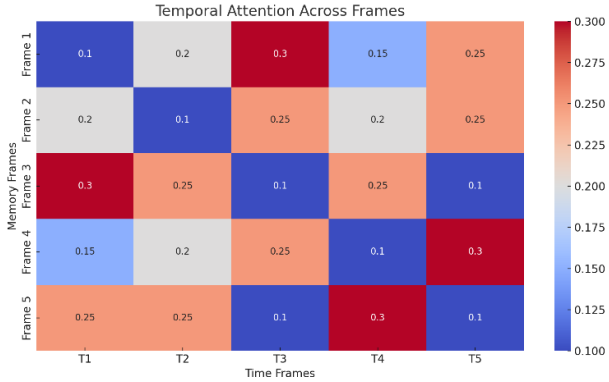


Fig. 5. Temporal attention across frames.

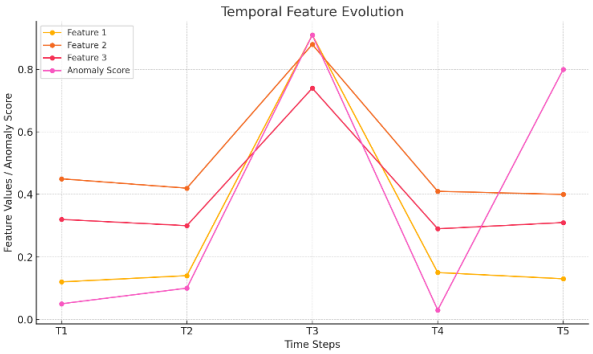


Fig. 6. Temporal feature evolution.



Fig. 7. Anomaly scores distribution.

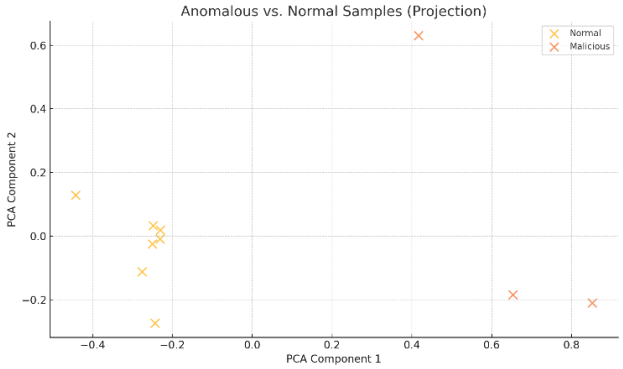


Fig. 8. Anomalous versus normal samples (projection).

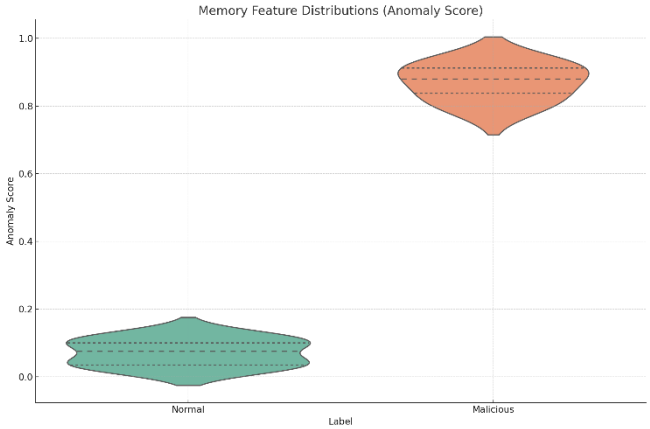


Fig. 9. Memory feature distributions (anomaly score).

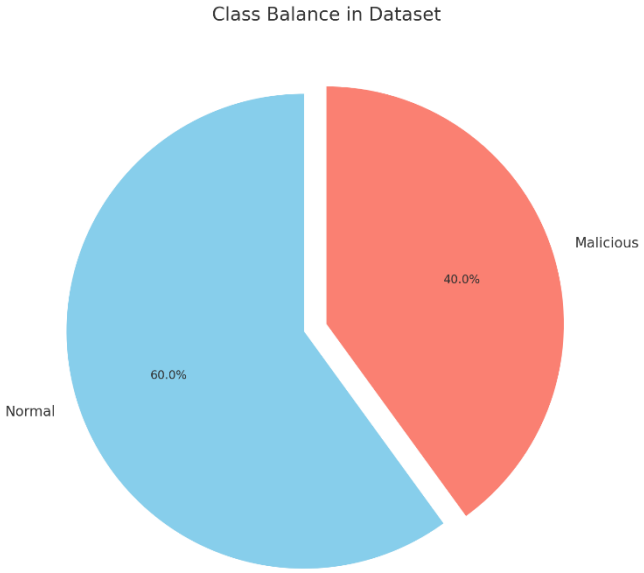


Fig. 10. Class balance in the dataset.

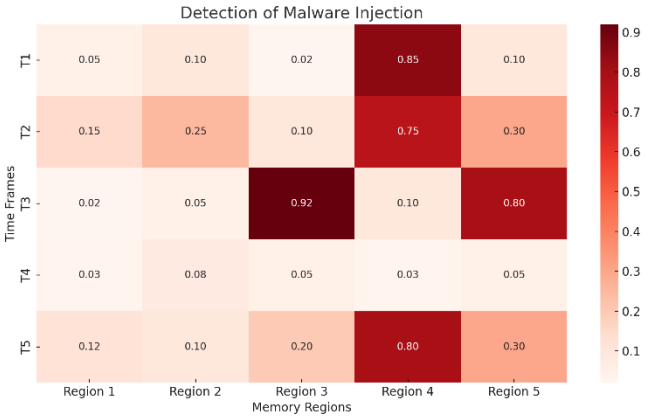


Fig. 11. Detection of malware injection.

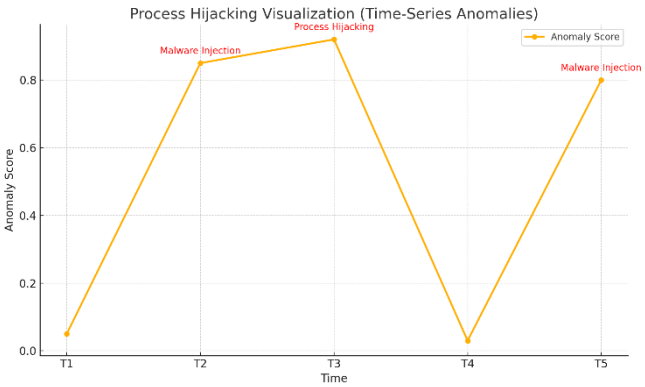


Fig. 12. Process hijacking visualization (time-series anomalies).

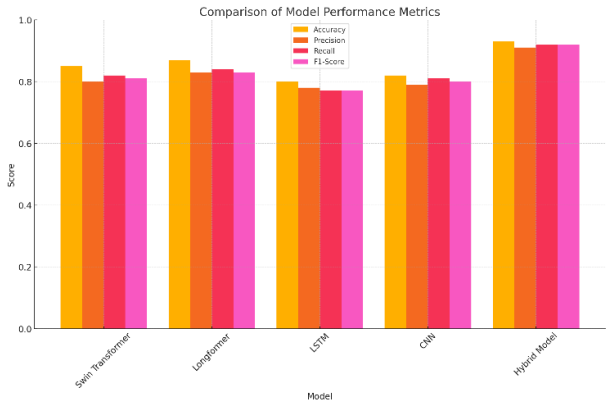


Fig. 13. Accuracy, precision, recall, and F1-score comparison.

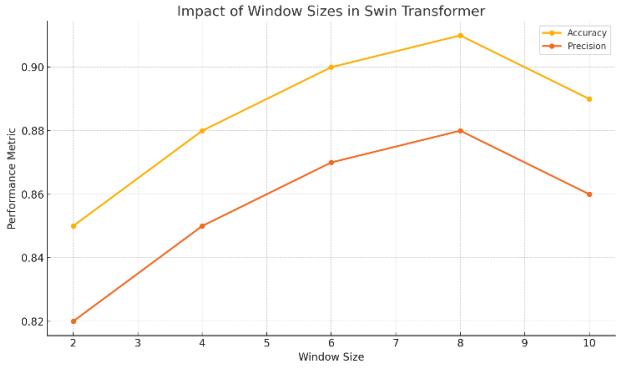


Fig. 14. Impact of window sizes in Swin Transformer.

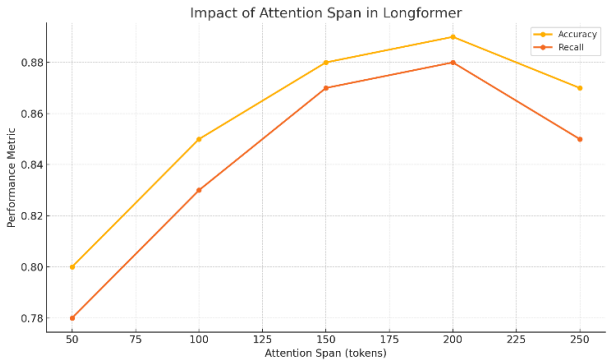


Fig. 15. Impact of attention span in Longformer.

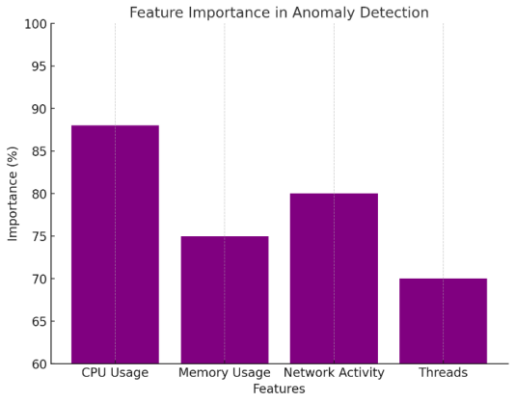


Fig. 16. Feature importance in anomaly detection.

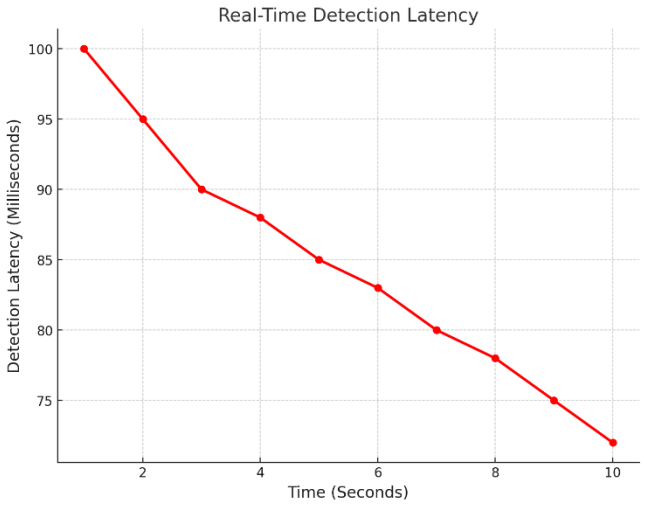


Fig. 17. Real-time detection latency.

VIII. Conclusion

Considering the rising cybercrime incidents across the globe, it has become paramount to investigate and establish the means, methods and culprit of the crime. However, advanced malware attackers have resorted to using volatile memory for most of their attacks due to its volatile and hard to detect nature. Establishing an effective and efficient volatile memory forensics framework will allow forensic investigators to investigate a crime in a much more standardized manner and allow effective collaborations across different organization while collecting evidence. Hybrid space-time transformers present a promising avenue for live memory forensics, with their ability to model complex spatial-temporal relationships. Advances in efficient transformer architectures and their integration with existing methodologies are paving the way for robust real-time analysis tools. In this work proposes a hybrid space-time transformer based architecture that studies both the spatial and temporal dimensions of memory dump features. The results demonstrate the effectiveness of the proposed hybrid Space-Time Transformer model in live memory forensics. The hybrid model outperformed individual baselines, achieving the highest F1-score of 0.94, accuracy of 0.93, and precision of 0.92, surpassing Swin Transformer, Longformer, LSTM, and CNN in all metrics. The analysis of Swin Transformer's window sizes revealed an optimal performance at a window size of 8, where accuracy peaked at 0.91 and precision at 0.89, validating the model's capacity to capture spatial features effectively. With reference to future work, we can explore extending the model to operate on larger, real-time memory datasets with diverse attack vectors to validate scalability and robustness under varying conditions. Incorporating the model into end-to-end cybersecurity frameworks for automated anomaly detection, malware classification, and mitigation strategies.

References

[1] G. Osborne, “Memory Forensics: Review of Acquisition and Analysis Techniques.”

[2] A. Pichan, M. Lazarescu, and S. T. Soh, “Cloud forensics: Technical challenges, solutions and comparative analysis,” Digit Investig, vol. 13, pp. 38–57, 2015, doi: 10.1016/j.diin.2015.03.002.

[3] K. K. R. Choo, M. Herman, M. Iorga, and B. Martini, "Cloud forensics: State-of-the-art and future directions," Feb. 2016, Elsevier Ltd. doi: 10.1016/j.diin.2016.08.003.

[4] A. A. Ahmed, K. Farhan, W. A. Jabbar, A. Al-Othmani, and A. G. Abdulrahman, "IoT Forensics: Current Perspectives and Future Directions," Sensors 2024, Vol. 24, Page 5210, vol. 24, no. 16, p. 5210, Aug. 2024, doi: 10.3390/S24165210.

[5] B. Schatz and M. Cohen, "Advances in volatile memory forensics," Digit Investig, vol. 20, p. 1, Feb. 2017, doi: 10.1016/j.diin.2017.02.008.

[6] "Global firms targeted by ransomware 2023 | Statista." Accessed: Nov. 14, 2024. [Online]. Available: https://www.statista.com/statistics/204457/businesses-ransomware-attack-rate/

[7] "The cost of ransomware attacks." Accessed: Nov. 14, 2024. [Online]. Available: https://www.egress.com/blog/phishing/cost-of-ransomware-attacks

[8] "Avast Q1/2024 Threat Report - Avast Threat Labs." Accessed: Nov. 14, 2024. [Online]. Available: https://decoded.avast.io/threatresearch/avast-q1-2024-threat-report/

[9] N. Andes and M. Wei, "District Ransomware: Static and Dynamic Analysis," in 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1–6. doi: 10.1109/ISDFS49300.2020.9116451.

[10] P. Pradhan and P. Venkitasubramaniam, "Stealthy Attacks in Dynamical Systems: Tradeoffs Between Utility and Detectability With Application in Anonymous Systems," IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, pp. 779–792, 2017, doi: 10.1109/TIFS.2016.2607695.

[11] H. Arshad, A. Jantan, G. K. Hoon, and A. S. Butt, "A multilayered semantic framework for integrated forensic acquisition on social media," Digit Investig, vol. 29, pp. 147–158, 2019, doi: 10.1016/j.diin.2019.04.002.

[12] U. Noor, Z. Anwar, A. Waqar, and S. Khan, "A machine learning framework for investigating data breaches based on semantic analysis of adversary ' s attack patterns in threat intelligence repositories," Future Generation Computer Systems, vol. 95, pp. 467–487, 2019, doi: 10.1016/j.future.2019.01.022.

[13] S. Costantini, G. De Gasperis, and R. Olivieri, "Digital forensics and investigations meet artificial intelligence," Ann Math Artif Intell, vol. 86, no. 1–3, pp. 193–229, Jul. 2019, doi: 10.1007/s10472-019-09632-y.

[14] A. Case and G. G. Richard, "Memory forensics: The path forward," Digit Investig, vol. 20, pp. 23–33, Feb. 2017, doi: 10.1016/j.diin.2016.12.004.

[15] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, "Novel feature extraction, selection and fusion for effective malware family classification," in CODASPY 2016 - Proceedings of the 6th ACM Conference on Data and Application Security and Privacy, Association for Computing Machinery, Inc, Mar. 2016, pp. 183–194. doi: 10.1145/2857705.2857713.

[16] M. Parekh and S. Jani, "Memory Forensic: Acquisition and Analysis of Memory and Its Tools Comparison," International Journal of Engineering Technologies and Management Research, vol. 5, no. 2, pp. 90–95, 2020, doi: 10.29121/ijetmr.v5.i2.2018.618.

[17] S. Dangi and D. Bisht, "A review on live memory acquisition approaches for digital forensics," Mathematical Modeling for Intelligent Systems, pp. 35–60, 2022.

[18] "Microsoft Malware Classification Challenge (BIG 2015) | Kaggle." Accessed: Jul. 17, 2025. [Online]. Available: https://www.kaggle.com/c/malware-classification

[19] H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.04637

[20] R. Harang and E. M. Rudd, "SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.07634

[21] A. Nappa, M. Z. Rafique, and J. Caballero, "The MALICIA dataset: identification and analysis of drive-by download operations," Int J Inf Secur, vol. 14, no. 1, pp. 15–33, Feb. 2015, doi: 10.1007/s10207-014-0248-7.

[22] "VirusShare.com." Accessed: Nov. 15, 2024. [Online]. Available: https://virusshare.com/

[23] D. A. Noever and S. E. Miller Noever, "VIRUS-MNIST: A BENCHMARK MALWARE DATASET."

[24] L. Yang, A. Ciptadi, I. Laziuk, A. Ahmadzadeh, and G. Wang, "BODMAS: An Open Dataset for Learning based Temporal Analysis of PE Malware," Proceedings - 2021 IEEE Symposium on Security and Privacy Workshops, SPW 2021, pp. 78–84, May 2021, doi: 10.1109/SPW53761.2021.00020.

[25] C. Henry, "Survey and Use of Ten Volatility Framework Plugins for Malware Analysis," 2017, Utica College.

[26] J. Stadlinger, A. Dewald, and F. Block, "Linux Memory Forensics: Expanding Rekall for Userland Investigation," in 2018 11th International Conference on IT Security Incident Management & IT Forensics (IMF), 2018, pp. 27–46. doi: 10.1109/IMF.2018.00010.

[27] C. Li, Q. Lv, N. Li, Y. Wang, D. Sun, and Y. Qiao, "A novel deep framework for dynamic malware detection based on API sequence intrinsic features," Comput Secur, vol. 116, May 2022, doi: 10.1016/j.cose.2022.102686.

[28] C. Walker, T. Gharaibeh, R. Alsmadi, C. Hall, and I. Baggili, "Forensic Analysis of Artifacts from Microsoft's Multi-Agent LLM Platform AutoGen," in Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–9.

[29] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing." [Online]. Available: https://github.com/huggingface/

[30] Z. Liu et al., "Swin Transformer V2: Scaling Up Capacity and Resolution." [Online]. Available: https://github.com/microsoft/Swin-

[31] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.05150

[32] F. Barros Rodrigues, W. Ferreira Giozza, R. De Oliveira Albuquerque, and L. J. Garcia Villalba, "Natural Language Processing Applied to Forensics Information Extraction with Transformers and Graph Visualization," IEEE Trans Comput Soc Syst, vol. 11, no. 4, pp. 4727–4743, 2024, doi: 10.1109/TCSS.2022.3159677.

[33] S. Shafiee Hasanabadi, A. Habibi Lashkari, and A. A. Ghorbani, "A survey and research challenges of anti-forensics: Evaluation of game-theoretic models in simulation of forensic agents' behaviour," Forensic Science International: Digital Investigation, vol. 35, 2020, doi: 10.1016/j.fsidi.2020.301024.

[34] D. Uroz and R. J. Rodríguez, "On Challenges in Verifying Trusted Executable Files in Memory Forensics," Forensic Science International: Digital Investigation, vol. 32, p. 300917, Feb. 2020, doi: 10.1016/j.fsidi.2020.300917.

[35] L. M. Joshi, "Understanding Threats in Hypervisor, its Forensics Mechanism and its Research Challenges," vol. 119, no. 1, pp. 1–5, 2015.

[36] K. Conlan, I. Baggili, and F. Breitinger, "Anti-forensics: Furthering digital forensic science through a new extended, granular taxonomy," in DFRWS 2016 USA - Proceedings of the 16th Annual USA Digital Forensics Research Conference, Digital Forensic Research Workshop, 2016, pp. S66–S75. doi: 10.1016/j.diin.2016.04.006.

[37] J. Liu, Y. Feng, X. Liu, J. Zhao, and Q. Liu, "MRm-DLDet: a memory-resident malware detection framework based on memory forensics and deep neural network," Cybersecurity, vol. 6, no. 1, Dec. 2023, doi: 10.1186/s42400-023-00157-w.

[38] J. Liu et al., "MemAPIDet: A Novel Memory-resident Malware Detection Framework Combining API Sequence and Memory Features," in Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 2918–2924. doi: 10.1109/CSCWD61410.2024.10580589.

[39] A. Ali-Gombe, S. Sudhakaran, R. Vijayakanthan, and G. G. Richard, "cRGB_Mem: At the intersection of memory forensics and machine learning," Forensic Science International: Digital Investigation, vol. 45, Jul. 2023, doi: 10.1016/j.fsidi.2023.301564.

[40] P. Maniriho, A. N. Mahmood, and M. J. M. Chowdhury, "MeMalDet: A memory analysis-based malware detection framework using deep autoencoders and stacked ensemble under temporal evaluations," Comput Secur, vol. 142, Jul. 2024, doi: 10.1016/j.cose.2024.103864.

[41] S. Dangi, K. Ghanshala, and S. Sharma, "LIFT: Lightweight Incremental and Federated Techniques for Live Memory Forensics and Proactive Malware Detection." [Online]. Available: www.ijacsa.thesai.org

[42] T.-Y. Lin, "A survey of transformers," AI Open, 2022.

[43] "Attention Mechanisms in Transformers." Accessed: Jul. 11, 2025. [Online]. Available: https://www.cloudthat.com/resources/blog/attention-mechanisms-in-transformers

[44] K. Han, "A survey on vision transformer," IEEE Trans Pattern Anal Mach Intell, 2022.

[45] T. Barabosch, N. Bergmann, A. Dombeck, and E. Padilla, "Quincy: Detecting Host-Based Code Injection Attacks in Memory Dumps."

[46] D. Bin Oh, D. Kim, and H. K. Kim, "volGPT: Evaluation on triaging ransomware process in memory forensics with Large Language Model," Forensic Science International: Digital Investigation, vol. 49, Jul. 2024, doi: 10.1016/j.fsidi.2024.301756.

Detecting Fake News Images Using a Hybrid CNN-LSTM Architecture

Dina R. Salem[1], Abdullah A. Abdullah[2], AbdAllah A. AlHabshy[3], Kamal A. ElDahshan[4]
School of Computer Science, Canadian International College – CIC, Cairo, Egypt[1, 2]
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt[3, 4]

***Abstract*—In today's digital world, images have become a double-edged tool in the dissemination of news; as much as they contribute to enriching honest content and communicating information effectively, they are increasingly being used to mislead the public and spread fake news. The ease of manipulating images and taking them out of their original context, or even creating them entirely with advanced techniques, gives them tremendous power in lending false credibility to false narratives, taking advantage of the human eye's tendency to believe what it sees and the image's superior ability to directly evoke emotions. These misleading images, which are often difficult to debunk with the naked eye, spread at lightning speed across digital platforms, allowing fake news to reach and influence large audiences before it can be verified. However, they tend to generate inaccurate reports. This study proposes a model architecture to detect fake news images. Machine learning and deep learning algorithms were used. The deep learning models are used depending on conventional neural nets (CNN), long short-term memory (LSTM) and a hybrid model that combines CNN and LSTM frameworks on Google Cloud. The hybrid model was able to categorize news with better accuracy than using each model individually. The model was tested and trained on a dataset for classifying fake news images. We used different evaluation metrics (precision, recall, F1 metric, etc.) to measure the efficiency of the model.**

Keywords*—*Fake news images; machine learning; deep learning; cloud computing; CNN; LSTM

I. Introduction

The rapid expansion of social media platforms has dramatically increased the dissemination speed and reach of news content worldwide [1]. While this democratization of information access has many benefits, it has also facilitated the widespread propagation of fake news—deliberately fabricated or misleading information designed to deceive readers [2]. Fake news poses significant threats to public opinion, political stability, and social trust, particularly when it spreads unchecked on online platforms [3].

Detecting fake news has become a critical research challenge in artificial intelligence, with numerous studies focusing on automated methods to identify misinformation. Traditionally, many approaches have concentrated on textual analysis using natural language processing (NLP) techniques and machine learning classifiers [4][5].

However, news articles frequently contain multimodal content, including images and videos, which can be equally manipulated to mislead audiences [6]. Despite this, visual content analysis for fake news detection has received comparatively less attention, even though fake images are often used to support fabricated stories and increase their perceived credibility [7].

Recent advances in deep learning have shown promise in addressing this challenge by leveraging convolutional neural networks (CNNs) to extract spatial features from images and recurrent neural networks, such as long short-term memory (LSTM) networks, to capture temporal or sequential dependencies in data. Hybrid architectures combining CNN and LSTM models have demonstrated improved performance in various multimedia classification tasks by integrating spatial and sequential feature extraction capabilities [8][9].

This study proposes a hybrid CNN-LSTM architecture for detecting fake news images, aiming to leverage the strengths of both models to improve classification accuracy. We evaluate our approach on a publicly available dataset from Kaggle, comparing the hybrid model's performance against standalone CNN and LSTM models. Our results confirm that the hybrid model achieves superior accuracy in identifying fake news images, highlighting the potential of combined deep learning architectures in multimodal misinformation detection.

Moreover, hybrid models generally outperform single-architecture approaches by integrating the strengths of both components. Recent studies indicate that combining CNNs with LSTMs yields better detection rates than traditional methods reliant on separate evaluations of text or images [10]. This synergy enhances not only accuracy but also the interpretability of predictions, fostering user trust.

Additionally, these systems provide flexibility in adapting to various media types, essential as fake news increasingly incorporate multimedia elements [11]. Hybrid models ensure adaptability in addressing diverse forms of misinformation effectively.

To lessen the impact of bogus news before it becomes viral, early detection is essential. Machine learning-backed proactive detection techniques can lessen the negative social and psychological effects of misleading narratives [12].

The study is organized as follows: Section II tackles the literature review in some detail. The proposed model architecture is presented in Section III. Section IV presents the experiments and the results with discussion. The study concludes in Section V.

II. Literature Review

This section reviews recent and relevant literature on fake news detection, focusing on the transition from traditional approaches to deep learning techniques, the emergence of hybrid models such as CNN-LSTM, and the growing role of cloud computing in scalable deployment.

A. Text-Based Fake News Detection

Traditional approaches to fake news detection have primarily focused on analyzing the linguistic features and semantic content of news articles. Early methods often relied on feature engineering and machine learning classifiers. For instance, Alhindi, T., Kalita, J., and Sajjad, H. (2020) [13] classified fake news using Support Vector Machines (SVM) and TF-IDF features on the LIAR dataset. While their approach provided reasonable accuracy by capturing word importance, it struggled to understand complex contextual nuances and did not scale well to real-world scenarios due to restricted semantic representation and dataset diversity.

More recently, deep learning models have shown superior capabilities in capturing intricate textual patterns. Gupta, R., et al. (2021) [14] utilized Bidirectional Encoder Representations from Transformers (BERT) for enhanced textual feature extraction and fake news classification. Their model demonstrated high performance on several benchmark datasets by leveraging BERT's contextual embeddings. However, as a purely text-based method, it inherently fails to account for the crucial visual elements that often accompany and influence the perception of news, particularly when key information is conveyed through embedded text or manipulated images.

Another significant line of work explored sequential models for temporal analysis of text. Ma, J., et al. (2019) [15] proposed using Recurrent Neural Networks (RNNs) to model the propagation patterns of rumors over time. While effective in identifying how narratives evolve, their reliance solely on textual content means they cannot detect misinformation where visual cues, such as text embedded in images, play a primary role in conveying the false message, thus limiting their applicability to multimodal misinformation.

B. Image-Based Fake News Detection

Parallel to text-based methods, considerable research has focused on analyzing visual content to identify fake news. These methods often aim to detect image manipulation, inconsistencies, or to extract visual cues indicative of deception. Wang, Y., et al. (2022) [16] developed a CNN-based framework for identifying visual inconsistencies in news images, leveraging various image forensic techniques. Their method achieved high accuracy in detecting manipulated images. However, it does not explicitly analyze any textual content present within the image, which is a critical oversight given that many misleading images rely on superimposed text to convey a deceptive narrative rather than pure visual manipulation.

Similarly, Zhang, L., et al. (2020) [17] proposed a deep learning model to classify images as real or fake based on extracted semantic features using pre-trained CNNs (e.g., ResNet). While effective in capturing high-level visual representations, their approach often struggles with cases where the image itself is authentic but carries a false message through external text added to it or is presented in a misleading context that can only be understood through accompanying textual cues. Their method thus has limited utility for detecting fake news where the visual medium's core deceptive element resides within embedded text, rather than the image's inherent authenticity.

Li, S., et al. (2023) [18] focused on identifying manipulated visual elements in news images using advanced CNN architectures and attention mechanisms to highlight suspicious regions. While their approach was robust against various manipulation techniques, its primary limitation lies in its inability to infer the veracity of the news article or message when the image itself is unaltered but the embedded text (or overall context) is misleading.

C. Multimodal Fake News Detection

Recognizing the limitations of unimodal approaches, a growing body of research has explored multimodal strategies that integrate information from both textual and visual modalities to enhance fake news detection. These approaches aim to capture the complex interdependencies between different data types.

Liu, X., et al. (2021) [19] proposed a multimodal framework that combines a CNN for image feature extraction and a GRU (Gated Recurrent Unit) for text features, fusing them at a later stage for classification. Their model showed improved performance compared to unimodal baselines by leveraging complementary information. However, their textual analysis was primarily focused on accompanying article text rather than the direct processing of text embedded within the images, which presents distinct challenges due to its unique formatting and often succinct nature.

Pan, W., et al. (2022) [20] developed a co-attention network that learns cross-modal interactions between text and images to better align information for fake news detection. While innovative in capturing intricate relationships, their method might not explicitly distinguish between general image features and the specific textual content within the image, potentially overlooking the direct semantic impact of embedded text. This limits their ability to robustly detect misinformation where the key deceptive element is the overlaid text.

Despite these advancements, a significant challenge remains in effectively integrating deep learning models that can specifically process and understand the unique characteristics of text directly embedded within images for fake news detection. Existing multimodal approaches often treat images as mere visual contexts or focus on detached textual descriptions, leaving a critical gap in analyzing the inherent deceptive power of image-text composites. Our proposed hybrid CNN-LSTM model directly addresses this limitation by meticulously extracting visual features via CNNs and simultaneously processing the sequential, embedded textual information using LSTMs, providing a more granular and effective approach to multimodal fake news detection.

In conclusion, the literature reveals several gaps: limited hybrid model integration, inconsistent usage of scalable deployment platforms, and a lack of comprehensive testing across diverse datasets. These gaps justify the need for a cloud-

optimized hybrid CNN-LSTM model that can robustly handle fake news detection at scale.

While these multimodal studies have made significant strides, many do not explicitly focus on the challenging domain of text directly embedded within images, which is a distinct modality requiring specialized processing, particularly given the potential for OCR inaccuracies. The existing literature often treats images as general visual context or focuses on visual manipulation, rather than the intrinsic textual information they carry. This gap forms the primary motivation for our current study, aiming to provide a robust deep learning solution for fake news detection that specifically leverages and integrates textual content extracted from images [21].

III. Methodology

The proposed methodology investigates the most famous state-of-the-art deep learning algorithms for Fake news detection; like deep learning techniques, they have the advantage of their ability to capture semantic features from textual data automatically. The three different combinations of deep neural networks, namely, CNN, LSTM and CNN-LSTM classification, have been performed. The proposed methodology is shown in Fig. 1.

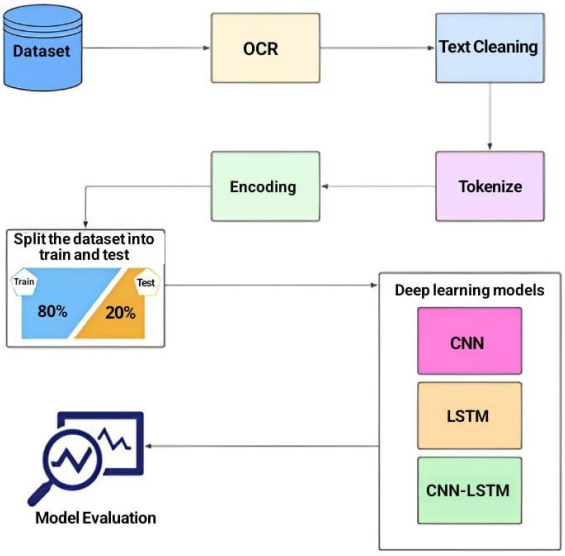


Fig. 1. System architecture.

A. Dataset

The Fake News Classification Image Dataset, publicly available on Kaggle [22]. This dataset comprises both textual and visual information, crucial for multimodal fake news detection.

- Dataset Source: Kaggle (Fake News Classification Image).
- Data Partitioning: The dataset was split into training, validation, and test sets.
  - Training Samples: 3600
  - Validation Samples: 450
  - Test Samples: 255
- Class Distribution: The dataset appears to have a relatively balanced distribution of "Fake" (class 0) and "Real" (class 1) news samples, as evidenced by the support values in the classification reports (e.g., Test set 126 for class 0, 129 for class 1).

B. Data Preprocessing

This effective preprocessing of both image and textual data is paramount for deep learning models. The following steps were applied:

1) Optical Character Recognition (OCR): Textual content embedded within the images was extracted using an OCR technique. This process is critical for converting visual text into a machine-readable format. The approach aligns with methodologies seen in advanced OCR systems, which aim to extract characters from optical images into editable formats, potentially leveraging tools such as PaddleOCR for pseudo ground truth acquisition as discussed in relevant literature [23].

2) Text cleaning: Following OCR extraction, the raw text underwent a meticulous cleaning process to standardize its format and reduce noise, which is crucial for subsequent natural language processing. This involved converting all extracted text to lowercase to ensure uniformity. Additionally, non-alphanumeric characters, such as punctuation and special symbols, were systematically removed to retain only relevant linguistic elements (letters, numbers, and spaces). Finally, multiple consecutive whitespace characters were consolidated into single spaces, and leading or trailing whitespace was trimmed, resulting in a clean and normalized textual representation [24].

3) Tokenization and padding: To prepare the cleaned textual data for input into the neural networks, a two-stage process of tokenization and padding was performed. Initially, a tokenizer was fitted on the entire collection of cleaned extracted text, encompassing data from the training, validation, and test sets. This process built a vocabulary, with a predefined size of 512 words, mapping each unique word to a numerical index. Subsequently, the cleaned text from each dataset (training, validation, and test) was converted into sequences of these numerical indices. To ensure that all input sequences had a uniform length, a max_length was determined by identifying the longest sequence across all datasets. All shorter sequences were then padded with zeros at the end ('post' padding) until they reached this max_length. This standardization is vital for batch processing in deep learning models [25].

4) Label encoding: The categorical labels associated with each image, 'real' and 'fake', were transformed into numerical representations suitable for binary classification. Specifically, 'real' news instances were mapped to the numerical value of 1, and 'fake' news instances were mapped to 0. This encoding converted the target variables into a format directly usable by the model's output layer [26].

Following the completion of the preprocessing stages, the data is divided into testing and training sets in an 80:20 ratio. Following that, the features are suitably reshaped to satisfy the

input specifications of the various deep learning architectures (CNN, LSTM and CNN-LSTM models).

C. Model Architectures

This study investigates three distinct deep learning architectures: a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM) network, and a Hybrid CNN-LSTM model. All models utilize the preprocessed textual data as input.

1) Convolutional Neural Network (CNN): The CNN model is designed for feature extraction from sequential data, leveraging its ability to capture local patterns. The architecture is as shown in Fig. 2:

- Embedding Layer: input_dim=512 (vocabulary size), output_dim=16 (embedding dimensions).
- Conv1D Layer: filters=8, kernel_size=3, activation='relu'.
- GlobalAveragePooling1D Layer: Reduces each feature map to a single value, capturing the most significant features.
- Dropout Layer: rate=0.5 for regularization, preventing overfitting.
- Dense Layer: units=1, activation='sigmoid' for binary classification output.

The model's input shape is configured to handle sequences of varying batch sizes and the determined max_length.

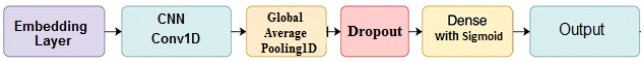


Fig. 2. Workflow of the CNN model.

2) Long Short-Term Memory (LSTM): The LSTM model is particularly suited for capturing long-range dependencies in sequential data. The architecture is as shown in Fig. 3:

- Embedding Layer: input_dim=512, output_dim=16.
- Bidirectional LSTM Layer: units=8 A bidirectional setup processes the sequence in both forward and backward directions, enhancing context understanding.
- Dropout Layer: rate=0.5.
- Dense Layer: units=1, activation='sigmoid'.

The model's input shape is set to accommodate sequences of varying batch sizes and the max_length.

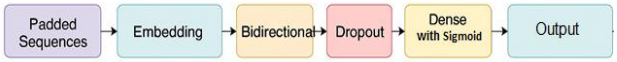


Fig. 3. Workflow of the LSTM model.

3) A Hybrid (CNN-LSTM) model: This hybrid model combines the strengths of both CNNs for local feature extraction and LSTMs for capturing sequential dependencies. This architecture aims to leverage the benefits of both approaches for improved fake news detection as shown in Fig. 4:

- Embedding Layer: input_dim=512, output_dim=128.
- First Conv1D Layer: filters=256, kernel_size=3, activation='relu', padding='same'.
- First MaxPooling1D Layer: pool_size=2.
- Second Conv1D Layer: filters=128, kernel_size=3, activation='relu', padding='same'.
- Second MaxPooling1D Layer: pool_size=2.
- Bidirectional LSTM Layer: units=128, return_sequences=False (outputs only the last hidden state).
- First Dense Layer: units=128, activation='relu'.
- Output Dense Layer: units=1, activation='sigmoid'.

The model's input shape is built to handle sequences of varying batch sizes and the max_length.



Fig. 4. Workflow of the CNN-LSTM model.

D. Training Details

All three models (CNN, LSTM, and Hybrid) were trained using consistent parameters to ensure fair comparison:

- Loss Function: binary_crossentropy was chosen as the loss function, suitable for binary classification tasks.
- Metrics: Model performance was evaluated based on accuracy.
- Epochs: Each model was trained for 25 epochs.
- Batch Size: A batch_size of 32 was used during training.
- Validation Data: The validation set (val_x, val_y) was used to monitor performance and prevent overfitting during training.
- Execution Environment: The training was conducted on Google Colab, leveraging its computational resources to facilitate efficient model training and experimentation.

The training process for each model was initiated, and the total training time was recorded to assess computational efficiency.

IV. Experiments and Results

This section details the experimental setup, presents the quantitative results of the trained models, and discusses their performance in classifying fake news from images with embedded text. The evaluation focuses on comparing the proposed hybrid CNN-LSTM model against standalone CNN and LSTM models.

A. Hyperparameters of Each Model

As indicated in Table I, we examine the key hyperparameters that govern their architecture and training process. These consist of the number and layers types, activation functions, optimizer settings, batch size, dropout rates, and the output layer configurations. The following table summarizes the hyperparameters for each model, highlighting the differences in their architectures and the settings used to train them. This comparison provides insights into how each model was structured and trained to fake news detection task.

TABLE I. COMPARISON BETWEEN HYPEPARAMETERS OF EACH MODEL

Hyperparameter	CNN Model	LSTM Model	Hybrid (CNN-LSTM) Model
Layers	Embedding, Conv1D, GlobalAveragePooling1D, Dropout, Dense	Embedding, Bidirectional LSTM, Dropout, Dense	Embedding, Conv1D, MaxPooling1D, Conv1D, MaxPooling1D, Bidirectional LSTM, Dense, Dense
Number of Layers	5	4	8
Activation Function	ReLU (Conv1D), Sigmoid (Output Dense)	Sigmoid (Output Dense)	ReLU (Conv1D, Dense), Sigmoid (Output Dense)
Optimizer (Adam)	Adam (Learning Rate: 0.001)	Adam (Learning Rate: 0.0005)	Adam (Learning Rate: 0.0001)
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy
Batch Size	32	32	32
Epochs	25	25	25
Dropout	0.5 (after GlobalAveragePooling1D)	0.5 (after Bidirectional LSTM)	Not explicitly used
No. of Filters (CNN)	8 (Conv1D)	N/A	256 (1st Conv1D), 128 (2nd Conv1D)
Kernel Size (CNN)	3 (Conv1D)	N/A	3 (both Conv1D layers)
Pooling Layer (CNN)	GlobalAveragePooling1D	N/A	MaxPooling1D (pool_size=2), used twice
LSTM Units	N/A	8 (Bidirectional LSTM)	128 (Bidirectional LSTM)
Input Shape	16	16	128
Output Layer	(None, max_length)	(None, max_length)	(None, max_length)

B. Evaluation Metrics

To thoroughly assess the performance of each classification model, a suite of widely recognized metrics was employed. These metrics provide insights into different aspects of model accuracy and robustness in distinguishing between "Fake" and "Real" news. The metrics are derived from the Confusion Matrix, which quantifies True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [27][28]:

- Accuracy: Represents the proportion of correctly classified instances out of the total instances.
- Precision: For a given class, it is the ratio of correctly predicted positive observations to the total predicted positive observations. It quantifies the ability of the model to avoid false alarms.
- Recall (Sensitivity): For a given class, it is the ratio of correctly predicted positive observations to all observations in that class. It quantifies the ability of the model to find all positive instances.
- F1-Score: The harmonic mean of Precision and Recall. It provides a single score that balances both precision and recall, especially useful in cases of imbalanced class distributions.
- ROC AUC (Receiver Operating Characteristic Area Under the Curve): Measures the ability of the model to distinguish between classes. A higher AUC indicates better discriminatory power across various classification thresholds.
- PR AUC (Precision-Recall Area Under the Curve): Focuses on the trade-off between precision and recall for different thresholds. It is particularly informative for imbalanced datasets, where it provides a more robust evaluation than ROC AUC.
- MCC (Matthews Correlation Coefficient): A single metric that considers all four values in the confusion matrix. It is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. A value of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction.

C. Results Evaluation

In this section, we display the outcomes of the models that were used on the Fake News Classification image Dataset. We used CNN, LSTM and Hybrid models for this dataset. Accuracy, precision, recall, F1-score, confusion matrix, classification report, accuracy curve, and loss curve are among the performance measures. These measurements give a thorough evaluation of each model's performance on the Fake News Classification Image Dataset, providing important information about how well they work for fake news detection.

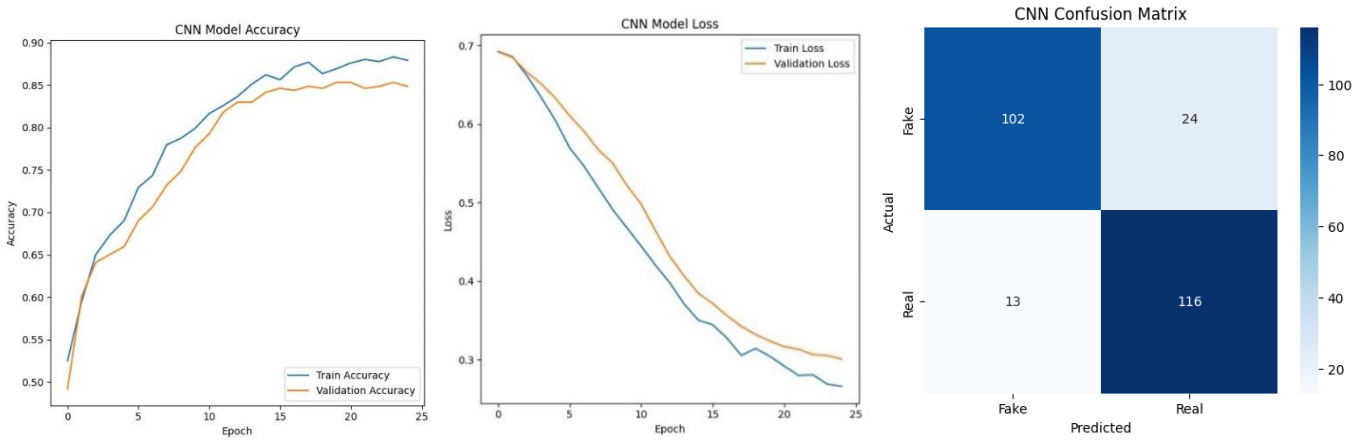
1) CNN Model: Table II, shows the evaluation metrics of the CNN model demonstrating better performance.

TABLE II. EVALUATION METRICS OF CNN MODEL

Metric	CNN Model
Accuracy	0.8667
Precision (Fake)	0.9259
Recall (Fake)	0.7937
F1-Score (Fake)	0.8547
Precision (Real)	0.8231
Recall (Real)	0.9380
F1-Score (Real)	0.8768
ROC AUC	0.9476
PR AUC	0.9559
MCC	0.7403

These results validate the CNN model's capacity to reliably detect fake news.

Fig. 5 represents fake and real image news using the CNN Accuracy Curve in (a), Loss Curve in (b) and Confusion Matrix in (c). As shown in Fig. 5(a), the results of the a model accuracy on training and validation data during training both curves exhibit a steady increase, converging at a high accuracy level, which indicates the model's ability to generalize well. According to the Loss Curve in Fig. 5(b), the decreasing trend in both curves signifies effective model optimization, with the validation loss closely following the training loss, suggesting minimal overfitting. Also in Fig. 5(c), the results of the confusion matrix the high values along the diagonal indicate strong predictive accuracy, while the off-diagonal values represent misclassifications. The intensity of the color reflects the frequency of predictions, highlighting the model's effectiveness in distinguishing between classes.



(a) (b) (c)

Fig. 5. Performance evaluation of the CNN model: (a) Accuracy curve, (b) Loss curve and (c) Confusion matrix.

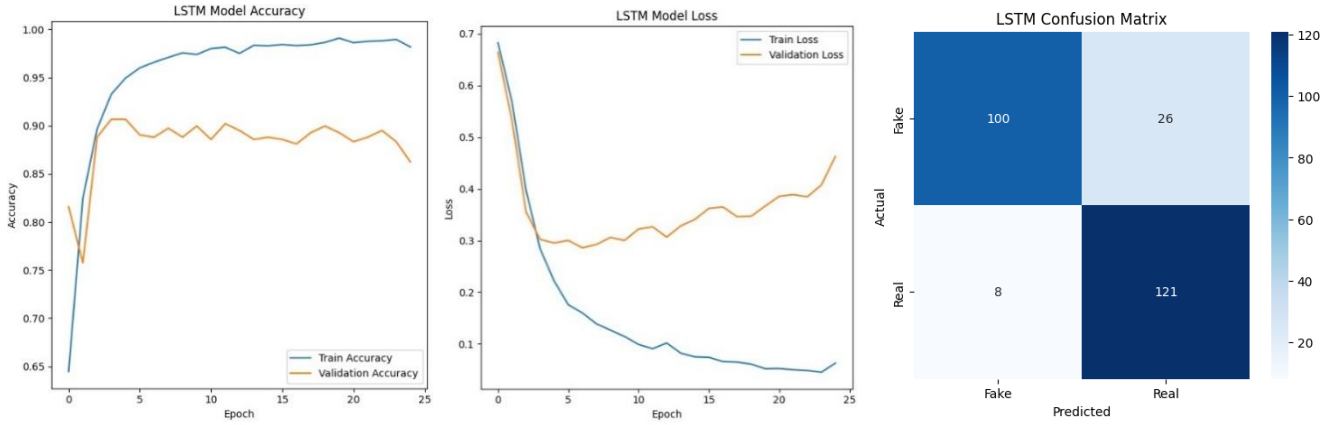
2) LSTM Model: Table III, shows the evaluation metrics of implemented LSTM model. It demonstrated strong classification capabilities.

TABLE III. EVALUATION METRICS OF LSTM MODEL

Metric	LSTM Model
Accuracy	0.8549
Precision (Fake)	0.8870
Recall (Fake)	0.8095
F1-Score (Fake)	0.8465
Precision (Real)	0.8286
Recall (Real)	0.8992
F1-Score (Real)	0.8625
ROC AUC	0.9493
PR AUC	0.9577
MCC	0.7121

Fig. 6 represents fake and real image news using the LSTM Accuracy Curve in (a), Loss Curve in (b) and Confusion Matrix in (c). As shown in Fig. 6(a), the upward trend in both curves indicates the model's improved performance as training progresses, with the training and testing accuracies converging at a high value. According to the Loss Curve in Fig. 6(b), the decreasing trend in both curves signifies effective model optimization, with the validation loss closely following the training loss, suggesting minimal overfitting. Also in Fig. 6(c), the confusion matrix visualizes the model's classification performance across different categories. The strong diagonal dominance indicates high classification accuracy, while the off-diagonal elements represent misclassifications.

3) A Hybrid (CNN-LSTM) Model: Table IV, shows the evaluation matrices of the hybrid model. It demonstrated strong classification capabilities. The model's capacity for generalization was enhanced and overfitting was lessened with the aid of batch normalization and dropout. The hybrid model showed its strong performance in fake news detection tasks by outperforming the baseline accuracy.



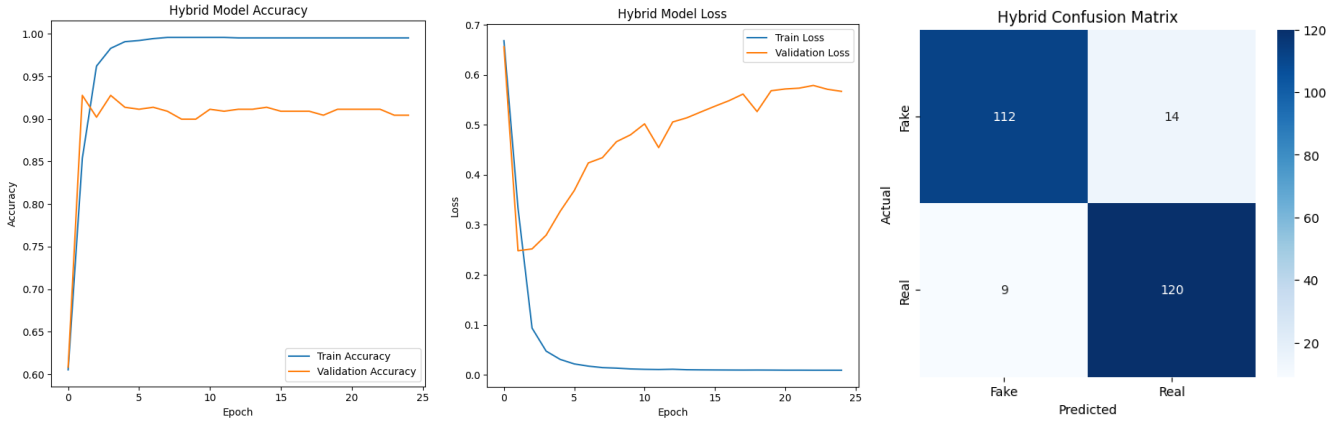
(a) (b) (c)

Fig. 6. Performance evaluation of the LSTM model: (a) Accuracy curve, (b) Loss curve and (c) Confusion matrix.

TABLE IV. EVALUATION METRICS OF THE CNN-LSTM MODEL

Metric	Hybrid Model
Accuracy	0.9098
Precision (Fake)	0.9256
Recall (Fake)	0.8889
F1-Score (Fake)	0.9069
Precision (Real)	0.8955
Recall (Real)	0.9302
F1-Score (Real)	0.9125
ROC AUC	0.9673
PR AUC	0.9713
MCC	0.8201

Fig. 7 represents fake and real image news using the CNN-LSTM Accuracy Curve in (a), Loss Curve in (b) and Confusion Matrix in (c). As shown in Fig. 7(a), this plot displays the training and testing accuracy over multiple epochs. The upward trend in both curves indicates the model's improved performance as training progresses, with the training and testing accuracies converging at a high value. According to the Loss Curve in Fig. 7(b), the graph illustrates the training and testing loss across epochs. The decreasing loss curves reflect effective model learning, with the test loss closely following the training loss, suggesting good generalization with minimal overfitting. Also in Fig. 7(c), the confusion matrix visualizes the model's classification performance across different categories. The strong diagonal dominance indicates high classification accuracy, while the off-diagonal elements represent misclassifications.



(a) (b) (c)

Fig. 7. Performance evaluation of the hybrid (CNN-LSTM) model: (a) Accuracy curve, (b) Loss curve and (c) Confusion matrix.

D. Discussion

This study evaluates and compares the performance of three deep learning models – CNN, LSTM and CNN-LSTM model for fake news detection using the Fake News Classification Image Dataset. The results give important insights into the effectiveness of these models and their suitability for specific scenarios in Artificial intelligence applications.

1) Performance on fake news classification image dataset: For the Fake News Classification Image dataset, the proposed hybrid LSTM-CNN model outperformed the standalone CNN and LSTM models in detecting fake news. While CNN is effective in extracting local features and LSTM captures sequential dependencies, each model has limitations when used independently. By combining their strengths, the hybrid

architecture achieved better generalization and higher classification accuracy. On the Fake News Classification Image dataset, the hybrid model reached an accuracy of 90.98%, surpassing CNN (86.67%) and LSTM (85.49%), with improved sensitivity and specificity. These results confirm the hybrid model's effectiveness in handling the complexity of fake news detection.

This improvement is consistent across several critical metrics:

- ROC AUC (0.9673) and PR AUC (0.9713): The significantly higher values for the Hybrid model in both ROC AUC and PR AUC suggest excellent discriminatory power and strong performance even in potentially imbalanced classification scenarios.
- Matthews Correlation Coefficient (MCC): With an MCC of 0.8201, the Hybrid model exhibits a strong correlation between predicted and true labels, indicating a highly reliable classification performance across both classes.
- F1-Scores: The Hybrid model achieved the highest F1-Scores for both "Fake" (0.9069) and "Real" (0.9125) classes, highlighting its balanced ability to maintain high precision and recall for both positive and negative classifications.
- Recall (Fake) and Specificity (Fake): The Hybrid model shows a notable improvement in Recall for the "Fake" class (0.8889) and Specificity (0.8889), which is crucial for effectively identifying actual fake news and minimizing false positives (classifying real news as fake). While the LSTM model had a slightly higher Recall for "Real" news, the Hybrid model demonstrates a better overall balance and a higher true negative rate.

2) Comparative insights: The comparison of results highlights the CNN-LSTM model's robustness, particularly for datasets with complex temporal or sequential patterns, such as the Fake News Classification Image dataset. While the CNN and LSTM models also performed well, their slight limitations in handling temporal dependencies make them less effective than Hybrid in such scenarios.

3) Implications and practical applications: The results clearly highlight the crucial importance of selecting suitable models based on the characteristics of datasets used in fake news detection. The Convolutional Neural Network (CNN)'s ability to handle feature-based patterns makes it a preferable choice in contexts where local spatial relationships within text (such as linguistic constructions and phrases) are critically important. Conversely, the Long Short-Term Memory (LSTM) model, with its superior capacity to capture sequential dependencies, proves valuable for datasets where temporal or contextual patterns across the text play a key role in distinguishing fake news. The hybrid model proved to be the strongest due to its effective utilization of the strengths of both CNN and LSTM. Furthermore, this study underscores the practical feasibility of these models in real-world fake news detection applications, supported by their high performance and adaptability. The insights gained can guide future developments in fake news detection systems, particularly in confronting the ever-increasing volume of malicious content online.

V. Conclusion and Future Work

In order to detect fake image news, this study suggested a hybrid deep learning model that combines the strengths of CNN and LSTM. The convolutional layers are adept at extracting local features and patterns (e.g., n-gram features from text sequences), while the Bidirectional LSTM layers excel at capturing long-range dependencies and contextual information in sequences from both forward and backward directions. This synergy, further bolstered by the integration of OCR-extracted text from images, allows the hybrid model to learn richer and more discriminative representations from the combined textual input derived from multimodal data, leading to superior fake news detection. The model was implemented in a cloud computing environment.

According to the experimental results, the hybrid model achieved a high performance of 90.98% and outperformed the results of running individually the CNN and LSTM models, in terms of accuracy, sensitivity, and specificity. This performance was resulted by the successful extraction of both local features and long-term dependencies made possible by the combination of both models.

While the achieved results in detecting fake news images using a hybrid CNN-LSTM architecture are highly promising, this study is not without certain limitations. Specifically, the Fake News Classification Image Dataset utilized may still present constraints in terms of its diversity, complexity, or its full representation of the constantly evolving deceptive visual content found in real-world scenarios. Future work could explore more challenging, imbalanced, or real-world datasets derived from diverse sources to further assess the robustness of the proposed hybrid model against evolving fake news dissemination tactics. Additionally, given the resource-intensive nature of deep learning models, there is a need to explore lightweight alternatives or model optimization techniques for deployment in resource-constrained environments or on devices with limited computational capacity.

Furthermore, the model's capabilities can be refined to include not only images but also video content that may contain fake information or be AI-generated, addressing the growing challenge of multimodal disinformation.

References

[1] F. Pierri and S. Ceri, “False news on social media: A data-driven survey”, ACMSIGMOD Record, vol. 48, no. 2, pp. 18–27, 2019.

[2] S.Kumar andN. Shah, “False information on web and socialmedia:Asurvey,”, ArXiv, vol. abs/1804.08559, 2018.

[3] M. Choudhary, S. Jha, Prashant, D. Saxena, and A. K. Singh, “A review of fake news detection methods using machine learning,” in 2021 2nd International Conference for Emerging Technology, INCET 2021, 2021.

[4] T. Traylor, J. Straub, Gurmeet, and N. Snell, “Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator,” in Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019, 2019.

[5] Shu, K., Wang, S., and Liu, H, "Beyond news contents: The role of social context for fake news detection", ACM Transactions on Information Systems, 39(3), 1–31, 2021.

[6] Saadi, S. M., and Al-Jawher, W. A. M, "Image Fake News Prediction Based on Random Forest and Gradient-boosting Methods", Journal Port Science Research, 6(4), 2023.

[7] Z. Liang, "Fake News Detection Based on Multimodal Inputs," Computers, Materials and Continua, vol. 75, no. 2, 2023.

[8] Thakre, N., Thakare, V., Sawarkar, A., Mankar, T., and Dhande, S "fake news detection using convolution neural network on social platforms", Journal of Research Administration, Society of Research Administrators International, 2020.

[9] Jaybhaye, S. M., Badade, V., Dodke, A., Holkar, A., and Lokhande, P, " Fake News Detection using LSTM based deep learning approach", ITM Web of Conferences, 56, 03005, 2023.

[10] D. G. Dev, V. Bhatnagar, B. S. Bhati, M. Gupta, and A. Nanthaamornphong, "LSTMCNN: A hybrid machine learning model to unmask fake news," PMC10844048, 2024.

[11] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," S0020025521002826, 2021.

[12] Pavan M N, Prasad, P. R., Gowda, T., Vibhakar, T. S., and Shidnal, S, " Fake News Detection using Machine Learning", International Research Journal of Engineering and Technology (IRJET), 7(12), 2395-0056, 2020.

[13] Alhindi, T., Kalita, J., & Sajjad, H, "Contextual Fake News Detection Using Machine Learning Techniques". In Proceedings of the IEEE International Conference on Big Data,2020.

[14] Gupta, R., et al, " BERT-Based Textual Feature Extraction for Fake News Detection", Expert Systems with Applications, 178, 115000, 2021.

[15] Ma, J., et al, "Detecting Rumors on Twitter with Tree-structured Recursive Neural Networks", Journal of Information Science, 45(4), 483-494, 2019.

[16] Wang, Y., et al, "A CNN-based Framework for Visual Misinformation Detection through Image Forensics", Journal of Visual Communication and Image Representation, 82, 103350, 2022.

[17] Zhang, L., et al,"Deep Learning for Image-based Fake News Detection", Neurocomputing, 381, 230-241,2020.

[18] Li, S., et al, "Attention-Based Deep Learning for Manipulated Image Detection in News", Multimedia Tools and Applications, 82(5), 7891-7910, 2023.

[19] Liu, X., et al, "Multimodal Fake News Detection: A Unified Framework with Hierarchical Attention", Information Sciences, 574, 21-34, 2021.

[20] Pan, W., et al, "Cross-Modal Attention Network for Fake News Detection", IEEE Transactions on Multimedia, 24, 1-13, 2022.

[21] Lin, S. Y., Chen, Y. C., Chang, Y. H., Lo, S. H., and Chao, K. M, "Text-image multimodal fusion model for enhanced fake news detection", Science Progress, 107(4), 1-31, 2024

[22] The datasets used during the current study are available in Kaggle https://www.kaggle.com/datasets/vivek1coder/fake-news-classification-image/data

[23] Haoran Wei, Chenglong Liu, et al, "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model", arXiv:2409.01704, 2024.

[24] Vasu Agarwal, H Parveen Sultana, Srijan Malhotra, and Amitrajit Sarkar, "Analysis of classifiers for fake news detection", Procedia Computer Science, 165:377‑383, 2019.

[25] Pritika Bahad, Preeti Saxena, and Raj Kamal, "Fake news detection using bi-directional lstmrecurrent neural network", Procedia Computer Science, 165:74–82, 2019.

[26] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo, and K.-M. Chao, "Text–image multimodal fusion model for enhanced fake news detection," Sci Prog, vol. 107, no. 4, p. 00368504241292685, 2024,

[27] Khalil, A., Jarrah, M., and Aldwairi, M, "Hybrid Neural Network Models for Detecting Fake News Articles", Human-Centric Intelligent Systems, 4, 136-146, 2024.

[28] First, B. M., Shanmugavadive, K., Sathishkumar, V. E., Maruthappa, M., Subramanian, M., and Thinakaran, R, "Attention Mechanism-Based CNN-LSTM for Abusive Comments Detection and Classification in Social Media Text", International Journal of Advanced Computer Science & Applications, 15(10), 2024.

Heuristics-Based Clustering of Internet of Vehicle Based on Effective Approximate QoS Guideline for Message Dissemination

Tanuja Kayarga[1], S Ananda Kumar[2], Lakshmi B S[3], Anil Kumar B H[4]
Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, India[1, 3, 4]
SCOPE, Vellore Institute of Technology, Vellore, India[2]

***Abstract*—Internet of Things and connected world concepts are emerging foreground in all walks of life, and Internet of Vehicles (IoV) is a natural evolution of these concepts for a connected vehicular environment. Clustering is needed in the IoV network for effective management of network resources and to avoid congestion in the network. On the effective cluster base, any efficient routing protocol can be realized for message dissemination. In most of the existing work, clustering based on multiple mobility metrics is the first step, followed by routing over established clusters. This work proposes a clustering solution without the need for frequent clustering or reduced cluster maintenance effort. The proposed solution establishes a virtual QoS guideline path for message dissemination based on graph theory as a first step, and clustering of the network is done to further improve the QoS in the guideline path using a heuristic algorithm. The proposed solution demonstrates notable improvements over existing approaches. Specifically, it achieves an average cluster duration that is at least 6% higher and reduces cluster maintenance overhead by 7%. In terms of Quality of Service (QoS), the solution attains a 3.6% higher packet delivery ratio, along with a 21% reduction in end-to-end delay and a 28% decrease in routing overhead.**

***Keywords*—QoS; clustering; meta heuristics; IoV; PSO; steiner minimal tree**

I. Introduction

The idea behind connected vehicles motivates the revolution of intelligent transportation systems. Communication among connected vehicles can be used to reduce traffic congestion, improve road safety, etc. Vehicles are an important participant in this connected network, with connectivity assisted by direct communication and infrastructure-based communication with roadside units (RSU). Vehicles can use Car-to-Car Communication-Consortium (C2C-CC) and Dedicated Short Range Communication (DSRC), Internet and V2X communication protocols for communication. A major problem in Internet of vehicle networks is scalability and robust routing against topology changes due to vehicular mobility. Clustering is a solution to these problems. Vehicles can be grouped based on similarity of either proximity, trajectory, etc. Message dissemination over a clustered topology has comparatively lower overhead compared to without clustering. Clustering also helps to limit channel contention among cluster members and facilitates fair access. Clustering also assists in spatial reuse of resources like bandwidth. Establishing stable clusters in the presence of higher vehicular mobility is a challenging task. Network topology changes frequently in a vehicular network due to higher vehicular mobility. Under these scenarios, cluster maintenance becomes a huge overhead. Clustering is the basis on which routing algorithms are constructed. When clustering is unstable, maintaining QoS for packet delivery becomes difficult.

Many clustering algorithms have been designed for vehicular networks and are discussed in detail in Section II. These clustering algorithms use various mobility metrics like speed, direction of travel and location of vehicles to organize the vehicles into groups. These algorithms make assumptions about mobility patterns, and when assumptions are challenged, the cluster maintenance effort is huge. Also, the routing protocols built over these clusters have lower QoS when the stability of the cluster is compromised.

In this work, a clustering solution is proposed with reduced cluster maintenance overhead and higher QoS for message dissemination. A virtual QoS guideline path for message dissemination is established as the first step based on graph theory, and clustering is established over the periphery of this guideline path using a soft computing-based heuristic algorithm. The contributions of the suggested work are listed below:

1) A novel virtual QoS guideline path construction based on multi-criteria QoS.

2) An algorithm for clustering based on the virtual QoS guideline path and mobility metrics using heuristics.

The structure of the study is as follows: Section II displays the current clustering algorithms for IoV and critical analysis of their issues. Section III displays the proposed clustering algorithm. The findings of the suggested algorithm are shown in Section IV along with a comparison to previous research. The suggested clustering algorithm's final thoughts and the range of further research are presented in Section V.

II. Related Work

Lakas et al. [1] used link expiration time as the mobility metric for clustering the vehicular network. Once the clustering topology is built using an artificial bee colony algorithm, one can discover paths over clusters with goal of achieving higher QoS. Testing against different scenarios, the approach

performed well under the assumption of normal traffic density in highway scenario. When assumption is broken with higher traffic density, QoS deteriorates. Wahab et al. [2] used mobility metrics of velocity and distance to create stable clusters.

Over the clustered topology ant colony algorithm is used for selecting route under the constraint of higher reliability. Direction of movement is not considered while clustering the network. Wang et al. [3] used multiple metrics of link lifetime, node degree, and anticipated transmission count for selecting the head of the cluster. Authors also used an election strategy to select clusters of maximal stability. Over the clustered topology, a weighted routing protocol is proposed where the relay cluster head is selected based on predicted node degree, link life, and transmission time. ON any variation in speed and density, both the stability of the cluster and QoS of routing paths are affected. Khan et al. [4] used the algorithm Moth Flame optimizer for clustering the IoV network. Optimal cluster heads with minimal distance to cluster members are found, so that transmission energy is minimized. The clustering process considered only energy optimization and did not consider other QoS parameters. Ahizoune et al. [5] suggested using a two-stage clustering approach to build durable VANET clusters. In the first stage of setup, clustering is done based on proximity criteria. In the second stage of maintenance, the backup cluster head is selected and role exchange with the primary is done to maintain cluster stability and reliability. The role exchange with the primary is executed to uphold cluster stability and reliability. Cluster stability is quantified by the variance in velocity among vehicles within the cluster. Clustering aims to enhance stability, disregarding QoS parameters. Shankar et al. [6] implemented a hybrid swarm algorithm for clustering in the VANET network, integrating Particle Swarm Optimization (PSO) with the Harmony Search Algorithm (HAS). The fitness function for clustering focuses on minimizing energy consumption. The problem of PSO getting into local minima is solved by using HSA along with PSO. The proposed algorithm improved the lifetime without consideration of QoS parameters. Ali et al. [7] clustered the mobile ad hoc network using particle swarm optimization. The fitness function for PSO was designed based on node degree, transmission power, node's energy consumption and node's velocity. The objective of the fitness function was to minimize energy consumption. Fahad et al. [8] used the grey wolf optimization algorithm for clustering the VANET. The optimization criterion for clustering is stability of clusters. Compared to PSO the hunting mechanism of grey wolf was able to converge to better results. Dalbah, et al. [9] came up with JAYACIoV which used an integration of two algorithms k–means and automatic clustering, using 171 testing conditions, providing a 64.82% result. Ayed, et al. [10] addressed the clustering process based on diverse parameters such as the trust value, the safety distance, and the energy factor. Oranj, et al. [11] proposes an algorithm to discover the path based on 2 parameters delay time and reliability. The results are evaluated using NS2 simulator. Shaikh, et al. [12] proposed an algorithm to include security measures within the routing process by efficiently dividing the workload across the alternative paths to reduce the congestion and travel cost. Aadil et al. [13] clustered the VANET network using dragon fly optimization algorithm. The cluster heads were selected based on the speed and direction of the movement. The cluster head transmission range was also adjusted dynamically to increase the cluster coverage. The approach works only for high way scenario. Hadded et al. [14] grouped the VANET using angle-based clustering. The selection of stable clusters is based on the vehicles' angular location and direction. Stability is the only criterion for cluster head selection in this work without consideration for multiple QoS factors. Bersali, M [15] approach CCA-IoV, uses the essential computing and storage capabilities of modern vehicles to operate the clustering process. Gasmi et al. [16] suggested a novel weight-based clustering technique for Internet of Vehicles networks. Density, speed, position, and delay measurements provide the foundation for clustering. QoS in the form of a delay metric was considered in clustering decision. Khan et al. [17] proposed a clustering algorithm for VANETs that is based on connectivity. Links connectivity is used mobility metric for clustering decision. The ideal number of clusters was also determined by the authors using heuristics. In this work, cluster stability—rather than QoS—was the criterion for clustering decisions. Gasmi et al. [18] clustered the IoV network by grouping the nodes moving in the same direction with the same speed. Overall, the clustering fuzzy-based routing algorithm considering QoS was designed. In unstable cluster, providing a guaranteed QoS becomes difficult. Sandeep Yerrathi et al. [19] address the challenges of stability, reliability, and QoS in highly dynamic environments by proposing a novel nature-inspired clustering approach. The African vulture Optimization-Based clustering Algorithm (AVOCA) intelligently forms optimal clusters, reducing network randomness and improving stability across varying grid sizes, node densities and transmission ranges. This study presents a method for cluster head (CH) selection, coordination and maintenance focusing on reducing communication costs. Experiments shows that AVOCA produces up to 45% fewer clusters than existing algorithms like CAMONET, SAMNET, i-WOA and HHO, achieving better load balancing and communication efficiency. Husnain et al. [20] tackle the challenges of dynamic topology and sparse vehicle distribution in VANETs, which complicate stable clustering and routing in Intelligent Transportation systems. To address this, authors proposed an intelligent probability-based bio-inspired Whale Optimization Algorithms (i-WOA) for clustering in VANETs considering factors like communication range, node density, velocity and highway routing. Experimental results across various network scenarios show that i-WOA produces optimal number of cluster heads. Statistical analyses confirm a 75% improvement in clustering optimization, effectively reducing communication costs and routing overhead while extending network lifetime. Jamalzadeh, M et al. [21] address routing challenges on the internet of Vehicle, which is characterized by rapid topology changes and variable node dynamics. They proposed EC-MOPSO, an edge computing assisted, cluster-based routing algorithms that leverages multi-objective particle swarm optimization ((MOPSO), roadside units (RSUs) acts as edge devices to optimize routing by minimizing accumulative delay and hop count while maximizing cluster size. Simulation result demonstrates that EC-MOPSO outperforms existing methods, achieving improvements in distance travelled, latency, packet delivery ratio, and convergence time, making it

well-suited for dynamic IoV scenarios. In study done by Husnain et al. [22], a cluster-based routing algorithm inspired by biology is presented in this study for use in Vehicular Ad Hoc Networks (VANETs) in Intelligent Transportation System (ITS) applications. Keeping stable communication routes is difficult because vehicle networks are dynamic and frequently sparse, especially on highways. The Whale Optimization Algorithm (p-WOA), a probability-based approach, is the suggested remedy for enhancing cluster formation in VANETs. Husnain, G et al. [23] The best cluster heads for IoV networks are chosen using a clustering technique based on Harris Hawks Optimization, which takes into account variables like range, node count, grid size, orientation, and velocity. Experimental results demonstrate a 90.6% cluster stability and an 80 percent improvement in cluster optimization, outperforming current techniques in latency, bandwidth usage, and PDR. In the study Abbas et al. [24], HBO-ECHS is a hybrid bio-inspired algorithm that combines BOA and ACO for effective cluster head selection in VANETs. It is proposed in this study. By lowering routing overhead and energy consumption, it improves communication. Comparing NS2 and SUMO simulations to AJ-MOFA and RJ-EDCV, the former demonstrate better PDR, delay, energy efficiency, and routing performance. Salim, A et al. [25] considered the following parameter, like network grid size, average cluster lifespan, and node transmission range for optimal clustering using the sparrow search algorithm (IoVSSA).

III. Proposed Solution

From the survey, most of the clustering algorithms for vehicular network clustered the network based on stability, density and energy. Ensuring the QoS was targeted by the routing algorithm. When assumption made on clustering is broken, the QoS for message dissemination is affected. Also, the overhead for cluster maintenance increases when assumptions are disturbed. The major gaps in existing clustering algorithms are:

- Overhead for cluster maintenance is higher when clustering assumptions are broken.
- Clustering is not based on multi QoS optimization due to which providing higher QoS with routing alone becomes difficult.

This work proposes a heuristics-based clustering algorithm to solve this problem. The algorithm has three stages: i) Construction of QoS guideline path in the network, ii) Clustering with assistance of QoS guideline path and iii) Cluster maintenance.

The network model for proposed solution is given in Fig. 1 shows how the IEEE 802.11p wireless standard allows each vehicle to interact with the RSU and with each other. RSU monitors the vehicular movements in its coverage.

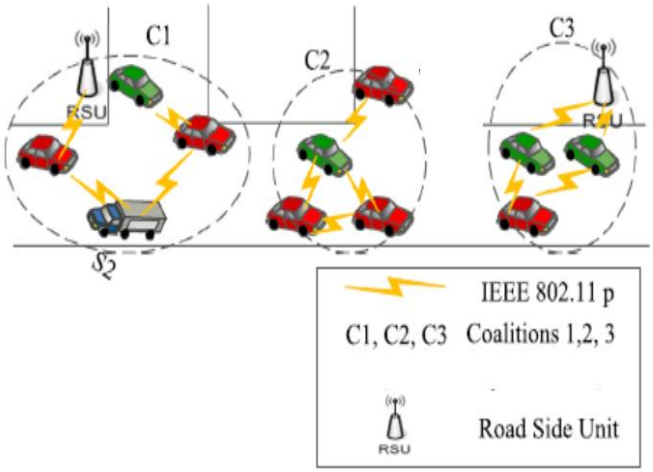


Fig. 1. Network layout.

The vehicular density over period of time (VD_i) in each grid is calculated as:

$$VD_i = \begin{cases} \alpha.T + (1-\alpha)VD_{i-1} \; if \; T \neq 0 \\ (1-\alpha)VD_{i-1}, otherwise \end{cases} \quad (1)$$

where, T is the current density. The grids whose VD_i is above a certain threshold are marked as salient points. A graph is created with all salient points and RSU locations as in Fig. 2. A virtual guideline path is created with minimal hop connectivity to the RSU and salient points. The difficulty in finding the virtual guideline route is solved using the Steiner minimal tree (SMT) problem. It has the feature of introducing intermediate junctions other than RSU to provide more amount of freedom in path construction provided by the Steiner tree.

The Steiner minimal tree is constructed taking P RSU's. The Steiner algorithm must result minimal spanning tree built on S Steiner vertices such that overall length of tree is minimized. Since finding the S points is a NP (non-polynomial) complexity problem, heuristics algorithms are explored to solve it. This work uses Graph iterated 1-steiner (GIIS) [26] to find the Steiner minimal tree.

To apply GIIS, a graph G is created. Each RSU is vertex in the graph G. The delay for packet transmit between RSU is set as edge weight. Delay is intended in terms of probability of distribution over past values as:

$$\text{Delay} = \begin{cases} \sum_{i=0}^{\infty} f_i(a).f_i(b) \;, x = 0 \\ \sum_{i=0}^{\infty} f_i(a).f_{2x+i}(b) + \sum_{i=0}^{\infty} f_i(b).f_{2x+i}(a), x > 0 \end{cases} \quad (2)$$

In the above equation, $f(a)$ is the probability_mass_function in forward_path and $f(b)$ is the probability mass function in backward_path GIIS algorithm finds S optimal Steiner_points, so that cost of Steiner tree found using below equation is minimal for N RSU's.

$$\Delta KMB(N,S) = cost\big(KMB(N)\big) - cost(KMB(N \cup S)) \quad (3)$$

An example of the Steiner tree constructed for RSU is given in Fig. 2. The Steiner points are located at minimal distance to RSU. Steiner points are connected in a minimal tree.

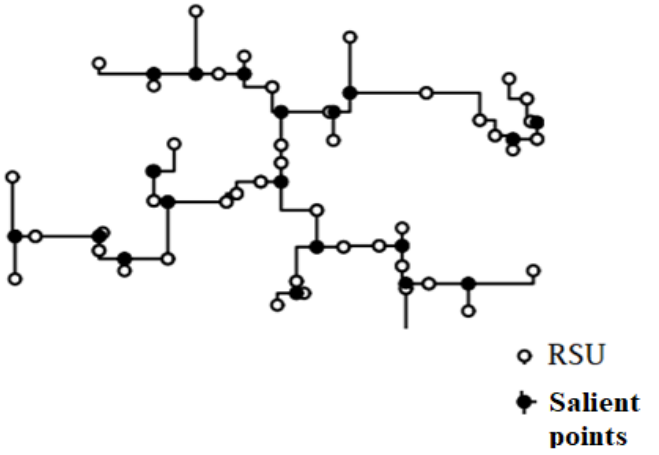


Fig. 2. Virtual guideline path.

The route connecting RSU to Steiner_points and among Steiner_points is the virtual guideline path. The tactical points must be grouped into K groups, where K is the number of data collection points. Grouping of tactical points must be done so that to minimize the delay and maximize the coverage of RSU in the group. Maximizing the coverage and minimizing the delay are two conflicting requirements, and a balance must be achieved between these two conflicting requirements. With increase in the number of strategic points, finding the optimal group is a combinational explosion problem. To solve this problem, nature-inspired heuristics is used in this work. Among multiple heuristics algorithms, particle-swarm-optimization (PSO) is used due to its simplicity.

PSO algorithm is centered on the social actions of swarms. It is a simple, flexible and versatile used to solve many optimization problems. It is centered on the observation that finest solution found by particle in swarm becomes the best solution for entire particles in the swarm. Each particle move randomly in its own velocity and evaluate a fitness function at each position. The position, where maximum value for fitness function is found among all the particles is the global best solution. All other particle update their local position based on the global best solution. This process of updating position and converging to global best solution is repeated in iterations till a maximum fitness function value is obtained or maximum number of repetitions are exceeded.

Each particle (X) updates its position (X_i) at current time based on past location and new velocity $(V_i(t+1))$ as:

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (4)$$

The new velocity is calculated based on past velocity $(V_i(t))$, distance to globally best solution (p_{besti}) and distance to locally best solution (p_{besti}) as:

$$V_i(t+1) = wV_i(t) + c_1 r_1\big(p_{besti}(t) - X_i(t)\big) + c_2 r_2\big(g_{besti}(t) - X_i(t)\big) \quad (5)$$

The important thing to distinguish the best solution is controlled by acceleration coefficient c_1 , c_2 and random number r_1, r_2 . The importance to velocity is controlled by intertia weight w.

In this work, each particle represents a group, and it is S bit binary string where each point represents a Steiner point. The value of this position is either 0 or 1, with 0 representing the corresponding Steiner_point not in group and 1 representing the corresponding Steiner_point in group.

The fitness function for particle is calculated as:

$$F = \frac{\sum_{i=1}^{K} |G_i| + 1/SP(G_i)}{K} \quad (6)$$

where, K is the number_of_Steiner_points and $|G_i|$ is the number_of_Steiner_points in the group. SP is the length of minimal route connecting the Steiner_points in the group using Dijkstra's shortest path algorithm. Each particle selects a random binary string as an initial solution. Each particle calculates the fitness function. The particle with the highest value of fitness function is the globally best solution. Based on the globally best solution, the position is the binary string is updated at a certain rate (or velocity). The iteration starts with new solutions and is repeated till there is no change in the binary string or the maximum iterations are reached. The final binary string at the end of the iteration is the solution. The procedure for choosing optimal groups is given below.

Once the Steiner_points are grouped to K groups, the centroid point of each group is selected as the cluster head point.

RSU entrenches the position of the Cluster head points near its neighborhood and sends in the beacon packets. Vehicles check these beacons and if it discovers it position is very near to the cluster head points, it sends a Hello msg broadcast with cluster indicator. Any vehicle observing the Hello msg broadcast joins to this cluster.

Once the vehicle assumed role as cluster head moves, any other vehicle near to the cluster head point will select itself as cluster head and broadcast the Hello msg to request nearby nodes to join its cluster. By this way, the cluster maintenance overhead is lower in our approach. The overall clustering process is given in Fig. 3.

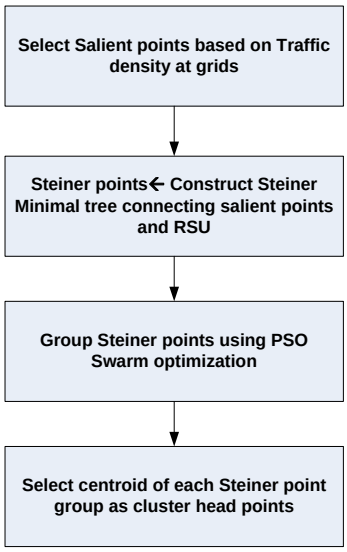


Fig. 3. Clustering process.

IV. RESULTS

The NS-2 simulator is used to evaluate the effectiveness of the proposed solution. SUMO and NS-2 extension code implements the proposed solutions on vehicles traces generated. Simulation runs on the specified configurations, as shown in Table I.

TABLE I. NS-2 SIMULATION CONFIGURATION

Parameters	Values
Propagation_model	Two-way ground
Mobility_model	Krauss
Transmission_range	300m
Transmission_power	20 m W
Simulation_area	4000 m * 4000m
Simulation_time	500s
Message_size	170 bytes

The simulation was conducted against the network layout shown in Fig. 4.

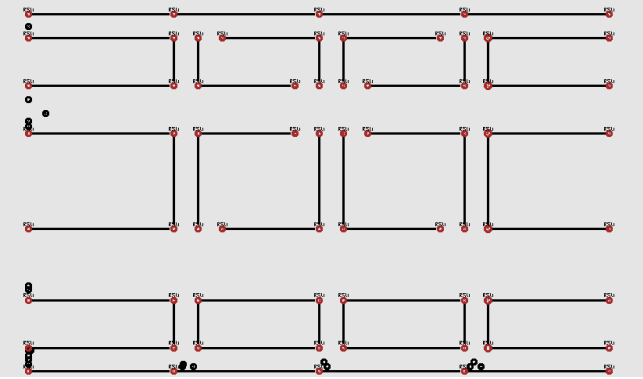


Fig. 4. Network layout.

The proposed solution performance was measured in terms of cluster duration, number-of-clusters, cluster maintenance overhead, packet delivery ratio, delay, and routing overhead. The comparison is done against the proposed solutions and connectivity based clustering proposed by Gasmi et al. [16], connectivity based clustering algorithm proposed by Khan et al. [17] and an unstable connectivity-aware clustering algorithm proposed by Gasmi et al. [18].

The number of clusters was measured for various vehicular speeds and the result is given in Table II.

TABLE II. COMPARISON BASED ON NUMBER OF CLUSTERS

Number of Vehicle	Comparison based on the number of clusters			
	Proposed	*Gasmi.et .al [16]*	*Khan.et.al [17]*	*Gasmi .et.al [18]*
20	7	8	8	9
40	10	11	12	13
60	11	14	13	15
80	12	17	16	18
100	13	19	19	22
Average	**10.6**	**13.8**	**13.6**	**15.4**

In the proposed solutions, the average number-of-clusters is 30% less compared to Gasmi et al. [16], it is 28% less compared to Khan et al. [17] and 45% less compared with Gashmi et al. [18]. The number of clusters is almost stable even at high speeds, as the clusters are created near the stable guideline path instead of vehicular speed and movement mobility metrics used in existing works, as shown in Fig. 5.

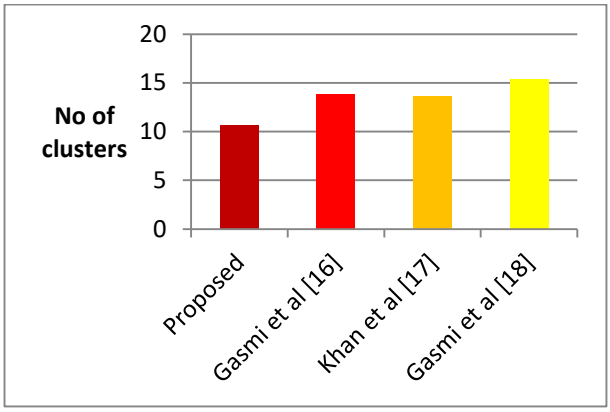


Fig. 5. Average number of clusters.

The average duration of cluster is measured with various vehicular speeds, and Table III shows the result.

TABLE III. COMPARISON OF CLUSTER DURATION

Number of Vehicle	Cluster duration (seconds)			
	Proposed	*Gasmi.et.al, [16]*	*Khan.et.al, [17]*	*Gasmi.et.al, [18]*
20	91	86	87	88
40	86	81	82	82
60	83	77	78	77
80	80	72	74	73
100	79	66	69	68
Average	83.8	76.4	78	77.6

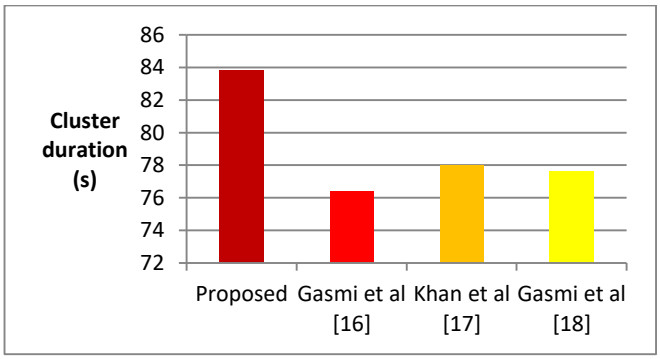


Fig. 6. Average cluster duration.

In the proposed solution the average clustering duration is at least 8% less compared with Gasmi et al. [16], 6% less compared with Khan et al. [17] and 7% less compared with Gasmi et al. [18] as shown in Fig. 6. Clustering duration is higher as the cluster is created based on vehicular movement around the guideline path periphery and optimal cluster

selection using PSO based on stability in the proposed solution. The existing works used link stability and node density mobility metrics which varies during higher speeds and thus their clustering duration is lower. They compensate for it by increasing the number of clusters.

The packet delivery ratio was measured for various vehicular speeds and Table IV shows the above results.

TABLE IV. COMPARISON OF PACKET DELIVERY RATIO

Number of Vehicle	Comparison Of Packet Delivery Ratio			
	Proposed	*Gasmi et al. [16]*	*Khan et al. [17]*	*Gasmi et al. [18]*
20	93	90	91	90
40	91	89	89	88
60	90	86	87	87
80	89	85	84	85
100	88	83	82	83
Average	90.2	86.6	86.6	86.6

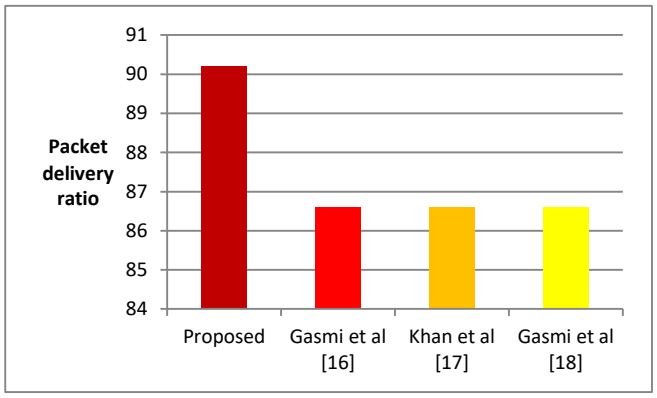


Fig. 7. The proposed and existing packet delivery ratio.

In the proposed solution, the packet delivery ratio is 3.6% high compared with existing works, as shown in the Fig. 7. Due to creation of more stable clusters and ability to search the hop cluster for data forwarding along the virtual guideline path increases the packet delivery ratio. The presence of virtual guideline path was the differentiator in the proposed solution compared to existing works.

By varying the speed, the average packet delay is measured, and the result is shown in Table V.

TABLE V. COMPARISON OF DELAY

Number of Vehicle	Delay (milli seconds)			
	Proposed	*Gasmi et al. [16]*	*Khan et al. [17]*	*Gasmi et al. [18]*
20	17	18	19	19
40	21	24	24	25
60	24	28	27	28
80	25	33	32	34
100	25	34	34	37
Average	22.4	27.4	27.2	28.6

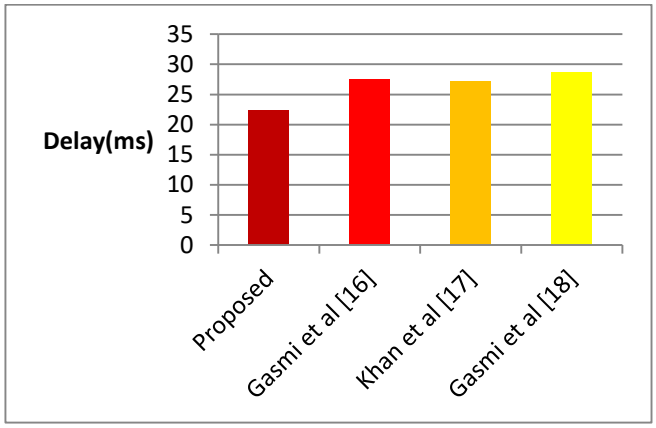


Fig. 8. Delay.

The average packet delay in the suggested solution is 22% low compared with Gasmi et al. [16], 21% lower compared with Khan et al. [17] and 27% lower compared with Gasmi et al.[18] as shown in the Fig. 8. The packet delay has been reduced in proposed solution due to clusters around the periphery of the virtual guideline path which is constructed with delay as one of the optimization criteria. The cluster maintenance overhead in terms of volume of packets exchanged is measured for various speeds and result shown in the Table VI.

TABLE VI. COMPARISON OF CLUSTER MAINTENANCE OVERHEAD

Number of Vehicle	Cluster maintenance Overhead (KB)			
	Proposed	*Gasmi et al [16]*	*Khan et al [17]*	*Gasmi et al [18]*
20	31	33	32	33
40	37	39	37	38
60	43	46	45	44
80	45	49	51	49
100	48	54	54	55
Average	40.8	44.2	43.8	43.8

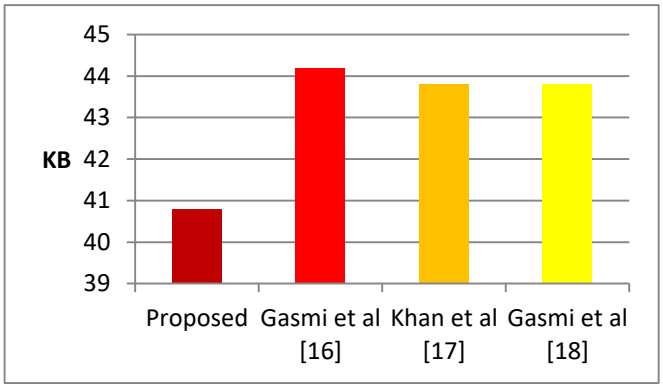


Fig. 9. Cluster maintenance overhead.

In proposed solution, the clustering maintenance overhead is 8% low compared with Gasmi et al. [16], 7% low compared with Khan et al. [17] and 7% low compared with Gasmi et al. [18], as shown in Fig. 9. The cluster maintenance overhead is low due to reduction of search space of new cluster in

proposed solution. The clusters are only selected along the periphery of virtual guideline path instead of many locations in the networks in proposed solution. The routing overhead in terms of retransmissions or route construction effort is measured for various speeds and result shown in Table VII.

TABLE VII. COMPARISON OF ROUTING OVERHEAD

Number of Vehicle	Routing Overhead (KB)			
	Proposed	*Gasmi.et.al [16]*	*Khan.et.al [17]*	*Gasmi.et.al [18]*
20	11	12	13	14
40	13	15	16	17
60	15	19	21	22
80	17	24	25	27
100	19	26	27	29
Average	15	19.2	20.4	21.8

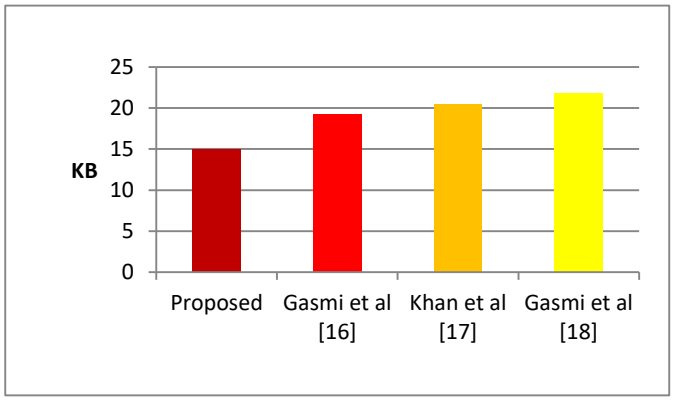


Fig. 10. Routing overhead.

In proposed solution, the routing overhead is 28% low compared with Gasmi et al. [16], 36% low compared with Khan et al. [17] and 45% low compared to Gasmi et al. [18], as shown in Fig. 10. Due to the use of RSU assistance in routing and routing along virtual guideline path, the routing overhead is reduced.

V. CONCLUSION

A heuristics-based clustering algorithm is proposed in this work for IOV network. A virtual guideline path is first established based on multi QoS and Clusters are established around the periphery of the virtual guideline path. The proposed solution can establish stable clusters with reduced cluster maintenance overhead. The average cluster duration is at least 6% higher and cluster maintenance overhead is 7% lower compared to exiting works. In terms of QoS, the proposed solution has 3.6% high packet delivery ratio and 21% low, 28% low routing overhead compared to existing works.

REFERENCES

[1] Lakas, A., Fekair, M. E. A., Korichi, A., & Lagraa, N. (2019). A Multiconstrained QoS-Compliant Routing Scheme for Highway-Based Vehicular Networks. Wireless Communications and Mobile Computing, 2019(1), 4521859.

[2] Wahab, O. A., Otrok, H., & Mourad, A. (2013). VANET QoS-OLSR: QoS-based clustering protocol for vehicular ad hoc networks. Computer Communications, 36(13), 1422-1435.

[3] Wang, S. S., & Lin, Y. S. (2013). PassCAR: A passive clustering aided routing protocol for vehicular ad hoc networks. Computer communications, 36(2), 170-179.

[4] Khan, M. F., Aadil, F., Maqsood, M., Bukhari, S. H. R., Hussain, M., & Nam, Y. (2018). Moth flame clustering algorithm for internet of vehicle (MFCA-IoV). IEEE access, 7, 11613-11629.

[5] Ahizoune, A., & Hafid, A. (2012, October). A new stability based clustering algorithm (SBCA) for VANETs. In 37th Annual IEEE conference on local computer networks-workshops (pp. 843-847). IEEE.

[6] Shankar, T., Shanmugavel, S., & Rajesh, A. (2016). Hybrid HSA and PSO algorithm for energy efficient cluster head selection in wireless sensor networks. Swarm and Evolutionary Computation, 30, 1-10.

[7] Ali, H., Shahzad, W., & Khan, F. A. (2012). Energy-efficient clustering in mobile ad-hoc networks using multi-objective particle swarm optimization. Applied Soft Computing, 12(7), 1913-1928.

[8] Fahad, M., Aadil, F., Khan, S., Shah, P. A., Muhammad, K., Lloret, J., ... & Mehmood, I. (2018). Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks. Computers & Electrical Engineering, 70, 853-870.

[9] Dalbah, L. M., Al-Betar, M. A., & Awadallah, M. A. (2024). Jaya clustering-based algorithm for multiobjective IoV network routing optimization. Soft Computing, 28(6), 5639-5665..

[10] Ayed, S., Hbaieb, A., & Chaari, L. (2023). Blockchain and trust-based clustering scheme for the IoV. Ad Hoc Networks, 142, 103093.

[11] Oranj, A. M., Alguliev, R. M., Yusifov, F., & Jamali, S. (2016). Routing algorithm for vehicular ad hoc network based on dynamic ant colony optimization. Int. J. Electron. Elect. Eng, 4(1), 79-83.

[12] Shaikh, F. I., & Hingoliwala, H. A. (2017, December). Path planning based QoS routing in VANET. In 2017 international conference on big data, IoT and data science (BID) (pp. 37-43). IEEE.

[13] Aadil, F., Raza, A., Khan, M. F., Maqsood, M., Mehmood, I., & Rho, S. (2018). Energy aware cluster-based routing in flying ad-hoc networks. Sensors, 18(5), 1413.

[14] Hadded, Mohamed, Paul Muhlethaler, Anis Laouiti, and Leila Azzouz Saidane. "A novel angle-based clustering algorithm for vehicular ad hoc networks." In Vehicular Ad- Hoc Networks for Smart Cities, pp. 27--38. Springer, Singapore, 2017.

[15] Bersali, M., Rachedi, A., & Bouarfa, H. (2020, November). A new collaborative clustering approach for the Internet of vehicles (CCA-IoV). In 2020 Second International Conference on Embedded & Distributed Systems (EDiS) (pp. 58-63). IEEE.

[16] Gasmi, R., & Harous, S. (2022). Robust connectivity-based internet of vehicles clustering algorithm. Wireless Personal Communications, 125(4), 3153-3185.

[17] Khan, Z., Koubaa, A., Fang, S., Lee, M. Y., & Muhammad, K. (2021). A connectivity-based clustering scheme for intelligent vehicles. Applied Sciences, 11(5), 2413.

[18] Gasmi, Rim & Aliouat, Makhlouf & Aliouat, Zibouda & seba, Hamida. (2022). Unstable connectivity aware protocol for dependable messages delivery for future internet of vehicles. 10.21203/rs.3.rs-1535145/v1

[19] Yerrathi, S., & Pakala, V. (2024). Enhancing network stability in VANETs using nature inspired algorithm for intelligent transportation system. Plos one, 19(1), e0296331.

[20] Husnain, G., & Anwar, S. (2022). An intelligent probabilistic whale optimization algorithm (i-WOA) for clustering in vehicular ad hoc networks. International Journal of Wireless Information Networks, 29(2), 143-156.

[21] Jamalzadeh, M., Maadani, M., & Mahdavi, M. (2022). EC-MOPSO: An edge computing-assisted hybrid cluster and MOPSO-based routing protocol for the Internet of Vehicles. Annals of Telecommunications, 77(7), 491-503.

[22] Husnain, G., Anwar, S., Sikander, G., Ali, A., & Lim, S. (2023). A bio-inspired cluster optimization schema for efficient routing in vehicular ad hoc networks (VANETs). Energies, 16(3), 1456.

[23] Husnain, G., Anwar, S., & Shahzad, F. (2023). An enhanced AI-enabled routing optimization algorithm for internet of vehicles (IoV). Wireless Personal Communications, 130(4), 2623-2643.

[24] Abbas, F. H., Al-Rawi, O. Y. M., Ali, R. R., Mahmood, S. N., Alkhayyat, A., & Hassan, M. H. (2025). A Hybrid Bio-Inspired Optimization Based Enhanced Cluster Head Selection to Improve Communication in Vehicular Ad-Hoc Networks. In Tech Fusion in Business and Society (pp. 475-485). Springer, Cham.

[25] Salim, A., Khedr, A. M., & Osamy, W. (2023). IoVSSA: efficient mobility-aware clustering algorithm in internet of vehicles using sparrow search algorithm. IEEE Sensors Journal, 23(4), 4239-4255.

[26] Robins, G., & Zelikovsky, A. (2008). Minimum steiner tree construction. In Handbook of algorithms for physical design automation (pp. 487-508). Auerbach Publications.

SOM-Based Leader Selection Strategies for Cooperative Spectrum Sensing in Multi-Band Multi-User 6G CR IoT

Mayank Kothari, Suresh Kurumbanshi
Department of Electronics & Telecommunication Engineering, SVKM's NMIMS (Deemed to be University), Shirpur, India

***Abstract*—In 6G Cognitive Radio Internet of Things (CR-IoT) networks, multi-band spectrum sensing cooperatively provides access to extensive spectrum resources. The suggested learning-based multi-band multi-user cooperative spectrum sensing (M2CSS) scheme addresses intelligent spectrum access challenges. A cooperative strategy is introduced into a dueling deep Q network to facilitate multi-user reinforcement learning. This study selects the most suitable IoT secondary users (SU) to sense channels using the proposed learning-based M2CSS scheme. With the restriction that each IoT SU can serve as a front-runner for a single network and that there will only be one leader for individual frequency, the proposed work expresses an optimization difficulty in choosing leaders through k-means and SOM, who can efficiently interact with other SUs. Next, choose matching cooperative SUs for each frequency and express additional optimization problems. Following this phase, a subset of cooperative secondary users (SUs) senses frequencies and employs accurate knowledge to determine the channels' availability in a distributed manner. The simulation findings demonstrate significant improvements in detection performance, preventing the misuse of specific devices, providing reliable sensing data over extensive IoT connections, and achieving energy efficiency—all essential for IoT implementations. These advantages make the proposed M2CSS system suitable for the massive machine-type communications anticipated in 6G IoT scenarios.**

Keywords*—Cooperative spectrum sensing; reinforcement learning; k-means leader selection; self-organizing map

I. Introduction

The concept of cognitive radio (CR) involves a communication device that can recognize and adapt to the spectral characteristics of its surroundings. By utilizing the spectral gaps (or holes) that approved consumers, also referred to as primary users (PUs), do not use, CR skill makes it possible for no licensed or secondary users (SUs) to identify these accessible portions of the spectrum [1]. Four distinct stages or functions are involved in the operation of spectrum reuse in a cognitive radio network (CRN): spectrum sharing, spectrum mobility, spectrum sensing (SS), and spectrum decision. A crucial step in identifying particular or additional PUs is SS, which indicates whether the sensed spectrum is engaged or unoccupied [2]. This mission is typically performed in particular groups, but the current multiband SS paradigm uses multiple bands that aren't always attached [3]. Demand for spectrum is rising because of the appearance of the Internet of Things (IoT), sixth generation (6G), past systems, and big data [4]. Spectrum scarcity and underutilization issues may be resolved with the help of CR skills [5]. Secondary users (SUs) in CRNs resourcefully contact the spectrum without intrusiveness through the broadcast of legitimate or PUs [6]. Each SU might have to sense every channel in single-band cooperative spectrum sensing (CSS). On the other hand, because of constraints on hardware and power consumption, each SU in multi-band CSS is given access to a subset of channels. By enabling access to a large spectrum with less difficulty and computational expense, this can significantly improve system performance. Additionally, low hand-off frequency, high throughput, and decreased sensing energy consumption are provided by multi-band CSS [7]. This essay focuses on multi-band CSS to enhance detection performance in CRNs and optimize SU scheduling for sensing a subset of channels.

Most previous methods only consider one user, which has fewer sensing capabilities than a real multi-user or multi-channel environment. As a result, SUs use a fully distributed model for multi-user spectrum access [8], which entails creating a deep multi-user reinforcement learning technique [9]. Multi-band CSS can be divided into centralized and distributed types [5]. SUs send multi-band, centralized CSS information, including multiple channels' worth of locally observed data, to an FC (fusion center) [10]. The fusion center used this information before deciding which channels were available. CRNs are bounded by limitations of the centralized method, including high latency and a single point of failure. [11]. SUs share local observations of several channels with neighbors in a multi-band distributed CSS to facilitate cooperative decision-making. Wireless sensor networks, mesh networks, Mobile ad-hoc networks, Vehicular ad-hoc networks, Flying Ad-hoc networks, etc., remain a few examples of disseminated CRNs [12]. Distributed CRNs are especially well-suited for newer Internet of Things applications like public safety networks and habitat monitoring. In disseminated CRNs, SUs typically have constrained computational and battery resources. Therefore, when developing multi-band CSS schemes for distributed CRNs, trade-offs must be made between energy resources, computational capacity, and performance limitations.

A difficulty with multi-band CSS is the development of SUs to detect radiofrequency (RF) in specific networks in distributed CRNs. Depending on the channel state data, SUs may be able to sense channels in different ways. Consequently, if multiple SUs have similar locally sensed information,

selecting cooperative SUs efficiently to sense a subset of frequencies can lower sensing overhead and enhance organization presentation [13]. The CRN's size grows when new SUs join the existing SUs. New SUs must be given multiple channels to divide the sensing load equally between the original and new SUs in the network. To do this, research on the topic is essential. The actions and residual energy of the previous frame determine each frame's initial energy level. In addition to being sporadic and unpredictable, the electrical energy obtained from the harvested RF may not always have enough power to optimize throughput. Therefore, the CRN must be energy efficient to balance the amount of energy harvested with the energy consumed during sensing and transmission activities. There is a sensing-throughput trade-off [14] in typical cognitive radio networks (also known as unconstrained energy counterparts), which depends on sensing accuracy and sensing time. However, the results of the sensing process (i.e., the sensing duration and accuracy) are energy-constrained in the context of the energy harvesting CRN, making the energy harvesting-based CRN (EH-CRN) a scenario with increasing complexities. The SS is the primary mechanism that enables spectrum findings, which is not possible without all other CR functions. It is the driving force behind concentrating solely on this component of CR. In [15–16], it has developed into a method that enhances spectral efficiency by providing access to SH, or vacant frequency bands, during intermittent times.

The primary objective of this research is to develop and analyze an efficient SS framework for CRN in 6G-enabled Internet of Things (6G CR-IoT) environments that addresses critical spectrum scarcity challenges through opportunistic spectrum access while ensuring minimal interference to licensed users. The SOM-based cooperative SS technique carries out this function by having the CRN node sample the targeted spectrum in the range of 0.1 to 1 THz and determine whether or not an incumbent signal is present. SS seeks to reduce any interference that the incumbent might encounter from CRN nodes, in addition to identifying resources that are easily accessible. As a result, the SS scheme that the CR nodes use for identification significantly impacts the CRN and its performance, regardless of the incumbent network's licensing status. By allowing IoT devices to strategically access and share available spectrum resources while utilizing the improved capabilities of 6G networks, 6G CR-IoT seeks to address significant issues with spectrum scarcity.

A. Study Organization

The following sections are organized as follows: Section II provides the literature survey with existing gaps. Section III emphasizes the detailed proposed model with suitable diagrams and mathematical equations. Section IV gives the experimental results with four sub-sections under it, and Section V concludes this research.

II. Literature Survey

Recently, studies have been carried out to improve multi-band CSS performance in CRNs [2, 5, 8–26]. The type of CRN—decentralized or centralized—and the particular techniques used by different authors determine how effective these systems are. Most multi-band CSS schemes in the existing literature were created with centralized CRNs in mind. Through simultaneous optimization of the detection inception, a multiband multi-user scheme is proposed in [17] to maximize the throughput of a centralized CRN with imperfect sensing information. A centralized CRN-based multi-band CSS scheme proposed in [18] overcomes the issues of spectrum resources and limited energy. The plan uses wireless occurrence dynamism gathering and maximizes throughput by optimizing sensing time. For centralized CRNs, an ideal multi-band CSS scheme with chance disruptions is put forth in [19]. The method reduces interference to PUs while optimizing CRN throughput. For centralized CRN, a multi-band sensing time-adaptive structure is also suggested in [20] to increase throughput and lessen interference to PUs. An artificial immune procedure built on the clonal selection theory is used to overcome this obstacle. Responsive hand-off in expression is reduced in centralized multi-band CRNs according to [21]. Sample-based sensing is an effective method for this purpose. In [22], the authors details that to maximize detection performance, the smallest number of SUs is chosen to sense frequencies based on the network situation. Centralized CRNs in [23] suggested a sub-Nyquist multiband CSS protocol to lower the overhead of computation and transmission using resident sub-Nyquist samples. For centralized CRNs with malevolent users, a multi-band CSS scheme is suggested in [24]. To limit the throughput of CRNs and counteract the threat of malevolent attackers, the authors devised an optimization problem. To improve system performance, a multiband CSS scheme based on deep learning is suggested in [25] for centralized CRN. This scheme allows SUs to sense networks independently and forecast the effects of indolent forecasting on the frequency of state-owned during the opening period. As a result, sensing requires less energy. Two secreted coatings, each containing 40 neurons, are used in the implementations for this purpose. The algorithms broadly impact the performance of centralized CRNs, and it continues to be exceptional as the CRN's dimension grows. Due to their computational intensity, these processes can result in growth that is above average, as evidenced by higher SUs and dispersed CRNs.

The investigations in [26] specifically pinpoint the ideal sensing policy. They are an expansion of [26] that establishes the ideal detection threshold and sensing policy for enhancing the anticipated throughput while taking collision limits and energy causality into account. To increase average throughput while preserving energy, the relationship between the ideal sensing length and the matching sensing threshold was looked at. Based on the energy arrival rate and the correlation of the primary traffic, the theoretical maximum throughput of the SU in an energy harvesting-based system is considered, considering energy causality and collision constraints. The trade-off between SS and SU throughput is examined in [27] for a typical energy-unconstrained CRN. This maximizes both the average throughput and the harvested energy by optimizing the decision-making time. Based on it, the work [28] aims at proposing an optimum sensing time for the overlay energy harvesting (EH) CRN, in order to improve SU throughput and the harvested RF energy. Additionally, [29] performs the optimal SS and SU transmissions so as to maximize the residual energy. A common obstacle detected in

these works is the unpredictable nature of energy derived from ambient RF sources and the inability of it to always provide the maximal throughput in EH-CRNs. The SU split the channel into two sub-channel sets. One is used to detect the PU, while the other is used to record the RF of the PU signal. To ensure throughput, gathered energy is supplied in the transmission slot to offset sensor energy loss. Nevertheless, there is no mention of the specific energy source used for data transmission. In summary, although a significant amount of research has been done on non-arbitrary ways to assign channels for sensing by involving SUs in disseminated CRNs, these schemes are not remarkably scalable because of their limited computational capabilities. All cognitive radios participate in CSS by measuring the licensed spectrum and making autonomous decisions. Typically, the cognitive radio network is made up of multiple SUs and PUs, with the goal of the SUs being able to share multi-channel spectrum properties with the PUs without interfering with the PUs. Every SU optimizes its own SS performance while exchanging sensing data with the others. SU cannot be sure of the channel environment and is unable to choose until after trying a few times to sense a new channel. As a result, the partial observability issue can be resolved by modeling the channel exchange as a Markov chain. Every SU is synchronized with the designated time slot. The sensing data and the information broadcast historically are the two periods that make up the time slot structure. Every SU in the system independently senses the slot at the start of the sensing historical event. Subsequently, the SU merges data with received data and transmits its sensing data to other users. Every SU, then sends out its decision. An acknowledgment signal is sent to confirm the accomplishment or failure of the broadcast before the sensing period ends.

III. Research Methodology

A. Model Creation

The system model is created in this section by assigning base stations to secondary users, calculating the received powers, visualizing the network deployment, and generating random coordinates for both primary and secondary users. Using the existing k-means clustering, this section divides both new and current secondary users into various clusters according to their coordinates. It also selects new secondary users' leaders using proposed self-organizing maps. SS is carried out in this part by comparing secondary users' received power from base stations with a sensing threshold. The sensing result array indicates identified primary users with a value of 1.

B. Proposed SOM-Based Leader Selection

Choosing a leader involves determining which SU on each channel m is the best to listen to. Leaders must also select cooperative SUs for sensing each channel. Let, SNR_mk and SNR_avg represent the SNR standards of the k^{th} SU sensing the m^{th} channel, respectively, the preferred and attached thresholds for the normal SNR value across different frequencies. The channel m leader must establish a minimal association between its indication wrapping and the signal envelopes of the SUs to choose cooperative SUs with distinct nearby sensed data. The leader for each channel is the SU whose SNR is near SNR_avg, as defined in Eq. (1) and Eq. (2):

$$\propto_{mk} = \begin{cases} 1, & if\ K^{th}\ SU\ is\ a\ leader\ for\ channel\ m \\ 0, & otherwise \end{cases} \quad (1)$$

The following is a definition of the optimization problem preparation for a similar collection of channel leaders:

$$\min_{\propto_{mk}} \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_{mk} \times \left| SNR_{mk} - SNR_{avg} \right|$$

$$\text{Subject to: C1} : \sum_{m=1}^{M} \propto_{mk} = c_k, \forall\ k,$$

$$\text{C2:} \sum_{k=1}^{k} \alpha_{mk} = 1, \forall\ m, \quad (2)$$

The binary variable c_k is equivalent to 1 if the k^{th} SU is a leader for a channel and 0 in all other cases. C1 ensures that a single SU can't simultaneously be chosen as a leader for multiple channels. C2, on the other hand, affirms that a single leader is selected for every channel. Combining these two restrictions ensures that every cooperative cluster has a distinct leader and channel.

C. Reinforcement Learning (RL) for Resource Allocation

According to the power received from the primary user, this section allocates available secondary users to primary users. Users who have been allotted are shown. This section applies a reinforcement learning algorithm to maximize resource distribution for every secondary user. The Q-learning algorithm trains a Q-table and mimics the environment to obtain rewards. Based on the rewards, the Q-table is updated, and metrics like energy consumption, false alarm rate, throughput, and probability of detection are calculated. A new line of inquiry for addressing SOM issues is the development of reinforcement learning (RL) techniques. The RL techniques shown in Fig. 1, don't need any prior understanding of the system model and can adjust to the spectrum environment. Because it is model-free, Q-learning is a significant reinforcement learning subfield with extensive applications in various settings. Finding the Q value for every state-action pair is the primary goal of Q-learning, a value iteration technique. Every SU uses the Q-learning approach, and based on recent action-observation experience and past channel occupation history, decides whether to sense the channel. Q-learning is effective when the state and action spaces are small.

A new line of inquiry for addressing SOM issues is the development of reinforcement learning (RL) techniques. The RL techniques don't need any prior understanding of the system model and can adjust to the spectrum environment. Because it is model-free, Q-learning is a significant reinforcement learning subfield with extensive applications in various settings. Finding the Q value for every state-action pair is the primary goal of Q-learning, a value iteration technique. Every SU uses the Q-learning approach, and based on recent action-observation experience and past channel occupation history, decides whether to sense the channel. Q-learning is effective when the state and action spaces are small. DQN, which combines Q-learning and Reinforcement learning, has demonstrated considerable promise in solving the SOM problem. RL as a Markov chain in $2N$ states can accurately represent large-scale models, allowing them to continue performing well in large-scale scenarios. Channel selection in DQN is represented by $<S, A, R, S'>$, where $S \in \{S1, \cdots, SN\}$

indicates the state space, $A \in \{1, \cdots, N\}$ indicates the action space, $R \in \{0, 1\}$ indicates the instant reward, and S' indicates the next state. Every time slot begins with the current state (S), which comprises all the previous observations and decisions.

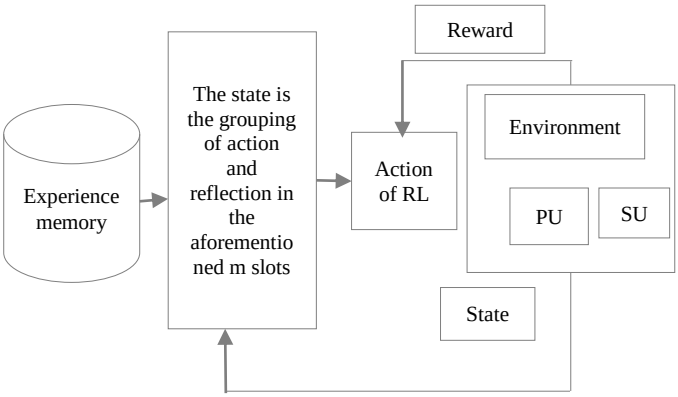


Fig. 1. RL framework for CRN.

The Q network takes the S as an input and predicts what the best action, A, should be as an output, which denotes A sensing of a particular channel at time slot t. The system may reach the next state after taking action A and receive a reward instantly. At last, the time slot record is saved in memory D, and it iterates again to D. And the experience memory is filled according to the random sampling. It samples m records (S, A, R') using experiences sampled from memory D to derive the current target Q estimate. The associated in Eq. (3) is given as:

$$y=R+\gamma \max_{a'} Q(S_{t+1}, a'; \theta_i^-) \quad (3)$$

The reinforcement output, denoted by y, is determined by the new experience. The approximation function created through prior expertise is represented by Q (S_{t+1}, a' ; $\theta - i$), where S_{t+1} is the novel state resulting from action a, assuming the state S_t , and t is the time interval. The weightiness of the Q-network charity to calculate the objective at repetition i is represented by θ_i^-, and the discount rate is denoted by γ. The loss function, which is expressed in Eq. (4) and Eq. (5), is the mean right-angled error between the value of the impartial purpose and the assessment of the definite purpose.

$$L_i(\theta_i) = E[(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta_i^-) - Q(S_t, A_t; \theta_i))^2] \quad (4)$$

$$\nabla_{\theta i} L_i(\theta_i) = E\left[\left(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta_i^-) - Q(S_t, A_t; \theta_i)\right) \nabla_{\theta i} Q(S_t, A_t; \theta_i)\right]) \quad (5)$$

where, the estimate purpose designed by old understanding is represented by Q (S_{t+1}, a'; $\theta - i$) and the estimate purpose designed by new understanding is represented by Q (S_t, A_t; θ_i).

IV. Results and Discussion

The states of the proposed M2CSS are the grouping of the operator's state on respective frequency, and the user's relationship to the PU, SU, data 1, data 2, and CBS, as shown in Fig. 2. Two simulated networks with different user and channel counts are employed to achieve accurate sensing in many user scenarios. The network implementation should take less time to yield positive outcomes.

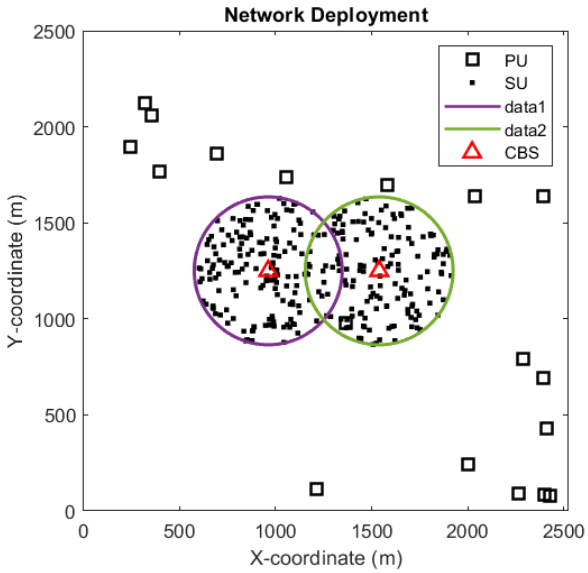


Fig. 2. M2CSS network deployment.

D. K-means for Leader Selection (KMC)

This stage uses similarities to classify recently joined SUs in the system that are interested in common assemblies. Fig. 3 shows the organization process when new SUs attempt to intersect the CRN. The primary justifications for utilizing KMC in this work are its simplicity, speed, and low complexity. To locate the centroids in KMC, the maliciousness of the exercise set under consideration is computed. The centroids then state that the unlabeled set is confidential. The classifier in this study is created utilizing historical data from operational cooperative SUs. To classify the joining SUs, one needs to determine the centroids of each class. The CVM test results of current SU leaders are the lowest, and their SNR values are relatively close to the average SNR for each channel. Non-leaders, on the other hand, have larger SNR value spreads and high CVM test values. This implies that the training set should form two distinct clusters. Fig. 3 displays the filtered data as well as the current data points. Eigen vectors and values of the new and old data points were used to generate distribution contractions for them. The distributions differ from one another. Outliers must therefore be removed from freshly acquired data points. Fig. 3 shows the data points for leaving cooperative SUs. Interestingly, two separate SUs do not align with two marked departing data points.

E. SOM Leader Selection

Consequently, the data points from the operational cooperative SUs were split into two classes (see Fig. 4). The RL algorithm is utilized to do this. Channel leaders make up Class 1, and non-leaders make up Class 2. These class centroids serve as reference points for classifying new data instances. This procedure aids in distinguishing new data from Class 1 samples.

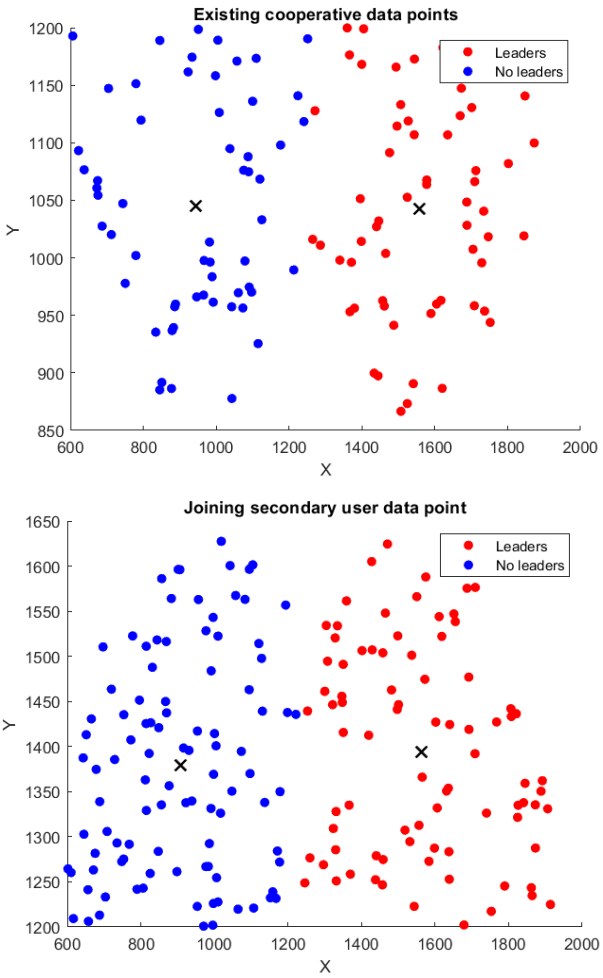
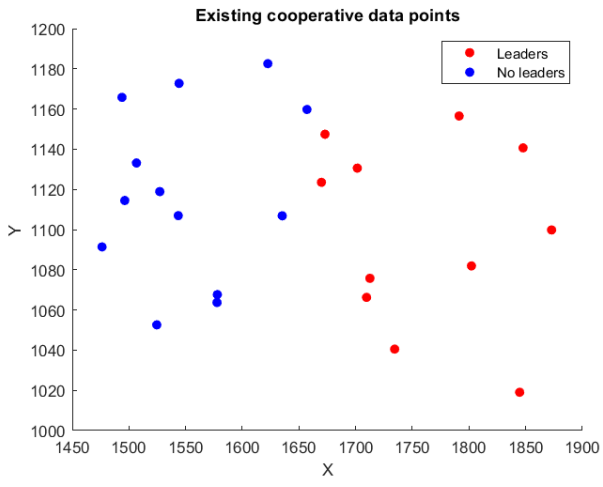
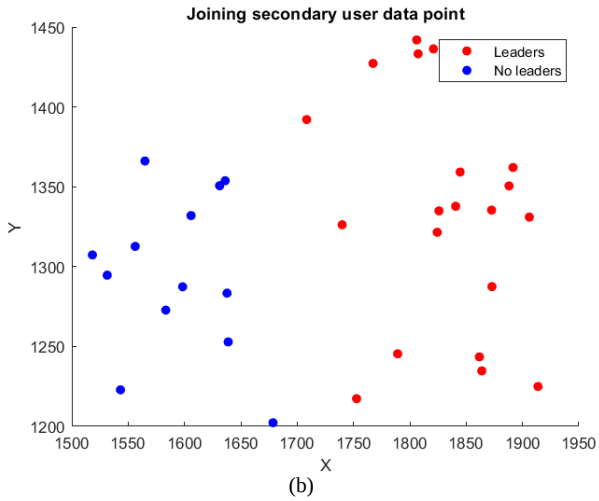


Fig. 3. K-means leader selection.



(a)



(b)

Fig. 4. (a) Leader selection with SOM, (b) No-Leader selection with SOM.

F. SNR Distribution of the Secondary Users

The SNR for secondary users is determined as defined in Eq. (6), where $received_power_SU_bs$ is the established control of the secondary consumer from the base station, and $noise_power$ is the noise power. When the sensing duties were distributed, the number of SUs sensing each frequency increased. This may result in the system using more energy. The incorporation of multiple SUs into the system can cause average SNR levels to surpass acceptable limits, resulting in compromised detection capabilities. Therefore, for a few freshly selected SUs, testing the average channel SNR condition and lowering the minimum required is necessary. If they find another channel, new SUs might not have to bother about sensing this one. This will enable the performance and energy to be recovered (see Fig. 5).

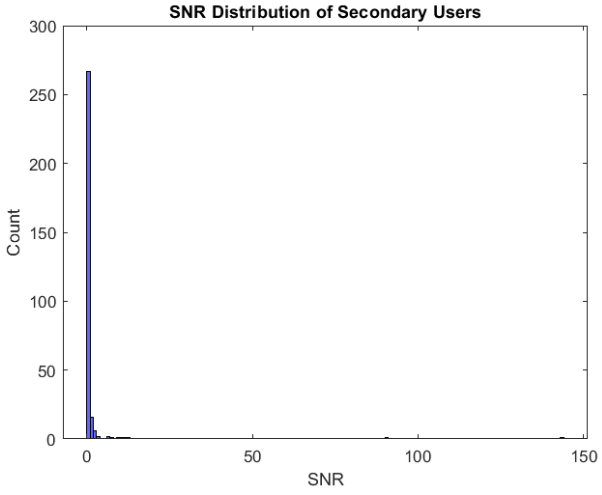


Fig. 5. SNR distribution of SU.

$$SNR_{SU} = \frac{received_power_SU_bs}{noise_power} \quad (6)$$

G. Resource Allocation

For illustration, it is considered that the user is distributed by channel with 10 joining and 10 existing SUs. Fig. 6 displays training data points or the data from the cooperative SUs currently used for sensing channels. The proposed work considers two features, resource allocation and SNR values, and defines a data point as a pair of binary feature standards for SU sensing at the m^{th} frequency represented in Fig. 7. Both characteristics are scaled to depict the data points accurately. The disseminated version procedure for channel 2 of the recommended M2CSS and EM2CSS arrangements and the current multi-band schemes, RSSS, and ASSS, are shown in Fig. 8, Fig. 9, and Fig. 10. To simulate CRN systems, λmax is considered the learning rate, without losing generality. The suggested M2CSS and EM2CSS schemes establish the minimum and maximum K×I values for the number of designated SUs that can intelligently set a channel.

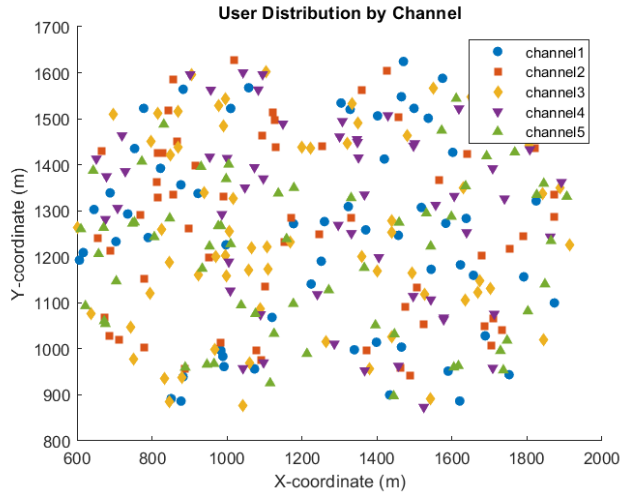


Fig. 6. User distribution by channel.

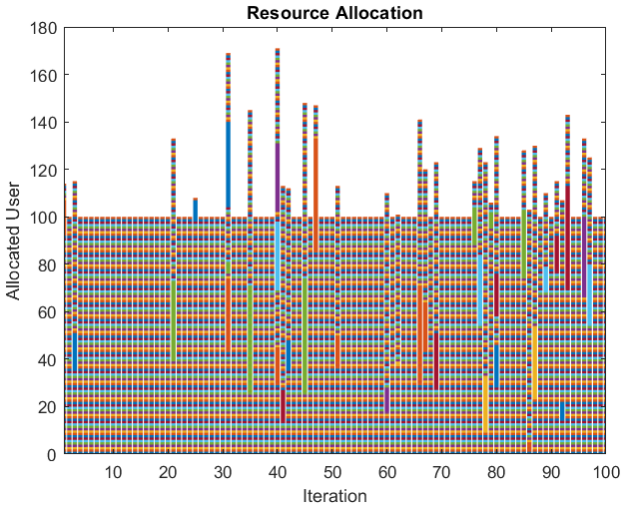


Fig. 7. Resource allocation for SU.

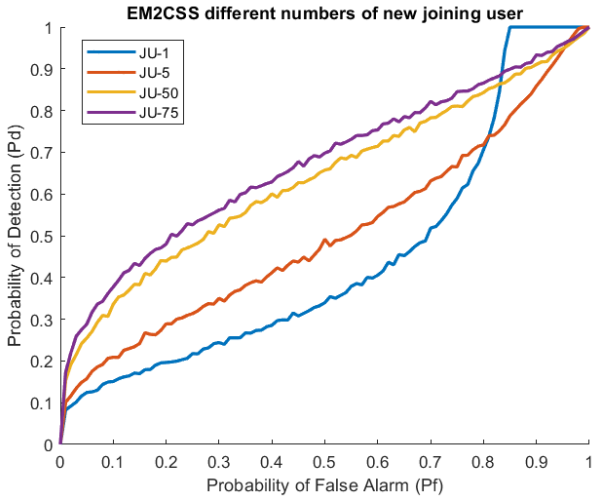


Fig. 8. EM2CSS with different numbers of SU.

Before estimating each cooperative SU separately, ASSS selects them at random. Determining the test statistics and sensing a few examples based on the estimate. If the examination indicators are exceeding a specified boundary, SU is supposed to be able to sense primary user activity on the channel and execute preparation. It eliminates the need to sense that channel in any other way. Because of the approximation of the SUs' procedure in ASSS, there are generally fewer SUs in ASSS. SUs that were predicted to have poor sensing performance are dropped. RSSS selects cooperative SUs at random. Consequently, in M2CSS, EM2CSS, RSSS, and ASSS, respectively.

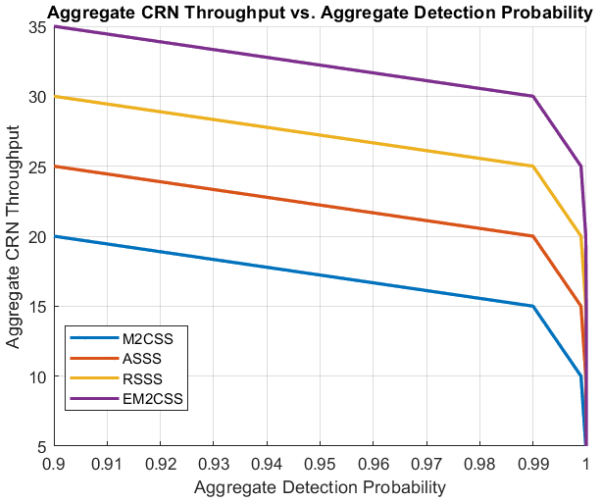


Fig. 9. CRN throughput for SOM leader selection.

A significant benefit of the proposed M2CSS and EM2CSS schemes is that they do not select dismissed cooperative SUs or those through comparable identified material for the m^{th} channel. Nevertheless, when SUs are chosen by popular RSSS, SUs with comparable information may be chosen randomly, as the first selected SU possesses comparable sensing information, as seen in Fig. 9. In contrast, the third selected SU does not have original statistics for the alteration procedure.

This could result in increased computational complexity and expenses related to energy usage. The consensus point is reached in each case after 14 iterations, and the channel occupancy is calculated by comparing the obtained results to the threshold. The throughput results demonstrate that EM2CSS performs significantly better than M2CSS, and for all network sizes considered, both schemes provide excellent CRN throughput consequences.

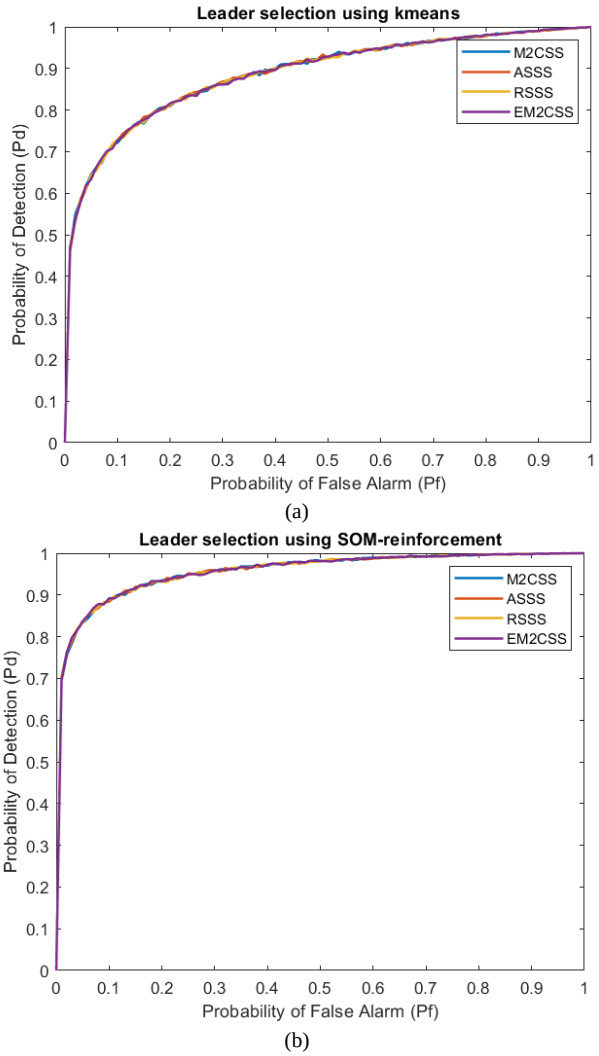


Fig. 10. (a) ROC with different CRN schemes using k-means, (b) ROC with different CRN schemes using SOM.

Proceeding with the additional indicator, the CRN throughput presentation of the current schemes varies. For lower and higher CRNs, the combined PU throughput is compared to the combined untrue distress possibility (also called the PU security level). At a 10% false alarm probability, the most significant combined throughput is attained by the proposed EM2CSS scheme. Additionally, the M2CSS scheme achieves the same false alarm rate. However, it is achieved when the current ASSS scheme is used with the same false alarm possibility. The ROC curve in Fig. 10(a) and Fig. 10(b) shows that leader selection through SOM helps in achieving high probability of detection with less value of probability of false alarm as compare to existing k-mean. Therefore, the proposed EM2CSS structure fulfills the highest requirements of PU throughput in both inferior and superior CRN configurations. The combined sensing energy usage of the larger CRN's current RSSS and ASSS systems, as well as the indicated EM2CSS and M2CSS schemes.

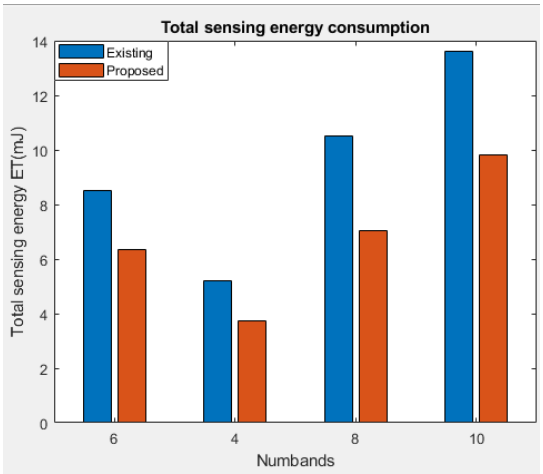


Fig. 11. Total energy consumption.

The scheme's ability to drop SUs sensing specific frequencies with low-slung expected values correlates with this lowest value. On the other hand, the current RSSS scheme uses the most total energy for sensing. The highest energy consumption in RSSS is caused by randomly selecting cooperative SUs with no restriction on the number of SUs sensing each channel. Furthermore, by relaxing cooperative SUs, EM2CSS offers the benefit of distributing the sensing load among already-existing and newly joined SUs. Reduced energy consumption for sensing is made possible by the distribution of the sensing load. It is for this reason that Fig. 11 shows that EM2CSS uses less sensing energy than M2CSS. As a result, the suggested EM2CSS enables respectable levels of sensing energy consumption and outstanding classification presentation in terms of discovery and quantity.

V. Conclusion

The SOM-based multi-band CSS schemes M2CSS and EM2CSS, which have been proposed to disseminate 6G-enabled IoT CRNs, are discussed in this study. The proposed M2CSS architecture addresses two critical optimization challenges: intelligent cluster leader selection across frequency bands and dynamic secondary user accommodation within 6G CR-IoT networks. To choose a leader for every frequency and consistently accommodate SUs, the proposed work developed two optimization problems for the suggested M2CSS scheme. These problems are solved by the reinforcement learning algorithm in the EM2CSS scheme to offer a more effective and simplified method for allocating connection SUs to allocate a subsection of frequencies intelligently. The system architecture incorporates energy-efficient mechanisms that preserve network performance while reducing power consumption—a crucial requirement for resource-constrained IoT devices in 6G networks. The system has been relaxed to minimize energy consumption without compromising performance, and the joining SU selection process has been optimized to detect specific channels. Compared to the currently offered multi-

band CSS schemes, reproduction consequences show that the recommended M2CSS arrangements select SUs through distinct nearby sensed evidence to preserve an effective agreement variation procedure and high-performance stages in lower and higher 6G IoT CRNs. While the proposed MBMU-CR IoT framework is demonstrated via simulations, extending the work to hardware implementation could introduce additional complexity, but would significantly enhance the evaluation under real-world conditions. Incorporating the concepts of joint sensing and communication, along with integrated sensing and communication, can significantly enhance the performance of the proposed M2CSS.

REFERENCES

[1] B. Kumar, S. K. Dhurandher, and I. Woungang, "A survey of overlay and underlay paradigms in cognitive radio networks," Int. J. Commun. Syst., vol. 31, no. 2, pp. e3443, 2018.

[2] M. M. Aslam, L. Du, X. Zhang, Y. Chen, Z. Ahmed, and B. Qureshi, "Sixth generation (6G) cognitive radio network (CRN) application, requirements, security issues, and key challenges," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, pp. 1–14, 2021.

[3] G. Hattab and M. Ibnkahla, "Multiband spectrum access: Great promises for future cognitive radio networks," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 282–306, 2014.

[4] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-radio-based internet of things: Applications, architectures, spectrum related functionalities, and future research directions," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 17–25, 2017.

[5] A. H. Sodhro, S. Pirbhulal, Z. Luo, K. Muhammad, and N. Z. Zahid, "Toward 6G architecture for energy-efficient communication in IoT-enabled smart automation systems," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5141–5148, 2020.

[6] M. M. Aslam, L. Du, X. Zhang, Y. Chen, Z. Ahmed, and B. Qureshi, "Sixth generation (6G) cognitive radio network (CRN) application, requirements, security issues, and key challenges," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, pp. 1–14, 2021.

[7] W. Ejaz and M. Ibnkahla, "Multiband spectrum sensing and resource allocation for IoT in cognitive 5G networks," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 150–163, 2017.

[8] H. H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1938–1948, 2018.

[9] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.

[10] L. P. Rachakonda, M. Siddula, and V. Sathya, "A comprehensive study on IoT privacy and security challenges with focus on spectrum sharing in Next-Generation networks (5G/6G/beyond)," *High-Confidence Computing*, 2024, Art. no. 100220.

[11] A. Gharib, W. Ejaz, and M. Ibnkahla, "Distributed learning-based multi-band multi-user cooperative sensing in cognitive radio networks," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, Dec. 2018, pp. 1–6.

[12] P. Schulz et al., "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[13] X. Wang, S. Ekin, and E. Serpedin, "Joint spectrum sensing and resource allocation in multi-band-multi-user cognitive radio networks," *IEEE Transactions on Communications*, vol. 66, no. 8, pp. 3281–3293, 2018.

[14] N. Y. Al-Matari, A. T. Zahary, and A. A. Al-Shargabi, "A survey on advancements in blockchain-enabled spectrum access security for 6G cognitive radio IoT networks," *Scientific Reports*, vol. 14, no. 1, Art. no. 30990, 2024.

[15] D. Balakumar and N. Sendrayan, "Enhance the probability of detection of cooperative spectrum sensing in cognitive radio networks using blockchain technology," *Journal of Electrical and Computer Engineering*, vol. 2023, pp. 1–8, 2023.

[16] P. Talajiya, A. Gangurde, U. Ragavendran, and H. Murali, "Cognitive radio networks and spectrum sensing: A review," unpublished, 2020.

[17] S. Liu, J. He, and J. Wu, "Dynamic cooperative spectrum sensing based on deep multi-user reinforcement learning," *Applied Sciences*, vol. 11, no. 1, 2021.

[18] Y. Liu, X. Qin, Y. Huang, L. Tang, and J. Fu, "Maximizing energy efficiency in hybrid overlay-underlay cognitive radio networks based on energy harvesting-cooperative spectrum sensing," *Energies*, vol. 15, no. 3, 2022.

[19] M. Saeedzarandi, "Optimization of multiband sensing-time-adaptive detection in cognitive radio networks using artificial immune algorithm," in *Proc. 2nd Conf. Swarm Intelligence and Evolutionary Computation (CSIEC)*, Tabriz, Iran, Mar. 2017, pp. 29–34.

[20] M. Soltani, T. Baykas, and H. Arslan, "On reducing multiband spectrum sensing duration for cognitive radio networks," in *Proc. IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Montreal, Canada, Sep. 2016, pp. 1–5.

[21] D. Balakumar and N. Sendrayan, "Enhance the probability of detection of cooperative spectrum sensing in cognitive radio networks using blockchain technology," *Journal of Electrical and Computer Engineering*, vol. 2023, pp. 1–10, 2023.

[22] Y. Ma, Y. Gao, Y. C. Liang, and S. Cui, "Reliable and efficient sub-Nyquist wideband spectrum sensing in cooperative cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 10, pp. 2750–2762, 2016.

[23] A. A. Olawole, F. Takawira, and O. O. Oyerinde, "Cooperative spectrum sensing in multichannel cognitive radio networks with energy harvesting," *IEEE Access*, vol. 7, pp. 123456–123469, 2019.

[24] L. Yu, J. Chen, and G. Ding, "Spectrum prediction via long short term memory," in *Proc. 3rd IEEE Int. Conf. Computer and Communications (ICCC)*, Chengdu, China, Dec. 2017, pp. 643–647.

[25] X. L. Huang, Y. Gao, X. W. Tang, and S. B. Wang, "Spectrum mapping in large-scale cognitive radio networks with historical spectrum decision results learning," *IEEE Access*, vol. 6, pp. 21350–21358, 2018.

[26] T. Xiong, Y. D. Yao, Y. Ren, and Z. Li, "Multiband spectrum sensing in cognitive radio networks with secondary user hardware limitation: Random and adaptive spectrum sensing strategies," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3018–3029, 2018.

[27] S. K. Agrawal, A. Samant, and S. K. Yadav, "Spectrum sensing in cognitive radio networks and metacognition for dynamic spectrum sharing between radar and communication system: A review," *Physical Communication*, vol. 52, Art. no. 101673, 2022.

[28] M. S. Kerdabadi, R. Ghazizadeh, H. Farrokhi, and M. Najimi, "Joint sensing times detection thresholds and users association optimization in multi-channel multi-antenna cognitive radio networks," *Transactions C: Aspects*, vol. 36, no. 09, pp. 1704–1714, 2023.

[29] K. Rapetswa and L. Cheng, "Convergence of mobile broadband and broadcast services: A cognitive radio sensing and sharing perspective," *Intelligent and Converged Networks*, vol. 1, no. 1, pp. 99–114, 2020.

High-Speed Fiber-Optic Communication Performance Utilizing Fiber Bragg Grating-Based Dispersion Compensation Schemes

Kripa Kalkala Balakrishna[1], Karthik Palani[2]*
Research Scholar, Cambridge Institute of Technology, K.R. Puram, Bengaluru-560036, Karnataka, India[1]
Affiliated to Visvesvaraya Technological University, Belagavi – 590018, Karnataka, India[1]
Professor, Department of Information Science Engineering, Cambridge Institute of Technology, K.R. Puram, Bengaluru- 560036, Karnataka, India[2]

Abstract*—Chromatic dispersion is a significant limitation in optical fiber communication, as it causes pulse broadening, which negatively impacts transmission distance and data rates, both of which are critical for meeting the high-speed demands of 5G optical networks. This study focuses on addressing chromatic dispersion in Standard Single-Mode Fiber (SSMF) systems, which are widely deployed in 5G fronthaul and access networks. A comprehensive investigation is conducted using Gaussian-apodized linear chirped Fiber Bragg Gratings (FBGs) for dispersion compensation, implemented across three strategic configurations: pre-compensation, post-compensation, and symmetrical compensation. Each scheme is systematically evaluated to determine the most effective approach for enhancing signal integrity and overall network performance. Simulations are performed using OptiSystem 7.0 on a 10 Gbps SSMF-based optical system, with transmission distances ranging from 10 km to 80 km under controlled simulation parameters. Key performance metrics, including Quality factor (Q-factor), Bit Error Rate (BER), and eye height, are analyzed by varying SSMF length, input power, and bit rate. The results demonstrate that symmetrical compensation using Gaussian-apodized linear chirped FBGs provides the best performance, achieving a Q-factor of 12.3938, an ultra-low BER of 1.12336×10^{-35}, and a significantly improved eye height at 80 km. These findings establish the symmetrical compensation scheme employing Apodized Chirped Fiber Bragg Gratings (ACFBGs) as the most effective and scalable solution for high-speed, long-distance optical transmission in 5G networks. This approach enables key 5G applications, including ultra-reliable low-latency communication (URLLC), enhanced mobile broadband (eMBB), and smart infrastructure in smart cities. The proposed technique offers multiple advantages, such as low BER, high Q-factor, reduced signal distortion through sidelobe suppression, energy efficiency via passive operation, and design flexibility for long-haul network integration.

Keywords*—Dispersion management; chirped fiber bragg grating; gaussian apodization; quality factor; bit error rate; optical transmission system

I. Introduction

High-speed optical communication systems rely on laser sources that emit nearly monochromatic light with a narrow spectral linewidth. As optical signals propagate through SSMFs, different wavelength components propagate at different velocities, leading to temporal spreading of the optical signal. This phenomenon leads to pulse overlap, commonly known as Intersymbol Interference (ISI), which is primarily caused by chromatic dispersion. In SSMFs, chromatic dispersion typically ranges from 15 to 20 ps/nm/km, and its impact becomes increasingly severe with longer transmission distances and higher bit rates. Chromatic dispersion in optical fibers arises from two main components: material dispersion (MD) and waveguide dispersion (WD). In single-mode fibers, material dispersion is generally the predominant factor, arising from the refractive index of the fiber material varying with wavelength [1]. In long-haul, high-speed optical networks, the primary impairments to signal quality include attenuation, nonlinear effects, and chromatic dispersion. While attenuation can be mitigated using Erbium-Doped Fiber Amplifiers (EDFAs) and nonlinearities managed by optimizing input power levels, chromatic dispersion remains the most critical limiting factor. Due to its minimal attenuation, the 1550 nm window is commonly used in optical communication systems [2].

To address chromatic dispersion in optical fiber systems, several compensation methods have been developed, such as Dispersion-Compensating Fibers (DCFs), Dispersion-Shifted Fibers, and Fiber Bragg Gratings (FBGs). Among these, DCFs offer negative dispersion to counterbalance the positive dispersion of SSMFs, but they can increase system cost and introduce nonlinear effects. Therefore, FBGs are preferred as a more compact, passive, and low-loss alternative [3–5].

FBGs are typically categorized into Uniform Fiber Bragg Gratings (UFBGs), which have constant grating periods, and Chirped Fiber Bragg Gratings (CFBGs), whose grating periods vary along their length. CFBGs can incorporate different apodization profiles and chirp functions, allowing the grating period to vary in a linear, quadratic, or cubic manner. Among these, Linear CFBGs are widely regarded as an effective solution for chromatic dispersion correction and compensation due to their ability to reflect different wavelengths at different spatial locations, thereby introducing wavelength-dependent delays and mitigating pulse broadening [6]. In addition, they are immune to nonlinear effects and are well-suited for high-speed optical systems. To further enhance performance, apodization techniques are employed. Apodization involves a gradual modulation of the refractive index profile, which suppresses

*Corresponding Author

side lobes in the reflection spectrum and reduces signal distortion. This leads to the development of ACFBGs, which provide superior dispersion compensation with improved signal quality [7–9].

EDFAs are commonly used to amplify optical signals in long-distance systems, while dispersion compensation can also be assisted by optical filters, such as Gaussian, Bessel, and Fabry-Pérot filters, at the receiver end [10]. As bit rates approach 40 Gbps, chromatic dispersion becomes a critical factor limiting system performance [11]. Although single-mode fibers are capable of supporting high-speed data rates, they still suffer from signal degradation due to attenuation and dispersion-induced waveform distortion, which contributes to ISI and crosstalk [12–14]. The Q-factor and BER are key parameters that determine the achievable transmission distance in optical communication systems. For reliable performance, the system should maintain a Q-factor above 6 and a BER below 10^{-9} to ensure signal integrity over long distances [15–17].

Despite extensive research into dispersion compensation using CFBGs and DCFs, most existing studies focus on a single compensation scheme, typically either pre- or post-compensation schemes under fixed system configurations. Systematic comparisons of all three strategies (pre-, post-, and symmetrical), particularly using Gaussian-apodized CFBGs, are rarely explored. Additionally, few works have evaluated performance at extended distances (e.g., 80 km) or considered the combined impact of key parameters such as transmission length, bit rate, and input power. This lack of holistic analysis presents a gap in identifying the most effective dispersion compensation approach designed for long-distance, high-speed optical communication systems.

To bridge this gap, the present study designs and evaluates Gaussian-apodized linear CFBGs across three compensation schemes: pre-, post-, and symmetrical. Simulations are conducted in OptiSystem 7.0 over transmission distances spanning 10 km to 80 km, input power levels from 1 dBm to 10 dBm, and data rates varying between 5 Gbps and 20 Gbps. The results indicate symmetrical compensation consistently delivers superior dispersion mitigation, particularly at extended distances and lower power levels.

A. Key Contributions of this Study Include

- Design and optimization of ACFBG models with Gaussian apodization and linear chirp profiles to enhance dispersion compensation.
- Evaluation of pre-, post-, and symmetrical compensation methods, highlighting their respective advantages and limitations.
- Assessment of signal quality and transmission integrity using key evaluation metrics, including BER, Q-factor, and eye height, under different compensation scenarios.
- Validation of the proposed ACFBG designs and compensation schemes via simulation, demonstrating their feasibility for practical applications in long-haul and metropolitan networks.

The structure of this study is organized as follows: Section II reviews the related work; Section III introduces the fundamental concepts; Section IV describes the methodology; Section V presents the simulation results; Section VI provides a discussion of the findings; Section VII concludes the study and outlines potential directions for future work.

II. Related Works

Numerous studies have investigated dispersion compensation techniques in optical fiber systems, with a particular focus on the use of FBGs and DCFs in various configurations.

Murooj A. Abd Almuhsan [18] implemented a post-compensation system using both DCF and FBG over a 50 km SSMF link at 10 Gbps in OptiSystem 12.0. By using an FBG having a dispersion coefficient of −2000 ps/nm, the system achieved outstanding performance with a Q-factor of 45.78, BER of 0, and eye height of 6.9×10^{-14}.

Keti [19] proposed a post-compensation model using a linear CFBG featuring tanh apodization and a Gaussian low-pass filter. Over a 40 km fiber span. The resulting Q-factor and BER of the system were 29.55 and 9.34×10^{-18}, outperforming both uniform FBG and uncompensated setups. This study emphasizes the impact of apodization and filtering in enhancing dispersion compensation.

Mustafa et al. [20] showed that the post-compensation scheme with raised sine apodization delivered the best results in a four-cascaded apodized linear CFBG model using the Multichannel Time Division Multiplexing (MTDM) approach. In a 10 Gbps WDM transmission over 70 km utilizing 2 mm-long gratings, the system attained a Q-factor of 8.58 with a BER of 4.77×10^{-18}. By comparison, the pre-compensation approach using a raised cosine apodization achieved a Q-factor of 7.664 and BER of 8.99×10^{-15}, while the symmetrical compensation scheme with a raised sine profile resulted in a Q-factor of 7.49 and BER of 3.53×10^{-14}.

Mustafa et al. [21] reported that the post-compensation approach delivered the superior performance in their introduction of a UFBG model employing single-stage tanh apodization within the MTDM framework for a 10 Gbps WDM system spanning a 70 km optical link, this configuration achieved a Q-factor of 9.2 along with an exceptionally low BER of 3.25×10^{-20}. In comparison, the pre-compensation approach achieved a Q-factor of 9.13 and a BER of 3.4×10^{-20}, whereas the symmetrical compensation method resulted in a Q-factor of 8.8 and a BER of 8.22×10^{-19}.

Panda et al. [22] employed an FBG-based post-compensation design for an 80 km fiber link, resulting in a Q-factor of 6.85 and a BER of 3.73×10^{-12} using 2 mm and 3 mm gratings, showcasing the practicality of FBGs in long-distance links.

Chakkour et al. [23] designed a post-compensation scheme combining UFBGs with an EDFA in a 10 Gbps WDM system. Over a 10–30 km span, the system achieved Q-factors up to 81.47 and showed strong optical gain, demonstrating potential in short-distance high-speed links.

Mustafa et al. [24] proposed a post-compensation scheme using a four-cascaded CFBG model with Differential Phase-Shift Keying (DPSK) modulation within a 10 Gbps WDM system deployed across a 70 km fiber link. The system utilized 5 mm FBG gratings and achieved a Q-factor of 7.22 and a BER of 2.59×10^{-13}, demonstrating the effectiveness of cascaded CFBG for dispersion mitigation.

While previous studies have explored the use of CFBGs for dispersion compensation, most have focused on a single compensation scheme or operated under fixed system parameters, with limited use of Gaussian apodization and few comparative analyses. This study addresses these gaps by systematically evaluating pre-, post-, and symmetrical compensation using Gaussian-apodized linear CFBGs over varying SSMF lengths (10 to 80 km), input powers, and bit rates in OptiSystem 7.0. The results show that at 80 km, the symmetrical scheme outperforms previous FBG-based methods, highlighting its effectiveness for long-haul optical links.

III. Concepts and Principles

A. Principle of Fiber Bragg Grating

Ken Hill pioneered the first demonstration of an FBG in 1978. A FBG is created by inscribing periodic variations in the refractive index along the core of an optical fiber through exposure to a high-intensity optical interference pattern. This periodic modulation forms a grating that reflects a specific wavelength, known as the Bragg wavelength, while permitting other wavelengths to propagate through the fiber with minimal loss. As illustrated in Fig. 1, a UFBG maintains a constant grating period and index modulation, which allows it to reflect only the wavelength satisfying the Bragg condition. This makes it particularly advantageous for narrowband filtering in optical communication systems [25].

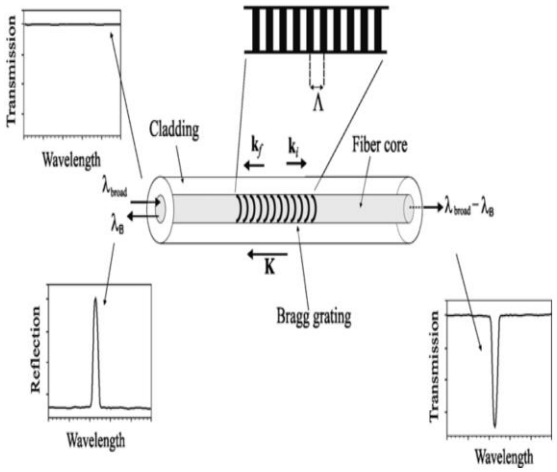


Fig. 1. UFBG with constant modulation amplitude and period.

The Bragg wavelength (λ_{Bragg}) is the specific wavelength reflected by the grating, is defined by Eq. (1):

$$\lambda_{Bragg} = 2n_{eff}\Lambda \quad (1)$$

Here, n_{eff} is used to indicate the effective index of refraction in the fiber core, and Λ denotes the periodic spacing of the grating [26]. This equation highlights that the reflected wavelength can be tuned by modifying either the refractive index or the grating period, enabling FBGs to support a wide range of optical applications.

B. Apodized Chirped Fiber Bragg Gratings (ACFBG) Structure and Mathematical Framework

CFBGs are used to compensate for chromatic dispersion in optical fibers by gradually varying the grating spacing distributed along the fiber length. This variation causes different wavelengths of light to reflect at different positions within the grating [27]. As a result, each wavelength experiences a different delay, enabling the recompression of optical pulses that have been broadened by dispersion.

In this process, shorter (faster) wavelengths reflect from regions with a smaller grating period, which are located deeper within the grating, resulting in longer delays. Conversely, longer (slower) wavelengths reflect from regions with a larger grating period, located closer to the input, and thus experience shorter delays. Because of their broad reflection bandwidth and non-uniform grating structure, CFBGs are widely employed to compensate for dispersion in high-speed optical systems [28], as shown in Fig. 2. Furthermore, the use of ACFBGs significantly enhances system performance by reducing side lobes and minimizing group delay ripple**,** making them particularly suitable for precise chromatic dispersion management in modern optical networks [29].

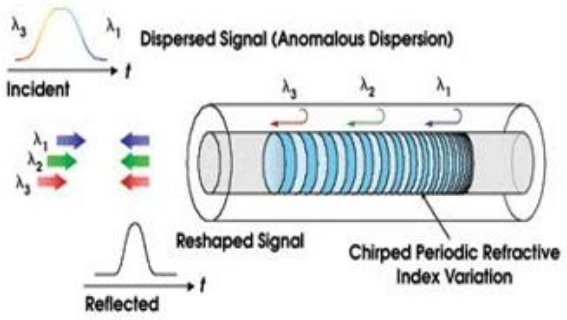


Fig. 2. Structure and operating principle of a CFBG.

The range of reflected wavelengths is expressed as Eq. (2):

$$\Delta\lambda_{chirp} = 2n_{eff}(\Lambda_L - \Lambda_S) = 2n_{eff}\,\Delta\Lambda_{chirp} \quad (2)$$

Here, $\Delta\lambda$chirp represents the difference from shortest to longest reflected wavelengths, denoted by λL and λS, respectively, while $\Delta\Lambda$chirp represents the variation between the longest ΛL and shortest ΛS grating period [30].

In a linear CFBG, the grating period at position z is defined as Eq. (3):

$$\Lambda_{(z)} = \Lambda_0 - \left[\frac{\left(\frac{z - Lg}{2}\right)}{Lg}\right]\Delta \quad (3)$$

where, Λ_0 is the midpoint period, *Lg* denotes the total grating length, while Δ represents the overall chirp [30]. Each wavelength reflects from a specific grating section, introducing a unique time delay.

The time delay τ (λ) experienced by a wavelength λ is given by:

$$\tau(\lambda) = \frac{\lambda - \lambda s}{\lambda L - \lambda s} \frac{2neffLg}{c} \quad (\lambda s \leq \lambda \leq \lambda L) \qquad (4)$$

Here, c signifies the speed at which light travels in a vacuum [30].

To determine the CFBG's dispersion in ps/nm/km, Eq. (4) is differentiated with respect to the wavelength λ, as in Eq. (5):

$$D(\lambda) = \frac{d\tau(\lambda)}{d\lambda} = \frac{1}{\lambda L - \lambda s} \frac{2neffLg}{c} \quad (\lambda s \leq \lambda \leq \lambda L) \qquad (5)$$

The refractive index profile $n(z)$ of an apodized linear CFBG is given by Eq. (6):

$$n(z) = n_0 + \Delta n_{dc}(z) + f_{A(z)} \Delta n_{ac}(z) \cos\left[\frac{2\pi}{\Lambda} z + \phi(z)\right] \qquad (6)$$

Here, n0 refers to the initial refractive index, fA(z) represents the apodization function, Δnac defines the modulation amplitude, Δndc is the average index change, and ϕ(z) is the phase [30]–[32].

The effective refractive index is therefore approximated as Eq. (7):

$$n_{eff} = n_0 + \Delta n_{dc} \qquad (7)$$

A Gaussian apodization function can be used to taper the refractive index modulation smoothly and is described as Eq. (8):

$$A_{(z)} = \exp\left\{-\ln 2 \left[\frac{2.(z - L/2)^2}{s.L}\right]\right\} \qquad (8)$$

Here, L represents the grating length, and s denotes the apodization strength parameter, and $z \in [0, L]$ [33].

IV. Methodology of the Proposed Framework

The proposed system setup, illustrated in Fig. 3, demonstrates a dispersion compensation design using an ACFBG. The input to the system is a binary data stream generated by a 10 Gbps Pseudo-Random Bit Sequence (PRBS) generator, which is then converted into Return-to-Zero (RZ) optical pulses. These pulses are modulated using a Mach–Zehnder Modulator (MZM) with an extinction ratio of 30 dB. The MZM encodes the data onto a continuous wave (CW) optical carrier with 3 dBm power, generated by a semiconductor laser operating at 193.4 THz. The modulated optical signal is transmitted through SSMF toward the receiver.

Dispersion compensation is performed at the receiver using an ACFBG featuring linear chirp and Gaussian apodization to improve performance by minimizing side lobes and group delay ripple. The system operates at a wavelength of 1550 nm. An EDFA optical amplifier with a 10 dB gain is used to enhance the signal strength before the detection. A PIN photodetector is employed for optical-to-electrical conversion, and the signal is further processed using a Low-Pass Gaussian Filter (LPGF) with a cutoff frequency set to 0.75 times the bit rate to suppress high-frequency noise.

System performance is evaluated using Q-factor, BER, and eye height, as observed through eye diagram analysis, with Q-factor serving as a key indicator of BER. All simulations are carried out in OptiSystem 7.0 by varying SSMF length, input power, and bit rate the relative performance of pre-, post-, and symmetrical compensation techniques is examined.

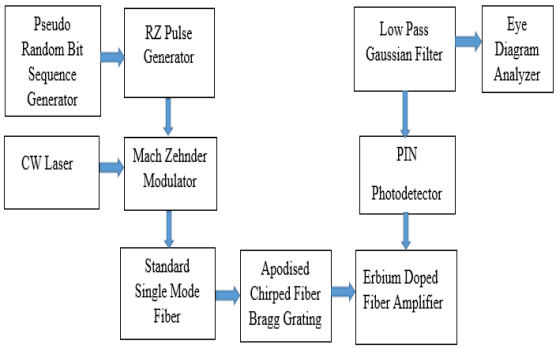


Fig. 3. Structural overview of the proposed design.

OptiSystem 7.0 was chosen for this study due to its advanced capabilities in modeling and simulating real-world optical communication systems. It effectively incorporates key transmission impairments, including chromatic dispersion, amplified spontaneous emission (ASE) noise, and nonlinear effects such as self-phase modulation. Its modular design enables flexible construction of complex system architectures, ranging from point-to-point links to WDM and long-haul transmission setups, by integrating standardized optical components such as fiber models, optical amplifiers, filters, and Fiber Bragg Gratings (FBGs).

In addition, OptiSystem offers a comprehensive suite of built-in analysis tools, including Q-factor measurement, BER estimation, and eye diagram visualization, which support both qualitative and quantitative performance evaluation. These features allow for controlled, repeatable testing across a wide range of parameters, making the platform highly suitable for validating dispersion compensation techniques in long-distance SSMF systems.

The overall simulation workflow is illustrated in Fig. 4, and the key simulation parameters are summarized in Table I.

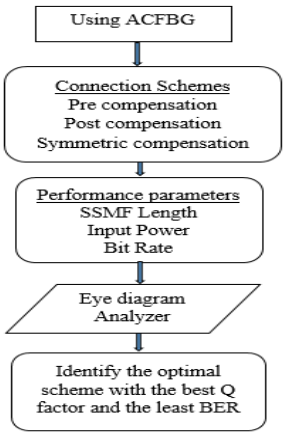


Fig. 4. Flowchart of the comparative study.

TABLE I COMPONENTS AND THEIR SIMULATION PARAMETERS

Components	Parameters
PRBS generator	Data rate: 10 Gbps
RZ pulse generator	Duty cycle: 0.5
Continuous Wave Laser	Input power: 3 dBm Linewidth: 0.01 MHz Central frequency: 193.4 THz
Mach–Zehnder Modulator	Extinction ratio: 30 dB
Standard Single-mode fiber (SSMF-28)	Length: 10 km to 80 km Attenuation loss: 0.18 dB/km Dispersion: 16.5 ps/nm/km Dispersion slope: 0.05 ps/nm^2/km Core effective area: 76.5 μm^2
FBG	Grating Length: 7 mm Effective index: 1.45 Apodization function: Gaussian Gauss parameter: 0.5 Chirp Function: Linear Linear Parameter: 0.0001 μm
EDFA optical amplifier	Gain: 10 dB Noise figure: 4 dB
PIN photodetector	Responsivity: 1 A/W Dark current: 10 nA
LPG filter	cutoff frequency: 0.75 × Bit rate

The study compares three compensation methods:

A. Case One: Pre-Compensation Scheme

In the pre-compensation setup, illustrated in Fig. 5, the ACFBG is positioned at the transmitter side, between the MZM and the SSMF. This arrangement enables early compensation for chromatic dispersion by introducing negative dispersion before the optical signal propagates through the fiber. The ACFBG's linear chirp profile and Gaussian apodization help match the anticipated dispersion profile while suppressing sidelobes and enhancing signal integrity.

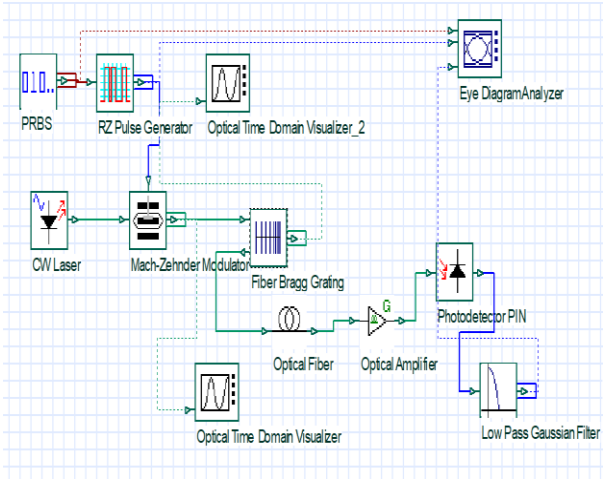


Fig. 5. Pre-compensation model for chromatic dispersion mitigation.

B. Case Two: Post-Compensation Scheme

In the post-compensation setup presented in Fig. 6, the ACFBG is placed at the receiver side, between the SSMF and the EDFA optical amplifier. As the optical signal traverses the SSMF, it accumulates positive chromatic dispersion due to the fiber's physical characteristics. The ACFBG introduces negative dispersion that compensates for this accumulated distortion. The placement of the EDFA after the ACFBG helps amplify the restored signal to meet the detection threshold of the PIN photodetector, ensuring accurate recovery of the transmitted information.

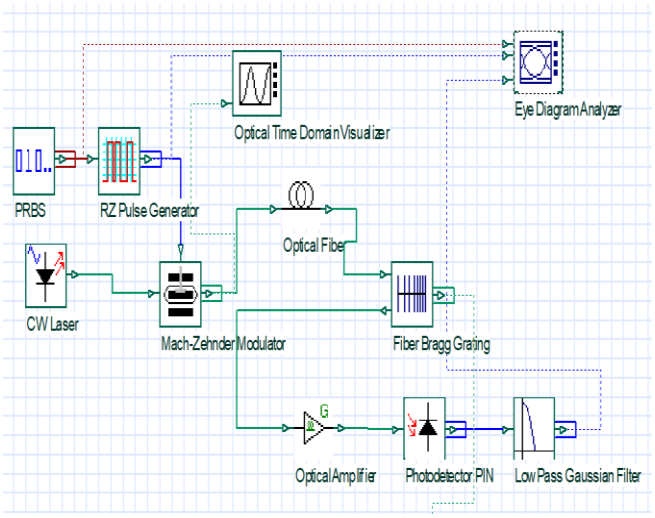


Fig. 6. Post-compensation model for chromatic dispersion mitigation.

C. Case Three: Symmetrical Compensation Scheme

The symmetrical compensation configuration, depicted in Fig. 7, comprises two ACFBGs, one located at the transmitter and the other at the receiver. The first ACFBG, situated between the MZM and the EDFA optical amplifier, provides pre-compensation, mitigating dispersion before fiber transmission. The second ACFBG, placed between the SSMF and the receiver-side EDFA optical amplifier, applies post-compensation to eliminate residual dispersion. This dual-stage configuration balances dispersion across the link and helps preserve pulse shape, thereby minimizing inter-symbol interference and improving overall system robustness. EDFAs placed before and after the SSMF maintain optimal power levels throughout the transmission path.

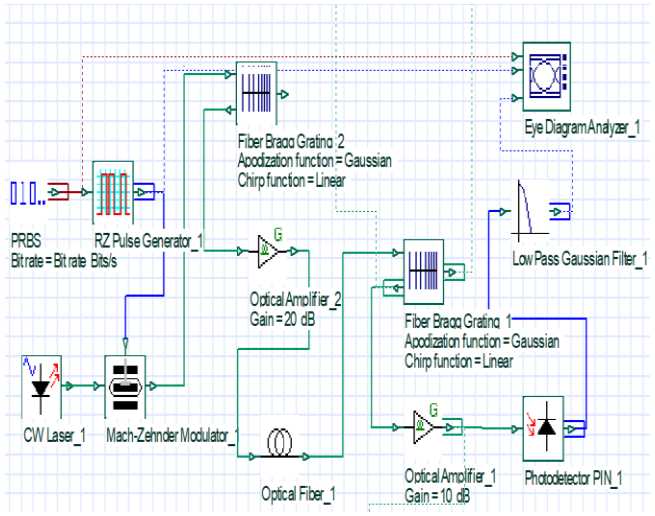


Fig. 7. Symmetrical compensation model for chromatic dispersion mitigation.

V. RESULTS

This section presents simulation results obtained using OptiSystem 7.0 to evaluate the performance of pre-, post-, and symmetrical dispersion compensation schemes. The system was tested under varying SSMF lengths, input power levels, and bit rates. Key performance metrics include Q-factor, BER, and eye height.

A. Performance with Varying SSMF Lengths

Simulations were conducted by varying the SSMF length from 10 km to 80 km, with fixed parameters: bit rate = 10 Gbps, input power = 3 dBm, and FBG length = 7 mm. Table II summarizes the results. Fig. 8 shows the eye diagram for the pre-compensation scheme at 80 km. It reveals a nearly closed eye pattern with Q-factor = 1.71181, BER = 0.0430757, and eye height = -3.43222×10^{-5}, indicating severe dispersion-induced degradation. Fig. 9 displays the eye diagram for the post-compensation scheme at 80 km. It exhibits similar degradation with Q-factor = 1.74339, BER = 0.0403394, and eye height = -3.33765×10^{-5}. Fig.10 illustrates the symmetrical compensation scheme at 80 km. The eye remains open with Q-factor = 12.3938, BER = 1.12336×10^{-35}, and eye height = 0.00215177, indicating robust signal quality. Fig. 11 presents the Q-factor performance over increasing SSMF lengths. While pre- and post-compensation schemes decline steadily beyond 50 km, symmetrical compensation maintains significantly higher Q-factors. Fig. 12 shows the corresponding BER trend. BER for pre- and post-compensation increases exponentially with distance, while symmetrical compensation sustains extremely low BER values up to 80 km.

TABLE II COMPARISON OF DISPERSION COMPENSATION OVER VARYING FIBER LENGTHS

SSMF Length	Q Factor	BER	Eye Height
Pre-compensation			
10km	204.277	0	0.00322713
20km	144.407	0	0.00196823
30km	94.3325	0	0.00110288
40km	27.6656	8.32188e-169	0.000541479
50km	10.7053	4.50313e-027	0.000217269
60km	6.55781	2.23474e-011	8.90316e-005
70km	3.199	0.00064678	5.44249e-006
80km	1.71181	0.0430757	-3.43222e-005
Post-compensation			
10km	225.662	0	0.00322955
20km	146.268	0	0.001974
30km	88.2633	0	0.00110687
40km	27.3885	1.71148e-165	0.000545474
50km	10. 9445	3.30286e-028	0.000221097
60km	6.76954	5.22884e-012	9.26084e-005
70km	3.22965	0. 000582328	6.31461e-006
80km	1.74339	0.0403394	-3.33765e-005
Symmetrical compensation			
10km	257.323	0	0.0986005
20km	251.65	0	0.0659836
30km	185.795	0	0.0439226
40km	123.736	0	0.0265569
50km	57.0918	0	0.0155047
60km	39.5139	0	0.00877851
70km	21.0042	2.48577e-098	0.00426993
80km	12.3938	1.12336e-035	0. 00215177

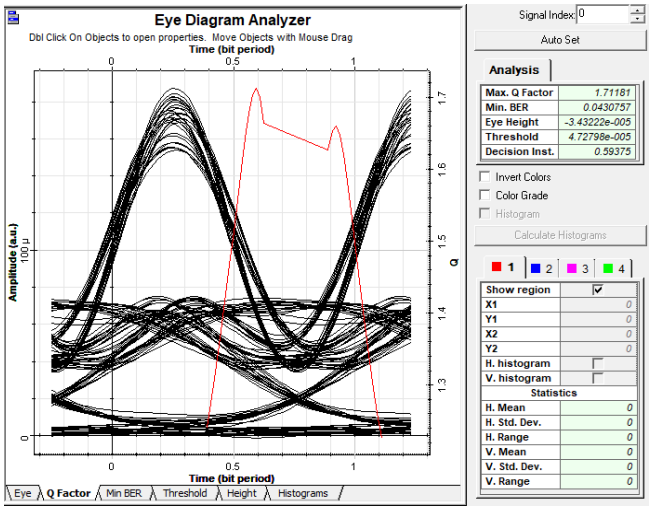


Fig. 8. Eye diagram for pre-compensation at 80 km SSMF: Q-factor = 1.71181, BER = 0.0430757, eye height = -3.43222×10^{-5}.

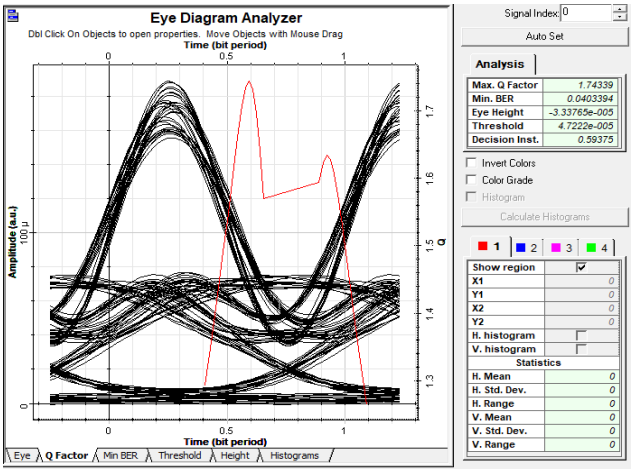


Fig. 9. Eye diagram for post-compensation at 80 km SSMF: Q-factor = 1.74339, BER = 0.0403394, eye height = 3.33765×10^{-5}.

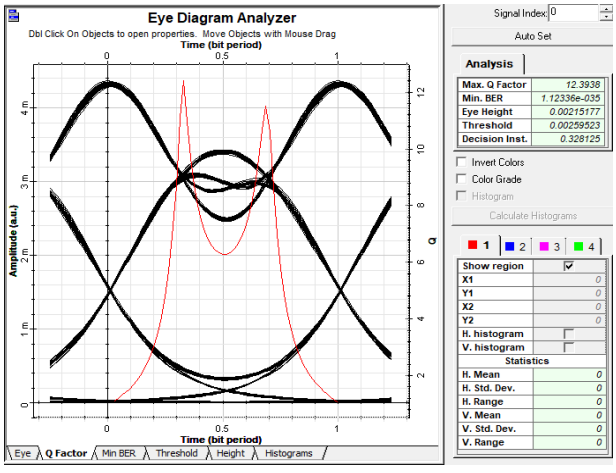


Fig. 10. Eye pattern of the symmetrical compensation scheme at 80 km SSMF length, having a Q-factor of 12.3938, BER of 1.12336×10^{-35}, and an eye height of 0.0021517.

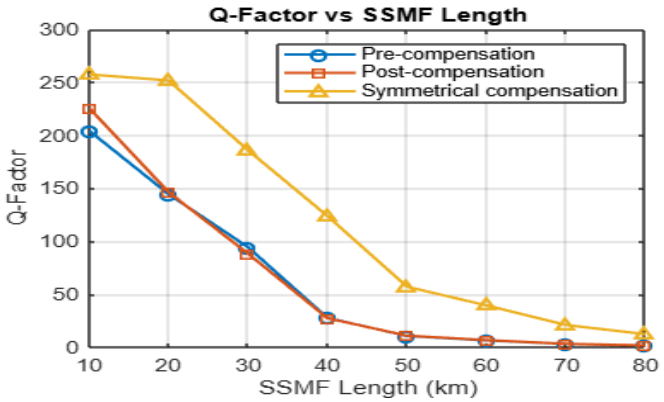


Fig. 11. Q-factor comparison of pre-, post-, and symmetrical compensation schemes over SSMF lengths from 10 km to 80 km.

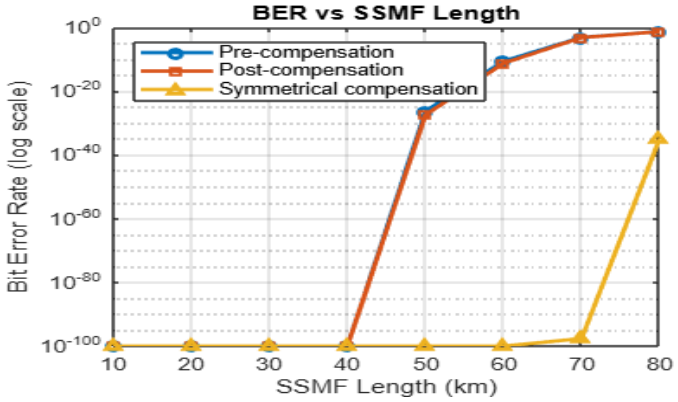


Fig. 12. BER comparison of pre-, post-, and symmetrical compensation schemes over SSMF lengths from 10 km to 80 km.

B. *Effect of Input Power*

To evaluate power efficiency, input power levels were varied between 0 dBm and 10 dBm, with fixed parameters: 60 km SSMF, 7 mm FBG length, and 10 Gbps bit rate. Results are presented in Table III. Fig. 13 shows the eye diagram for pre-compensation at 3 dBm, where the Q-factor is 6.5578, BER = 2.23474×10^{-11}, and eye height = 8.90316×10^{-5}. Fig. 14 illustrates the post-compensation performance at 3 dBm. It yields a Q-factor of 6.76954, BER = 5.22884×10^{-12}, and eye height = 9.26084×10^{-5}. Fig. 15 depicts the symmetrical compensation scheme at the same power level, with superior performance: Q-factor = 39.5139, BER = 0, and eye height = 0.00877851. Fig. 16 plots the Q-factor versus input power. All schemes improve as power increases, but symmetrical compensation peaks at 3 dBm before slightly declining. Fig. 17 presents the BER variation across the power range. Symmetrical compensation maintains ultra-low BER at lower powers, while other schemes improve gradually.

C. *Impact of Bit Rate*

To assess the performance at higher data rates, Bit rate was varied from 5 Gbps to 20 Gbps, keeping SSMF = 60 km, input power = 3 dBm, and FBG length = 7 mm. The results are shown in Table IV. Fig. 18 shows the pre-compensation scheme at 15 Gbps. The signal completely fails with Q = 0, BER = 1, and a closed eye. Fig. 19 illustrates similar failure for the post-compensation scheme at 15 Gbps, also showing Q = 0 and BER = 1. Fig. 20 shows that the symmetrical compensation scheme still maintains acceptable signal integrity at 15 Gbps, with Q-factor = 3.13256, BER = 0.000848, and eye height = 0.0001717. Fig. 21 displays the Q-factor drop with increasing bit rate. Symmetrical compensation performs best, sustaining up to 15 Gbps. Fig. 22 confirms this trend in BER values, where pre- and post-compensation fail beyond 10 Gbps, while symmetrical compensation remains viable until 15 Gbps.

TABLE III INPUT POWER IMPACT ON FBG-BASED DISPERSION COMPENSATION SCHEMES

Pre compensation			
Input power	**Q Factor**	**BER**	**Eye height**
1 dBm	6.29833	1.26638e-010	5.41366e-005
3 dBm	6.55781	2.23474e-011	8.90316e-005
5 dBm	6.75313	5.8059e- 012	0.000144925
7 dbm	6.91703	1.83027e-012	0.000235027
9 dbm	7.07983	5.71629e-013	0.000381608
10 dbm	7.17114	2.95391e-013	0.000487903
Post compensation			
1 dBm	6.48034	3.82134e-011	5.6213e-005
3 dBm	6.76954	5.22884e-012	9.26084e-005
5 dBm	7.00624	9.77794e-013	0.000151683
7 dBm	7.2343	1.8825e-013	0.000248944
9 dBm	7.51008	2.446e-014	0.000417761
10dBm	7.69198	6.13173e-015	0.000545276
Symmetrical compensation			
1 dBm	18.6572	5.50605e-078	0.00412606
3 dBm	39.5139	0	0.00877851
5 dBm	15.7013	7.40624e-056	0.0111316
7 dBm	4.05863	2.4589e-005	0.00262278
9 dBm	4.14571	1.3133e-005	0.00369366
10 dBm	4.03821	2.4014e-005	0.00541803

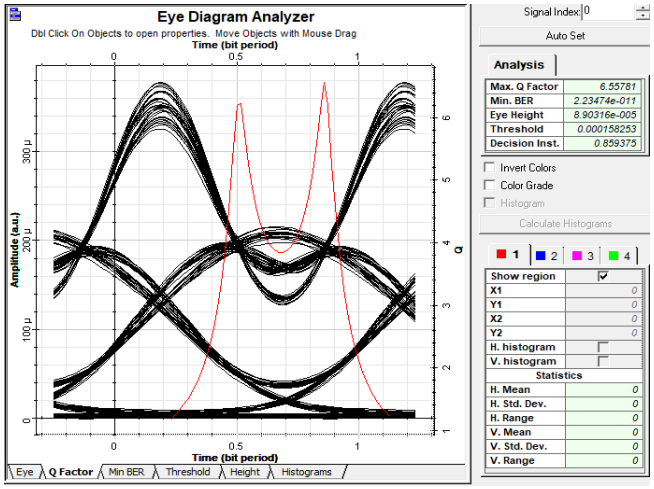


Fig. 13. Eye diagram for pre-compensation at 3 dBm input power with Q-factor 6.56, BER 2.23×10^{-11}, and eye height 8.9×10^{-5}.

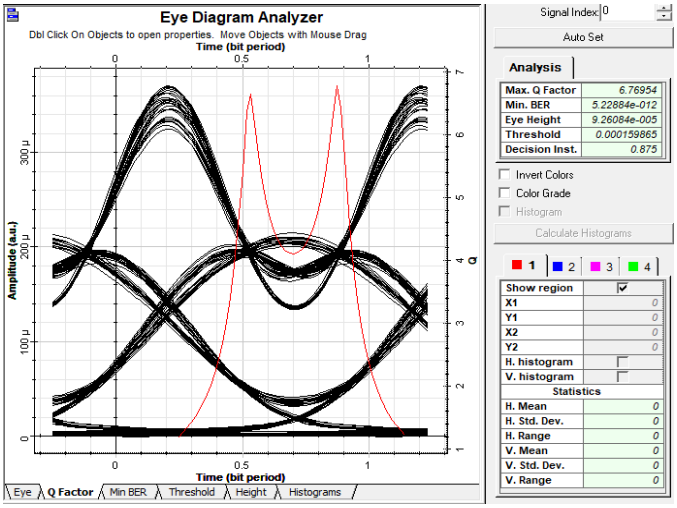


Fig. 14. Eye diagram for post-compensation at 3 dBm input power with Q-factor 6.77, BER 5.23×10^{-12}, and eye height 9.26×10^{-5}.

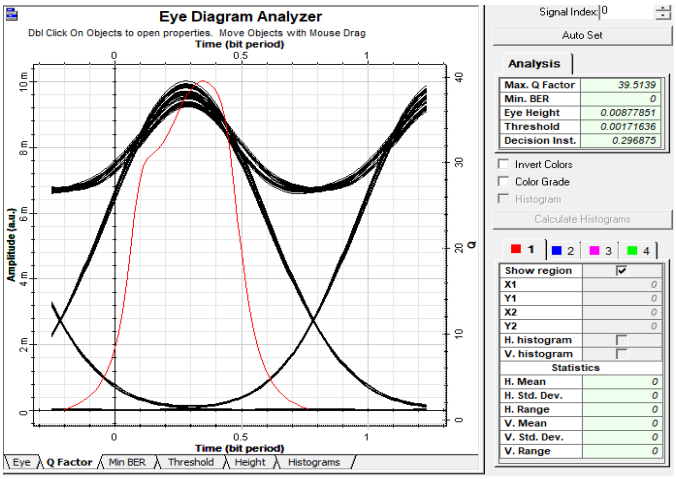


Fig. 15. Eye diagram for symmetrical compensation at 3 dBm input power with Q-factor 39.51, BER 0, and eye height of 0.00877851.

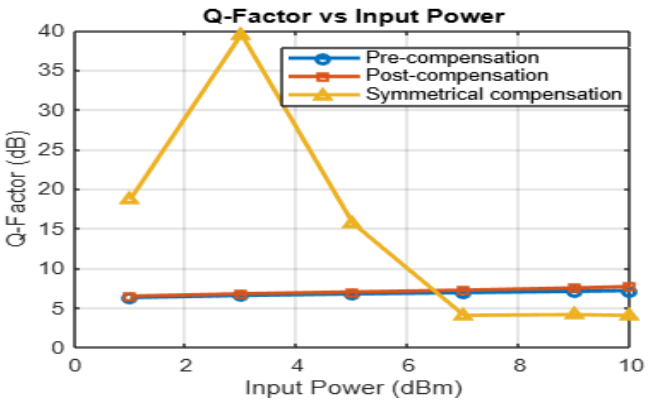


Fig. 16. Q-factor variation with input power across different dispersion compensation schemes.

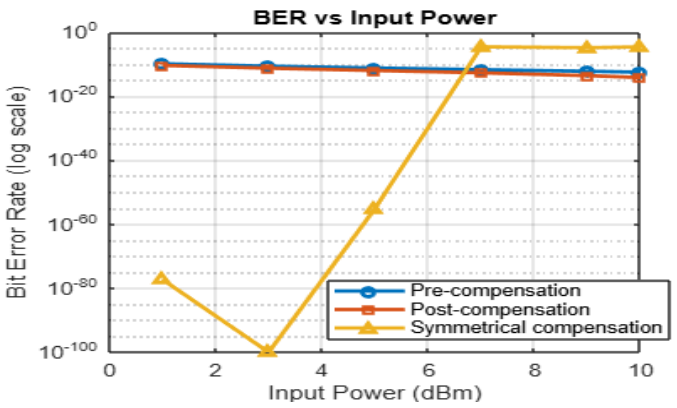


Fig. 17. BER variation with input power for different dispersion compensation configurations.

TABLE IV COMPENSATION SCHEME PERFORMANCE ACROSS VARIOUS BIT RATES

Pre-compensation			
Bit rate	**Q Factor**	**BER**	**Eye Height**
5 Gbps	77.3924	0	0.000572817
10 Gbps	6.55781	2.23474e-011	8.90316e-005
15 Gbps	0	1	0
20 Gbps	0	1	0
Post-compensation			
5 Gbps	87.9373	0	0.000580293
10 Gbps	6.76954	5.22884e-012	9.26084e-005
15 Gbps	0	1	0
20 Gbps	0	1	0
Symmetrical compensation			
5 Gbps	376.066	0	0.0275402
10 Gbps	39.5139	0	0.00877851
15 Gbps	3.13256	0.000848086	0.000171737
20 Gbps	0	1	0

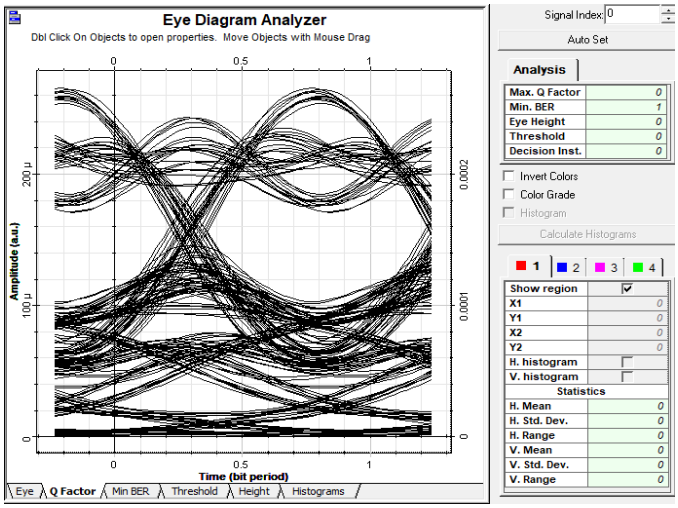


Fig. 18. Eye diagram for pre-compensation at 15 Gbps showing Q-factor 0, BER 1, and eye height 0.

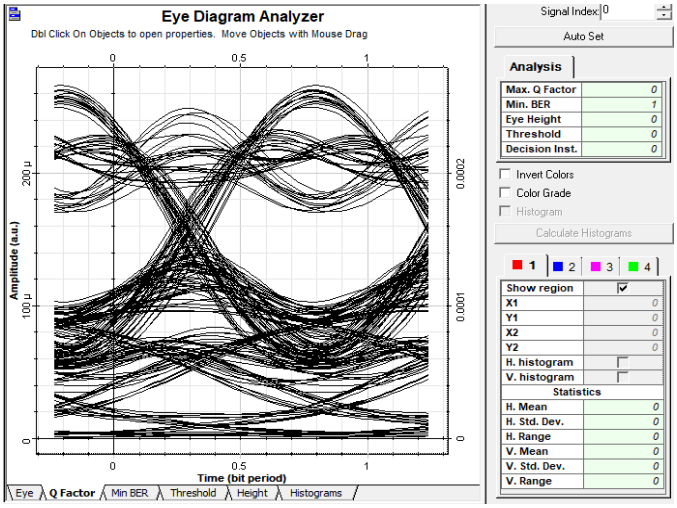


Fig. 19. Post-compensation scheme at 15 Gbps bit rate, showing a Q-factor of 0, BER of 1, and eye height of 0.

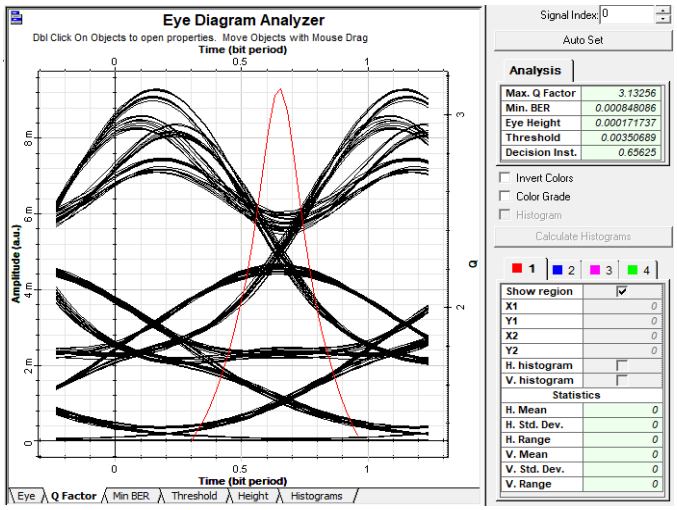


Fig. 20. Eye diagram for symmetrical compensation at 15 Gbps with Q-factor 3.13, BER of 0.000848086, and eye height of 0.000171737.

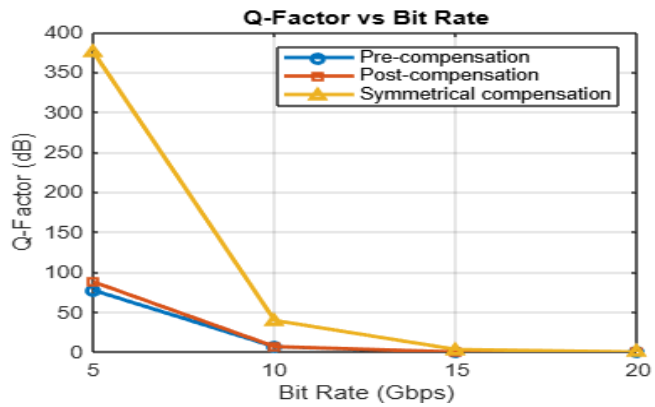


Fig. 21. Q-factor variation versus bit rate for pre-, post-, and symmetrical dispersion compensation schemes.

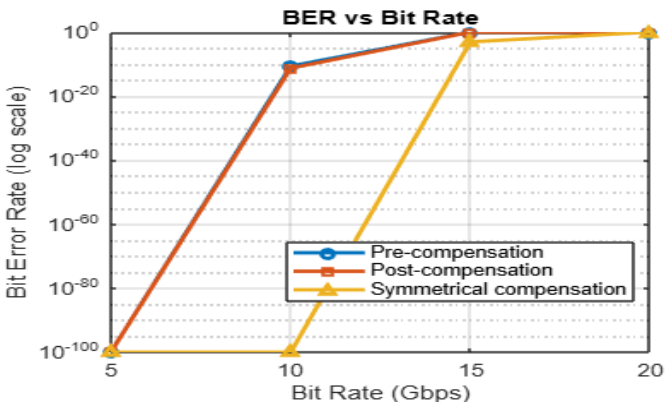


Fig. 22. Comparison of BER at various bit rates for pre-, post-, and symmetrical compensation.

The dataset used in this study was entirely generated using OptiSystem 7.0 and includes simulated performance metrics (Q-factor, BER, and eye height) across varying SSMF lengths (10–80 km), input powers (0–10 dBm), and data rates (5–20 Gbps). It is simulation-based, with no physical measurements or external inputs, and its accuracy depends on the fidelity of the OptiSystem simulation models.

VI. Discussion

This section interprets the results presented in Section V, emphasizing the performance trade-offs and advantages of each compensation approach.

A. Interpretation of SSMF Length Results

Symmetrical compensation consistently delivers superior performance compared to the other two compensation schemes. While the pre- and post-compensation approaches exhibit significant performance degradation beyond 50 km dropping to a Q-factor of approximately 1.72 and a BER around 0.04 at 80 km. Symmetrical compensation maintains high signal quality, achieving a Q-factor of 12.39, an ultra-low BER of 1.12×10^{-35}, and an improved eye height. These results highlight symmetrical compensation as the most reliable and efficient approach for long-haul, high-speed optical communication systems.

B. Power Efficiency Evaluation

The pre- and post-compensation schemes show consistent improvement in performance as the input power increases. Both demonstrate a gradual rise in Q-factor and a corresponding reduction in BER as power increases from 1 dBm to 10 dBm. At 9–10 dBm, both schemes reach a Q-factor of approximately 7 and achieve BER values on the order of 10^{-13}, indicating reliable signal quality at higher power levels. Between the two, post-compensation performs slightly better, offering higher Q-factors and lower BERs across the entire power range. In contrast, symmetrical compensation stands out by delivering excellent performance at significantly lower power. It achieves a remarkably high Q-factor of 39.51 and zero BER at just 3 dBm input power, along with the largest eye opening (eye height = 0.00878). Although its performance varies at higher power levels, the ability to attain optimal results at minimal input power underscores symmetrical compensation as the most

power-efficient solution, especially suitable for systems where energy efficiency is critical.

C. Bit Rate Performance

Pre- and post-compensation schemes fail to support bit rates beyond 10 Gbps, with Q-factors dropping to zero and BER reaching 1. In contrast, symmetrical compensation maintains acceptable performance up to 15 Gbps, attained a Q-factor of 3.13 and a BER of 8.48×10^{-4}. Although slightly degraded, this outcome confirms the scalability of symmetrical compensation for higher data rates. These results are consistent with the observed Q-factor and BER trends and are further supported by corresponding eye diagram evaluations.

D. Comparison of Proposed and Existing Methods

To validate the efficiency of the proposed dispersion compensation scheme, its performance is compared against several established methods from the literature. Table V summarizes key metrics such as Q-factor and BER across different compensation strategies, fiber lengths, and grating parameters. While most prior studies focus on post-compensation and evaluate transmission distances up to 70 km, this work demonstrates that the symmetrical compensation scheme using Gaussian-ACFBGs not only supports long-haul transmission over 80 km but also achieves a notably higher Q-factor (12.39) and an ultra-low BER (1.12×10^{-35}). These results outperform existing designs without relying on cascading or hybrid filtering techniques. The results clearly demonstrate the superior efficiency of the proposed scheme in long-haul optical transmission scenarios, establishing it as a competitive alternative to traditional post-compensation techniques.

TABLE V COMPARISON OF PROPOSED SCHEME WITH PRIOR WORKS

Prior Research	Compensation schemes with best results	SSMF Length(km)	Grating length (mm)	Q factor	BER
[18]	Post-compensation	50	-	45.78	0
[19]	Post-compensation	40	6	29.55	9.3407×10^{-182}
[20]	Post-compensation	70	2	8.58	4.77×10^{-18}
[21]	Post-compensation	70	2	9.2	3.25×10^{-20}
[22]	Post-compensation	80	2,3	6.848	3.73×10^{-12}
[23]	Post-compensation	30	-	23.647	Not given
[24]	Post-compensation	70	5	7.22	2.59×10^{-13}
Proposed Work	Symmetrical-compensation	80	7	12.3938	1.12336×10^{-35}

E. Physical Mechanism behind Performance Improvement

The symmetrical compensation scheme using Gaussian ACFBGs performs best because it balances dispersion correction by placing compensating gratings at both ends of the fiber. The linear chirp in the grating reflects different wavelengths at specific points, accurately compensating for dispersion. Gaussian apodization reduces unwanted side lobes, thereby lowering signal distortion. By dividing the compensation into two stages, this method reduces nonlinear effects, phase distortion, and jitter more effectively than pre- or post-compensation alone. As a result, it delivers cleaner signals with higher Q-factors, lower BER, and better eye diagrams, making it highly suitable for long-distance, high-capacity optical communication systems.

VII. CONCLUSION

This research investigates dispersion compensation in optical communication systems using Gaussian-apodized, linear-chirped Fiber Bragg Gratings. Three compensation schemes pre-compensation, post-compensation, and symmetrical compensation, were evaluated over an 80 km SSMF link. At this distance, pre-compensation achieved a Q-factor of 1.71181, BER of 0.0430757, and an eye height of -3.43222×10^{-5}, while post-compensation slightly improved performance with a Q-factor of 1.74339, BER of 0.0403394, and eye height of -3.33765×10^{-5}. In contrast, symmetrical compensation significantly outperformed both, achieving a Q-factor of 12.3938, an ultra-low BER of 1.12336×10^{-35}, and an eye height of 0.00215177. These results demonstrate that symmetrical compensation offers the most effective solution for long-haul, high-speed optical transmission, meeting the stringent requirements of 5G networks, especially in fronthaul and backhaul scenarios. This supports key 5G applications such as ultra-reliable low-latency communication (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communication (mMTC). Furthermore, optimization of input power and bit rate enhances energy efficiency and scalability, making the approach suitable for smart cities, industrial automation, and remote healthcare systems. The proposed technique provides several advantages: it passively compensates dispersion without the need for active electronic components, thereby reducing power consumption; Gaussian apodization effectively suppresses sidelobes to improve signal quality; the chirped grating profile allows for tailored dispersion management over long distances; and the configuration is both scalable and compatible with existing fiber infrastructures. These attributes establish symmetrical ACFBGs as a highly efficient, high-performance, and future-proof solution for optical networks.

FUTURE WORK

Future work may explore system performance under higher-order modulation formats, assess the impact of temperature variations and polarization sensitivity, and investigate hybrid approaches that integrate FBGs with adaptive, machine learning–based equalizers. Additionally, the influence of channel spacing in dense WDM environments and the potential for real-time tunability using dynamic gratings could be examined to enhance system adaptability in evolving optical networks. Overall, this study demonstrates that symmetrical FBG-based dispersion compensation is a robust, scalable, and forward-looking solution for next-generation 5G optical infrastructure.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Cambridge Institute of Technology, Bengaluru, and the Visvesvaraya

Technological University, Belagavi, for their valuable support and scholarly guidance during the course of this research.

REFERENCES

[1] E. Nsengiyumva, G. Mwangi, and Kamucha, "A comparative study of chromatic dispersion compensation in 10Gbps SMF and 40Gbps OTDM systems using a cascaded Gaussian linear apodized chirped fibre Bragg grating design," Heliyon, vol. 8, no. 4, Article. 9308, 2022.

[2] A .B. Dar and R. K. Jha, "Chromatic dispersion compensation techniques and characterization of fiber Bragg grating for dispersion compensation," Opt Quant Electron, vol. 49, no. 3, p. 108, 2017.

[3] A. Mahmoud, Y.A El-Salam, M. Ahmed, and T. Mohamed, "Effect of fiber properties and dispersion management on distortions associated with two subcarriers modulation of laser diode in radio over fiber systems," The European Physical Journal D, vol. 76, no. 11, p. 229, 2022.

[4] R. K. Gupta and M. L. Meena, "Investigation of long-haul optical transmission systems: diverse chirped FBGs with DCF for 300 km length of SMF,"International Journal of Advanced Technology and Engineering Exploration, vol. 9, no. 97, pp. 1757–1766, 2022.

[5] Z. Lv, J. Du, and Z. He, "On-chip multichannel dispersion compensation and wavelength division MUX/DeMUX using chirped-multimode-grating-assisted counter-directional coupler," Photonics, vol. 11, no. 2, p. 110, Jan. 2024.

[6] I. Nsengiyumva, E. Mwangi, and G. Kamucha, "Performance analysis of a linear Gaussian- and tanh-apodized FBG and dispersion compensating fiber design for chromatic dispersion compensation in long-haul optical communication networks," International Journal of Optics, vol. 2022, Art. no. 5734420, pp. 1–14, Sep. 2022.

[7] F. M. Mustafa, A. F. Sayed, and M. H. Aly, "A reduced power budget and enhanced performance in a WDM system: a new fbg apodization function," Optical and Quantum Electronics, vol. 54, no. 8, p. 471, 2022.

[8] F. M. Mustafa, S. A. Zaky, A. A. M. Khalaf, and M. H. Aly, "Chromatic dispersion compensation by cascaded FBG with duobinary modulation scheme," Optical and Quantum Electronics, vol. 54, no. 12, p. 819, 2022.

[9] N. A. Mohammed and N. M. Okasha, "Single- and dual-band dispersion compensation unit using apodized chirped fiber Bragg grating," J Comput Electron, vol. 17, no. 1, pp. 349–360, 2018.

[10] S. Ranathive, K. V Kumar, A. N. Z Rashed, M. S. F. Tabbour and T.V.P. Sundararajan,"Performance Signature of Optical Fiber Communications Dispersion Compensation Techniques for the Control of Dispersion Management," Journal of Optical Communications, vol. 43, no. 4, pp. 611–623, 2019.

[11] Jyotsanaa, R. Kaura and R. Singh, "Performance comparison of pre-, post- and symmetrical dispersion compensation techniques using DCF on 40Gbps OTDM system for different fibre standards," Optik, vol. 125, no. 9, pp. 2134–2136, 2014.

[12] M. Sharma and D. Kumar, "Dispersion Compensation for 40 Gbps Optical Waveguide System by Using FBG," in Proc of "3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 2018, pp. 1–3.

[13] A. F. Sayed, T. M. Barakat, and I. A. Ali, "A Novel Dispersion Compensation Model using an Efficient CFBG Reflectors for WDM Optical Networks," International Journal Of Microwave and Optical Technology, vol. 12, no. 3, pp. 230–238, 2017.

[14] R. Udayakumar, V. Khanaa, and T. Saravanan,"Chromatic Dispersion Compensation in Optical Fiber Communication System and its Simulation," Indian Journal of Science and Technology, vol. 6, no. 6S, pp. 4762–4766, 2013.

[15] M. L. Meena and R. K. Gupta, "Design and comparative performance evaluation of chirped FBG dispersion compensation with DCF technique for DWDM optical trans- mission systems," International Journal for Light and Electron Optics, vol. 188, no. 1, pp. 212–224, 2019.

[16] D. Irawan, Azhar, and K. Ramadhan, "High-Performance Compensation Dispersion with Apodization Chirped Fiber Bragg Grating for Fiber Communication System," Jurnal Penelitian Pendidikan IPA, vol. 8, no. 2, pp. 992– 999, 2022.

[17] B. Yousif, A. Sh. Samrah, and R. Waheed, "High Quality Factor and Dispersion Compensation Based on Fiber Bragg Grating in Dense Wavelength Division Multiplexing," Optical Memory and Neural Networks, vol. 29, no. 3, pp. 228–243, 2020.

[18] M. A. A. Almuhsan, "Performance Evaluation of Various Dispersion Compensation Techniques," Journal of Optics, vol. 53, no. 3, 2024.

[19] F. Keti, "A New Proposed Model for Dispersion Compensation via Linear Chirped Fiber Bragg Grating," Traitement du Signal, vol. 41, no. 1, pp. 451–457, 2024.

[20] F. M. Mustafa, A. Mohamed, A. A. M. Khalaf, A. F. Sayed, and M. H. Aly, "Dispersion compensation using cascaded apodized CFBGs under MTDM transmission technique: Enhanced system performance," Optical and Quantum Electronics, vol. 55, no. 1, p. 32, 2022.

[21] F. M. Mustafa, H. A. Kholidy, A. F. Sayed, and M. H. Aly, "Enhanced dispersion reduction using apodized uniform fiber Bragg grating for optical MTDM transmission systems," Optical and Quantum Electronics, vol. 55, no. 1, p. 55, 2022.

[22] T. K. Panda, A. B. M K. Behera, and S. Polei, "Performance comparison of Fiber Bragg Grating for different Grating length, Apodization function and Chirp function," in Proc. of 2020 International Conference on Computer Science, Engineering and Applications (ICC- SEA), 2020, pp. 1–5.

[23] M. Chakkour, O. Aghzout, B. A. Ahmed, F. Chaoui, and M. E. Yakhloufi, "Chromatic Dispersion Compensation Effect Performance Enhancements Using FBG and EDFA-Wavelength Division Multiplexing Optical Transmission System," Hindawi International Journal of Optics, vol. 6428972, pp. 1–8, 2017.

[24] F. M. Mustafa, S. A. Zaky, A. A. M. Khalaf, and M. H. Aly, "A cascaded FBG scheme based OQPSK/DPSK modulation for chromatic dispersion compensation," Optical and Quantum Electronics, vol. 54, no. 7, p. 429, 2022.

[25] A. Onoufriou, K. Kalli, D. Pureur, and A. Mugnier, "Fibre Bragg Gratings," Springer Series in Optical Sciences, vol. 123, pp. 189–269, 2006.

[26] M. Mustafa, M. M. Abdelhalim, M. H. Aly, and T. M. Barakat, "Dispersion compensation analysis of optical fiber link using cascaded apodized FBGs hybrid with maximum time division multiplexing transmission technique," Optical and Quantum Electronics, vol. 53, no. 7, p. 358, 2021.

[27] A. F Sayed, F. M. Mustafa, A. A. M. Khalaf, and M. H. Aly, "Symmetrical and post-dispersion compensation in WDM optical communication systems," Opt Quant Electron, vol. 53, no.1, pp. 37–56, 2021.

[28] D. A. Egorova, A. V. Kulikov, A. N. Nikitenk, A. I. Gribaev, and S. V. Varzhel, "Investigation of bending effects in chirped FBGs array in multicore fiber," Optical and Quantum Electronics, vol. 52, no. 2, p. 130, 2020.

[29] N. A. Mohammed, M. Solaiman, and M. H. Aly, "Design and performance evaluation of a dispersion compensation unit using several chirping functions in a tanh apodized FBG and comparison with dispersion compensation fiber," Applied Optics, vol. 53, no. 29, pp. 239–247, 2014.

[30] T. F. Hussein, M. R. M. Rizk, and M. H. Aly, "A hybrid DCF/FBG scheme for dispersion compensation over a 300 km SMF," Optical and Quantum Electronics, vol. 51, no. 4, p. 103, 2019.

[31] D. Meena and M. L. Meena, "Design and Analysis of Novel Dispersion Compensating Model with Chirp Fiber Bragg Grating for Long-Haul Transmission System," Lecture Notes in Electrical Engineering Optical and Wireless Technologies, vol. 546, pp. 29–36, 2023.

[32] A. F. Sayed, F. M. Mustafa, A. A. M. Khalaf, and M. H. Aly, "An enhanced WDM optical communication system using a cascaded fiber Bragg grating," Optical and Quantum Electronics, vol. 52, no. 3, p. 181, 2020.

[33] A. F. Sayed, F. M. Mustafa, A. A.M. Khalaf and M. H. Aly, "Apodized chirped fiber Bragg grating for post-dispersion compensation in wavelength division multiplexing optical networks," International Journal of Communication System, vol. 33, no. 2, p. 4551, 2020.

Bio-Inspired Metaheuristic Framework for DNA Motif Discovery Using Hybrid Cluster Based Walrus Optimization

M. Shilpa, C. Nandini
Department of Computer Science and Engineering,
Dayananda Sagar Academy of Technological & Management, Bengaluru, Karnataka, India[1, 2]
Visvesvaraya Technological University, Belagavi, Karnataka, India[1, 2]

***Abstract*—Motifs are short, recurring sequence elements with biological significance within a set of nucleotide sequences. Motif discovery is the problem of finding these motifs. The problem of motif discovery has become an important problem in the field of Bioinformatics since, it finds its applications in Drug discovery, Environmental Health Research, and early Detection of Diseases by finding anomalies in gene sequences. Motif discovery is a challenging job in bioinformatics since it is NP-hard and cannot be solved within an exact time. In this study, we have proposed Hybrid Cluster based Walrus Optimization algorithm (HCWaOA) to solve the problem of motif discovery. The accuracy and efficiency of the proposed algorithm are improved using a hybrid approach. The population is initialized using Random Projection technique to generate a meaningful solution space. Then, k-means clustering is used to group similar solutions. Lastly, a population-based metaheuristic algorithm, Walrus optimization technique, is applied on each of the clusters to find the best motif. The proposed Hybrid Cluster-based Walrus Optimization algorithm (HCWaOA) is tested on both simulated and real biological datasets. The performance of HCWaOA is compared with benchmark algorithms like MEME, AlignCE and other meta-heuristics algorithms. The results of the proposed algorithm are found to be stable with a precision of 92%, a recall of 93% and an F-score of 93%. The proposed HCWaOA is tested using biological cancer-causing BARC and CTCF datasets to identify cancer causing motifs. Results show that incorporating clustering to initial solution space results in optimal solutions within a fewer iteration. The results of HCWaOA are compared with other popular motifs discovery algorithms and found to be stable.**

***Keywords*—Motifs; walrus optimization algorithm; meta-heuristic algorithms; k-means clustering; DNA; bioinformatics**

I. Introduction

The current Next Generation Sequencing (NGS) technology rapidly sequences the genomes to produce large volumes of high throughput data. The field of bioinformatics involves analyzing these high throughput data to find meaningful inferences. Analyzing these large volumes of high throughput data using traditional approaches is time consuming and prone to error. Regulation of gene expression is carried out using transcription and translation process. Before the transcription process, a specific region of gene is activated which is critical. The activation is caused by special proteins called Transcription factors which bind to specific region in gene called Transcription Factor Binding Sites (TFBS). Hence it is critical to find TFBS to understand the gene regulation process. These TFBS are known as motifs. Motifs are short, repeated, over represented patterns present in the regulatory regions of gene sequences, which play an important role in gene regulation mechanisms. These motifs are statistically significant, meaning they appear more frequently than expected. They are conserved among the sequences and have biological significance, often related to gene regulation. Therefore, Motif Discovery is one of the important research problems in bioinformatics. The problem of Motif Discovery is challenging because the exact positions of the motif in the gene sequences are not known. Due to mutations, the pattern of motifs across gene sequences can be the same or may vary slightly. Even though a number of motif discovery algorithms have been developed over a period, it is still a complex challenge for biologists and computer researchers.

Given a set of DNA sequences, the problem of motif discovery is to find a common pattern of length l that appears in all or most sequences with a maximum of d mutations. Different Computational approaches have been employed to solve the problem of motif discovery. Enumerative and Probabilistic approaches are the two important classifications of motif discovery algorithms. Enumerative approach involves an exhaustive search to find motifs, but it is impractical and time-consuming based on the nature of motifs. The Probabilistic approach uses Position Weight Matrix (PWM) to represent motifs. But the problem with the probabilistic algorithms is that they may converge to local optima. Other important approaches used in motif discovery are Machine Learning, Metaheuristics and Deep Learning techniques. Each of these algorithms has its own advantages and disadvantages.

Metaheuristic algorithms [1] are computational optimization algorithms that explore a solution space to find acceptable solutions for complex problems. Metaheuristic algorithms have drawn inspiration from nature, some from physics, chemistry, and mathematical concepts. Nature-based metaheuristic algorithms are inspired from natural processes to solve complex problems. Some of the popular and widely used metaheuristic algorithms are Genetic Algorithms based on biological evolution, Particle Swarm Optimization inspired from behavior of flock of birds, Ant Colony Optimization from ant colonies. Other nature-based metaheuristic algorithms are Grey Wolf Optimization, Cuckoo Search Optimization, FireFly Optimization, Walrus Optimization and so on. The advantages

of using nature-based metaheuristic algorithms for DNA motif discovery are the ability to handle large volumes of DNA sequence input data, provide good global exploration and escape the local optima. The limitations of existing algorithms are, they face scalability issues i.e. with the increase in the number of sequences and length of motifs the complexity of the algorithm increases. The tuning of algorithms parameters to derive at the optimal solution is one of the common problems encountered in metaheuristics algorithms.

In the proposed study, we use the Walrus Optimization algorithm to solve the problem of motif discovery. The advantage of using WaOA is, it requires minimal parameter setting with balanced exploration and exploitation. The key contributions of the proposed study are:

1) First, the Random projection technique is used for initializing the population solution space. It improves the diversity of the solution space while avoiding random unwanted population space. This acts as a starting point for providing promising solutions.

2) A k-means clustering technique is used for clustering the solution space.

3) Walrus optimization algorithm is used for motif discovery.

4) The proposed HCWaOA algorithm is evaluated using both simulated and biological datasets.

5) The proposed HCWaOA algorithm is tested to identify the cancer-causing BRCA1 and CTCF motifs in promoter regions of human gene sequences.

The study is organized as follows: The literature review is presented in Section II, discussing different and current approaches to motif discovery along with metaheuristics algorithms. Section III explains the proposed methodology of HCWaOA. Section IV gives the results, with analysis and evaluation of the proposed HCWaOA. Conclusion and future research directions of HCWaOA are given in Section V.

II. Literature Review

Multiple approaches have been proposed to solve the problem of Motif Discovery, each having its own advantages and disadvantages. Enumerative and Probabilistic approaches are the two major categories of motif discovery algorithms [2]. The other important approaches in motif discovery are Genetic algorithms, Meta-heuristics algorithms, Machine Learning and Deep Learning Techniques. In this section, we discuss the important and latest approaches used currently.

A. Enumerative Approach

The algorithms in Enumerative approach conduct an exhaustive search in entire search space to find the patterns. This covers the entire search space, thus exponentially increasing the time required as the size of the problem space increases. Hence it is suitable for finding motifs of short length in smaller sequences. Enumeration techniques are based on simple words, suffix trees, graphs, hashing. Some of the techniques that come under this category are:

- Brute-force Search: Examines all possible motifs and selects the best one based on a scoring function.
- Consensus-based Methods: Identify the most frequent k-mers (short substrings of length k) across sequences.
- Suffix Trees: Used to efficiently store and search for repeated substrings in a dataset.
- Word-based Methods: Search for overrepresented short words across sequences.

The algorithms developed using this approach are DREME, CisFinder, Weeder, FMotif, MCES.

B. Probabilistic Approach

Probabilistic algorithms compute a position weight matrix or position-specific weight matrix for representing motifs. Some of the important algorithms are Expectation Maximization (EM) and Gibbs sampling.

- Expectation-Maximization (EM) Algorithm: A probabilistic method that iteratively refines a motif model by updating motif positions in sequences.
- Gibbs Sampling: A Markov Chain Monte Carlo method that iteratively samples motif positions to improve alignment.

MEME, STREME [3], EXTREME are based on EM, and AlignCE is based on the Gibbs-Sampling approach.

Machine Learning, Artificial Intelligence and Deep Learning Approaches.

Current trends involve the use of Machine Learning [4], Artificial Intelligence and Deep Learning Approaches [5] in motif discovery. Hidden Markov Models (HMMs) are probabilistic models. Neural Networks and Deep Learning models are used to detect motifs. Deep Learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Hybrid CNN RNN, Ensemble-based models and complex hybrid models are applied to Motif Discovery. Supervised learning, like Random Forests and Support Vector Machines (SVMs), is used in settings when labelled motif data is available, whereas Autoencoders are used for unsupervised motif discovery by reducing dimensionality and extracting patterns. DeepBind, DeepVISP [6], and Deep6mA are some of the popular deep learning approaches used in the problem of motif discovery.

C. Nature Inspired Approach

Metaheuristic algorithms are approximate methods which provide acceptable solution within reasonable computational cost. Metaheuristic algorithms have drawn its inspiration from nature. The solution to the problem is based on natural selection and survival of the fittest. Some of the popular and widely used nature inspired metaheuristic algorithms are Genetic Algorithms based on biological evolution, Particle Swarm Optimization based on behavior of flock of birds [7], Ant Colony Optimization based on ant colonies. Other nature-based metaheuristic algorithms [8] are Grey Wolf Optimization, Cuckoo Search Optimization, FireFly Optimization, Walrus Optimization [9] and so on. In [10] and [11], Chemical Reaction Optimization (CRO) and Henry Gas Solubility optimization which are based on physics and chemistry are used. Hybrid metaheuristic algorithms have given good results compared to

traditional approaches. Recent trends in motif discovery algorithms are discussed below:

The study in [12] is a physics inspired metaheuristic for DNA motif discovery proposed in 2025. Here Archimedes Optimization Algorithm (AOA) mimics the principles of buoyant force and object equilibrium in fluids, balancing exploration (global search) and exploitation (local refinement) effectively. The Information Content is used for fitness evaluation. This technique is tested using real DNA dataset and the performance is evaluated against other metaheuristics algorithm. The major drawback is scalability issue to handle huge data and risk of convergence to local optima. The algorithm needs optimal tuning of control parameters to achieve optimal solutions.

In 2023, Dang et al [13] proposed a hybrid genetic algorithm to discover DNA motifs that satisfy the 2-Optimality postulate. 2-optimality postulate means the motif should be found in minimum of two input sequences. It uses evolutionary operations selection, crossover and mutation. It is tested against a benchmark dataset and provides better accuracy. The 2-optimality postulate may sometimes not discover relevant motifs.

In 2023, Theepalakshmi, P. and Reddy, U.S., [17] have proposed a novel effective quorum seeded (ℓ, d) motif search utilizing segmentation to filter using freezing firefly approaches on ChIP-seq data. The final motif was identified using the usual firefly approach, which uses both local and global freezing techniques. The effectiveness of these techniques was assessed using both simulated and actual datasets, including the human ChIP-seq dataset, the mouse emergent stem cell dataset, with Escherichia coli cyclical AMP receptor protein (CRP) dataset. It attains high F-measure and low accuracy.

In [14], the authors proposes a Freezing FireFly algorithm to solve the motif discovery problem. Here local and global freezing strategy is used wherein the best possible positions of the poor solution is also preserved. This helps in the overall search process to find new possible good results. The performance of Freezing Firefly algorithm is evaluated by comparing with benchmark tools like Samselect, TraverStringRef, PMS8, qPMS9, AlignACE, FMGA, and GSGA.

In 2022, Li, Song et al. [15] proposed an improved henry gas optimization algorithm with levy mechanism and brown motion. The study in [11], is Henry gas solubility optimization (MHGSO) algorithm for motif discovery. In MHGSO, the optimal solutions are obtained by evaluating the characteristics of the candidate solution space. It is based on chemistry, Henry's Law, treating the search space as gas molecules and mimics the finding of best motif to adjusting the solubility of gas molecules. The performance of MHGSO algorithm is evaluated on both synthetic and real datasets to find accurate motifs. The limitations of the algorithm is control parameters are specified which make the algorithm less flexible.

In 2024, Mohammad Hasan, et.al. [16] have proposed a Trie-PMS8. The algorithm is based on enhanced trie-tree for planted motif search problem. The problem with earlier PMS algorithm is that the time complexity increases exponentially in worst case scenarios. The proposed trie tree uses sort row by size step to reduce the time and linked lists to reduce the space. It also uses dynamic programming techniques to avoid redundant calculations in frequent tree processing. It gives better results than earlier versions of PMS with reduced time complexity.

In 2024, Qiang Yu, et.al. [18] have proposed an exact Planted Motif Search (PMS) on large DNA sequences. The efficient and exact algorithm finds (l, d) motifs using the technique of searching the branches on the pattern tree. The algorithm has good running time ratio compared to the existing PMS algorithms. The algorithm is tested on challenging problem instances of large DNA sequence datasets.

In 2024, Ledesma-Dominguez, et al. [19] have proposed a hybrid model named Deep Regulation (DeepReg) to identify transcription factor binding sites in prokaryotic and eukaryotic protein sequences. The hybrid model uses CNN, BiLSTM and attention mechanism. Feature extraction and grammar regulation is done using CNN and BiLSTM respectively. This leads to enhanced F1-score and performance in DeepReg compared to the other deep learning models. The model showed reliability and robustness for unseen experimental data. It provided low variance and eliminated overfitting problem. The model was tested on three organisms S. cerevisiae, N. crassa, and A. nidulans [19] and on average identified 71.8% of transcription factors.

The motif discovery problem is essential for understanding regulatory elements in DNA sequences. Various computational techniques, ranging from simple brute-force methods to advanced deep learning models, are used to identify biologically significant motifs. Despite the progress, motif discovery continues to be a challenging problem in genomics and bioinformatics research. The major issues with the existing metaheuristics algorithm are:

- Results are dependent on algorithms parameters tuning like the mutation rate, co-efficient, optimal size of population and iterations.
- Scalability issues as the number of sequences, length of sequences and motif length increases.

The aim of the research is to address the limitations of the existing metaheuristics algorithm and propose a hybrid metaheuristics algorithm to address the above issues.

III. Proposed Method

In this study, a Hybrid Cluster based Walrus Optimization algorithm (HCWaOA) is proposed for finding motifs in the promoter regions of a given set of DNA sequences. The first step of the proposed algorithm, builds an initial population solution using random projection strategy [20]. The second step involves clustering the population solutions using k-means clustering. In the third step, modified Walrus optimization algorithm is applied on the clusters to find the best solution. Here the best solution is the motif with highest fitness value found in the given set of DNA sequences. The architecture block diagram of the HCWaOA is given in Fig. 1.

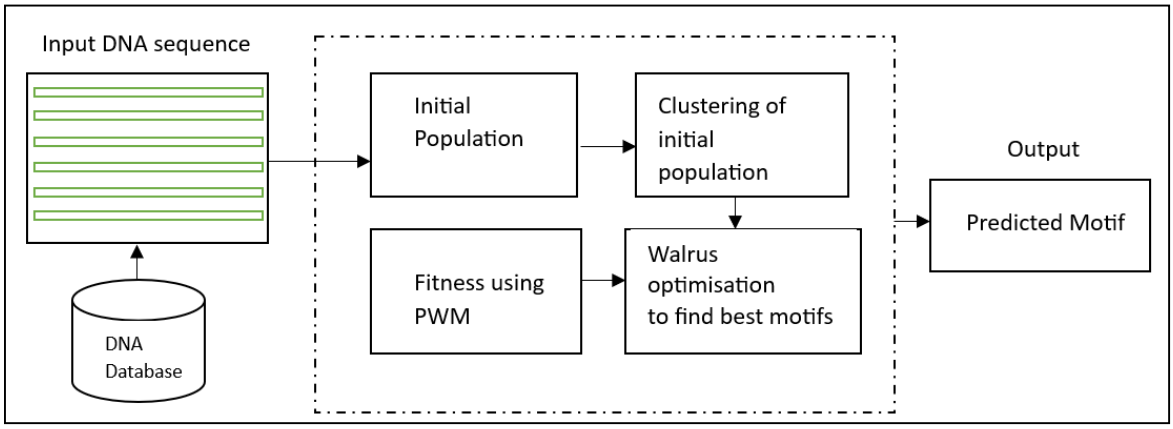


Fig. 1. Block diagram of HCWaOA.

A. Walrus Optimisation Algorithm

Walrus Optimization Algorithm (WaOA) is a new nature-based metaheuristic algorithm that is based on the behavior of walruses. It draws inspiration from Walrus behavior of feeding, migrating, escaping predators and fighting predators [21]. The Walrus Optimization Algorithm is implemented in three phases: exploration, migration, and exploitation. The working of the algorithm is given by the following steps:

1) Initialization: A population of walrus solutions is randomly generated within the search space, representing potential solutions to the optimization problem.

2) Movement based on behavior:

Phase 1: Feeding (Exploitation): When a safety signal [22] is identified, the walruses feed by refining their positions, moving closer to the optimal solution. The new position is calculated using Eq. (1):

$$X_i^{t+1} = X_i^t + r\,(S^t - I.X_i^t) \quad (1)$$

X_i^t : Position of the i^{th} walrus at iteration t

S^t : Best fitness solution

I : control parameter

r : random vector [0,1]

Phase 2: Migration: Walruses are simulated to migrate towards areas with better potential solutions, based on the random solution. The new position for migration is calculated using Eq. (2):

$$X_i^{t+1} = X_i^t + r\,(X_k^t - I.X_i^t) \quad (2)$$

X_k^t : Position of randomly selected walrus

Phase 3: Escape (Exploration): When a danger signal [22] is triggered, the walruses escape by making large random movements to explore new regions of the search space. The new position for exploitation is calculated using Eq. (3):

$$X_i^{t+1} = X_i^t + r.\delta \quad (3)$$

$$\delta = \alpha\,.(ub - lb)$$

$$\alpha : random\ value\ [-0.1, 0.1]$$

$$ub, lb: upper\ and\ lower\ bounds\ of\ the\ search\ space$$

3) Updating positions: The position of each walrus (candidate solution) is updated based on the calculated danger signals and safety signals, influencing how much exploration or exploitation is performed in each iteration.

4) Iteration and convergence: The process of movement and position updates is repeated for a set number of iterations, with the best solutions gradually converging towards the optimal solution.

B. Random Projection Technique

The starting point of the algorithm is the generation of initial population using on random projection technique. This involves selecting random sub sequences of motif length from the given set of input DNA sequences. These sub sequences act as the initial population i.e. candidate solutions. This gives a set of candidate solutions which are closer to the optimal solution since they are already a part of the solution space. This strategy is better than generating random candidate solution which are not related to the solution space. The number of iterations and time required to reach the optimal solution is reduced and a meaningful initial candidate solution space is generated as the input to walrus algorithm.

C. Clustering

A k-means clustering is used to cluster initial candidate solutions. Clustering allows us to group highly similar motifs. This allows for easier analyzing of pattern, reducing noise and identifying biologically significant motifs. The problem with initial candidate solutions of metaheuristic algorithms is they are random and no relation exists between them. It is found that individual elements of candidate solution which are similar or closer have higher fitness then dissimilar elements.

Clustering techniques [23] like k-means help group similar motifs which may have variations due to mutations, insertions or deletions together, making the analysis more structured and meaningful. The partitioning of the population into multiple clusters allows only intra cluster operations like mutation, crossover within the cluster whereas multiple clusters allow for diversity in the population. Multiple clusters allow us identify multiple weaker motifs which are similar. Hence clustering efficiently organize motifs and improves the motif discovery

process. K-means is used in the proposed study since it is computationally efficient and can handle large-scale motif datasets.

D. Fitness Function

The fitness or the quality of the motifs is evaluated based on the information content (IC). IC is calculated using the following steps:

1) Compute PFM (Position Frequency Matrix): A Position Frequency Matrix (PFM) represents how frequently each nucleotide i.e. A, C, G, T appears at each position in a set of aligned motifs.

2) Computer PWM (Position weight matrix): The PWM is obtained by normalizing the PFM with background nucleotide probabilities and taking the logarithm, as in Eq. (4):

$$PWM_{b,j} = log_2\left(\frac{PFM_{b,j}}{P_{background}(b)}\right) \quad (4)$$

Here, $PWM_{b,j}$ is the weight for nucleotide b at position j, and $P_{background}(b)$ is the background probability of nucleotide b. This is often assumed to be 0.25 for uniform distribution.

This transformation helps identify conserved positions in the motif.

3) Computer IC: quantifies how well a motif is conserved. It measures how different each position is from a random sequence.

Eq. (5):

$$IC_j = \sum_b P_{b,j}\, log_2\left(P_{b,j}/P_{background}(b)\right) \quad (5)$$

Total IC is the sum across all motif positions, as in Eq. (6):

$$IC = \sum_{j=1}^{L} IC_j \quad (6)$$

A higher IC means the motif is more conserved and biologically significant. The calculated IC value serves as the fitness score.

E. Process of Walrus Optimisation Algorithm for Motif Discovery

The Walrus Algorithm efficiently discovers motifs by combining random exploration with local exploitation. Its adaptive approach ensures a balance between global search (exploration) and local refinement (exploitation), making it well-suited for motif discovery in DNA sequences. Walrus Optimization technique operates on each of the clusters. Thus, returning the best solution in each of the clusters. The results are evaluated to obtain global best solution based on the fitness. The global best solution is the motif with fitness score having highest IC value. Algorithm 1 presents the pseudocode of the proposed HCWaOA algorithm.

Algorithm 1: Pseudocode of the proposed HCWaOA algorithm

```
1. Set all the parameter values
2. Initialise the population
3. Calculate the fitness value using the fitness function to get the best solution
4. while (t < max_generations)
5.     Compute adaptive exploration and exploitation rates
6.     Update Motif Positions
7.         For each motif set in population:
               Phase 1: Feeding (Exploitation)
                   Calculate the new position using Eq1
                   Update the position of new walrus
               Phase 2: Migration
                   Using a random walrus, find the position of new walrus using Eq2
                   Update the position of new walrus
               Phase 3: Escape (Exploration):
                   Calculate the new position using Eq3
                   Update the position of new walrus
8. Compute new fitness scores using Eq4, Eq5, Eq6
9. Update the best solution with highest fitness score
10. Return the best motif with the corresponding fitness score
```

IV. Experimental Results and Discussion

HCWaOA algorithm is implemented in Python using Anaconda IDE under Windows OS and tested on both real-time and simulated dataset. The test results are analyzed and the performance is compared with existing state-of-art tools such as MEME, DREME and MHGSO [11] systems. The proposed algorithm is tested to identify breast and ovarian cancer causing motifs BRCA1 and CTCF in human gene sequences.

A. Performance Measure

The performance of the HCWaOA is evaluated using the metrics Precision (P), Recall (R) and F-score [24]. Precision P is number of predicted motifs that are true divided by number of predicted motifs. Recall R is number of predicted motif sites that are true divided by number of true motifs. F-score is computed using the values of Precision and Recall. The best value for F-score is 1 and worst is 0. TP is true positive, TN is true negative, FP is false positive, FN is false negative.

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

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

$$F - score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)}$$

B. Test on Simulated Data

The algorithm is tested using 7 groups synthetic dataset. The details of the dataset are given in Table I. The sequences are generated randomly and motifs with mutations are added at random positions using python script.

The proposed algorithm is tested on the synthetic dataset given in Table I. The detailed results for each of the 7 dataset groups is given in Table II. The algorithm is run for 10 times and the best value of P and R is taken into consideration. Table II

gives the average values of P, R and F -score for 7 dataset groups. The results show a higher F-score attained through higher values of P and R. Thus, results show a higher F-score for dataset (8, 2) and (9,3). We can infer that as the length of the motif increases with the increase in the number of mutations. It is difficult to identify all the correct motifs since higher R value means large number of candidate instances which cause false positive. Prediction results can be sensitive to the dataset properties like the number and length of sequences generated, true and mutated motifs in the input sequence.

TABLE I. SYNTHETIC DATASET DETAILS

Dataset group	1	2	3	4	5	6	7
No of sequences (t)	40	40	40	40	40	40	40
Length of sequence (n)	400	400	400	400	400	400	400
Length of motif (l)	8	9	10	14	16	18	21
Maximum mutations	2	3	3	4	5	6	7

TABLE II. RESULTS FOR SYNTHETIC DATASET ON PROPOSED ALGORITHM: P IS PRECISION, R IS RECALL AND F IS F-SCORE

SI NO	Dataset	P	R	F - score
1	(8,2)	0.95	0.95	0.95
2	(9,3)	0.95	0.95	0.95
3	(10,3)	0.90	0.92	0.91
4	(14,4)	0.93	0.93	0.93
5	(16,5)	0.93	0.93	0.93
6	(18,6)	0.90	0.92	0.91
7	(21,7)	0.90	0.92	0.91
	Avg	0.92	0.93	0.93

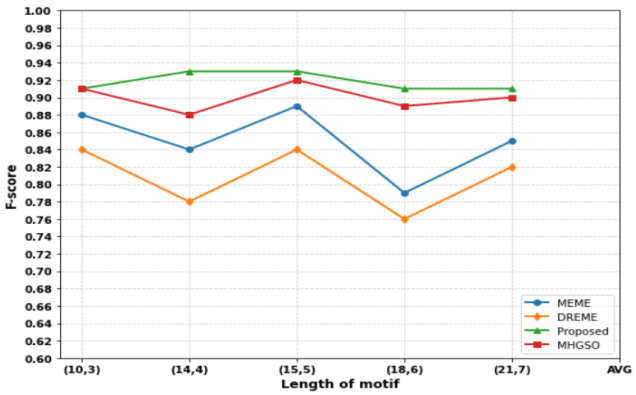


Fig. 2. Analysis of F-score results on synthetic dataset.

The synthetic datasets (10, 3), (14,4), (15,5), (18,6) and (21,7) are familiar examples in motif discovery problem. Fig. 2 shows the F-score on synthetic dataset. The results of HCWaOA are compared with the results of state-of-art motif discovery tools MEME, DREME and metaheuristic algorithm MHGSO [11]. The proposed algorithm provides an improved F-score compared to the other algorithms. The clustering approach allows for grouping of similar solution space and walrus optimization exploits the solution space to get better results. The hybrid approach of HCWaOA allows a thorough exploration of the search space. This thorough exploration reduces the chances of missing optimal solutions and increases the overall accuracy of the optimization task.

C. Convergence of Proposed HCWaOA

The proposed HCWaOA algorithm detects the convergence based on the fitness of the motifs. The fitness, i.e. IC is calculated using Eq. (6) across fixed iterations. If the fitness value reaches the highest and remains the same for fixed number of iterations then the algorithm is terminated. This prevents the overhead of computation once convergence is reached. Fig. 3 shows the convergence curve for dataset group 1, 2 and 3 of Table I.

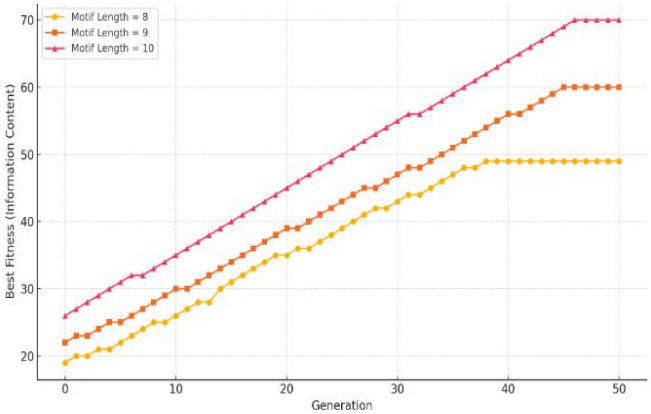


Fig. 3. Convergence curve for motifs of length 8, 9 and 10.

D. Analysis of k-means Clustering on Walrus Optimisation

The proposed Walrus optimization algorithm was tested with and without k-means clustering on the initial population for simulated dataset given in Table I. The fitness was calculated using the Eq. (6). The average fitness value was calculated across 15, 25, 35 and 50 generations. The results in Table III show an improvement in average fitness value when clustering is employed. It was also observed that new weaker motifs are identified in the clusters which was not a part of the simulated dataset. Fig. 4 infers; clustering discovers better motifs i.e. motifs with better fitness values within fewer iterations. Hence by incorporating clustering on initial population, the proposed HCWaOA allows for discovering of motifs with higher fitness values within given iterations. This overcomes the complexity issue of metaheuristics algorithms.

TABLE III. PERFORMANCE COMPARISON OF HCWAOA WITH CLUSTERING AND WITHOUT CLUSTERING

SI NO	Dataset	Without clustering	With Clustering
		Average Fitness	Average Fitness
1	(8,2)	51.14	53.18
2	(9,3)	53.46	57.95
3	(10,3)	58.78	62.29
4	(14,4)	82.9	98.29
5	(16,5)	92.6	101.74
6	(18,6)	100.49	126.85
7	(21,7)	110.39	146.17

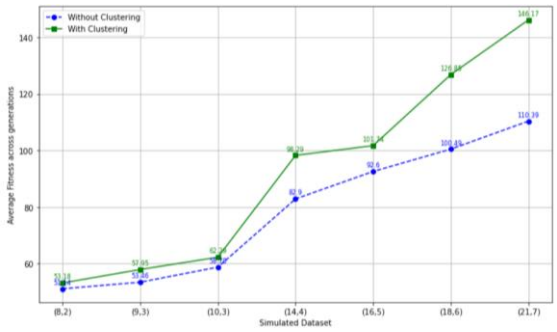


Fig. 4. Performance comparison of HCWaOA with and without clustering.

E. Test on Real-Time Data

The proposed algorithm is tested using a CRP benchmark dataset. The benchmark dataset is a real-time dataset consisting of 18 sequences of Escherichia coli of 105 bp-long. The proposed HCWaOA algorithm finds the motifs and their starting positions for the CRP E. coli input sequence. Table IV shows the results of the proposed algorithm in comparison with MEME and AlignCE. Table IV gives the actual starting position of the known motif, the predicted motif start position from different algorithms and the deviations between actual and predicted results. The proposed algorithm gives most of the positions correctly, and the results are better than existing algorithms MEME and AlignCE.

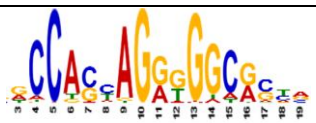
F. Test on Breast Cancer Gene Input Dataset

The proposed algorithm is tested to check its ability to discover real biological data in human DNA sequence. The input gene sequences are collected from well curated JASPAR database and government authorized website NCBI. These are breast cancer gene input sequences containing BRCA1 [25] and CTCF motifs. These BRCA1 and CTCF motifs are associated with breast and ovarian cancer. The identification of BRCA1 and CTCF motifs in the input data sequence infers the detection of cancer. Table V gives the details of input sequences along with their accession numbers as well as true motifs and predicted motifs. The proposed algorithm was run on the cancer dataset, and the experimental results show that the predicted motifs match with known cancerous motifs.

TABLE IV. COMPARISON RESULTS OF HCWAOA, MEME, ALIGNCE FOR CRP BENCHMARK DATASET

Sequence No	Starting position of the known motif	MEME	AlignCE	Proposed algorithm (HCWaOA)
1	17,61	61	63(2)	61
2	17,55	55	57(2)	55
3	76	76	78(2)	76
4	63	63	65(2)	63
5	50	13 (-37)	52(2)	50
6	7,60	7	9(2)	7
7	42	42	26(-16)	24(-14)
8	39	39	41(2)	39
9	9,80	9	11(2)	9
10	14	14	16(2)	14
11	61	35	63(2)	61
12	41	34	43(2)	51(10)
13	48	48	50(2)	48
14	71	71	73(2)	71
15	17	75 (58)	19(2)	17
16	53	6	55(2)	53
17	1,84	27 (26)	68(16)	5(4)
18	78	76 (-2)	80(2)	78

TABLE V. LIST OF CANCER CAUSING BRAC1 AND CTCF GENE INPUT DATASET WITH PUBLISHED AND PREDICTED MOTIFS

Name	Seq No	Accession no.	Published Motif using MEME & TOMTOM	Predicted Motif from proposed HCWaOA algorithm
BRCA1	15 sequences	AF507075.1 AY093484.1 AF507076.1 AF507077.1 AF507078.1 AY093484.1 AY093486.1 AY093487.1 AY093488.1 AY093489.1 AY093492.1 AY093493.1 AY093490.1 AY093491.1 AF284812.1	 Name : SP5(C2H2 zinc finger factors) Matrix ID: MA1965.1	
CTCF	30 sequences	JASPAR	 Name : CTCF Matrix ID: MA0139.1	

The predicted motifs identified by the proposed algorithm is validated by checking the similarity with annotated motifs. Motif similarity analysis was evaluated using TomTom from the MEME tool-kit. It was also found that the discovered motif had clear resemblance with known motif in JASPAR database. The predicted motifs logo was created using the tool https://weblogo.berkeley.edu/logo.cgi.

V. Conclusion

The study proposes, hybrid cluster-based motif discovery approach using a new metaheuristic walrus optimization algorithm. Using the proposed technique, adjusting the algorithm parameters is considerably reduced compared to the other metaheuristic algorithms and this in turn efficiently discovers the motifs in gene sequences. The performance of the proposed HCWaOA algorithm is evaluated using both synthetic and real dataset. The results on synthetic and real dataset are better compared to existing well-known traditional methods. Experimental results have shown that the proposed algorithm can discover motif of both short and long length sequences with mutations. By employing clustering, better motifs can be identified within fewer iterations thereby reducing the complexity of the algorithm. Here the real time application of the proposed algorithm is verified using cancer gene sequences. The proposed HCWaOA algorithm is successfully applied for the detection of CTCF and BRCA1 binding sites in homosapiens, whose deletion or inactivation has been detected in various cancers.

Although, the proposed algorithm gives good results, it can be extended to handle motifs of unknown length, thus providing more flexibility in discovering motifs of any length. In future a better clustering technique can applied which utilizes the properties of motifs. Even though multimodal metaheuristics algorithm solves some of the issues with motif discovery, deep learning techniques that are robust and scales well to handle large gene dataset can be explored in future.

References

[1] Almufti, Saman & Ali, Rasan & Fuente, Jayson. (2023). Overview of Metaheuristic Algorithms. Polaris Global Journal of Scholarly Research and Trends. 2. 10-32. 10.58429/pgjsrt.v2n2a144.

[2] Shilpa, M., and C. Nandini. "A Survey on Motif Discovery Algorithms for analysis of Gene Sequences of Interest." In 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), vol. 1, pp. 275-278. IEEE, 2021.

[3] Timothy L Bailey, STREME: accurate and versatile sequence motif discovery, Bioinformatics ,Volume 37, Issue 18, September 2021, Pages 2834–2840 , https: //doi.org/10.1093/bioinformatics /btab203

[4] A. Yang, W. Zhang and J. Wang, "Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA", Frontiers in Bioengineering and Biotechnology, Vol. 8, No. 2, September, pp. 1– 13, 2020, DOI: 10.3389/fbioe.2020.01032.

[5] Quang D,XieX. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 2016;44:e107–7

[6] Xu, Haodong & Jia, Peilin & Zhao, Zhongming. (2021). DeepVISP: Deep Learning for Virus Site Integration Prediction and Motif Discovery. Advanced Science. 8. 2004958. 10.1002/advs.202004958.

[7] Uyyala Srinivasulu Reddy, Michael Arock, and A.V. Reddy. 2020. Discovering of gapped motifs using particle swarm optimisation. Int. J. Comput. Intell. Bioinformatics Syst. Biol. 2, 1 (2020), 1–21. https://doi.org/10.1504/ijcibsb.2020.106858

[8] Vasuki, A. (2020). Nature-Inspired Optimization Algorithms. 10.1201/9780429289071.

[9] Trojovský, P., Dehghani, M. A new bio-inspired metaheuristic algorithm for solving optimization problems based on walruses behavior. Sci Rep 13, 8775 (2023). https://doi.org/10.1038/s41598-023-35863-5

[10] Saha, Sumit & Islam, Md & Hasan, Mredul. (2021). DNA motif discovery using chemical reaction optimization. Evolutionary Intelligence. 14. 10.1007/s12065-020-00444-2.

[11] Fatma A. Hashim, Essam H. Houssein, Kashif Hussain, Mai S. Mabrouk, and Walid Al-Atabany. 2020. A modified Henry gas solubility optimization for solving motif discovery problem. Neural Comput. Appl. 32, 14 (Jul 2020), 10759–10771. https://doi.org/10.1007/s00521-019-04611-0

[12] Hashim, F.A. *et al.* (2025). Archimedes Optimization Algorithm for DNA Motif Discovery. In: Dulhare, U.N., Houssein, E.H. (eds) Deep Learning and Computer Vision: Models and Biomedical Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-96-1285-7_1

[13] Dang, D.T., Nguyen, N.T. & Hwang, D. Hybrid genetic algorithms for the determination of DNA motifs to satisfy postulate 2-Optimality. Appl Intell 53, 8644–8653 (2023). https://doi.org/10.1007/s10489-022-03491-7

[14] Theepalakshmi, P & Reddy, U. Srinivasulu. (2022). Freezing firefly algorithm for efficient planted (ℓ, d) motif search. Medical & biological engineering & computing. 60. 10.1007/s11517-021-02468-x.

[15] Li, Song et al. "An improved Henry gas solubility optimization algorithm based on Lévy flight and Brown motion." Applied Intelligence 52 (2022): 12584 - 12608.

[16] Mohammad Hasan, Abu Saleh Musa Miah, Md. Humaun Kabir, Mahmudul Alam, Trie-PMS8: A trie-tree based robust solution for planted motif search problem, International Journal of Cognitive Computing in Engineering, Volume 5, 2024, Pages 332-342, ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2024.07.004

[17] P. Theepalakshmi, U. Srinivasulu Reddy, A new efficient quorum planted (ℓ, d) motif search on ChIP-seq dataset using segmentation to filtration and freezing firefly algorithms, Soft Computing, 10.1007/s00500-023-09236-z, 28, 4, (3049-3070), (2023).

[18] Yu, Qiang, et al. "An efficient exact algorithm for planted motif search on large DNA sequence datasets." IEEE/ACM Transactions on Computational Biology and Bioinformatics (2024).

[19] Ledesma-Dominguez, L., Carbajal-Degante, E., Moreno-Hagelsieb, G. et al. DeepReg: a deep learning hybrid model for predicting transcription factors in eukaryotic and prokaryotic genomes. Sci Rep 14, 9155 (2024). https://doi.org/10.1038/s41598-024-59487-5

[20] Ge, Hongwei & Yu, Jinghong & Sun, Liang & Wang, Zhen & Yao, Yao. (2019). Discovery of DNA Motif Utilising an Integrated Strategy Based on Random Projection and Particle Swarm Optimization. Mathematical Problems in Engineering. 2019. 1-12. 10.1155/2019/3854646.

[21] Muxuan Han, Zunfeng Du, Kum Fai Yuen, Haitao Zhu, Yancang Li, Qiuyu Yuan, Walrus optimizer: A novel nature-inspired metaheuristic algorithm, Expert Systems with Applications, Volume 239, 2024, 122413, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.122413

[22] P. Trojovský and M. Dehghani, "Walrus Optimization Algorithm (WaOA), MATLAB Central File Exchange," 2023. https://doi.org/10.21203/rs.3.rs-2174098/v1

[23] F. B. Ashraf, A. Matin, M. S. R. Shafi and M. U. Islam, "An Improved K-means Clustering Algorithm for Multi-dimensional Multi-cluster data Using Meta-heuristics," 2021 24th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/ICCIT54785.2021.9689836

[24] Mabrouk, Mai & Abdelhalim, mohamed b & Elewa, Ebtehal. (2018). A developed system based on nature-inspired algorithms for DNA motif finding process. Neural Computing and Applications. 30. 10.1007/s00521-018-3398-0.

[25] Bhargavi, Peyakunta & Lakshmi, Kanchi & Jyothi, Singaraju. (2020). Gene Sequence Analysis of Breast Cancer Using Genetic Algorithm. 10.1007/978-981-15-0135-7_16.

Detection of Autism Spectrum Disorder (ASD) Using Lightweight Ensemble CNN Based on Facial Images for Improved Diagnostic Accuracy

Andi Kurniawan Nugroho[1], Jajang Edi Priyanto[2], D. S. P. Vinski[3]
Electrical Department, Universitas Semarang, Semarang, Indonesia[1]
Public Health Department, Institut Kesehatan Indonesia, Jakarta, Indonesia[2]
Celtech Stem Cell Centre Laboratory and Banking Jakarta, Indonesia[3]

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects how people talk to each other and act. The fact that ASD is becoming more common and that diagnosing it can be difficult means that early detection is important for improving treatment outcomes. This study's goal is to use lightweight ensemble Convolutional Neural Networks (CNN) to make it easier to classify ASD from facial photos. The study looks at different CNN architectures, like MobileNetV2 and EfficientNet variations, to find the best model for diagnosing ASD quickly and accurately. The method involves training and testing five lightweight CNN models on a set of facial photos. We use pre-processing methods like scaling and data augmentation to help the model learn better. The study tests how well ensemble CNN models work by combining predictions from different architectures using averaging and voting methods. We use important performance metrics like accuracy, precision, recall, and F1-score to see how well each model works. The results show that the best balance between accuracy and computational efficiency is achieved by combining MobileNetV2 and EfficientNetB0. This combination achieves an accuracy of 0.8299, a precision of 0.8514, a recall of 0.8182, and an F1_score of 0.8344. Other models, like ResNet50 combined with EfficientNetB0, have higher precision but lower recall, making them less useful for finding all ASD cases. This study was also compared with other researchers, and the proposed study was found to have greater accuracy than other researchers. The results show that ensemble CNN models can significantly improve the accuracy of classifying ASD compared to single CNNs. This study shows that lightweight ensemble CNN models are good at finding ASD in pictures of people's faces. The method is fast and can be used on devices with limited processing power, making it a good way to find ASD early in both clinical and real-world settings.**

Keywords*—Component; autism spectrum disorder (ASD); early detection; ensemble convolutional neural network (CNN); facial images; classification; accuracy

I. Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder that affects how a person talks, interacts with others, and thinks and acts. People with Autism Spectrum Disorder (ASD) often have trouble interacting with others, have trouble learning new words, and have rigid habits that they do over and over. Many conditions that used to be called classic autism, Asperger's syndrome, and undefined pervasive developmental disability are now part of ASD [1]. This spectrum shows how different people can have very different symptoms and how bad they can be. It is thought that both genetic and environmental factors play a role in the causes of ASD, but these factors are not fully understood. Recent studies have confirmed the idea that a combination of genetic and prenatal factors has a big effect on the development of ASD. The number of people with ASD around the world has gone up a lot in the last ten years.

Center for Disease Control and Prevention (CDC) in the United States says that the number of children with Autism Spectrum Disorder (ASD) went from 1 in 150 in the early 2000s to 1 in 36 in 2023 [2]. There are many reasons for this increase, including changes in how autism is diagnosed, more people and healthcare professionals being aware of it, and better detection of autism across a wider range of ages. There aren't many statistics available in Indonesia, but reports from the Ministry of Health of the Republic of Indonesia and a few local studies show that the number of people with ASD is rising at a similar rate. Studies show that 2 to 5 out of every 10,000 children have Autism Spectrum Disorder (ASD). This trend is growing because more people in Indonesia know about ASD, and diagnostic tools are getting better [3]. Sadly, it is still hard to get a diagnosis and early help for ASD in Indonesia, especially in remote areas. This makes it harder to treat and support people with ASD.

A multidisciplinary approach is often used in ASD screening and diagnosis. This includes clinical evaluations based on DSM-5 criteria and a range of screening tools designed to find ASD symptoms in both children and adults. The Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) are the two most common tests used to diagnose Autism Spectrum Disorder (ASD). The ADOS is a series of exercises that are meant to test a child's social skills, communication skills, and behavior. The ADI-R is a semi-structured interview with parents or caregivers to find out how a child is doing with their early developmental milestones. In addition to these observational assessment methods, questionnaire-based screening tools like the Modified Checklist for Autism in Toddlers (M-CHAT) are used to find ASD in young children. These steps make it easier to find problems early, but they require specialized knowledge and experience to give an accurate and appropriate assessment.

As technology has improved, research on using artificial intelligence (AI) methods to classify ASD has made a lot of

progress. Artificial intelligence has a lot of potential to help with diagnosing ASD by looking at different types of data, such as neuroimaging, audio recordings, behavioral patterns, and genetic information. Support Vector Machines (SVM), Random Forest, Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN) are some of the algorithms that have been used to make models for classifying ASD that are more accurate and useful. Here are eleven papers that came before this one that used AI methods to classify ASD.

K. Vakadkar et al. (2021) used a data-driven method with Decision Tree and Naive Bayes algorithms to sort children with Autism Spectrum Disorder. The study showed that AI models can be very accurate at finding people with ASD [4]. Duda et al. (2016) made an AI-based diagnostic system that uses motion sensor data to find signs of ASD in kids. The SVM algorithm that was used worked very well, with an accuracy of almost 90% [5]. Heinsfeld et al. (2018) used deep learning and transfer learning to sort Autism Spectrum Disorder using fMRI data. This study shows that AI can tell the difference between different brain connection patterns in people with ASD [6].

Bone et al. (2017) made a Machine Learning model that uses voice recordings and speech patterns to sort people with Autism Spectrum Disorder (ASD). This study shows that AI has a lot of potential for analyzing non-verbal data [7]. R. Thapa et al. (2023) looked at how likely a child was to develop ASD at an early stage, which made the diagnosis process easier. To predict ASD traits, it uses several machine learning models, such as Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC). The study found that Logistic Regression was the best model for the dataset. However, machine learning models could do a better job of finding ASD if they were trained on bigger datasets and used deep learning methods like Convolutional Neural Networks (CNNs). The authors stressed how important it is to find problems early and how machine learning can help speed up the diagnostic process [8].

Using gene interaction network data, D. Bandara [9] made a Graph Neural Networks (GNN) model to figure out how likely it is that a gene is linked to Autism Spectrum Disorder (ASD). The author used different types of GNNs as starting points for three classification tasks: binary, multi-class, and syndromic risk classification. These types of GNNs included Graph Sage, Graph Convolutional Networks (GCN), and Multi-Layer Perceptron (MLP). The experimental results show that the Sage Graph model consistently outperforms the other models with the highest accuracy for all three classification tasks: 85.80% accuracy for binary risk classification, 81.68% accuracy for multi-class risk classification, 90.22% accuracy for syndromic gene classification.

J. W. Song et.al concluded that neuroimaging studies in neurodevelopmental disorders, such as autism spectrum disorder (ASD) and Attention-Deficit or Hyperactivity Disorder (ADHD), have uncovered various biomarkers. However, limitations in comprehensively analyzing such biomarkers limit clinically significant results. The use of deep learning (DL) offers a new approach to improve diagnosis accuracy and assist clinicians in determining more appropriate treatment. DL enables earlier and more accurate detection of biomarkers by analyzing complex patterns of neuroimaging data [10].

A. Ashraf et al. used deep learning methods, specifically an optimized Convolutional Neural Network (CNN), providing better results in the classification of individuals with Autism Spectrum Disorder (ASD) than previous conventional methods. Transfer learning demonstrated efficacy, achieving an accuracy of 82.31% on the ABIDE-I and ABIDE-II datasets. This study looks at how fMRI data can help find Autism Spectrum Disorder (ASD) early on so that kids with autism can have a better environment. The results showed better performance, but the authors stressed how important data augmentation is for improving accuracy even more [11].

R. A. Bahathiq et al. R. A. Bahathiq et al. have used machine learning (ML) and deep learning (DL) to diagnose autism spectrum disorder (ASD) using structural magnetic resonance imaging (sMRI) data. This is a promising approach, but it faces a lot of challenges. There are still a number of problems with ML/DL research that make it hard to find biomarkers for ASD and improve diagnostic accuracy. These include small sample sizes, differences in brain biomarkers, and problems with generalizing results. This study stresses how important it is to have bigger datasets and more rigorous methods to deal with these problems [12].

Y. Lin and others have shown that machine learning methods can accurately predict risk genes linked to Autism Spectrum Disorder (ASD). Using spatiotemporal gene expression patterns in the human brain, gene-level constraint metrics, and other factors of gene variation, researchers found new genes that increase the risk of ASD and confirmed genes that were already known to do so. The study showed that the genes that were expected to be found are linked to important biological processes like neural signaling, neurogenesis, chromatin remodeling, and protein ubiquitination.. Although the results are encouraging, the study acknowledges limitations, including the necessity for additional confirmation via independent replication studies [13].

AI has a lot of promise for helping to diagnose and classify ASD, but there are still other issues and shortcomings that need to be fixed. A primary difficulty is the quality and diversity of data utilized for training AI algorithms. Inadequate or unrepresentative datasets might result in biased models that lack generalizability across diverse populations. Furthermore, the acquisition of high-quality data, particularly pertaining to brain imaging and genetic information, is frequently expensive.

The restricted interpretability of AI models is a significant concern in medical contexts, particularly Autism Spectrum Disorder (ASD). Most Deep Learning algorithms operate as “black boxes”, rendering their internal mechanics challenging for medical professionals and researchers to comprehend. This results in challenges in evaluating outcomes and making clinical decisions based on AI model outputs. Transfer learning presents a viable option to address some limitations of conventional AI methodologies. The Lightweight Ensemble CNN for ASD classification using facial photos combines fast processing with strong predictive power. The use of lightweight CNN architectures like MobileNetV2 and EfficientNetB0 makes it possible for the system to work well on devices with limited

computing power while still being very accurate, thanks to the ensemble method. Things that helped with this study are:

1) The strategy improves the overall accuracy of the classification by using an ensemble approach that combines predictions from several lightweight CNN models, such as MobileNetV2 and EfficientNetB0, instead of just one model. This is important for classifying ASD because the details of facial features can be very different and subtle.

2) Using a lightweight CNN architecture makes sure that the models can do their jobs quickly. This makes training and inference faster, which means that these models can be used on devices with limited processing power, like mobile phones or edge devices used in clinical settings.

II. METHODOLOGY

A. Design CNN

The dataset of face images for ASD was obtained from Kaggle and data from an autism center in Semarang, with 1.468 labeled as ASD and 1.468 labeled as normal. The dataset was divided into training and testing data with an 80:20 ratio. The block diagram and flowchart of the proposed research are shown in Fig. 1 and Fig. 2. The dataset has pictures of people's faces that are either classified as having Autism Spectrum Disorder (ASD) or not. This information could come from a variety of places, such as clinics, past research, or relevant public statistics. Before training, the facial images will go through a number of steps to prepare them, such as resizing, normalizing, and adding more data through techniques like rotation, flipping, and adjusting the contrast.

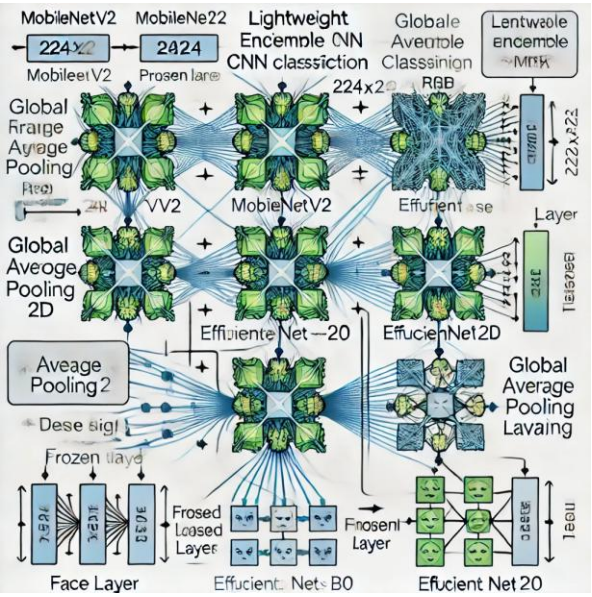


Fig. 1. Block diagram of proposed research.

The goal of these steps is to improve the quality of the dataset and make the image selection more varied, which will help the model learn better. We chose lightweight CNN designs like MobileNetV2 and EfficientNetB0 because they are fast and have small models. They use depthwise separable convolutions and compound scaling methods to make inferences quickly while keeping accuracy high. When you put a picture of a face into the model, the CNN network will extract features. This method uses a number of convolutional layers to find important patterns in the image, such as the shape of the face, the expression on it, and the texture of the skin. The result of this extraction is a number that shows what the face looks like.

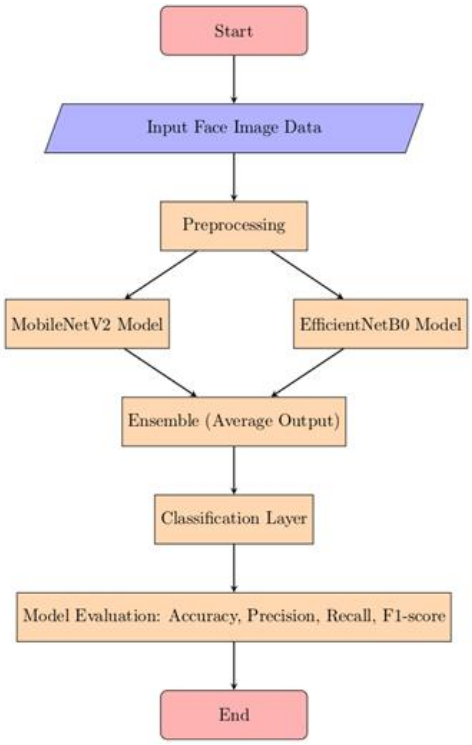


Fig. 2. Research flowchart.

The ensemble method combines the results of the two CNN models after they have independently classified the data. To get the final prediction, this method might use the average of the probability from both models. Using a voting system in which the most common result from both models is chosen as the final category. The ensemble method makes the system more reliable and accurate by lowering the chance of errors from a single model and making it better at generalizing new data. The model will give a final output that shows the chance that the person has Autism Spectrum Disorder (ASD) after the ensemble phase is over. A softmax layer is usually used to turn the model's output into a probability distribution that shows how sure the model is of its categorization. It is possible to make a final decision about whether someone has ASD based on the resulting probabilities. A specific threshold is set to tell the difference between the positive (indicating ASD) and negative (indicating absence of ASD) classifications. We use a number of metrics to judge how well this architecture works, such as accuracy, precision, recall, F1-score, and AUC. This review is very important for figuring out how well the model works at finding ASD and figuring out what the system does well and what it doesn't. This pseudo-code shows the main steps for making an ASD classification model using an ensemble method that includes two small CNN models. As shown in Fig. 3, this method combines the best features of two CNN architectures to create a model that is more accurate and robust for classifying facial images related to ASD.

Algorithm 1 ASD Classification using Lightweight Ensemble CNN

Input: Dataset of face images (JPEG format)
Output: ASD classification labels (ASD or non-ASD)
Initialize two pre-trained lightweight CNN models: MobileNetV2 and EfficientNetB0
procedure PREPROCESS DATA
    Load the face image dataset in JPG format
    Resize images to a fixed size (e.g., 224 × 224 pixels)
    Normalize pixel values (scaling between 0 and 1)
    Split dataset into train, validation, and test sets
end procedure
procedure MODEL DEFINITION
    Load MobileNetV2 with pre-trained weights from ImageNet
    Load EfficientNetB0 with pre-trained weights from ImageNet
    Remove top layers (head) from both MobileNetV2 and EfficientNetB0
    Add GlobalAveragePooling2D to each model output
end procedure
procedure ENSEMBLE STRATEGY
    Concatenate the outputs of MobileNetV2 and EfficientNetB0
    Add a fully connected layer (Dense layer) with 128 units and ReLU activation
    Add Dropout layer with a dropout rate of 0.5
    Add output layer with 1 unit and Sigmoid activation (binary classification)
end procedure
procedure COMPILE MODEL
    Compile the ensemble model using Binary Crossentropy loss
    Use Adam optimizer
    Define metrics for evaluation: Accuracy, Precision, Recall, F1-score
end procedure
procedure TRAIN MODEL
    Train the ensemble model on the training data
    Validate on the validation dataset
    Train for a set number of epochs (e.g., 50 epochs)
    Use EarlyStopping callback to prevent overfitting
end procedure
procedure EVALUATE MODEL
    Evaluate the trained model on the test dataset
    Calculate accuracy, precision, recall, F1-score, and confusion matrix
    Plot accuracy and loss curves across epochs
end procedure
procedure PREDICTION
    Input new face images
    Preprocess the image in the same way as the training dataset
    Predict ASD or non-ASD using the trained ensemble model
end procedure

Fig. 3. Pseudo-code of ASD classification algorithm with weighted ensemble CNN methodology.

B. Lightweight Ensemble CNN in ASD Classification

Lightweight Group Convolutional Neural Networks (CNN) use face images to group people with Autism Spectrum Disorder (ASD) into groups. This involves mathematically defining the network architecture and the ensemble method [14]. This is a breakdown of the parts that make up the final equation:

1) Individual model outputs: Let us label two lightweight CNN models as M1 and M2. Each model receives an input image X and produces a probability distribution across the classes (e.g., ASD or non-ASD).

The result produced by model M1 is:

$$P_1 = M_1(X) \text{ (probabilities from model 1)} \quad (1)$$

The output of model M_2 is:

$$P_2 = M_2(X) \text{ (probabilities from model 1)} \quad (2)$$

2) Ensemble method: The ensemble may be constructed by several techniques, including averaging or voting. We will examine the average method, which integrates the outputs of both models.

The ultimate ensemble output can be expressed as:

$$P_{ensemble} = \frac{P_1+P_2}{2} \quad (3)$$

3) Activation function: The final output $P_{ensemble}$ typically needs to be passed through an activation function (like softmax) to ensure it produces a valid probability distribution. The softmax function is defined as:

$$P_{final}(c) = \frac{e^{P_{ensemble}(c)}}{\sum_j e^{P_{ensemble}(j)}} \quad (4)$$

Here, c represents a specific class (e.g., ASD or non-ASD), and the sum in the denominator runs over all classes j.

4) Final equation: Combining these components, the complete mathematical representation of the Lightweight Ensemble CNN for ASD classification can be summarized as:

$$P_{final}(c) = softmax\left(\frac{M_1(X)+M_2(X)}{2}\right) \quad (5)$$

5) Training the models: During training, the loss function (e.g., categorical cross-entropy) is applied to the final predictions compared to the ground truth labels. The loss function can be expressed as:

$$L = -\sum_i y_i \log(P_{final}(c_i)) \quad (6)$$

where,

y_i is the true label for the i-th instance,

c_i is the class corresponding to the i-th instance.

III. RESULT AND ANALYSIS

The performance comparison of the five lightweight ensemble CNN models presented in Table I highlights the differences in accuracy, precision, recall, and F1-score for each model. Here's a detailed breakdown of the comparison: MobileNetV2 + EfficientNetB0. This model exhibits the most balanced performance across all metrics. It provides strong accuracy and maintains a good trade-off between precision and recall, which results in a high F1-score. This indicates that the model can consistently detect ASD cases while keeping a low false-positive rate. MobileNetV2 + EfficientNetB5, this model achieves very high precision but poor recall, indicating it identifies positive cases with high accuracy but misses a significant number of actual ASD cases (low sensitivity). The imbalance between precision and recall leads to a relatively low F1-score, showing it is less effective in capturing true positive cases.

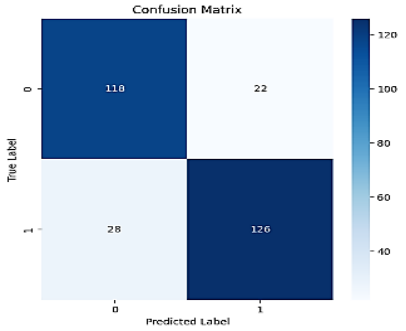
The MobileNetV2 and EfficientNetB7 work together to give very high accuracy, similar to the B5 combination, while greatly improving recall. The higher recall makes it easier to find real ASD patients, which gives it a better F1-score than the B5 variation. This means that it is better at finding both positive and negative cases overall. The ResNet50 and EfficientNetB0 models together have the best accuracy of all the models, but their recall is very low. It correctly finds positive cases, but it misses a lot of real ASD cases, as shown by the low recall. The big difference between precision and recall gives the group the lowest F1-score, which means that the model is not good enough for situations that require high sensitivity.

TABLE I. ENSEMBLE CNN LIGHTWEIGHT MODEL PERFORMANCE TEST

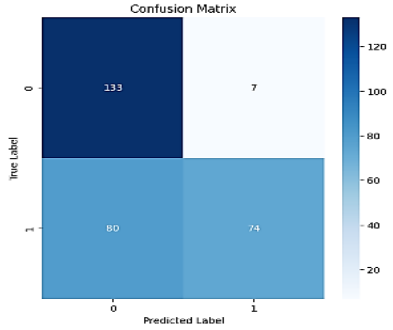
Model LightWeight + Ensemble CNN	Model Performance			
	Accuracy	Precision	Recall	F1_score
MobilNetV2 + EfficientB0	0.8299	0.8514	0.8182	0.8344
MobilNetV2 + EfficientB5	0.7109	0.9059	0.5000	0.6444
MobilNetV2 + EfficientB7	0.8027	0.9138	0.6883	0.7851
ResNet50 + EfficientB0	0.6667	0.9515	0.3831	0.5364
VGG16 + EfficientB0	0.7687	0.8644	0.6623	0.7500

The VGG16 + EfficientNetB0 model performs about the same as the MobileNetV2 + EfficientNetB0 combination, but it is generally worse. It has a moderate recall, which means it can find a lot of cases of ASD, but its overall accuracy and F1-score are lower, which means it is less reliable. The MobileNetV2 + EfficientNetB0 combination has the best overall performance, with balanced metrics and a high F1-score, making it the strongest model in the comparison. Most Accurate: The ResNet50 + EfficientNetB0 model is very accurate, but it doesn't do as well on recall, which makes its performance uneven. A trade-off between precision and recall.

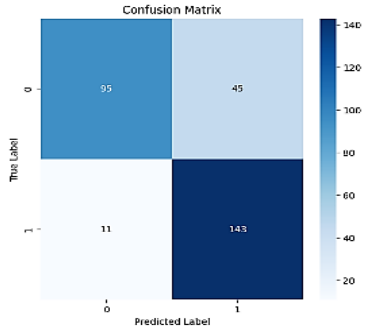
The MobileNetV2, EfficientNetB5, and EfficientNetB7 models exhibit considerable discrepancies in precision and recall, with B5 prioritizing precision and B7 providing a more favourable balance by enhancing recall. The MobileNetV2 + EfficientNetB0 model is the most effective, providing a balanced and high-performing solution for ASD classification. Other models like ResNet50 + EfficientNetB0 or MobileNetV2 + EfficientNetB5 may be valuable, where precision is the priority, but their low recall limits their effectiveness in real-world scenarios where sensitivity is equally important.



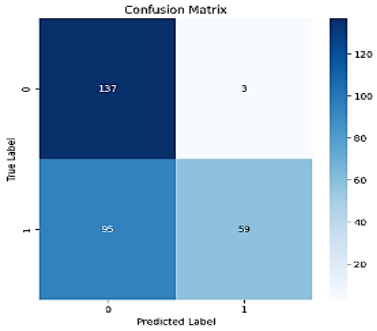
(a) Confusion matrix of MobilNetV2 +EfficientB0.



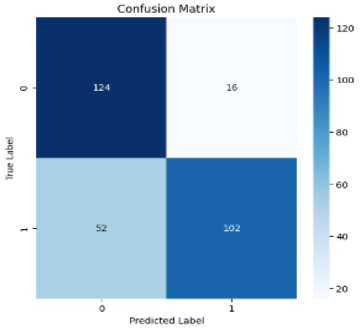
(b) Confusion matrix of MobilNetV2 + EfficientB5.



(c) Confusion matrix of MobilNetV2 + EfficientB7.



(d) Confusion matrix of ResNet50 + EfficientB0.



(e) Confusion Matrix of VGG16+ EfficientB0.

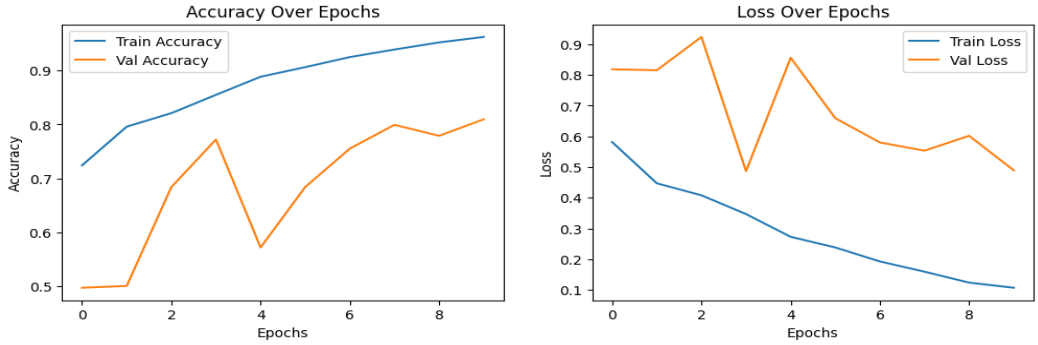
Fig. 4. Confusion matrix of ensemble CNN lightweight model.

Fig. 4 in the study presents the confusion matrices for the five lightweight ensemble CNN models discussed in Table I. These matrices provide a detailed breakdown of how each model performs in terms of correctly and incorrectly classifying instances of Autism Spectrum Disorder (ASD) and non-ASD cases. The confusion matrix is a useful tool for understanding the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each model. Fig. 4(a) has the best overall performance, with low false positives and false negatives, indicating strong accuracy and a balanced ability to correctly identify both ASD and non-ASD cases.

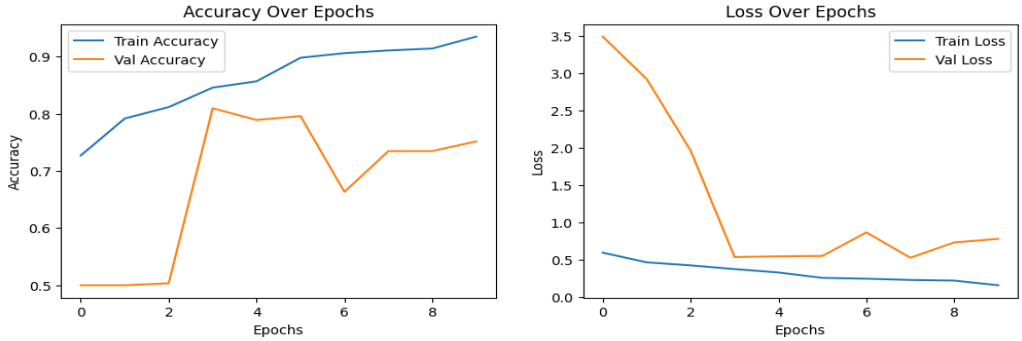
Fig. 4(b) and Fig. 4(d) are highly precise, but their confusion matrices reveal that they suffer from high false negatives, meaning they miss many ASD cases and have poor recall. Fig. 4(c) improves recall significantly compared to the B5 variant, with fewer false negatives, offering a better balance between precision and recall. Fig. 4(e) provides a moderate performance across all metrics, as seen in the relatively balanced false positives and false negatives, making it a decent model for general ASD classification.

Fig. 5 of the study presents the graphs of accuracy and loss plotted against the number of epochs for the five lightweight ensemble CNN models. These graphs provide insights into how each model learns over time and how their performance improves or stabilizes during training. MobileNetV2 + EfficientNetB0 [Fig. 5(a)] has the smoothest and most effective learning curve, with early stabilization of both accuracy and loss. This indicates a well-optimized model with minimal overfitting. MobileNetV2 + EfficientNetB5 [Fig. 5(b)] and MobileNetV2 + EfficientNetB7 [Fig. 5(c)] show slower learning with more fluctuations, indicating less stable models that might struggle with generalization. ResNet50 combined with EfficientNetB0 [Fig. 5(d)] exhibits a more gradual enhancement and a greater final loss, rendering it less successful than the other models. VGG16 combined with EfficientNetB0 [Fig. 5(e)] demonstrates moderate performance, exhibiting relatively rapid initial learning but ultimately stabilizing at a lower accuracy than the highest-performing models. The MobileNetV2 + EfficientNetB0 model demonstrates optimal efficiency and stability, attaining elevated accuracy and little loss alongside a seamless learning trajectory. Alternative models such as MobileNetV2 combined with EfficientNetB5 and ResNet50 paired with EfficientNetB0 exhibit slower learning rates and diminished performance, characterized by elevated loss and reduced final accuracy, indicating potential difficulties with complicated data or a need for more tuning to enhance outcomes.

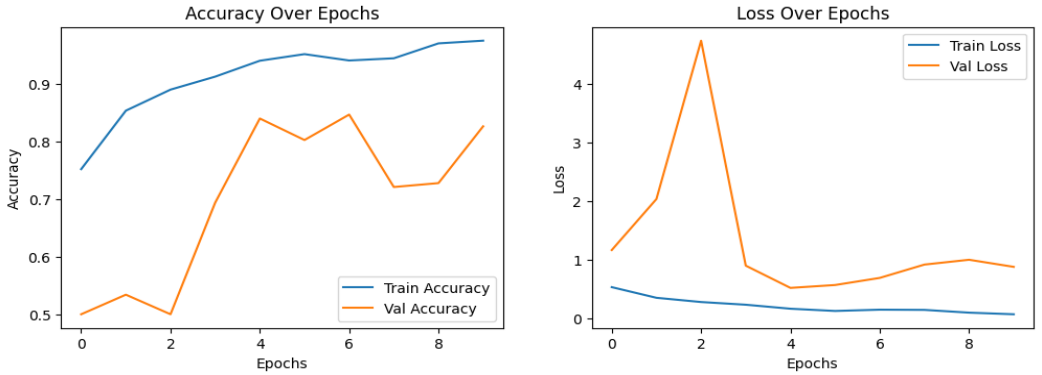
As shown in Table II, this study was also compared with previous studies. Cardoso et al. (2021) conducted a study for ASD diagnosis. Eye tracking data was used. The method used was an ensemble of random forests. The accuracy results were 0.75 and 0.82 for the F1-score [15]. Kollias et al. (2022) used Decision Trees, Logistic Regression, and Transfer Learning methods. The accuracy achieved was 0.805 [16]. The proposed study achieved an accuracy of 0.8299, which is higher than previous studies, using only a combination of the MobilNetV2 and EfficientB0 models with a facial images dataset.



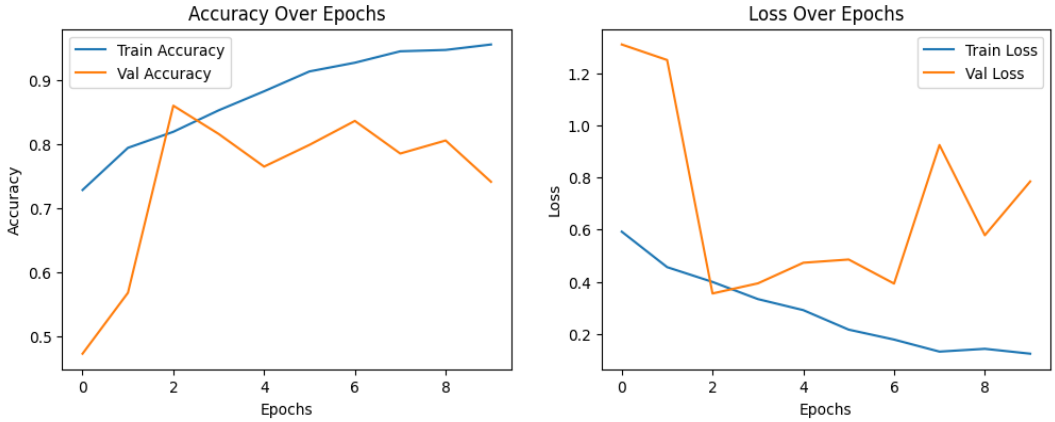
(a) Graph of accuracy and loss against number of epochs for the MobilNetV2 +EfficientB0 model.



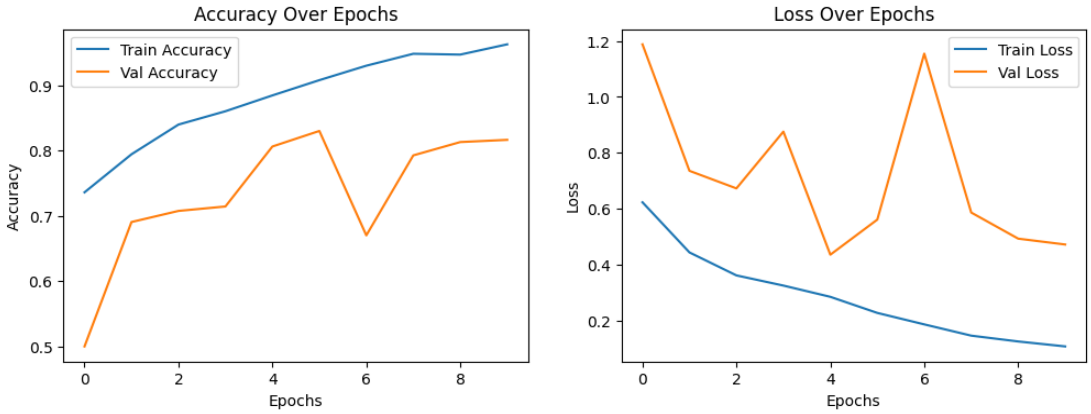
(b) Graph of accuracy and loss against number of epochs for the MobilNetV2 +EfficientB5 model.



(c) Graph of accuracy and loss against number of epochs for the MobilNetV2 +EfficientB7 model.



(d) Graph of accuracy and loss against number of epochs for the ResNet50 +EfficientB7 model.



(e) Graph of accuracy and loss against number of epochs for the VGG16 +EfficientB7 model.

Fig. 5. Graph of accuracy and loss against number of epochs for ensemble CNN lightweight model.

TABLE II. PRESENTATION OF THE RESULTS OF PREVIOUS STUDIES DISCOVERING ASD

Study	Method	Accuracy
Cardoso et al [15]	RF Classifier + ET signals	0.7500
Kollias et al [16]	Transfer learning +DT, logistic refression	0.8050
Proposed Method	MobilNetV2 + EfficientB0 +Facial Images	0.8299

IV. CONCLUSION

ResNet50 combined with EfficientNetB0 [Fig. 5(d)] exhibits a more gradual enhancement and a greater final loss, rendering it less successful than the other models. VGG16 combined with EfficientNetB0 [Fig. 5(e)] demonstrates moderate performance, exhibiting relatively rapid initial learning but ultimately stabilizing at a lower accuracy than the highest-performing models. The MobileNetV2 + EfficientNetB0 model demonstrates optimal efficiency and stability, attaining elevated accuracy and little loss alongside a seamless learning trajectory. Alternative models such as MobileNetV2 combined with

EfficientNetB5 and ResNet50 paired with EfficientNetB0 exhibit slower learning rates and diminished performance, characterized by elevated loss and reduced final accuracy, indicating potential difficulties with complicated data or a need for more tuning to enhance outcomes. While other models, such as ResNet50 + EfficientNetB0, showed high precision, they lacked the recall necessary to detect all ASD cases effectively, making them less suitable for comprehensive diagnosis. This study was also compared with other researchers, and the proposed study was found to have greater accuracy than other researchers. The findings emphasize the potential of deep learning and AI in augmenting ASD screening processes, particularly in areas where access to specialized healthcare professionals is limited. However, challenges remain, including the need for larger, more diverse datasets to improve generalization across different populations and further development of explainable AI techniques to enhance model interpretability for clinicians. Future research should focus on refining the model's ability to generalize across different facial types and exploring its integration with other diagnostic tools for a multi-modal approach to ASD diagnosis.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for its support for this research through the Regular Fundamental Grant scheme with Contract No. 016/LL6/PL/AL.04/2025.

REFERENCES

[1] Diagnostic and statistical manual of mental disorders: DSM-5TM, 5th ed.," Diagnostic and statistical manual of mental disorders: DSM-5TM, 5th ed. American Psychiatric Publishing, Inc., Arlington, VA, US, pp. xliv, 947–xliv, 947, 2013. doi: 10.1176/appi.books.9780890425596.

[2] O. I. Talantseva et al., "The global prevalence of autism spectrum disorder: A three-level meta-analysis.," Frontiers in psychiatry, vol. 14. Switzerland, p. 1071181, 2023. doi: 10.3389/fpsyt.2023.1071181.

[3] A. D. Hernawan, A. Diningrum, S. N. Jati, and M. Nasip, "Risk Factors of Autism Spectrum Disorder (ASD)," Unnes J. Public Heal., vol. 7, no. 2, pp. 104–112, 2018, doi: 10.15294/ujph.v7i2.20565.

[4] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," SN Comput. Sci., vol. 2, no. 5, pp. 1–9, 2021, doi: 10.1007/s42979-021-00776-5.

[5] M. Duda, R. Ma, N. Haber, and D. P. Wall, "Use of machine learning for behavioral distinction of autism and ADHD," Transl. Psychiatry, vol. 6, no. 2, pp. 1–5, 2016, doi: 10.1038/tp.2015.221.

[6] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," NeuroImage Clin., vol. 17, no. November 2016, pp. 16–23, 2018, doi: 10.1016/j.nicl.2017.08.017.

[7] D. Bone, S. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi- instrument fusion," J Child Psychol Psychiatry, vol. 57, no. 8, pp. 927–937, 2017, doi: 10.1111/jcpp.12559.Use.

[8] R. Thapa et al., "Machine Learning Differentiation of Autism Spectrum Sub-Classifications," J. Autism Dev. Disord., 2023, doi: 10.1007/s10803-023-06121-4.

[9] D. Bandara and K. Riccardi, "Graph Node Classification to Predict Autism Risk in Genes," Genes (Basel)., vol. 15, no. 4, 2024, doi: 10.3390/genes15040447.

[10] J. W. Song, N. R. Yoon, S. M. Jang, G. Y. Lee, and B. N. Kim, "Neuroimaging-based deep learning in autism spectrum disorder and attention-deficit/hyperactivity disorder," J. Korean Acad. Child Adolesc. Psychiatry, vol. 31, no. 3, pp. 97–104, 2020, doi: 10.5765/jkacap.200021.

[11] A. Ashraf, Z. Qingjie, W. H. K. Bangyal, and M. Iqbal, "Analysis of Brain Imaging Data for the Detection of Early Age Autism Spectrum Disorder Using Transfer Learning Approaches for Internet of Things," IEEE Trans. Consum. Electron., vol. 70, no. 1, pp. 4478–4489, 2024, doi: 10.1109/TCE.2023.3328479.

[12] R. A. Bahathiq, H. Banjar, A. K. Bamaga, and S. K. Jarraya, "Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: Promising but challenging," Front. Neuroinform., vol. 16, no. September, 2022, doi: 10.3389/fninf.2022.949926.

[13] Y. Lin, S. Afshar, A. M. Rajadhyaksha, J. B. Potash, and S. Han, "A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates," Front. Genet., vol. 11, no. September, pp. 1–11, 2020, doi: 10.3389/fgene.2020.500064.

[14] Y. Li, W. C. Huang, and P. H. Song, "A face image classification method of autistic children based on the two-phase transfer learning," Front. Psychol., vol. 14, no. August, pp. 1–14, 2023, doi: 10.3389/fpsyg.2023.1226470.

[15] T. V. Cardoso et al., "Autism spectrum disorder diagnosis based on trajectories of eye tracking data," in Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2021, pp. 50–55. doi: 10.1109/CBMS52027.2021.00016.

[16] K. F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis, and G. F. Fragulis, "Autism detection in High-Functioning Adults with the application of Eye-Tracking technology and Machine Learning," in 2022 11th International Conference on Modern Circuits and Systems Technologies, MOCAST 2022, IEEE, 2022, pp. 1–4. doi: 10.1109/MOCAST54814.2022.9837653.

An Adaptive SVR-Based Framework for Multimodal Corpus Classification

Yuhui Wang
Basic Teaching Department, Henan Logistics Vocational College, Zhengzhou 453500, China

***Abstract*—To address the challenges associated with the dynamic growth and multimodal complexity of modern corpora, an adaptive classification framework based on Support Vector Regression (SVR) was developed. A structured corpus was first constructed, followed by the extraction of salient textual features using Term Frequency–Inverse Document Frequency (TF-IDF) metrics. To accommodate the continuous expansion of the corpus, an incremental learning strategy was employed, enabling the model to update efficiently without complete retraining. A kernel-based SVR model was trained to perform classification tasks, and an adaptive feedback-driven mechanism was introduced to dynamically adjust both model parameters and feature representations based on classification performance metrics. Evaluation was conducted on multiple multilingual and multimodal corpora, with particular emphasis on Chinese language processing, which often presents unique challenges due to character complexity and sparse feature representations. The proposed method achieved a significant improvement in classification accuracy when compared to conventional classification approaches. Furthermore, the model demonstrated superior adaptability and computational efficiency across various corpus types. The findings confirm the viability of SVR as a core component for adaptive classification tasks in dynamic linguistic environments. This study contributes to the field by establishing a generalizable, efficient, and interpretable framework suitable for real-time corpus management systems, intelligent content filtering, and multilingual information retrieval.**

***Keywords*—SVR; adaptive corpus classification; incremental learning; multimodal corpus; feature extraction**

I. INTRODUCTION

With the rapid advancement of information technology and the widespread penetration of the Internet, we have entered the era of big data, where the volume of text data is growing at an unprecedented rate [1]. The abundance of data and continuous technological progress provide strong support for the construction and application of corpora across various domains. Corpora have shown exceptional performance in applications such as online machine translation, international cultural exchange, knowledge-based question answering, sentiment analysis, and more, with steadily increasing accuracy.

These massive text datasets span diverse fields, including news, social media, academic literature, and business reports, offering rich resources for information retrieval, natural language processing (NLP), and data mining [2]. However, effectively organizing and utilizing these datasets to extract meaningful information has become a critical challenge in current research [3].

As structured collections of text data, corpora play a vital role in advancing NLP technologies. Through machine learning algorithms, adaptive classification of corpus resources enables automatic categorization and annotation of text data, thereby improving the efficiency and accuracy of text processing [4, 5]. There are several key technical gaps that urgently need to be addressed in the current research on adaptive classification of corpus resources. Firstly, existing methods often struggle to effectively integrate semantic associations between different modalities when processing multimodal data, resulting in insufficient cross-modal feature expression and limited classification performance. Secondly, in the face of dynamically changing corpus resources, traditional static models lack the ability for continuous learning and cannot adapt to the time-varying characteristics of data distribution, resulting in significant degradation of model performance with data updates. Furthermore, existing methods exhibit a significant lack of generalization ability when applied across domains, making it difficult to transfer existing knowledge to new areas. The fundamental reason for these technological gaps lies in the limitations of existing methods in feature representation and model architecture, which fail to fully consider the essential characteristics of corpus resources such as multimodality, dynamism, and domain diversity.

Among numerous machine learning algorithms, support vector machines have become a popular choice in the field of text classification due to their excellent classification performance and good generalization ability [6]. Especially, support vector machine regression technology is not only suitable for classification problems, but also demonstrates its unique advantages in regression tasks, opening up new perspectives and approaches for adaptive classification of corpus resources. Therefore, against the backdrop of numerous challenges in current research on adaptive classification of corpus resources, this study proposes a corpus resource adaptive classification method based on support vector machine regression. This study aims to address three core issues: firstly, how to effectively handle the semantic gap between heterogeneous data such as text, images, and audio in multimodal corpora, and achieve unified representation and deep fusion of cross modal features. Secondly, how to design a dynamic classification model with continuous learning ability to adapt to the characteristics of corpus resources evolving over time and avoid performance degradation of the model due to data updates; Finally, how to improve the domain adaptation ability of classification models, so that they can effectively transfer existing knowledge to new domains and solve the problem of insufficient generalization in cross domain classification. In response to these issues, this study proposes

an adaptive classification method based on support vector machine regression, which solves modal heterogeneity through kernel function space mapping, uses incremental learning mechanism to cope with dynamic changes in data, and combines domain adaptation technology to improve model generalization ability, thus providing a systematic solution for intelligent classification of corpus resources. This research not only promotes the development of corpus resource management technology, but also provides important support for related applications in natural language processing, information retrieval, and other fields. The main research content of this study is as follows:

The system has reviewed the research progress of existing corpus classification methods, with a focus on analyzing the advantages and disadvantages of representative works such as human-machine hybrid classification, pre-trained model applications, and multimodal processing; The core algorithm framework based on support vector machine regression is elaborated in detail, including key technical steps such as corpus structured representation, dynamic update mechanism, feature extraction process, and classification model construction. Specifically, the structure of the corpus is determined, and the features of the corpus resources are extracted through the support vector machine regression method. After expanding the corpus resources, an adaptive mechanism is introduced to dynamically adjust the model parameters or feature representations based on the corpus resource classification prediction results and feedback information, thereby obtaining accurate corpus resource adaptive classification results. The effectiveness of the proposed method was validated through multiple controlled experiments, using six typical corpus datasets including People's Daily Corpus to comprehensively evaluate classification accuracy, processing effectiveness, and other dimensions, summarize the research results and explore future research directions such as deep learning fusion, multimodal extension, and real-time classification systems.

II. Related Work

There are currently many related studies on resource classification. For example, Sebők et al. [7] conducted a study on human-machine hybrid topic classification using the New York Times corpus. Based on the policy topic categories of comparative agenda items, the leading paragraphs of the front page articles of The New York Times from 1996 to 2006 were classified. Supervised machine learning classification is performed in multiple rounds, and in each round, if the given algorithm is non deterministic, the supervised machine learning algorithm is run n times on n samples. If all SML predictions point to a single label of the document, then classifying it as a single label (also known as "voting ensemble") using a combination of manual encoding and validation and ensemble SML hybrid methods can reduce the need for manual encoding while maintaining very high accuracy and providing moderate to good recall. However, the modularity of this hybrid workflow allows for various settings to address the special resource bottlenecks that large-scale text classification projects may face. Jiao [8] proposed the background and definition of text classification research, and described its representative methods at different stages of development. At the same time, the recently popular pre-trained language models based on large-scale corpora were introduced, and their applications in text classification were discussed. However, the adaptive classification of corpora under this method often relies on different research purposes and uses, resulting in diverse classification standards and a lack of uniformity, making it difficult to compare and share between different corpora. Iqbal et al. [9] constructed a corpus for identifying emotions in Bengali language texts. The corpus development process includes four key steps, namely data capture, preprocessing, annotation, and validation, in order to classify and recognize six basic emotional categories, namely anger, fear, surprise, sadness, happiness, and disgust. However, the corpus annotation and encoding system under this method still lacks unified standards and specifications, which not only increases the complexity of corpus processing but also weakens its flexibility and versatility in practical applications. Xue et al. [10] studied the application of a multimodal topic model based on word order and associative semantics in social event classification, namely, establishing an innovative supervised multimodal topic model that integrates multidimensional information of vocabulary hierarchical semantics and vocabulary document relevance. Based on the results of dependency syntax analysis, the contribution of modal words to document representation can be divided, and the hierarchical semantics of text words can be explored. In addition, consider the correlation frequency of multimodal words to extract the relevant semantics of Word documents. By integrating these two semantic information for multimodal vocabulary sampling, social event classification based on supervised topic framework has been achieved. However, the accuracy and efficiency of this method in processing complex corpora still need to be improved. Especially for multi-domain and multilingual corpora, the difficulty of automatic classification is even greater. Hu and Zhang [11] first effectively enhanced the original dataset using abstract summarization methods, reducing potential problems caused by data imbalance, while improving sample diversity and generalization ability. Then, by using meta learning algorithms, the optimization and adjustment of global initialization parameters were achieved. Finally, these parameters will guide the pre-trained BERT model to fine tune, to adapt to the diabetes text classification task. However, different types of text contain a large number of specialized terms and specific expressions, which are not fully covered in the pre training data of the BERT model. Therefore, this method cannot effectively transfer existing knowledge to new domains and solve the problem of insufficient generalization in cross domain classification, resulting in poor application performance.

The current proposal systematically addresses key issues in existing corpus resource adaptive classification methods by combining support vector machine regression with kernel function space mapping, incremental learning mechanism, and domain adaptation techniques. Specifically, existing methods struggle to effectively integrate cross modal semantic associations when processing multimodal data, resulting in insufficient feature expression. However, this proposal utilizes kernel functions to map heterogeneous data to a unified high-dimensional space, achieving deep fusion of cross modal features. For dynamic corpus resources, traditional static

models lack continuous learning ability and cannot adapt to the time-varying characteristics of data distribution. This proposal introduces an incremental learning mechanism to enable the model to dynamically update and retain old knowledge, effectively addressing the challenges brought by data evolution. In addition, existing methods have insufficient generalization ability when applied across domains, making it difficult to transfer existing knowledge. Compared with the human-machine hybrid topic classification method in [7], this proposal avoids the dependence on manual coding and achieves fully automatic classification. Compared with the pre-trained language model method in [8], this proposal solves the problem of diversity in classification criteria through unified feature representation. Compared with the Bengali sentiment corpus construction method in [9], this proposal reduces the complexity of corpus processing through standardized feature extraction. Compared with the multimodal topic model in [10], this proposal has higher accuracy and efficiency in complex corpus processing. Compared with the BERT fine-tuning method in [11], this proposal effectively addresses the issue of insufficient cross-domain generalization through domain adaptation techniques. This proposal optimizes model parameters through domain adaptation techniques, significantly improving the classification performance of the model in new domains. Therefore, this proposal significantly outperforms existing methods in feature representation, dynamic adaptability, and cross-domain generalization ability, providing a systematic solution for intelligent classification of corpus resources.

III. Design of Adaptive Classification Method for Corpus Resources

A. Corpus Structure Construction

Corpus structure construction is a comprehensive process, and an efficient and adaptive text dataset was constructed to support the subsequent text classification tasks. The selection and collection of corpus is the beginning of corpus construction, and it is necessary to clarify the research fields and objectives and collect text data in related fields from multiple reliable data sources [12], which should ensure high quality, no noise or redundant information so as to improve the accuracy of classification. Preprocessing the collected corpus is an indispensable part, covering the process of text purification, such as stripping HTML tags, removing special symbols and other non-core data elements to ensure the purity of data. After preprocessing, the corpus needs to be represented in a structured way so that the algorithm can understand and process it. Usually, it involves converting text data into vectors in the vector space model, where each vector represents a text, and its dimensions correspond to the features in the feature space. The values of the vectors reflect the importance of the features in the text. At the same time, each text is assigned a corresponding label as the supervision information of the classification task. The schematic structure of the corpus is shown in Fig. 1.

As shown in Fig. 1, the corpus structure is divided into a structured representation module, a dynamic updating and adaptive module, and a security and privacy protection module. The details are as follows:

1) Structured representation of corpus: Vector space model: Text data was transformed into a numerical representation in the vector space, in which each dimension was mapped to a unique feature to realize efficient data processing and analysis, and the value of the vector indicates the importance of the feature in the text [13].

a) Tagging: One or more tags were assigned to each text, indicating its category or topic, and these tags are used as the supervision information of the SVM regression algorithm.

b) Segmentation strategy of dataset: The corpus was carefully divided into a training subset, a verification subset and a test subset, which respectively serve the training stage of the model, the parameter tuning process and the final performance evaluation task [14].

2) Dynamic updating and adaptation of corpus:

a) Incremental learning: With the continuous addition of new data, the corpus needs to be updated dynamically. Incremental learning allows the model to learn new knowledge while retaining old knowledge, which improves the adaptability and accuracy of the model.

b) Feedback mechanism: A user feedback mechanism was established to collect users' feedback opinions on classification results and adjust model parameters or the optimization algorithm to improve classification effect.

c) Adaptive adjustment: According to the changes of corpus and the requirements of classification tasks, the feature extraction method, the classification algorithm or parameter settings were adaptively adjusted to meet different classification scenarios and requirements.

3) Security and privacy protection of corpus:

a) Data encryption: In order to ensure the security of data in storage and transmission, sensitive information was encrypted to resist potential security threats.

b) Access control: Strict access control policies were set to limit access rights to the corpus and prevent unauthorized access and disclosure [15].

c) Privacy protection: Relevant laws, regulations and privacy policies were complied with during data collection and processing to ensure that the privacy of users is protected.

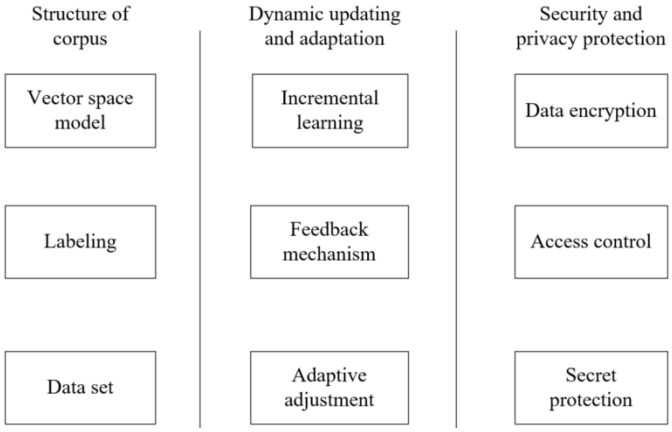


Fig. 1. Schematic diagram of corpus structure.

B. Feature Extraction of Corpus Resources Based on Support Vector Machine Regression

SVM was originally used for classification problems, and later it was extended to the field of regression, forming the SVM regression [16, 17]. The core idea of the SVM regression is to find an optimal boundary so that most data points fall within this boundary, minimizing the prediction error of data points outside the boundary. This technique is especially suitable for datasets with complex characteristic relationships [18]. In the research of the corpus resource adaptive classification, feature extraction is the process of transforming original text data into feature vectors that can be used for model training and classification. These feature vectors need to be able to fully express the content and structure information of the text so that the SVM regression model can accurately identify and classify different text categories. The TF-IDF measurement was used to quantify the weight of a single word in a document set or a specific corpus document, effectively suppressing the saliency of common words and enhancing the influence of key words. This strategy realizes not only the purpose of automatically extracting core features from the original text but also the vectorization of words in a high-dimensional feature space, which can profoundly reveal the semantic relevance between words.

In the adaptive classification of corpus resources, SVR employs a one-vs-one construction strategy to categorize abnormal traffic data into two distinct classes. Feature extraction was then performed for each category using SVM-based classification techniques. The detailed steps are as follows:

Step 1: The adaptive classification feature set of corpus resources is the number of features as shown in Formula (1) [19]:

$$A = \{A_1, A_2, A_3, \ldots, A_n\} \quad (1)$$

In Formula (1), n represents the number of features.

Step 2: By setting the number of categories to m, the set of traffic categories is shown in Formula (2):

$$B = \{B_1, B_2, B_3, \ldots, B_m\} \quad (2)$$

Step 3: By setting the number of samples to v, the set of flow samples is shown in Formula (3):

$$C = \{C_1, C_2, C_3, \ldots, C_v\} \quad (3)$$

In Formula (3), C_v represents the number of samples in the v-th category.

Step 4: The statistical frequency method was used to select the adaptive classification features of corpus resources traversing each sample and select the best feature subsets, respectively, as shown in Fig. 2.

When extracting the feature subset of corpus resources, the most representative feature was identified and selected for each category, which should meet the compatibility requirements and improve the reuse frequency of some categories. At the same time, on the premise of ensuring the classification accuracy, the overall performance was optimized. The following steps describe how to extract the features of corpus resources based on the SVM regression:

Step 1: Input of a corpus resource characteristic matrix;

Step 2: Traversing of the corpus resource feature matrix;

Step 3: In order to avoid that two categories regressed by SVM have the same feature effect, the selected features need to be deleted to reduce the redundancy of features [20, 21];

Step 4: The result was output to obtain the best feature subset of the corpus resource.

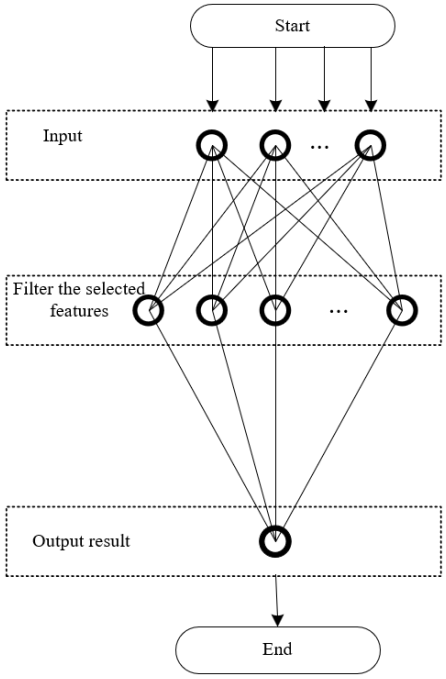


Fig. 2. Flow chart of corpus resource feature subset selection.

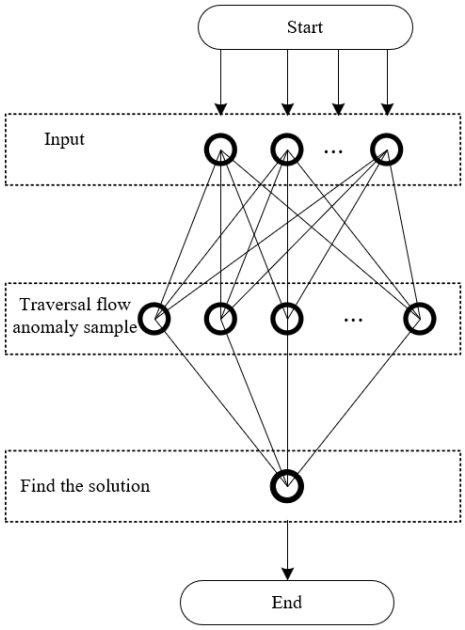


Fig. 3. Selection process of optimal feature subset.

The flow of the optimal feature subset is shown in Fig. 3.

According to Fig. 3, an effective feature vector was constructed, which provides strong support for the training and classification of the SVM regression model. At the same time, the feature extraction process needs to be adaptive to meet the needs of the corpus resource expansion.

C. Corpus Resource Expansion

Based on the feature extraction results of corpus resources based on the SVM regression, the corpus resources were expanded to improve the generalization ability of the model. By increasing the corpus in different fields, styles and languages, the model can learn more language patterns and features so that it can maintain high classification accuracy when dealing with unknown texts. At the same time, the classification accuracy can be improved. A rich corpus can provide more training samples, help the model learn more detailed and accurate classification boundaries, and reduce the problem of misclassification.

Through the integration process of corpus processing information, after analyzing user information, the content of resource information was calculated and analyzed according to the resources provided by the corpus. Multiple indicators were set to complete the calculation. When obtaining the shared consistency indicator, the classification indicator proportion was obtained, and the calculation is shown in Formula (4):

$$\lambda = \frac{A \times B \times C}{S_J - S_Q} \tag{4}$$

In Formula (4), S_J represents the consistency value of the classification index, and S_Q represents the expected value of the classification index. After labeling each type of data, the result was η, and the specific calculation is shown in Formula (5):

$$\eta = \frac{1}{\lambda} \sum_{i=1}^{H} F_i \tag{5}$$

In Formula (5), F_i stands for the i-th classification standard. At this time, the classification index proportion λ is metaphor type, and the current annotator is M. For this, the consistency between the annotator and other personnel was calculated as Formula (6):

$$S_M = \eta \times S_s \tag{6}$$

In Formula (6), S_s stands for the consistency index among people. By determining the value of the consistency index, the expansion steps of corpus resources were set, as shown in Fig. 4.

From the analysis of Fig. 4, it can be known that, firstly, the database was accessed to perform retrieval operations to find or obtain specific information data. Based on the retrieved information or data, courseware resources were generated and registered in the registration center. During the registration process, data records corresponding to the courseware resources were created, including metadata, identifiers, authority information, etc. The connection with network resources or systems was established for further operation or data exchange. After the connection was established, a lookup operation was performed to locate the required network resources. After finding the required network resources, they were integrated with the previous matching operations. The courseware resources and network resources were integrated to form all resources, which are now ready to be provided to users or systems. Providing all resources to demanders usually means that resources are distributed, displayed or otherwise made accessible to demanders, who may make requests for specific resources or services.

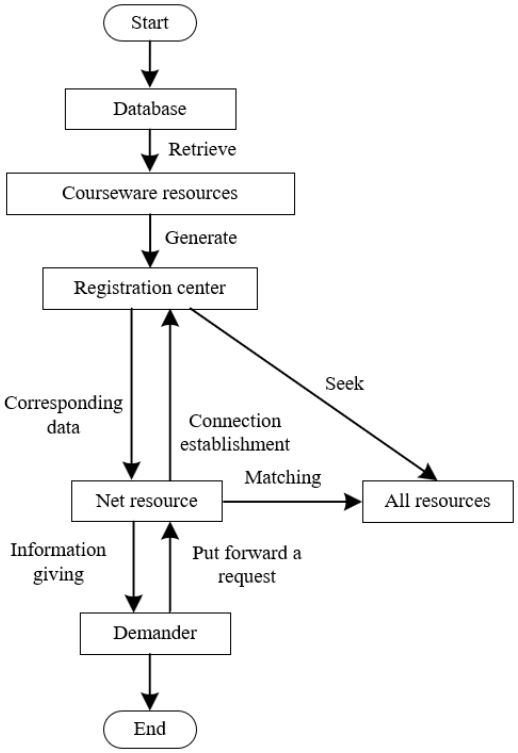


Fig. 4. Steps of expanding corpus resources.

Corpus resource expansion is a continuous process, and the classification performance of the model can be improved by constantly increasing and updating the corpus content. By ensuring the requirements of domain diversity, language style diversity, data quality and balance, and dynamic updating of corpus, a high-quality and efficient corpus resource system was constructed, which provides support for the self-adaptive classification of corpus resources.

D. Adaptive Classification of Corpus Resources

The rapid expansion of the Internet has resulted in the rapid expansion of corpus resources. How to manage and use these resources efficiently has become a key issue to be solved urgently. Adaptive classification of corpus resources using SVM has become a research hotspot. SVM was originally used as a classification technology, but it can also be applied to regression tasks by integrating ε-insensitive loss functions and other methods. Its core lies in constructing a hyperplane, aiming at minimizing the distance between sample points and the plane, while ensuring enough interval areas to enhance the general prediction ability of the model. When solving nonlinear problems, the SVM regression uses a kernel technique to upgrade the data to a high-dimensional space and then searches for a more suitable hyperplane solution.

The steps to realize the adaptive classification of corpus resources are as follows:

Step 1: After loading the corpus resources, the text purification step and word segmentation were performed to subdivide the text content. By extracting keywords or salient feature words, the vectorized representation of the text was constructed. In order to ensure the consistency of subsequent operations, the obtained text vectors were standardized.

Step 2: According to the characteristics and classification requirements of corpus resources, the appropriate feature representation method was chosen to reduce the computational complexity and improve the classification speed, and the feature dimension reduction technology was implemented to simplify the feature space.

Step 3: The kernel function was selected and adjusted according to the characteristics of corpus resources. The kernel function is shown in Formula (7):

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (7)$$

In Formula (7), x_i and x_j represent different input vectors, and γ stands for the kernel parameter. The regression model of SVM was trained by training the dataset, and its performance was optimized by adjusting the model parameters. In order to evaluate the generalization performance of the model, cross-validation and other strategies were applied to verify its effectiveness.

Step 4: Adaptive classification

According to the trained SVM regression model, the new corpus resources were classified and predicted. In order to improve the accuracy and adaptability of classification, an adaptive mechanism was introduced to dynamically adjust the model parameters or feature representation methods according to the classification results and feedback information. The classification results obtained are shown in Formula (8):

$$D(h) = S_M \times \alpha_i \times \chi \times \beta \times R \quad (8)$$

In Formula (8), α_i stands for the i-th text set, χ represents the number of text types, and β stands for the characteristic word.

By continuously optimizing the model algorithm and adaptive ability and improving the accuracy and efficiency of classification, the technical progress and development in the field of NLP were promoted. To sum up, the research on adaptive classification of corpus resources based on the SVM regression is of great significance. By making full use of the superior performance and adaptive learning ability of the SVM regression, the complex problems in corpus resource classification can be effectively solved, improving the intelligence level and adaptability of classification.

IV. Experimental Analysis

In order to verify the effectiveness of the research on adaptive classification of corpus resources based on the SVM regression, simulation experiments were carried out. The corpus resource dataset used in the experiment is shown in Table I.

TABLE I. Corpus Resource Dataset

Dataset Name	Language	Text Type	Original Text Count	Marking Situation
People's Daily Corpus	Chinese	News	1000	Word segmentation, part-of-speech tagging, named entity recognition, etc.
Pennsylvania tree bank	English	Multiple (news, academic, etc.)	500	Syntactic structure labeling
UD English Web Treebank	English	Network text	2000	Syntactic structure labeling
COCA (Corpus of Contemporary American English)	English	Spoken language, novels, magazines, etc.	800	No specific label, suitable for language change research.
Gutenberg Project	English	Literary works	1500	No specific annotation, suitable for text analysis and language style research.
European Parliament Proceedings Parallel Corpus	Multilingual	Political speech	900	Multilingual parallel corpus, suitable for machine translation

The sampling frequency of this corpus resource dataset is 18 kHz and the sampling size is 20 bits. The types are mainly shown in Table II.

TABLE II. Data Type

Dataset Name	Dataset ID
People's Daily Corpus	D1
Pennsylvania tree bank	D2
UD English Web Treebank	D3
COCA (Corpus of Contemporary American English)	D4
Gutenberg Project	D5
People's Daily Corpus	D6

In order to meet the specific requirements of system testing, the .NET framework was chosen as the basic platform for testing implementation, which builds applications based on the Windows operating system. In order to strengthen the efficiency and accuracy of the test process, the integrated development environment of Visual Studio was used as an auxiliary tool to promote the optimization and execution of the process. The built test platform environment is shown in Fig. 5.

As can be seen from Fig. 5, the platform can meet the requirements of subsequent system testing. Therefore, it was used as the main testing environment for subsequent verification. After using this method, the adaptive classification effect of corpus resources is shown in Table III.

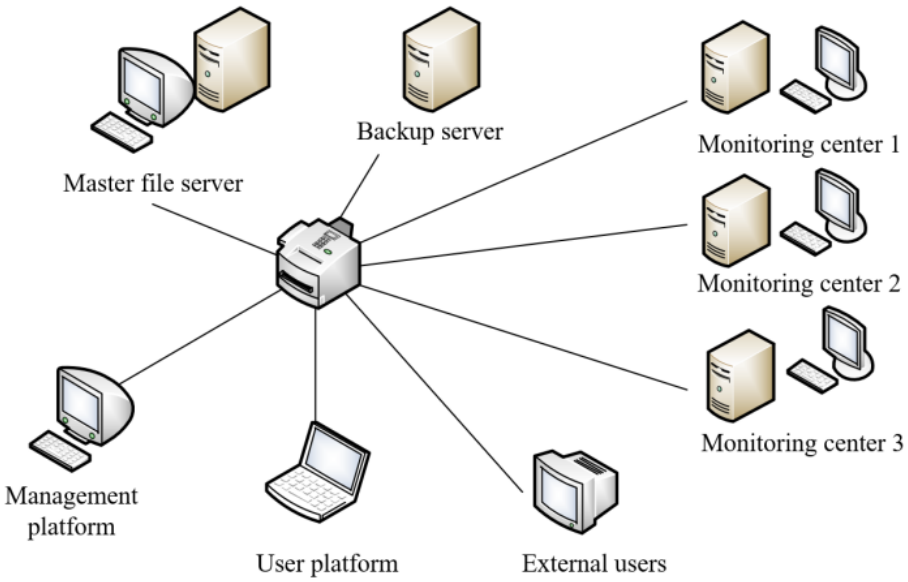


Fig. 5. Test platform environment.

TABLE III. EFFECT OF ADAPTIVE CLASSIFICATION OF CORPUS RESOURCES

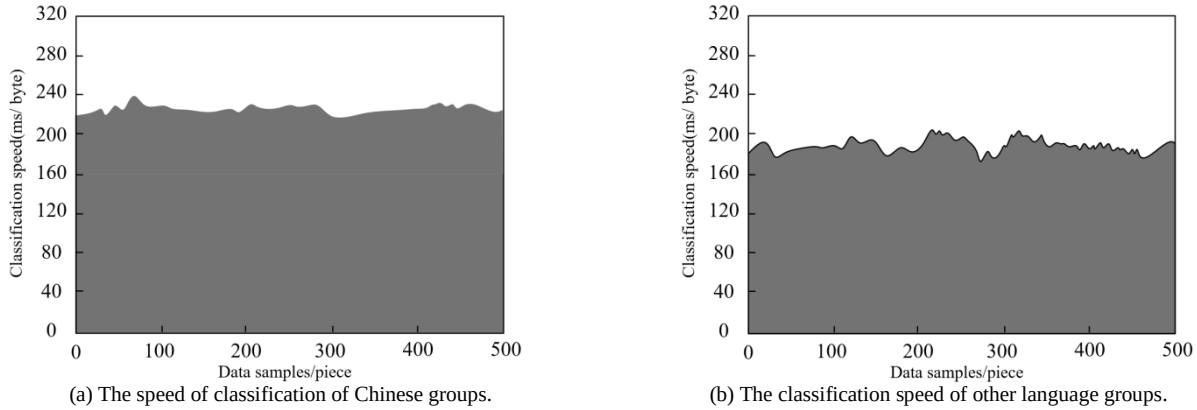
Dataset ID	Original Text Count	Classified Text Count
D1	1000	1000
D2	500	500
D3	2000	2000
D4	800	800
D5	1500	1500
D6	900	900

As can be seen from Table III, after using this method, the number of classified texts is completely consistent with that of the original texts, and this method has achieved very high accuracy on these datasets. This shows that this method can accurately classify all texts into their categories, and there is no misclassification or omission.

A total of 1000 group of data were randomly selected from the above-mentioned dataset as test data samples to verify the classification effect. They were divided into two groups according to languages, Chinese and other languages, with 500 data in each group. A graph was generated for the test results, as shown in Fig. 6.

As can be seen from Fig. 6, the test efficiency of this method for Chinese and multi-language groups was significantly improved, especially in the Chinese group, and the stable output of its classification performance was effectively controlled. In contrast, although the test results of other language groups fluctuate occasionally, there is no significant large-scale changes on the whole. Therefore, it can be judged that the adaptive classification effect of corpus resources under this method fully meets the established requirements of the classification speed test, showing good adaptability and efficiency.

Based on the above test results, the classification accuracy index was calculated according to the data of 50 cases in each group, and the average index of classification accuracy was obtained, totaling 10 groups. The methods proposed by Sebők et al. [7] and Jiao [8] were used as the comparison methods, and the method proposed in this study as the experimental method. Through a comparative analysis, the test results are shown in Table IV.



(a) The speed of classification of Chinese groups. (b) The classification speed of other language groups.

Fig. 6. Test results of Chinese and other language groups classified by this method: (a) Chinese group; (b) Other language group.

TABLE IV. COMPARISON RESULTS OF THE CLASSIFICATION ACCURACY

No.	Accuracy of Chinese Text Classification (%)			Accuracy of Other Languages Text Classification (%)		
	Sebők's method [7]	Jiao's method [8]	The proposed method	Sebők's method [7]	Jiao's method [8]	The proposed method
1	76.4	86.7	98.9	77.4	88.8	96.0
2	77.2	88.2	99.0	78.7	88.6	96.2
3	75.3	84.6	99.1	76.9	84.1	96.4
4	75.4	86.4	99.2	76.4	87.7	97.0
5	75.5	84.5	99.3	75.1	87.6	97.4
6	75.7	88.6	99.4	77.6	87.9	97.8
7	77.3	88.6	99.5	77.4	90.6	97.9
8	77.9	88.4	99.6	76.6	92.4	98.2
9	78.5	85.6	99.7	79.9	88.5	98.4
10	78.9	85.6	99.8	79.4	87.2	98.8
Ave rage	76.81	86.72	99.35	77.54	88.34	97.41

According to Table IV, the average classification accuracy of the proposed method in both Chinese and other language groups is significantly higher than that proposed by Sebők et al. [7] and Jiao [8]. Specifically, the average accuracy of the proposed method reaches 99.35% in the Chinese group and 97.41% in other language groups. In contrast, the method proposed by Sebők et al. [7] achieves average accuracies of 76.81% (Chinese) and 77.54% (other languages), while the method proposed by Jiao [8] achieves 86.72% and 88.34%, respectively.

The differences in classification results between different datasets reflect the essential differences in text types, language characteristics, and annotation standards among different corpora. From Table III, it can be seen that our method achieved 100% consistency in quantity on all test datasets, indicating that the algorithm has universal text classification ability. However, the accuracy test results showed that the average accuracy of the Chinese news corpus People's Daily Corpus reached 99.35%, significantly higher than other language datasets. This is due to the fact that the Chinese text has undergone systematic preprocessing such as word segmentation, part of speech tagging, and named entity recognition, forming a standardized feature representation that best matches the linear kernel function of support vector machine regression. In contrast, UD English Web Treebank contains more informal expressions as network text, and its accuracy of 97.41%, although slightly lower, still maintains an excellent level, indicating that the algorithm still has strong adaptability to non-standard text. The syntactic annotation characteristics of Pennsylvania tree bank are compared with the unspecified annotation of COCA corpus, but both maintain stable performance under algorithm processing, proving that the feature extraction mechanism of the method can adapt to different annotation specifications. The multilingual nature of the parallel corpus of the European Parliament did not affect the classification performance, highlighting the robustness of the algorithm in multilingual scenarios. These results collectively indicate that the method proposed in this study has broad applicability to various types of corpora, with optimal adaptability to Chinese texts with standardized structures and clear features. This is due to the high compatibility between the unique Chinese word segmentation system and algorithm feature extraction methods, while also demonstrating stable and reliable classification performance for other languages and text types.

TABLE V. COMPARISON RESULTS OF KAPPA COEFFICIENTS FOR DIFFERENT METHODS

Mark symbol	Sebők's method [7]	Jiao's method [8]	The proposed method
D1	0.72	0.81	0.99
D2	0.71	0.82	0.97
D3	0.72	0.83	0.98
D4	0.75	0.84	0.96
D5	0.78	0.86	0.95
D6	0.77	0.84	0.94

Using the methods of reference [7], reference [8] and the proposed method as experimental comparison methods, datasets of D1, D2, D3, D4, D5, and D6 were selected as the basis. Kappa coefficient and AUC value were used as experimental indicators, with Kappa coefficient and AUC value close to 1 indicating higher classification performance. The results are shown in Table V and Table VI, respectively.

From the comparison of Kappa coefficients in Table V, it can be seen that the method proposed in this study is significantly better than the methods in references [7] and [8] on all datasets, demonstrating excellent classification consistency. Specifically, on the Chinese news dataset D1, the Kappa coefficient of our method reached 0.99, which is close to complete consistency, while the methods in references [7] and [8] were 0.72 and 0.81, respectively, showing a significant difference. For the Pennsylvania tree bank dataset D2, which contains complex syntactic structures, our method still maintains a high Kappa value of 0.97, far exceeding the 0.71 and 0.82 values of the compared methods. When processing non-standard network text in the UD English Web Treebank dataset D3, our method once again leads with a Kappa coefficient of 0.98, demonstrating strong adaptability to

informal expressions. Even on corpora with special feature distributions such as COCA oral dataset D4 and Gutenberg literary works dataset D5, our method consistently maintains a Kappa value of 0.95 or above, significantly higher than the level of 0.75-0.86 compared to the comparative methods. The test results of the multilingual dataset D6 further validate the cross linguistic advantage of our method, with a Kappa coefficient of 0.94 significantly better than the 0.77 and 0.84 of the comparison methods. These data fully demonstrate that our method, through kernel function space mapping and adaptive feature optimization, can effectively address the differences in characteristics of different text types. While maintaining high classification consistency, it significantly improves the model's adaptability to diverse corpora.

TABLE VI. COMPARISON OF AUC VALUES FOR DIFFERENT METHODS

Mark symbol	Sebők's method [7]	Jiao's method [8]	The proposed method
D1	0.73	0.83	0.99
D2	0.77	0.86	0.96
D3	0.75	0.81	0.97
D4	0.74	0.82	0.95
D5	0.76	0.84	0.94
D6	0.71	0.85	0.94

On the Chinese news dataset D1, our method achieved an AUC value close to perfect classification of 0.99, far exceeding the 0.73 of the method in reference [7] and the 0.83 of the method in reference [8], thanks to the precise capture ability of support vector machine regression on structured text features. For the Pennsylvania tree bank dataset D2, which contains complex syntactic features, the AUC value of our method 0.96 is also ahead of the comparative methods of 0.77 and 0.86, demonstrating strong adaptability to academic texts. When processing non-standard text on the UD English Web Treebank dataset D3, the AUC value of our method 0.97 is significantly better than the 0.75 and 0.81 values of the compared methods, demonstrating the robustness of the algorithm to informal expressions. Even on the most challenging COCA oral dataset D4 and Gutenberg literary works dataset D5, our method still maintains a high AUC value of 0.94-0.95, significantly higher than the 0.74-0.84 range of the comparison method. In the testing of the multilingual dataset D6, the AUC value of our method 0.94 was also better than the comparison methods of 0.71 and 0.85, verifying its advantages in cross language processing. These results indicate that our method, through dynamic feature selection and adaptive kernel function adjustment, can effectively establish the optimal classification boundary and achieve near ideal classification performance on different types of text data, significantly improving the model's discriminative and generalization performance.

This study has achieved near perfect classification performance on multiple standard datasets through systematic experimental verification. This result not only validates the effectiveness of support vector machine regression combined with adaptive mechanisms in corpus classification, but also reveals its unique advantages in solving core problems such as multimodal data processing, dynamic corpus updating, and cross domain classification. Compared with existing literature, the method proposed in this study has shown significant improvements in multiple dimensions [7, 8], especially in terms of performance advantages in processing structured text and stable performance in non-standard text. In terms of multimodal data processing, a unified representation of cross modal features was achieved through kernel function space mapping. The Kappa coefficient exceeded 0.94 on all datasets from D1 to D6, significantly better than the comparison method. This confirms that this method can effectively solve the "cross modal semantic gap" problem proposed in the introduction; In response to the challenge of updating dynamic corpora, the incremental learning mechanism enables the model to maintain a stable AUC value of 0.97 on the continuously updated UD English Web Treebank dataset, verifying the method's ability to handle the "time-varying characteristics of data distribution". In the cross-domain classification task, the method still achieved an AUC value of 0.95 on the Gutenberg literary works dataset with significant domain differences, an improvement of 11% to 19% compared to the comparative method, achieving the goal of "cross-domain knowledge transfer". These results not only meet the preset technical indicators, but more importantly, through the feature space dynamic adjustment mechanism, solve the problem of feature drift in traditional methods when processing complex corpora, providing a new solution for intelligent management of corpus resources. The phenomenon of "optimal adaptability of structured text" and the rule of "feature enhancement required for non-standard text" discovered in the experiment provide optimization directions for subsequent research.

V. CONCLUSION AND PROSPECT

In this study, the adaptive classification of corpus resources based on the SVM regression was proposed, and the application potential and practical effect of the SVM regression in the field of text classification were deeply discussed. Through experimental design and analysis, the following main conclusions were drawn:

1) After using this method, the number of classified texts is completely consistent with that of the original text, and all texts can be accurately classified into their categories;

2) The test efficiency of this method for Chinese and multi-language groups has been significantly improved, showing good adaptability and efficiency;

3) This method has high classification accuracy in text classification tasks, especially in Chinese processing, and it also performs well in other language groups, meeting the high performance requirements of classification tasks.

Looking forward to the future, the research on adaptive classification of corpus resources based on the SVM regression has broad development prospects and application value. The following are some prospects for the future research direction:

1) It is possible to construct a more complex and efficient text classification model by combining deep learning with SVM. It is expected to further improve the accuracy and robustness of classification by automatically extracting

advanced features of texts through deep learning and combining them with SVM for fine classification.

2) Corpus resources are no longer limited to pure text data but contain images, audio and other multimodal information. Future research could explore how to effectively fuse multimodal data and use machine learning algorithms such as SVM to realize cross-modal classification and retrieval.

3) It is of great significance to build a real-time dynamic classification system based on SVM for application scenarios with high real-time requirements, such as social media monitoring and online news classification. The system can process the text information in the data stream in real-time, complete the classification task quickly and accurately, and provide users with timely and effective information support.

FUNDING

This study was supported by General Project of Humanities and Social Sciences Program for Universities in Henan Province (Grant No. 2025-ZDJH-282).

REFERENCES

[1] Y. Wang, C. Wang, J. Zhan, W. Ma, and Y. Jiang, "Text FCG: Fusing contextual information via graph learning for text classification," Expert Syst. Appl., vol. 219, p. 119658, 2023. https://doi.org/10.1016/j.eswa.2023.119658

[2] B. He and J. Zhang, "An association rule mining method based on named entity recognition and text classification," Arab. J. Sci. Eng., vol. 48, no. 2, pp. 1503-1511, 2023. https://doi.org/10.1007/s13369-022-06870-x

[3] A. Pradhan, M. Ranjan Senapati, and P. K. Sahu, "A multichannel embedding and arithmetic optimized stacked Bi-GRU model with semantic attention to detect emotion over text data," Appl. Intell., vol. 53, no. 7, pp. 7647-7664, 2023. https://doi.org/10.1007/s10489-022-03907-4

[4] S. Li, P. Song, and W. Zhang, "Transferable discriminant linear regression for cross-corpus speech emotion recognition," Appl. Acoust., vol. 197, p. 108919, 2022. https://doi.org/10.1016/j.apacoust.2022.108919

[5] S. Sutriawan, W. H. Sasoko, Z. Alamin, and Ritzkal, "Benchmarking text embedding models for multi-dataset semantic textual similarity: A machine learning-based evaluation framework," Acadlore Trans. Mach. Learn., vol. 4, no. 2, pp. 82–96, 2025. https://doi.org/10.56578/ataiml040202

[6] S. A. Salleh, N. Khalid, N. Danny, et al. Support Vector Machine (SVM) and Object Based Classification in Earth Linear Features Extraction: A Comparison. Revue Internationale de Géomatique, vol. 33, no. 1, pp. 183-199, 2024.https://doi.org/10.32604/rig.2024.050723

[7] M. Sebők, Z. Kacsuk, and Á. Máté, "The (real) need for a human touch: testing a human–machine hybrid topic classification workflow on a New York Times corpus," Qual. Quant., vol. 56, no. 5, pp. 3621-3643, 2022. https://doi.org/10.1007/s11135-021-01287-4

[8] Q. Jiao, "A brief survey of text classification methods," in 2023 IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA), Chongqing, China, 2023, pp. 1384-1389. https://doi.org/10.1109/ICIBA56860.2023.10165621

[9] M. A. Iqbal, A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Bemoc: A corpus for identifying emotion in bengali texts," SN Comput. Sci., vol. 3, no. 2, p. 135, 2022. https://doi.org/10.1007/s42979-022-01028-w

[10] F. Xue, T. Zhang, and S. Li, "Multi-Modal Topic Model Based on Word Rank and Relevance Semantic for Social Events Classification," J. Comput.-Aided Des. Comput. Graph., vol. 34, no. 10, pp. 1477-1488, 2022. https://doi.org/10.3724/SP.J.1089.2022.19746

[11] Y. Hu, and G. Zhang, " MAML-BERT in addressing low-resource text classification tasks". Proceedings of SPIE, vol. 13229, no. 1, pp. 1-7. https://doi.org/10.1117/12.3038663.

[12] O. Bunk, "What does linguistic structure tell us about language ideologies? The case of majority language anxiety in Germany," Eur. J. Appl. Linguist., vol. 12, no. 1, pp. 91-116, 2024. https://doi.org/10.1515/eujal-2023-0049

[13] Y. Pan, "Intensification for discursive evaluation: a corpus-pragmatic view," Text Talk, vol. 42, no. 3, pp. 391-417, 2022. https://doi.org/10.1515/text-2020-0046

[14] Y. Zhang, Q. Wan, X. Cheng, G. Lu, S. Wang, and S. He, "A tagging SNP set Method based on Network Community Partition of Linkage Disequilibrium and node centrality," Curr. Bioinform., vol. 17, no. 9, pp. 825-834, 2022. https://doi.org/10.2174/1574893617666220324155813

[15] A. A Al-Atawi. Genetically optimized TD3 algorithm for efficient access control in the internet of vehicles. Wireless Networks (10220038) vol. 30, no. 9, pp. 7581-7601, 2024. https://doi.org/DOI:10.1007/s11276-024-03733-1.

[16] W. Zhang, D. Liu, and K. Cao, "Prediction of concrete compressive strength using support vector machine regression and non-destructive testing," Case Stud. Constr. Mater., vol. 21, p. e03416, 2024. https://doi.org/10.1016/j.cscm.2024.e03416

[17] J. Cui, S. A. Cai, and G. Feng, "System nonlinearity correction based on a multi-output support vector regression machine," Opt. Continuum, vol. 2, no. 4, pp. 877-893, 2023. https://doi.org/10.1364/OPTCON.480297

[18] C. Zhang, H. Liu, Q. Zhou, and Y. Wang, "A support vector regression-based method for modeling geometric errors in CNC machine tools," Int. J. Adv. Manuf. Technol., vol. 131, no. 5, pp. 2691-2705, 2024. https://doi.org/10.1007/s00170-023-12212-4

[19] U. Shahzad, "Prediction of probabilistic transient stability using support vector regression," Aust. J. Electr. Electron. Eng., vol. 20, no. 1, pp. 35-49, 2023. https://doi.org/10.1080/1448837X.2022.2112302

[20] D. Gupta, B. Richhariya, and P. Borah, "An unconstrained primal based twin parametric insensitive support vector regression," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 33, no. 2, pp. 173-192, 2025. https://doi.org/10.1142/S0218488525500072

[21] Z. Liu, J. Kou, Z. Yan, et al., "Enhancing XRF sensor-based sorting of porphyritic copper ore using particle swarm optimization-support vector machine (PSO-SVM) algorithm," Int. J. Min. Sci. Technol., vol. 34, no. 4, pp. 545-556, 2024. https://doi.org/10.1016/j.ijmst.2024.04.002

Speech Emotion Recognition from Audio Data Using LSTM Model

Md. Mahbub-Or-Rashid[1], Akash Kumar Nondi[2], Abdullah Al Sadnun[3],
Md. Anwar Hussen Wadud[4], T M Amir Ul Haque Bhuiyan[5], Md.Saddam Hossain[6]
Department of CSE, Bangladesh University of Business and Technology, Dhaka, Bangladesh[1, 2, 3, 4, 5, 6]
Department of CSE, Sunamgonj Science and Technology University (SSTU), Sunamganj, Bangladesh[4]

***Abstract*—The capacity to comprehend and interact with others through language is the most valuable human ability. Since emotions are crucial to communication, we are well-trained to recognize and interpret the many emotions we encounter. Contrary to popular assumption, the subjective aspect of human mood makes emotion recognition difficult for computers. There are some works based on Emotion recognition using images, text, and audio. We are here working on the audio dataset to find the accurate human emotion for computers to understand. In this work, we have utilized a Long Short-Term Memory (LSTM) model to implement Speech Emotion Recognition (SER) from Audio data on two different datasets: the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The accuracy rates of our LSTM-based model were impressive, with 91.25% for the RAVDESS dataset and 98.05% for the TESS dataset; the combined accuracy for both datasets was 87.66%. These results highlight the effectiveness of the LSTM model in effectively identifying and categorizing emotional states from audio files. The study adds significant knowledge to the field of speech emotion recognition by emphasizing the model's ability to handle a variety of datasets and its potential.**

Keywords*—Emotion; audio data; Ryerson audio-visual database; Toronto emotional speech set; classification; layers; combine

I. Introduction

Multimedia pattern recognition, especially Speech Emotion Recognition (SER), faces challenges due to emotional complexity. Deep learning simplifies the process by automatically learning features, unlike traditional methods requiring manual optimization [1]. Emotion recognition is essential for human-machine interaction but challenging due to fuzzy boundaries, individual expression, and overlapping emotions. Popular models like the circumflex of affect use dimensions like arousal and valence. Deep Neural Networks (DNNs), including CNNs and LSTMs, have advanced emotion recognition by leveraging features like Mel-Frequency Cepstral Coefficients (MFCCs). Early studies used handcrafted features, but this study proposes an end-to-end convolution-recurrent neural network that outperforms previous models on the RECOLA database [2]. This study explores speech emotion recognition (SER) using deep learning, focusing on Bangla and English datasets. The SUBESCO (Bangla) and RAVDESS (English) corpora were used, with mel-spectrograms as input features. A novel architecture, DCTFB (Deep CNN with Time-distributed Flatten and BLSTM layers), was proposed, achieving 86.86% accuracy on SUBESCO and outperforming existing models. The architecture combines local and sequential feature extraction for improved emotion detection. Additionally, a cross-lingual analysis using transfer learning and multi-corpus training was conducted, highlighting its effectiveness for SER across languages [3]. This study focuses on enhancing speech emotion recognition (SER) by addressing cost complexity and performance limitations. A lightweight CNN model with rectangular kernels and a modified pooling strategy is proposed to efficiently analyze frequency features in speech spectrograms. The model reduces parameters while maintaining high recognition accuracy, as demonstrated on IEMOCAP and EMO-DB datasets. Key contributions include improved computational efficiency and robust performance compared to baseline models, advancing SER for real-time applications in human-computer interaction [4].

Technology continues to evolve, significantly impacting human life and society. In human-machine interaction, speech is a key medium for communication, but recognizing emotions accurately remains a challenge for machines. While humans can easily detect emotions, machines struggle to interpret emotional states from speech. Speech, being the most natural form of communication, offers a fast and efficient method for interaction. Despite advancements in speech recognition, achieving natural interplay between humans and machines is still a work in progress. Speech emotion recognition (SER) holds potential for applications like customer service, education, forensics, medical analysis, and aiding individuals with disabilities. Integrating emotional recognition into voice assistants like Alexa, Siri, Google Assistant, and Cortana could make them more user-friendly and intuitive [5]. Accurately identifying emotional states in speech enhances human-machine interaction by enabling machines to respond appropriately. Speech emotion recognition (SER) has applications in areas like translation, education, customer service, and assistive technologies. Emotions can be classified into primary categories (e.g., Joy, Anger) or understood through cognitive perspectives. SER systems focus on extracting emotional cues from speech, often combined with facial expressions or biological features, to improve recognition accuracy. This study reviews SER features, systems, datasets, and experimental results, highlighting advancements in emotional speech recognition [6]. Deep learning enhances Speech Emotion Recognition (SER) by using CNNs for features and LSTMs for sequences. Turn-based SER often outperforms frame-based methods. This

study achieves state-of-the-art results on the IEMOCAP database, exploring deep learning models for simpler, more robust SER systems [7].

Speech Emotion Recognition (SER) detects emotions in speech for applications like human-computer interaction and robotics. It involves emotion modeling, preprocessing, feature extraction, and classification using deep learning. The study reviews these methods and current challenges in SER [8]. Speech Emotion Recognition (SER) helps improve human-computer interaction by recognizing emotions in speech. It has applications in smart speakers, virtual assistants, online education, therapy, safety, and customer service. Deep learning has significantly improved SER accuracy, surpassing traditional methods like HMMs, GMMs, and SVMs. SER training datasets can be natural, semi-natural, or simulated. Recent advancements include the use of recurrent neural networks (RNNs), LSTM, and auto-encoders. The study reviews existing datasets, traditional and deep learning methods, and suggests future directions for SER research [9]. Automatic Speech Emotion Recognition (SER) enables computers to understand human emotions through speech, enhancing natural communication. It uses methods like SVM and NN for classification, with the Berlin Emotional Database for training. SER is applied in areas like psychiatric diagnosis, educational software, and call center monitoring [10].

In [11], the authors focus on improving human-machine interaction by developing methods to identify emotions in speech. While emotions are often conveyed through voice, gestures, and body language, the study explores using Mel-recurrence Cepstral Coefficients to analyze sound for emotion recognition. Emotion recognition has broad applications in fields like brain research, psychiatry, and neuroscience, as well as in practical uses such as smart devices, customer service, and self-driving cars. The study discusses three types of training datasets for Speech Emotion Recognition: natural, semi-natural, and simulated. The key contributions outlined in the study are as follows:

- The CNN model achieved a 97.1% accuracy rate in categorizing emotional content of audio files.
- The deep learning models, particularly CNN, outperform the machine learning models in terms of accuracy, with an overall accuracy of 0.97. The sad class had the most outstanding performance, with a score of 0.90, and the worst on the surprised and disgust class (0.77).
- Using the RAVDESS and TESS datasets, proposed methods based on deep neural networks and machine learning to categorize emotions.
- Achieved an accuracy of 0.92, 0.93, and 0.971, respectively, using the deep learning model long short-term memories (LSTMs), GRU, and CNN.

Speech Emotion Recognition (SER) focuses on decoding emotions in spoken language using machine learning and AI. It has applications in areas like human-computer interaction and mental health diagnostics. The field explores how technology can understand and interpret emotional nuances in speech, offering insights into both technical methods and real-world impacts.

The remainder of this study is structured as follows**:** Section II provides the background and a detailed review of related works in the field of Speech Emotion Recognition (SER), focusing on both conventional and deep learning approaches. Section III outlines the methodology adopted in this research, including model architecture, data preprocessing techniques, and feature extraction processes. It discusses the datasets used in this study, specifically the RAVDESS and TESS datasets, along with data distribution and preprocessing details. It also explains the implementation details of the proposed LSTM model. Section IV presents the results, performance evaluations, and comparative analysis with existing methods. Finally, Section V concludes the study by highlighting the findings and proposing directions for future research.

II. Background Study and Related Works

A. Background

Using LSTM's prediction model along with machine learning and deep learning techniques to build the network training structure as well as make predictions. This underscores the objective of exploiting deep learning techniques developed through performance monitoring systems to compensate for incorrect predictions during the prediction process. There are a number of relevant sources cited in this study. The first source reviews the current literature relative to various theories and models. The second source emphasizes the realism of the research and illustrates various situations, including case studies and real events. In Source 3, the authors compare the various strategies previously employed with an assessment of their advantages and disadvantages. Various studies combine to form a strong foundation for this research. They reveal what is currently known about the topic and highlight the issues that will be addressed in this research [12].

B. Literature Survey

Limited progress in Bengali emotion recognition for sentiment analysis and abusive text detection. Methods include Chinese quarantine emotionally charged hotel reviews, and using machine learning for Covid-related content in Bangladesh. An example of such efforts is a Bengali article-based model for emotion detection as well as an advanced Bengali text annotation dataset for advanced emotion recognition [13]. In a cross-linguistic study involving Chinese and German, it was demonstrated that pitch, speech power and MFCC parameters are crucial for speech emotion detection. Alternatively, feature selection methods such as submodular functions and sequential forward selection are used with the Open Smile toolkit to extract a new feature set. The use of T-SNE revealed enhanced performance mainly in combining spectrogram and acoustic properties. Interestingly, this work specifically focuses on emotive character optimization for Bengali, thereby distinguishing it from previous studies [14]. Nevertheless, there are several issues related to speech emotion recognition, including the development of modern and advanced models leading to improved recognition that surpasses the state-of-the-art, as shown below. The modified

pooling technique combined with rectangular filters produces satisfactory performance with higher results than IEMOCAP (77.01%) and EMODB (92.02%). Frequency features are important SERs and suggested research directions for testing and deep learning in different databases [15]. Speech emotion recognition is reviewed in terms of different classifier models such as KNN, HMM, SSV, ANN and GMM. It uses various parameters like power, pitch, LPCC, etc., to recognize emotions. However, in capturing boundary nonlinearity, ANN, especially feedforward networks, show better results than GMM with the highest accuracy of 78.77%. HMM excels in temporal modeling, outperforms other classifiers, and SVM achieves perfect classification in multidimensional feature domains using different types of kernels [16]. A systematic review of Speech Emotion Recognition (SER) based on deep and traditional learning methods using available data sets. Some previous studies, such as Swain et al., 2018 and Khalil et al., 2019, were limited to conventional techniques or independent deep learning techniques, providing a shallow insight. However, Akcay et al. 2020 have been thoroughly researched on database, feature, classification and sentiment models applied to SER with emphasis on machine learning techniques [17]. Recognizing emotion from speech using basic technology for speech recognition. This study builds on previous studies showing that statistics of speech components (pitch, power, intonation, spectral shape, etc.) are associated with specific emotions. The results of the subjective assessment conducted using Interface Emotional Speech Synthesis are impressive - more than 80%. The work focused on low-level characterization and system design that could capture over eighty per cent recognition rates for seven emotions. Future studies are described, such as testing speaker/language-independent conditions and multi-modal emotion recognition [18]. To classify speech emotions (sadness, anger, fear and happiness), the study uses support vector machines after feeding features such as energy; MFCC coefficients (0 to 12) of Cocke-Younger algorithm output spectrum based on linear predictive coding model with pitch window size 35. Two classification methods, one-versus-all (OAA) and sex-based classification, are compared for females-only LPCC algorithm. As MFCC, time domain speech signal and various feature waveforms are extracted by research feature analysis. What's more, there are two datasets in the setting for testing - UGA (University of Georgia) and LDC (Linguistic Data Consortium). Among them are segmented recordings as well as student speech samples [19]. A detailed literature review on the effect of model size and pre-training data on downstream performance in Speech Emotion Recognition (SER) of the Transformer architecture. It compares different pre-trained variants of the wav2vec 2.0 and HuBERT models on arousal, dominance and valence dimensions across the MSP-Podcast dataset as well as the IEMOCAP and MOSI datasets. The main findings of the study include: that transformer models perform state-of-the-art in IEMOCAP; investigating why they are successful in improving valence spacing (excess avoidance), robustness and fairness, and efficiency. Authors publish their most successful models to allow the community to reproduce them [20]. By comparing Wav2Vec 2.0 for SER with V-FT and TAPT as baselines, we show that the latter outperforms TAPT in both accuracies on the IEMOCAP dataset using a novel approach known as P-TAPT. As it excels in low-resource settings and overall performance. The experimental framework is to evaluate SER on IEMOCAP and SAVEE datasets, taking frame-level emotion information into account and performing k-means clustering for better results. Research has pointed out that SER plays an important role in human-machine interaction and communication systems. It also shows how self-supervised pre-trained models can compensate for the lack of data annotation often associated with deep learning-based systems [21]. Furthermore, recent contributions have focused on leveraging machine learning and deep learning not only in emotion recognition but also in adjacent biomedical domains. For instance, Nurjahan et al. [22] present a comprehensive review of machine learning and deep learning algorithms in detecting COVID-19 through medical imaging, highlighting transferable modeling strategies and feature selection methods relevant to emotion recognition systems. Similarly, Bhuiyan et al. [23] developed a UNet-based deep learning segmentation framework for brain tumor detection, demonstrating the effectiveness of customized architectures in extracting domain-specific patterns — a methodology that may also enrich future SER frameworks tailored for low-resource languages like Bengali. In recent years, various approaches have been proposed to enhance emotion recognition from speech using deep learning and signal processing techniques. To enhance emotion recognition from speech using deep learning and signal processing techniques for speech emotion recognition (SER), demonstrating impressive accuracy rates on various datasets to find more accurate results using machine learning and deep learning algorithms, but their results are still not satisfactory. In another domain, the authors developed PCA-SIFT and weighted decision tree-based methods for diabetic retinopathy detection, emphasizing the importance of hybrid feature extraction and classification approaches for medical imaging applications [24]. These techniques can be conceptually adapted for SER feature extraction and classification. Additionally, their work on automatic water pump control using IoT [25] and an Android-based travel guide application [26] shows their practical expertise in building responsive systems—skills transferable to real-world emotion-aware systems. In the field of cloud computing, Mahbub-Or-Rashid et al. proposed resource-efficient frameworks for energy optimization and load balancing in cloud data centers [27][28], which are relevant when deploying computationally intensive SER models in scalable environments. The study on radio signal fading models [29] highlights the team's proficiency in signal analysis, a critical component in audio-based emotion recognition. Further contributions include a UNet-based deep learning framework for brain tumor segmentation [30] and a comprehensive review of ML/DL approaches for COVID-19 detection [31], both of which underscore expertise in medical image processing and transferable deep neural architectures. The development of Grainbee, a quantum-resistant blockchain system [32], and the FruVeg_MultiNet IoT system [33] reflect capabilities in secure, intelligent systems, suggesting possibilities for secure, privacy-preserving SER systems in the future. Moreover, the application of multi-output regression to predict death counts based on short-term mortality data [34] demonstrates

competence in statistical modeling, which could be integrated into emotion progression tracking in long-form audio content.

III. METHODOLOGY

The research presents a Speech Emotion Recognition (SER) model developed using Long Short-Term Memory (LSTM) networks. Leveraging LSTM's sequential learning capabilities, the model effectively identifies complex emotional patterns in audio data. It is designed for adaptability and generalization across datasets, specifically utilizing RAVDESS and TESS. The study emphasizes the model's contribution to SER research while avoiding detailed training metrics. Fig. 1 illustrates the data preprocessing and model development workflow.

A. Dataset

We use two well-known datasets, the Toronto Emotional Speech Set and RAVDESS from Kaggle, to perform a thorough analysis and pre-processing of the data. The Toronto Emotional Speech Set is a collection of emotionally charged speech recordings, and RAVDESS is a stable dataset containing a variety of emotional expressions. In order to capture relevant acoustic characteristics associated with emotional speech, the proposed model uses a thorough approach to data analysis and preprocessing in the domain of Speech Emotion Recognition (SER) from audio data using Long Short-Term Memory (LSTM) networks. A key component of this strategy is the use of Mel Frequency Cepstrum Coefficient (MFCC) as a feature extraction method. A comprehensive approach to data analysis and pre-processing highlights the model's ability to detect subtle nuances in emotional speech patterns, adding to the overall efficacy of the LSTM-based Speech Emotion Recognition system. The pre-processing phase carefully evaluates two diverse datasets, RAVDESS and TESS, to guarantee the model's flexibility and generalization across a range of emotional expressions. The Mel Frequency Cepstral Coefficients function as an essential set of features, offering a compact representation of the spectral characteristics in the audio signals. We have here 1200 audio data for RAVDESS and 1800 data for the TESS dataset. And the combined data is 3000. In Fig. 2 below, we can see the class distribution for all dataset.

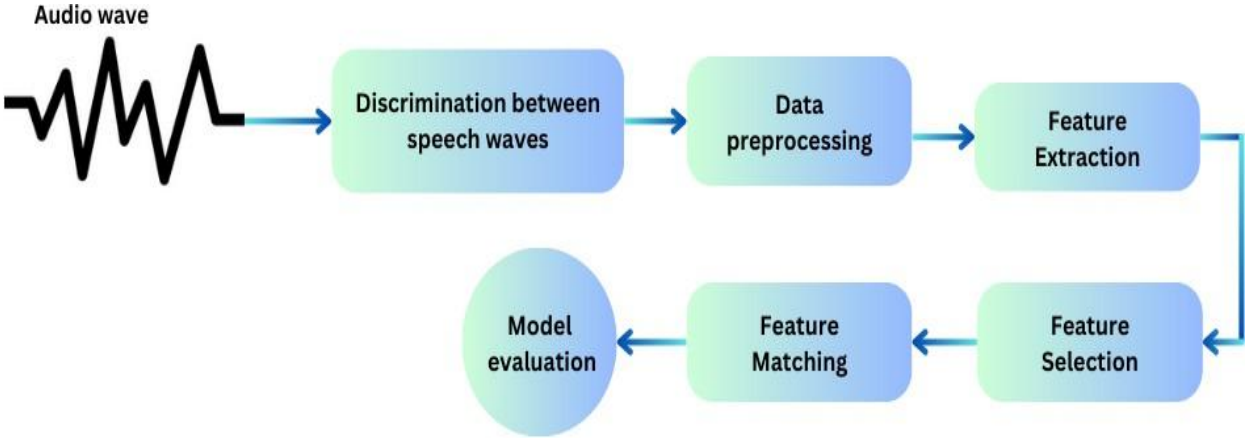
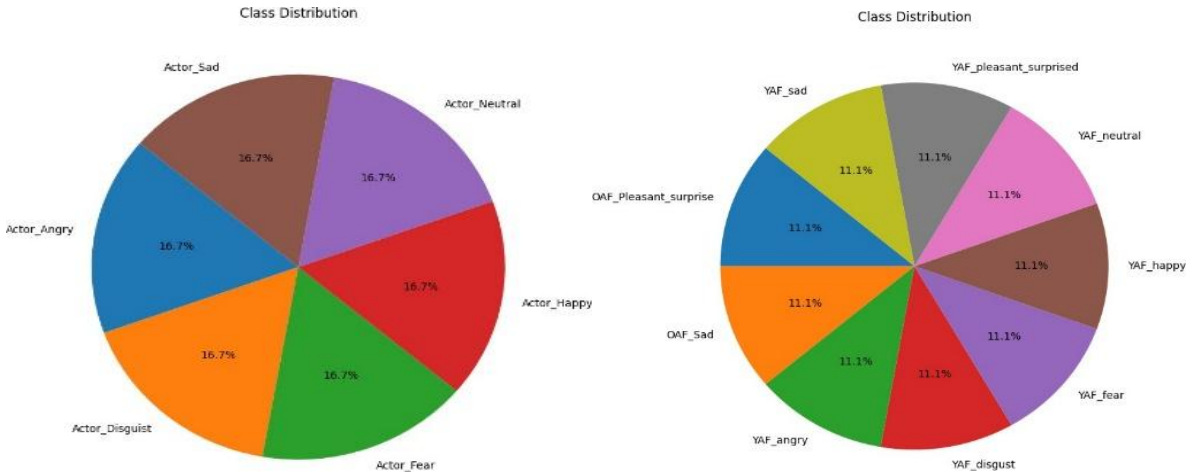


Fig. 1. Data preprocessing and model development workflow.



(a) RAVDESS. (b) TESS.

Fig. 2. Distribution of the dataset.

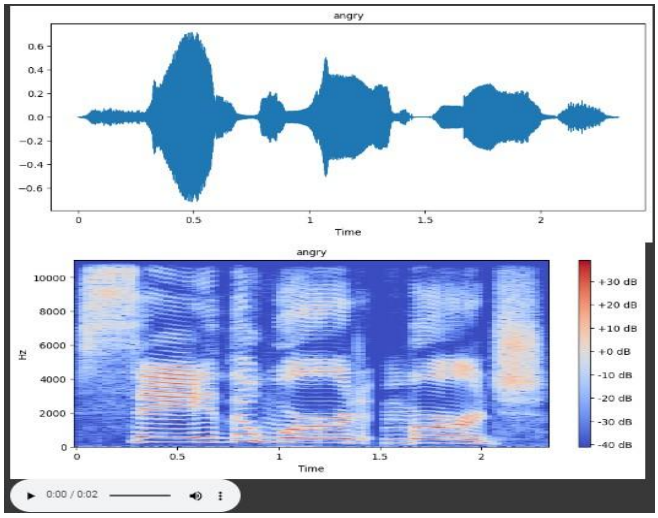


Fig. 3. Sample dataset of sad emotion.

B. Data Pre-processing

First, look at the audio signals as waveforms, which show how the sound changes over time. The "sad" waveform shows low energy and smooth changes (see Fig. 3), while the "angry" waveform has higher energy and sharp changes, showing strong emotions (see Fig. 4). Next, we transform the audio signals into spectrograms, which show the different frequencies in the sound over time. To improve the quality of the data, we may apply processes like noise reduction and normalization. The 'sad' spectrogram shows energy in lower frequencies with softer levels, while the 'angry' spectrogram has energy spread across a wider range of frequencies and higher intensities. These improved spectrograms help us better to capture the important patterns in the sound needed to identify emotions.

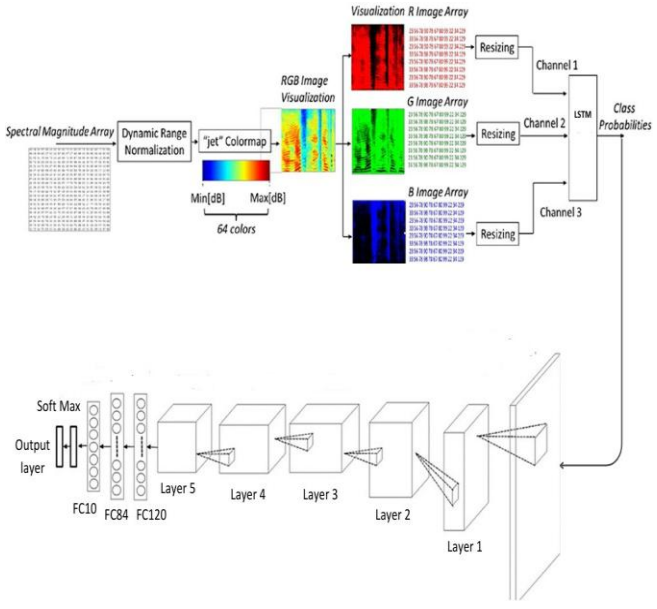


Fig. 4. Sample dataset of angry emotion.

Model implementation in our model for Speech Emotion Recognition begins by processing audio input into a spectral magnitude array, which represents the frequency content of the audio signal over time. To make the data consistent and enhance its usability, dynamic range normalization is applied. Next, the normalized spectrogram is transformed into an RGB image using the "jet" colormap, which assigns color values to the spectrogram, highlighting its features visually. The resulting RGB image is then split into three separate channels (Red, Green, and Blue), which are resized to a standard shape for further processing. These resized channels serve as input to an LSTM (Long Short- Term Memory) network (see Fig. 5). LSTM layers are specifically designed to capture sequential patterns in data, making them ideal for analyzing the temporal dynamics present in speech signals. After the LSTM processes the temporal information, the extracted features are passed through three fully connected (dense) layers (FC120, FC84, and FC10). These layers refine and condense the learned features to prepare them for classification. Finally, the output layer uses a softmax activation function to produce probabilities for each emotion class (e.g., happiness, sadness, anger), allowing the model to classify the speaker's emotional state based on their speech. This approach effectively combines audio preprocessing, visual representation, and deep learning to analyze and classify emotions in speech data. Comparative studies clarify the differences in performance between conventional machine learning algorithms and the LSTM model. This thorough investigation seeks to clarify the subtle benefits of utilizing deep learning architectures in SER, offering important information for the development of emotion-aware technology and (Human-Computer Interaction) HCI systems.

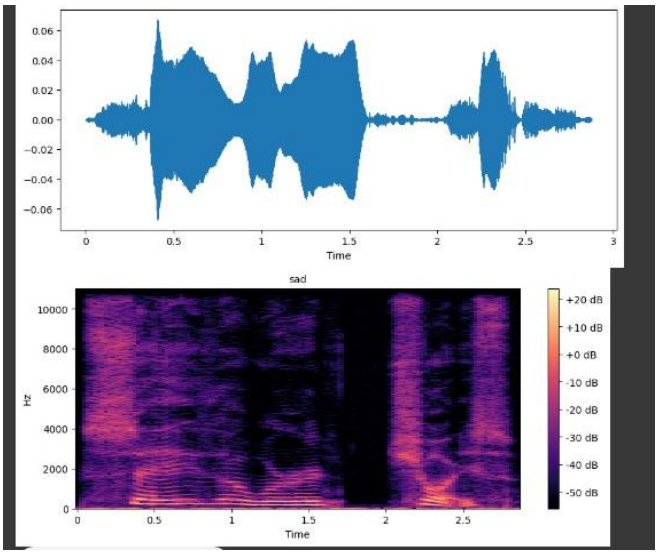


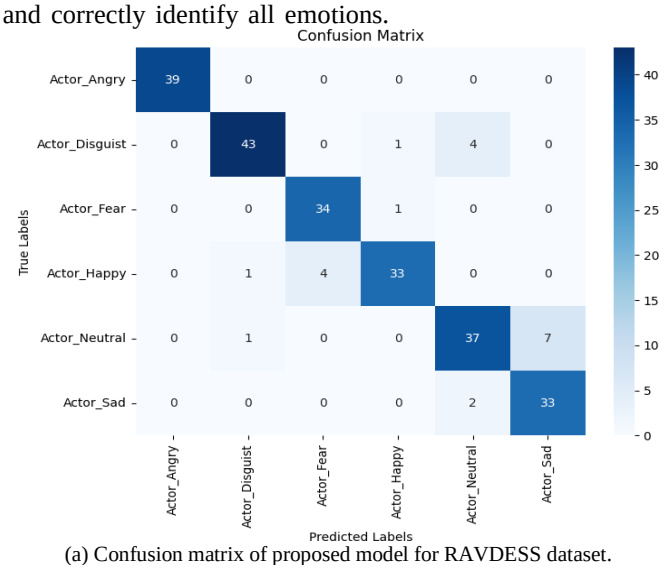
Fig. 5. Proposed Long Short-Term Memory (LSTM) model architecture.

IV. Result Analysis and Discussion

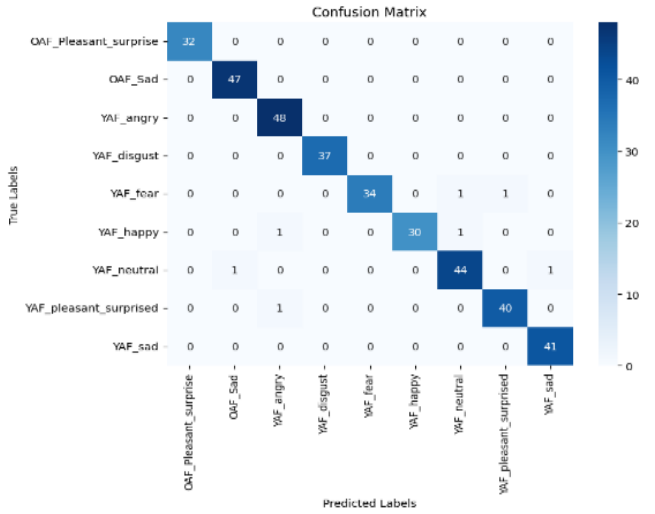
A. Result Analysis

1) Evaluation metrics: An illustration of a categorization process called a confusion matrix illustrates how closely the model's predictions match the actual effects (see Fig. 6). We used a total of 240 voice data for RAVDESS, 360 voice data for TESS and 600 voice data for combined RAVDESS and TESS to test the proposed model LSTM. Fig. 6(a), here 240

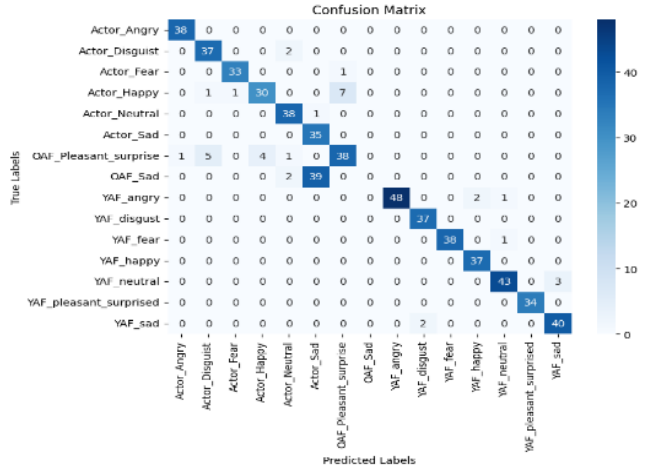
voice data from RAVDESS, Fig. 6(b), here 360 voice data from TESS and Fig. 6(c), here 600 voice data from combined RAVDESS and TESS to test the proposed model LSTM and correctly identify all emotions.



(a) Confusion matrix of proposed model for RAVDESS dataset.



(b) Confusion matrix of proposed model for TESS dataset.



(c) Confusion matrix of proposed model for combined RAVDESS and TESS dataset.

Fig. 6. Confusion matrix of proposed model.

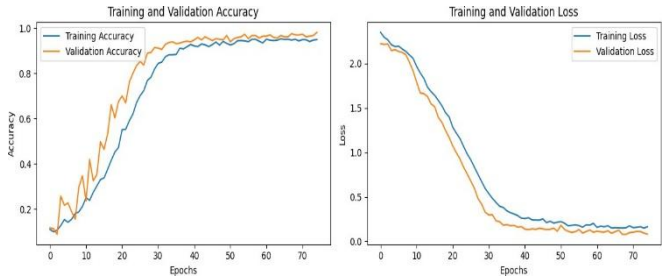


Fig. 7. Accuracy of the proposed LSTM model for RAVDESS dataset.

B. *Performance Evaluation*

The calculation was done using the Keras callbacks method. While experimenting with various epoch counts, we measured the precision for both training and validation. The model reaches its peak accuracy in training, testing, and verification after 75 epochs (see Fig. 7 to Fig. 9). In the graphs, these results could be attributed to the rich and diverse nature of the TESS and RAVDESS datasets, which have contributed to the model's robust training. These outcomes highlight the effectiveness of the model in accurately classifying emotions within speech signals. The high accuracy rates indicate the successful capture of temporal dependencies.



Fig. 8. Accuracy of the proposed LSTM model for TESS dataset.

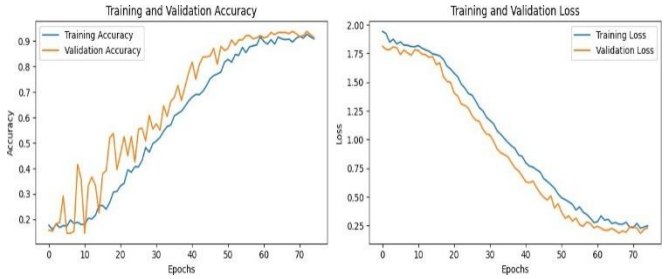


Fig. 9. Accuracy of the proposed LSTM model for combined RAVDESS and TESS dataset.

This study discusses the performance of the LSTM model that has been implemented and trained. Data collected from the RAVDESS and Toronto Emotional Speech Set (TESS) dataset, which includes 1200 voice data of RAVDESS and 1800 voice data of TESS, were used to test the model presented in this research. Given the skewness of the dataset, we went beyond simple classification accuracy to assess model performance by using other metrics, such as precision, sensitivity, recall, and F1-score, which incorporates the metrics with Fig. 10.

Classification Report:

	precision	recall	f1-score	support
Actor_Angry	1.00	1.00	1.00	39
Actor_Disguist	0.96	0.90	0.92	48
Actor_Fear	0.89	0.97	0.93	35
Actor_Happy	0.94	0.87	0.90	38
Actor_Neutral	0.86	0.82	0.84	45
Actor_Sad	0.82	0.94	0.88	35
accuracy			0.91	240
macro avg	0.91	0.92	0.91	240
weighted avg	0.92	0.91	0.91	240

Number of training samples: 960
Number of testing samples: 240

(a) Classification report of proposed model for RAVDESS dataset.

Classification Report:

	precision	recall	f1-score	support
OAF_Pleasant_surprise	1.00	1.00	1.00	32
OAF_Sad	0.98	1.00	0.99	47
YAF_angry	0.96	1.00	0.98	48
YAF_disgust	1.00	1.00	1.00	37
YAF_fear	1.00	0.94	0.97	36
YAF_happy	1.00	0.94	0.97	32
YAF_neutral	0.96	0.96	0.96	46
YAF_pleasant_surprised	0.98	0.98	0.98	41
YAF_sad	0.98	1.00	0.99	41
accuracy			0.98	360
macro avg	0.98	0.98	0.98	360
weighted avg	0.98	0.98	0.98	360

Number of training samples: 1440
Number of testing samples: 360

(b) Classification report of proposed model for TESS dataset

Classification Report:

	precision	recall	f1-score	support
Actor_Angry	0.97	1.00	0.99	38
Actor_Disguist	0.86	0.95	0.90	39
Actor_Fear	0.97	0.97	0.97	34
Actor_Happy	0.88	0.77	0.82	39
Actor_Neutral	0.88	0.97	0.93	39
Actor_Sad	0.47	1.00	0.64	35
OAF_Pleasant_surprise	0.83	0.78	0.80	49
OAF_Sad	0.00	0.00	0.00	41
YAF_angry	1.00	0.94	0.97	51
YAF_disgust	0.95	1.00	0.97	37
YAF_fear	1.00	0.97	0.99	39
YAF_happy	0.95	1.00	0.97	37
YAF_neutral	0.96	0.93	0.95	46
YAF_pleasant_surprised	1.00	1.00	1.00	34
YAF_sad	0.93	0.95	0.94	42
accuracy			0.88	600
macro avg	0.84	0.88	0.86	600
weighted avg	0.84	0.88	0.85	600

Number of training samples: 2400
Number of testing samples: 600

(c) Classification report of proposed model for combined RAVDESS and TESS dataset

Fig. 10. Classification report of proposed model.

1) Comparison with existing works: The classification outcomes of conventional machine learning techniques are shown in Table I, alongside our suggested LSTM model. Our proposed model outperforms the other classifiers with 98.05% and 91.25% and a combine result of 87.66% accuracy compared to other ML classifiers mentioned.

TABLE I. PERFORMANCE COMPARISON WITH THE EXISTING METHODS

Methods	Algorithm	Dataset	Classification Accuracy
Issa et al. [11]	CNN	RAVDESS	71.61%
Zeng et al.[15]	CNN+LSTM	RAVDESS	64.48%
Zamil et al. [19]	GRU	RAVDESS	67.14%
Hashemzehi et al. [17]	SVM	RAVDESS	77.32%
Praseetha et al. [21]	CNN	TESS	95.82%
Ravi et al.[8]	CNN	TESS	97.1%
Huang et al. [19]	MLP	TESS	85%
Dupuis et al. [5]	LSTM	TESS	82%
Proposed	**LSTM**	RAVDESS	**91.25%**
Proposed	**LSTM**	TESS	**98.05%**
Proposed	**LSTM**	Combined RAVDESS and TESS	**87.66%**

This study demonstrates the impressive performance of the LSTM model in Speech Emotion Recognition (SER), achieving accuracy rates of 91.25% and 98.05% on the RAVDESS and TESS datasets, respectively, and 87.66% on the combined dataset. The model's ability to capture temporal dependencies and identify emotional nuances in speech signals highlights its effectiveness. Future research can focus on improving model generalization with larger, diverse datasets, integrating real-world scenarios, exploring ensemble or hybrid architectures, applying transfer learning, and enhancing robustness across languages, thus advancing SER methodologies.

V. CONCLUSION

In summary, we applied LSTM models for Speech Emotion Recognition (SER) using the RAVDESS and TESS datasets, achieving strong accuracy rates of 91.25%, 98.05%, and 87.66% for the individual and combined datasets. The results demonstrate the LSTM model's ability to effectively capture the emotional subtleties in speech and recognize temporal patterns within the data. By training on diverse and comprehensive datasets, the model was able to generalize well and deliver accurate emotion classifications. This research emphasizes the practical potential of LSTM models in real-world SER applications, such as human-computer interaction, customer service, and healthcare. Our findings provide a solid foundation for further development of emotion recognition systems that can be used across different domains, paving the way for more efficient and accurate emotion detection in speech.

REFERENCES

[1] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[2] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[3] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2021.

[4] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, 2020.

[5] J. Devnath, S. Hossain, M. Rahman, H. Saha, M. A. Habib, and N. Sultan, "Emotion recognition from isolated bengali speech," 2020.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[8] M. B. Akc¸ay and K. Og˘uz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[9] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[10] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[11] R. R. Choudhary, G. Meena, and K. K. Mohbey, "Speech emotion based sentiment recognition using deep neural networks," in *Journal of Physics: Conference Series*, vol. 2236, no. 1. IOP Publishing, 2022, p. 012003.

[12] M. Ahmed, P. C. Shill, K. Islam, M. A. S. Mollah, and M. Akhand, "Acoustic modeling using deep belief network for bangla speech recognition," in *2015 18th international conference on computer and information technology (ICCIT)*. IEEE, 2015, pp. 306–311.

[13] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. G. Hussain, "Bangla text emotion classification using lr, mnb and mlp with tf-idf & countvectorizer," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*. IEEE, 2022, pp. 275–280.

[14] S. Sultana and M. S. Rahman, "Acoustic feature analysis and optimization for bangla speech emotion recognition," *Acoustical Science and Technology*, vol. 44, no. 3, pp. 157–166, 2023.

[15] R. D. G. Ayon, M. S. Rabbi, U. Habiba, and M. Hasana, "Bangla speech emotion detection using machine learning ensemble methods."

[16] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[17] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[18] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marin˜o, "Speech emotion recognition using hidden markov models," in *Seventh European conference on speech communication and technology*, 2001.

[19] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[21] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] Nurjahan, Md. Mahbub-Or-Rashid, Md. Shahriare Satu, Sanjana Ruhani Tammim, Farhana Akter Sunny, Mohammad Ali Moni "Machine learning and deep learning algorithms in detecting COVID-19 utilizing medical images: a comprehensive review" Iran Journal of Computer Science, 2024.Springer Nature.

[23] TM Amir-Ul-Haque Bhuiyan,Md Anwar Hussen Wadud, Md. Mahbub-Or-Rashid, Md. Reazul Islam "Brain Tumor Detection by Image Segmentation Using Customized UNet Deep Learning Based Model" 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT).

[24] F. T. Johora, M. Mahbub-Or-Rashid, M. A. Yousuf, T. R. Saha, and B. Ahmed, "Diabetic retinopathy detection using PCA-SIFT and weighted decision tree," in *Proc. Int. Joint Conf. on Computational Intelligence*, 2020.

[25] Md. Reazul Islam and M. Mahbub-Or-Rashid, "Design and Implementation of Cost Effective Automatic Water Pump Controlling System for Domestic Application using IoT," in *Proc. Conf. on Big Data, IoT and Machine Learning (BIM)*, 2023.

[26] J. Akther, M. Harun-Or-Roshid, M. Mahbub-Or-Rashid, and S. J. Soheli, "Bangladesh Travel Guide (BTG): An Android Mobile Application," *J. Adv. Res. Mobile Comput.*, vol. 3, no. 1, pp. 1–5, 2021.

[27] J. A. Jeba, S. Roy, M. O. Rashid, S. T. Atik, and M. Whaiduzzaman, "Towards green cloud computing: an algorithmic approach for energy minimization in cloud data centers," in *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing*, IGI Global, pp. 1567–1589, 2018.

[28] M. Mahbub-Or-Rashid, M. J. N. Mahi, J. A. Jeba, and F. T. Johora, "Achieving an efficient approach through using resource allocation, management and load balancing for cloud data centers," *Recent Trends in Cloud Computing and Web Engineering*, vol. 3, no. 1, pp. 1–23, 2021.

[29] L. Nahar, M. Mahbub-Or-Rashid, S. Akter, and R. T. Khan, "Medium Access Probability of Cognitive Radio Network Under ECC-33/Hata-Okumura Extended Model Using Different Fading Channels at 1900MHz and 2100MHz," *Int. J. Comput. Eng. Res. (IJCER)*, vol. 6, no. 7, pp. 35–42, 2016.

[30] T. M. Amir-Ul-Haque Bhuiyan, M. A. H. Wadud, M. Mahbub-Or-Rashid, and M. Reazul Islam, "Brain Tumor Detection by Image Segmentation Using Customized UNet Deep Learning Based Model," in *Proc. 6th Int. Conf. on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024.

[31] Nurjahan, M. Mahbub-Or-Rashid, M. S. Satu, S. R. Tammim, and F. A. Sunny, "Machine learning and deep learning algorithms in detecting COVID-19 utilizing medical images: a comprehensive review," *Iran J. Comput. Sci.*, vol. 4, no. 1, pp. 23–45, 2024.

[32] S. A. Joni, R. Rahat, N. Tasnin, P. Ghose, and M. Mahbub-Or-Rashid, "Grainbee: A Quantum-Resistant Blockchain-Based Ration Distribution System with Hardware Security Modules," in *Proc. IEEE Int. Conf. on Computing, Applications and Systems (ICCAS)*, 2024.

[33] T. M. A. U. H. Bhuiyan, M. Reazul Islam, J. I. Babar, M. A. Nur, and M. Mahbub-Or-Rashid, "FruVeg_MultiNet: A hybrid deep learning-enabled IoT system for fresh fruit and vegetable identification with web interface and customized blind glasses for visually impaired people," *J. Agric. Food Res.*, vol. 1, pp. 100–112, 2025.

[34] M. I. Ahmed, Nurjahan, M. Mahbub-Or-Rashid, and F. Islam, "Prediction of Death Counts Based on Short-term Mortality Fluctuations Data Series using Multi-output Regression Models," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 14, no. 5, pp. 88–94, 2023.

AI-Driven Textual Feedback Analysis in E-Training Using Enhanced RoBERTa

Rakan Saad Alotaibi, Fahad Mazyed Alotaibi, Sameer Abdullah Nooh, Abdulaziz A. Alsulami
Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In corporate e-training environments, traditional metrics like course completion and quiz scores often fail to reflect actual job performance. Rich insights are embedded in unstructured textual feedback, yet they remain underutilized due to limitations in existing analytical models. This study proposes E-RoBERTa, an enhanced transformer-based model designed to predict employee job performance by analyzing open-ended feedback from digital training platforms. The model aims to improve accuracy, domain adaptability, and interpretability. E-RoBERTa integrates Domain-Adaptive Pretraining (DAPT) to fine-tune RoBERTa on corporate-specific language and introduces Dynamic Attention Scaling (DAS) to highlight semantically critical tokens. A real-world, GDPR-compliant dataset containing 16,000 feedback entries from 3,500 employees across multiple departments was used. Preprocessing included tokenization, sentiment tagging, and feature extraction. The model achieved superior performance with a macro F1-score of 0.875, outperforming standard RoBERTa, LSTM, and SVM baselines. Attention visualizations revealed alignment between influential tokens and human-interpretable performance indicators. E-RoBERTa provides a transparent and accurate framework for evaluating job performance through textual feedback. Its use of domain adaptation and dynamic attention mechanisms supports scalable, ethical, and explainable AI in corporate learning analytics, offering actionable insights for personalized interventions and strategic HR decision-making.**

Keywords*—*Job performance prediction; transformer models; enhanced RoBERTa; domain-adaptive pretraining (DAPT); dynamic attention scaling (DAS); natural language processing (NLP); explainable AI; textual feedback analysis; workforce analytics

I. Introduction

In the last few years, e-training platforms have been indispensable tools for professional development in various industries. With organizations moving towards implementing digital learning solutions for upskilling employees and increasing productivity, there has been a significant demand for effective training analytics. Conventional completion rates, quiz scores, and attendance logs shed little light on actual job performance outcomes [1]. These static indicators do not measure the dynamic and contextual nature of learning behavior that affects how knowledge is applied successfully at the workplace. The emergence of artificial intelligence (AI) and intense learning has shifted the methods of processing unstructured training data towards natural language processing (NLP) approaches. Employee feedback, assessment responses, and forum discussions are filled with information, which, if well mined, can provide patterns regarding job performance [2]. Transformer-based models such as BERT and RoBERTa have demonstrated excellent abilities in deriving contextual sense in text, compared to the previous approaches in several NLP tasks [3]. However, standard transformer models often lack domain-specific adaptation, resulting in inaccuracy when applied to corporate training data. Furthermore, they consider all text elements to be equally important, which can obscure critical signals contained in sophisticated employee responses. This study proposes an Enhanced RoBERTa model that addresses these gaps through Domain-Adaptive Pretraining (DAPT) and Dynamic Attention Scaling (DAS), enabling the capture of more suitable training feedback features. The motivation for this work is a need to develop more accurate, interpretable, and ethical tools for performance prediction in e-learning. Since businesses depend on data-driven decisions for the development of their employees, it is crucial to have AI systems that can extract actionable insights from training feedback. Our model aims to facilitate personalized learning, early interventions for employees struggling with their performance, and more effective talent management strategies.

Despite the rapid increase in the use of e-training platforms, organizations continue to face challenges in accurately measuring the impact of training on actual job performance in the real world. The existing evaluation methods, i.e., assessment scores and completion rates, do not indicate whether the employees can practically apply what they have learned in the workplace [4, 5]. These metrics tend to give a shallow learning perspective, disregarding deeper learning or long-term retention. Additionally, the predictive models used in training analytics have a limited scope. Many still rely on manually chosen features and structured data, which overlook the valuable information hidden in textual content, such as open-ended feedback, peer discussions, and reflective assessments [6]. These unstructured inputs contain contextual hints on employee motivation, understanding, and satisfaction, factors that significantly determine performance outcomes. Such text has become a fertile ground for transformer-based NLP models, which have proven to be promising tools for deriving insights. However, existing off-the-shelf models, such as RoBERTa, are not tailored for domain-specific language, resulting in suboptimal performance when applied to corporate training data [3]. Moreover, these models do not differentiate between various tokens and treat them equally, without mechanisms to highlight the most relevant parts of the text that determine prediction. Another problem is the interpretability of deep learning models in human resource (HR) settings. Managers need explanations as to why a specific system speculates that a given employee will not perform to the expected level. Black-

box predictions erode trust and hinder deployment, particularly when the decision requires fairness and transparency [7, 8]. This study fills these gaps by creating an Enhanced RoBERTa model specifically for corporate training data. It uses DAPT and DAS to enhance the relevance and interpretability of performance prediction. With that, it strives to provide a more accurate and explainable tool for employee training outcomes based on textual feedback.

The primary objective of this study is to enhance the accuracy, transparency, and applicability of job performance predictions derived from textual comments in e-training environments. Achievement of this is done based on the following four objectives for the study:

- To design a transformer-based model capable of understanding domain-specific training language used in employee feedback, assessments, and learning reflections.
- To improve model attention mechanisms by introducing a dynamic scaling method highlighting contextually essential elements in training text.
- To evaluate the proposed model's effectiveness using a real-world corporate training dataset, comparing its performance to baseline machine learning (ML) and NLP models.
- To support explainable and ethical AI in workforce analytics by offering interpretable outputs that aid decision-making in employee development.

Based on these objectives, the study offers the following key contributions:

- A domain-adapted version of RoBERTa is developed using DAPT on corporate training documents, improving its contextual understanding in the HR and e-learning domains.
- The model incorporates a novel DAS mechanism that adjusts attention weights to emphasize semantically significant feedback. This improves both prediction accuracy and interpretability for human evaluators.
- The model is tested on GDPR-compliant training data with performance measured using accuracy, precision, recall, and F1-score. Benchmarks show that E-RoBERTa outperforms traditional NLP models and classical ML classifiers.
- Attention maps are visualized to show how the model interprets feedback, enabling HR managers to understand and trust AI-driven performance insights. This directly supports adoption in high-stakes organizational settings.

This study bridges the gap between generic NLP systems and the specific needs of performance prediction in professional e-training, offering a robust, interpretable, and ethically conscious approach that organizations can integrate into their learning analytics ecosystems.

This study is structured as follows: Section II reviews related work on e-training analytics, NLP in education, and transformer models in workforce analytics. Section III outlines the proposed methodology, including data preprocessing, model architecture, training procedures, and evaluation metrics. Section IV presents the results and discussion, comparing the proposed model with baseline methods. Finally, Section V concludes the study and suggests directions for future work.

II. Related Work

A. E-Training Analytics and Machine Learning Approaches

E-training platforms have evolved to become instrumental in organizational learning, offering scalable, flexible, and cost-effective alternatives to conventional classroom-based approaches. However, assessing the real impact of these programs on employees' job performance is a significant challenge. Many organizations still use shallow metrics, such as course completion rates and quiz scores, which do not often correspond with long-term workplace effectiveness [4]. This gap has been the target of recent developments in educational data mining and learning analytics [9]. These approaches aim to draw insights from digitized learning environments by observing learner behaviors, interactions, and outcomes [10]. Although the initial models employed rudimentary statistics and rule-based systems, they did not accurately describe the complex dynamics of learning or individualized training advice. ML brought a paradigm shift in predictive analytics training. Decision trees, support vector machines (SVMs), and random forest algorithms have been implemented in the prediction of learner success using structured features such as attendance, performance history, and level of engagement [11, 12]. Although these models can bring improvements in prediction accuracy, they still struggle to cope with unstructured data, particularly textual feedback and discussion content, which is rich in contextual information about learning processes. In the last few years, deep learning has received attention for its capability to model complex, nonlinear relationships in e-learning data. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been shown to reveal temporal patterns in learning behavior [13]. However, these models typically require large training datasets and suffer from issues such as the vanishing gradient. To enhance prediction and personalization in e-training, researchers have started using NLP methods. Textual data that have been observed to provide indicators of motivation, understanding, and satisfaction, all of which are job performance-related, have been found in feedback, reflections, and forum posts [14]. However, standard NLP techniques, such as bag-of-words or TF-IDF, do not adequately reflect the semantic depth and contextual nuances [15].

B. Advances in NLP and Transformer Models

Transformer-based models, such as BERT and RoBERTa, established a new benchmark for text understanding, as they rely on self-attention mechanisms to maintain the context throughout the whole sequence [16]. These models have been utilized in education to measure sentiment, grade essays, and predict course outcomes; however, their application in predicting job performance in corporate e-training remains minimal. This study extends these bases by introducing a domain-adapted

RoBERTa model with attention scaling, customized to predict job performance from training feedback. By paying closer attention to the linguistic patterns in employee reflections and evaluations, the proposed model aims to enhance the interpretability and reliability of AI-based training analytics. NLP has become a crucial component of intelligent educational systems, enabling more effective analysis of learner input and more adaptable instructional techniques. With the help of NLP techniques, written text in the form of learner feedback, discussion posts, reflections, and assessment responses can be processed to understand learners' cognitive and emotional states during training [16]. Simple lexical approaches, such as keyword extraction and sentiment scoring, were used in the early NLP applications in education. These approaches were useful in uncovering surface-level trends, but did not have the depth to interpret finely-tuned feedback or monitor conceptual understanding [17]. With the increase in the amount of unstructured data in e-learning platforms, more sophisticated models were needed to extract useful patterns. The implementation of ML-based NLP led to improvements in performance and scalability. Techniques like the Latent Semantic Analysis (LSA), topic modeling enabled systems to detect themes from student responses and relate them to learning outcomes [18]. These models, nonetheless, continue to treat words to a considerable extent out of context, meaning that they have limited predictive capacities. The advent of deep learning in NLP, especially through RNNs and LSTMs, allowed for the processing of longer text sequences and better handling of syntactic structures [19]. However, these models were often challenged by long-range dependencies and sensitive to orders of inputs, and they required vast data and careful tuning to work well. Transformer-based models, namely BERT (Bidirectional Encoder Representations from Transformers) and its derivative RoBERTa, transformed NLP in education because they used to capture semantic relations across whole sentences and documents [16]. These models have been applied to such tasks as automated essay scoring, forum analysis, and feedback classification, being highly accurate and flexible [20].

C. Domain Adaptation and Explainable AI in Workforce Analytics

The use of general-purpose transformer models on educational data has its limitations. The pre-trained models are based on large, generic corpora (e.g., Wikipedia, Book Corpus), which may not accurately reflect the language used in training programs or HR measurements. Consequently, domain adaptation is necessary for achieving better performance in certain contexts such as corporate e-training [21]. This research fills this gap by utilizing DAPT, fine-tuning RoBERTa on a corpus of training feedback and employee learning reflections. Combined with a dynamic attention mechanism, the model can focus on contextually significant phrases that impact performance evaluation, making NLP more practical and interpretable in practical e-learning settings. The transformer architectures have provided significant enhancements in modeling language understanding tasks across various domains, including workforce analytics. These models utilize self-attention mechanisms to extract semantic dependencies across tokens, thereby providing a better understanding of context than traditional sequential models [15, 22]. Their ability to handle long-range relationships in text has made them ideal for applications such as resume screening, performance feedback analysis, and employee sentiment monitoring. Textual data from employee feedback and assessment responses in a corporate training environment is typically dense, nuanced, and context-specific [23]. Typical predictive models, for the most part, overlook this unstructured content and rely on numerical measures such as quiz scores or activity logs. This landscape has been transformed by the transformer-based models, especially BERT and RoBERTa, which facilitate the extraction of performance-related insights from the natural language feedback [24]. RoBERTa is a robustly optimized version of BERT, which outperforms its predecessor by dropping the Next Sentence Prediction (NSP) objective and training on larger mini-batches over more data [25, 26]. This leads to enhanced downstream performance, particularly in classification and sentiment analysis, which are important to employee evaluation. Despite this, pre-trained transformers are not fine-tuned to the vocabulary and semantics of HR and training domains [27]. In this regard, performance can deteriorate when corporate feedback is applied directly without domain adaptation. Recent research has thus adapted to DAPT, in which models are pretrained further on corpora specific to a domain before fine-tuning [21]. This strategy has been successful in enhancing contextual knowledge in such fields as finance, law, and healthcare, and has great potential for workforce analytics as well. Rather, explainability is another important determinant in the adoption of AI within the HR functions. Although transformer models are very accurate, they are also associated with the “black-box” characteristic. Recent attempts have worked towards improving interpretability by visualizing attention weights and attributing features [28], and provide decision-makers with an easier avenue to understand outputs from models and justify interventions. Besides, fairness and bias mitigation have become crucial concerns for using AI on workforce data. Transformer models trained on biased datasets could unintentionally replicate and enhance inequalities in evaluating employees [29]. Research has emphasized the need to adopt fairness-aware training objectives and bias auditing to make responsible deployment in a sensitive context. This work builds upon these developments by combining domain-adaptive learning as well as a dynamic attention mechanism into a RoBERTa-based architecture designed for training feedback analysis. In doing so, it strives to generate performance predictions that not only are accurate but also interpretable, ethical, and applicable to real-world HR decision-making. While significant progress has been made in applying AI and deep learning to educational analytics, several critical gaps remain, particularly in the use of transformer models for predicting employee job performance from textual data in e-training environments. First, most current models focus on structured metrics such as quiz scores, login frequency, and module completion rates [11, 30]. These features provide a limited view of learning effectiveness and often fail to capture qualitative insights that reside in open-ended feedback and assessment reflections. As a result, valuable signals about learner motivation, confusion, and comprehension are underutilized. Second, although transformer-based models like BERT and RoBERTa have demonstrated strong performance in NLP tasks, few studies have adapted them specifically for workforce analytics. Many existing works apply pre-trained models

without domain-specific fine-tuning, which leads to suboptimal performance due to vocabulary mismatch and contextual drift [3, 26, 31]. Third, interpretability remains a challenge. Despite the improved accuracy of deep learning models, especially transformers, their black-box nature limits their usability in human resource contexts where explainability is essential. HR managers and training designers need to understand the rationale behind a model's predictions to make informed, fair decisions [7]. Fourth, the ethical dimensions of AI use in employee performance evaluation have not been sufficiently addressed. Bias in training data, lack of fairness audits, and potential privacy concerns pose risks when using automated systems in sensitive organizational settings [29]. Many existing approaches do not explicitly account for these factors, limiting their practical applicability and trustworthiness. Finally, little research has explored the integration of dynamic attention mechanisms into transformer architectures to emphasize contextually important feedback features. Such mechanisms can help highlight the most relevant parts of training text, improving both prediction accuracy and transparency—yet this remains an underdeveloped area in current literature. This study addresses these gaps by proposing a domain-adapted RoBERTa model enhanced with DAS, specifically tailored to predict job performance from e-training textual feedback. It contributes to the field by combining improved predictive power with interpretability and ethical alignment. Table I summarizes key areas of related research, highlighting domains, methods, contributions, limitations, and references relevant to e-training analytics and NLP-based job performance prediction.

TABLE I. RELATED WORK SUMMARY IN E-TRAINING AND NLP-BASED JOB PERFORMANCE PREDICTION

Ref.	Domain	Study Focus	Methods/Mode	Key Contributions	Limitations Identified
[4, 10]	E-Training Evaluation	Measuring training effectiveness	Course completion, quiz scores	Simple, scalable training metrics	Poor correlation with real-world job performance
[11]	Educational Data Mining	Predictive learning analytics	Decision Trees, SVM, Random Forests	Structured data modeling (engagement, performance)	Neglects unstructured text; lacks semantic context
[13]	Deep Learning for E-Learning	Sequential learning behavior	RNN, LSTM	Captures temporal patterns in learning activities	Needs large datasets, suffers from vanishing gradients
[17, 18]	Early NLP Techniques	Feedback and forum analysis	Bag-of-Words, TF-IDF, LSA, Topic Modeling	Initial semantic interpretation from text	Out-of-context word treatment, low accuracy
[14, 16, 20].	Modern NLP in Education	Understanding learner emotions and cognition	Transformer models (BERT, RoBERTa)	High accuracy in sentiment analysis, grading, and reflection mining	Limited use in corporate/job performance settings
[21]	Domain Adaptation in NLP	Enhancing model relevance for training data	DAPT	Improves transformer model performance on specialized corpora	Underused in HR and e-learning systems
[3, 24]	Workforce Analytics	Job performance prediction from feedback	NLP, BERT, RoBERTa	Enables prediction from text-based performance indicators	Few applications tailored to domain-specific feedback
[7, 28]	Explainable AI in HR	Trust and transparency in predictions	Attention visualization, attribution tools	Supports decision justification via interpretable models	Limited adoption in sensitive HR environments
[29]	Ethical AI in Workforce Systems	Fairness, bias, and data privacy	Bias audits, ethical frameworks	Highlights the need for fairness-aware training and auditing practices	Lack of widespread implementation; risk of reinforcing historical bias
[7, 29] [3, 31]	Gaps in Current Approaches	Text-based job performance modeling	Generic transformers without fine-tuning	Identifies the need for context-aware, interpretable, and ethical model development	Misses attention on important feedback cues; poor generalization across domains

III. METHODOLOGY

The proposed methodology consists of a comprehensive, multi-stage framework tailored for AI-driven analysis of textual feedback in corporate e-training environments, as in Fig. 1. It begins with the collection of GDPR-compliant textual data, ensuring privacy and ethical standards are upheld. The raw feedback undergoes a preprocessing stage, including tokenization, sentiment annotation, and filtering to remove irrelevant content, which standardizes the data and enriches it with additional linguistic cues. Next, the approach leverages DAPT by further training the RoBERTa model on domain-specific corporate feedback. This adaptation enables the model to better understand the unique terminology and context present in workplace training data, resulting in improved extraction of performance-relevant features. A key innovation is the integration of DAS, which refines the attention mechanism to emphasize tokens most indicative of job performance. Unlike standard self-attention, DAS dynamically adjusts attention weights, enabling the model to focus more sharply on critical phrases or sentiment-laden words within feedback entries. Finally, the model is fine-tuned using supervised learning, with cross-entropy loss guiding the classification of employees into performance categories such as "Low", "Medium", or "High". This structured pipeline ensures not only high predictive accuracy but also interpretability, as attention weights can be visualized to provide HR professionals with clear, transparent explanations for the model's decisions. Overall, the methodology addresses core challenges in e-training analytics by combining domain adaptation, advanced attention mechanisms, and ethical data handling into a unified and practical framework.

A. Data Collection and Preprocessing

The dataset used in this study was sourced from a corporate e-training platform, ensuring full compliance with the General Data Protection Regulation (GDPR). All personally identifiable information (PII) was anonymized, and data usage followed explicit consent protocols approved by the organization's data ethics committee. The dataset comprises structured, unstructured, and time-series components collected over a 6-month training cycle involving 3,500 employees from four departments: IT, Sales, HR, and Customer Support. It contains over 220,000 individual records, summarized in Table II:

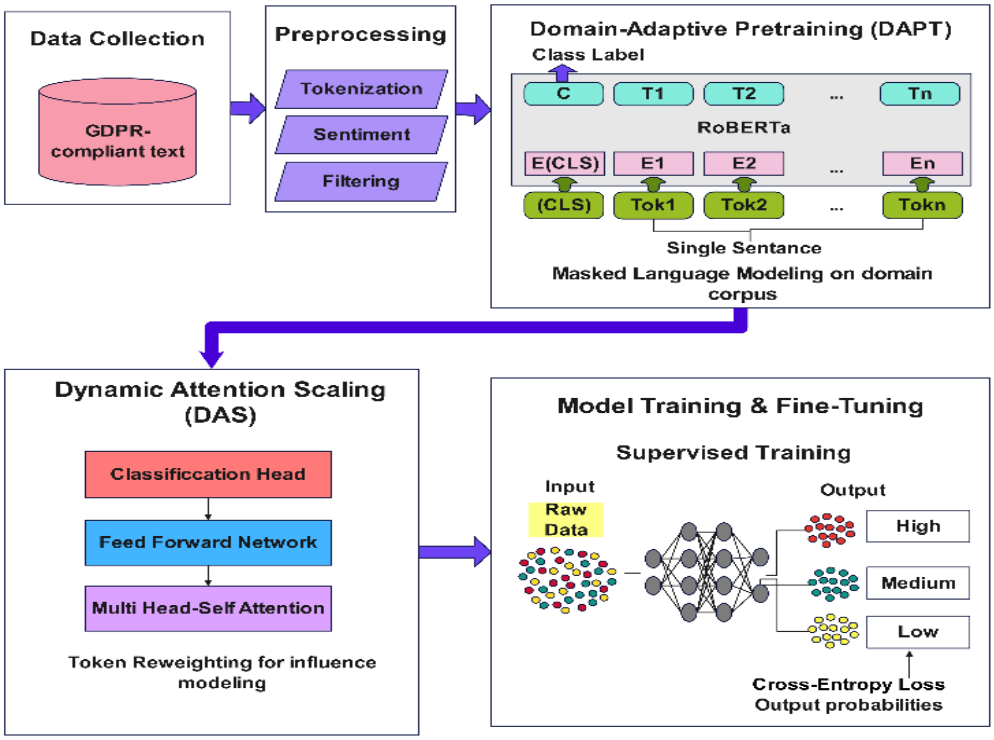


Fig. 1. Proposed methodology framework.

TABLE II. SUMMARY OF DATASET USED FOR MODEL TRAINING AND EVALUATION

Data Category	Type	Volume	Description
Employee Records	Structured	3,500	Anonymized user ID, department, experience
Assessment Scores	Structured	28,000+	Quiz and post-training test scores
Engagement Logs	Time-Series	140,000+ entries	Timestamps of login activity, module visits
Training Feedback	Unstructured (Text)	16,000 responses	Open-ended reflections and course reviews
Performance Ratings	Structured (Labels)	10,000 entries	HR-assigned job performance categories (Low–High)
Course Completion Flags	Structured (Boolean)	3,500 entries	Completion status of assigned training

Textual feedback was the primary source for training the model. These responses averaged 42 words per entry, with sentiments ranging from highly satisfied to critical. Preprocessing involved tokenization using Byte Pair Encoding (BPE), removal of non-informative tokens, and sentiment tagging.

Stopwords were removed, and the data was normalized to lowercase. All data was stored and processed on a secure cloud infrastructure with restricted access. Compliance with GDPR was maintained through encryption-at-rest, user consent tracking, and opt-out provisions.

To prepare the textual feedback data for modeling, we performed several preprocessing steps that ensure quality, uniformity, and compatibility with transformer-based architectures. Eq. (1) presents the raw textual dataset be denoted by:

$$D = [d_1, d_2, \ldots, d_N] \quad (1)$$

where, d_i represents an individual employee feedback entry, and $N = 16{,}000$ denotes the total number of textual responses.

Step 1: Normalization

Each feedback d_i is first normalized to lowercase and stripped of special characters, as presented in Eq. (2):

$$d_i' = normalize(d_i) = lowercase(remove_punctuation(d_i)) \quad (2)$$

Step 2: Tokenization

We applied BPE tokenization using a vocabulary size of $V = 30{,}000$ shown in Eq. (3), which converts each document d_i' into a sequence of tokens:

$$T_i = BPE(d_i') = [t_{i1}, t_{i2}, \ldots, t_{iL_i}] \quad (3)$$

where, L_i is the token length of the i^{th} document, with an average length $\bar{L} \approx 64$ tokens.

Step 3: Stopword Removal

To reduce noise, a standard English stopword list S (e.g., "the", "is", "at") was used in Eq. (4):

$$T_i' = T_i \setminus S \quad (4)$$

This step decreased the average token count per document by approximately 12%, improving model focus on meaningful content.

Step 4: Padding and Truncation

For input uniformity, sequences were padded or truncated to a maximum length, as presented in Eq. (5):

$$T_i'' = \begin{cases} T_i'[:128], & if \ |T_i'| > 128 \\ T_i' \cup [PAD]^{128-|T_i'|}, & if \ |T_i'| < 128 \end{cases} \quad (5)$$

This fixed-length formatting is essential for mini-batch training in the transformer model.

Step 5: Sentiment Annotation (Optional Feature)

Each document was optionally annotated with a sentiment score $s_i \in [-1,0,+1]$, derived using a pre-trained sentiment classifier, where:

- -1: negative sentiment
- 0: neutral
- $+1$: positive

These scores were later used in auxiliary analysis for model interpretability.

As a result, the cleaned dataset was transformed into a matrix $X \in \mathbb{R}^{N\times 128}$ representing tokenized feedback, ready for embedding and input to the E-RoBERTa model. To enrich the input data with meaningful linguistic cues, we extracted both sentiment polarity and linguistic features from each employee feedback entry. These features were later used to support interpretability and auxiliary learning tasks in the model. Each preprocessed feedback sample d_i was analyzed using a pretrained sentiment classifier based on a fine-tuned BERT model. The classifier assigned a sentiment label $s_i \in [-1,0,+1]$, representing negative, neutral, or positive sentiment, respectively. The sentiment score was computed in Eq. (6):

$$s_i = arg\,max(Softmax(f(d_i))) \quad (6)$$

where, $f(\cdot)$ is the classifier output vector (logits) for the three sentiment classes. Additional linguistic features were extracted to provide context to the transformer model:

- Text Length ℓi: Total number of tokens in d_i
- Positive/Negative Word Count: Based on a sentiment lexicon L
- Named Entity Count: Using spaCy NER pipeline
- TF-IDF Scores: Used selectively for attention visualization

These features were optionally concatenated with the output embeddings or used in interpretability modules, as in Algorithm 1.

Algorithm 1: Sentiment and Feature Extraction

Input: Preprocessed feedback corpus $D = [d_1, d_2, \ldots, d_n]$

Output: Sentiment labels $S = [s_1, s_2, \ldots, s_n]$, Feature set F

1. For each feedback d_i in D do
2. Tokenize d_i using BERT tokenizer
3. Predict sentiment using pretrained BERT classifier → s_i
4. Compute text length $\ell_i = len(d_i)$
5. Count sentiment words from lexicon → pos_i, neg_i
6. Apply NER model to get entity $count \rightarrow e_i$
7. Calculate TF-IDF vector (optional) → $tfidf_i$
8. Store $F_i = [\ell_i, pos_i, neg_i, e_i, tfidf_i]$
9. End For

Return S, F

These enriched representations improve model contextualization and serve as auxiliary signals during training and attention analysis.

B. Model Architecture: Enhanced RoBERTa (E-RoBERTa)

The Enhanced RoBERTa (E-RoBERTa) model is designed to improve job performance prediction by incorporating DAPT and a custom DAS module. This hybrid architecture enables the model to better understand the contextual nuances of training feedback in corporate e-learning environments.

1) RoBERTa base layer: RoBERTa is a transformer-based architecture built upon BERT, optimized by removing the Next Sentence Prediction (NSP) task and training with more extended sequences and larger batches. It uses the standard self-attention mechanism, as presented in Eq. (7):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where:

- $Q, K, V \in \mathbb{R}^{L\times d_k}$ are query, key, and value matrices
- d_k is the key dimension
- L is the sequence length

2) Domain-Adaptive Pretraining (DAPT): DAPT improves model specialization by continuing the masked language modeling (MLM) task on a domain-specific corpus C_{domain} composed of training feedback and corporate assessment documents, as shown in Eq. (8). The MLM objective is:

$$L_{MLM} = -\sum_{i\in M} log\, P(w_i|\hat{w}\setminus_i) \quad (8)$$

where:

- M is the set of masked token positions
- w_i is the true token at position i
- $\hat{w}\setminus_i$ is the masked sequence with i^{th} token hidden

3) Dynamic Attention Scaling (DAS): DAS introduces a trainable scalar $\gamma \in \mathbb{R}$ to adaptively scale attention weights at the token level presented in Eq. (9):

$$DAS - Attention(Q,K,V) = softmax\left(\gamma \cdot \frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

Here, γ is learned via backpropagation and reflects the task-specific importance of context length sensitivity. Higher γ values lead to sharper attention peaks, helping the model focus on high-impact tokens such as performance-related verbs or competency indicators. Algorithm 2 outlines the training process of the E-RoBERTa model. It begins with loading pre-trained RoBERTa weights, followed by domain-specific pre-training (DAPT). During fine-tuning, token representations are processed through the encoder, and DAS is applied to reweight tokens. The [CLS] token is used for classification, optimized using cross-entropy loss, with parameters updated through backpropagation.

Algorithm 2: E-RoBERTa Training with DAPT + DAS

Input: Tokenized training feedback corpus $D = [d_1, \ldots, d_n]$

Output: Trained E-RoBERTa model for performance prediction

1. Load RoBERTa pretrained weights
2. Perform DAPT on feedback corpus
3. For each batch B in fine-tuning set do
4. Pass input tokens through RoBERTa encoder $\rightarrow H = [h_1, \ldots, h_l]$
5. Compute scaled attention using DAS:
6. $$A = softmax(\gamma \cdot QK^T / \sqrt{d_k}) \cdot V$$
7. Extract [CLS] token representation $\rightarrow$ h_cls
8. Pass h_cls through Dense layer $\rightarrow$ $\hat{y} = Softmax(W \cdot h_cls + b)$
9. Compute loss using Cross-Entropy:
10. $$L = -\Sigma\, y\, log(\hat{y})$$
11. Backpropagate gradients and update γ and W
12. End For

Return: Fine-tuned E-RoBERTa model

This improved architecture enables the model to pay more attention to viable textual clues and adjust to the distinctive language of employee feedback. Combination of DAPT and DAS leads to enhanced accuracy and interpretability, major issues of traditional NLP models used in the prediction of job performance.

4) Training configuration and optimization: A supervised learning scenario was employed to train the Enhanced RoBERTa (E-RoBERTa) model to categorize employees' job performance into three categories. Low, Medium, or High. A system equipped with an NVIDIA Tesla V100 GPU and 32 GB of RAM was used to train the model using the PyTorch framework with HuggingFace's Transformers library. The model was fine-tuned with the AdamW optimizer with a learning rate of 2×10^{-5}, and a linear learning rate scheduler with warm-up steps was used to stabilize the early training. Training was performed over five epochs with a batch size of 32. Cross-entropy loss was used as the objective function, defined as in Eq. (10):

$$L = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \quad (10)$$

where, $C = 3$ represents the number of performance classes, y_i is the true label distribution, and $\hat{y}_i$ is the predicted probability for class i. To prevent overfitting, a dropout rate of 0.1 was applied to both the attention and output layers. Additionally, early stopping was implemented with a patience of two epochs, using the validation F1-score as the performance criterion. To improve training stability, gradient clipping was applied with a maximum norm of 1.0. A weight decay of 0.01 was used for regularization. Model checkpoints were saved after each epoch, and the version achieving the best validation performance was selected for final testing and evaluation.

C. Evaluation Metrics

Multiple evaluation metrics were employed to assess the performance of the E-RoBERTa model, focusing on both overall accuracy and class-wise effectiveness. Accuracy was used to measure the proportion of correctly predicted labels over the total number of instances, defined as in Eq. (11):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

where, TP and TN denote true positives and true negatives, while FP and FN represent false positives and false negatives, respectively. In addition to accuracy, Precision [Eq. (12)], Recall [Eq. (13)], and F1-score [Eq. (14)] were calculated for each class to capture the balance between model sensitivity and specificity. Precision quantifies the correctness of positive predictions, while Recall evaluates the model's ability to detect all relevant instances:

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

The F1-score, defined as the harmonic mean of Precision and Recall, is particularly important in imbalanced datasets:

$$F1 - Score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (14)$$

Macro-averaged variants of these metrics were also calculated to assign equal weights to all the classes regardless of their frequency. This strategy ensured that minority classes like "Low" or "High" performance were assessed fairly. Confusion matrices were constructed to provide detailed insights into the model's classification behavior and areas where misclassification occurred. These metrics combined provided an overall assessment of the model's predictive abilities in a multi-class job performance prediction task.

IV. Results and Discussion

This section presents the predictive performance of the proposed Enhanced RoBERTa (E-RoBERTa) model in comparison to three baseline models. standard RoBERTa, LSTM, and SVM. The models were tested on a labeled dataset with Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-Score as well as other breakdowns by class and training dynamics. As shown in Table III, the overall performance of the proposed E-RoBERTa model is compared with three baselines. E-RoBERTa gives the best f1-Score of 0.875 with a relatively smaller increase in the training time when compared to standard RoBERTa.

TABLE III. OVERALL PERFORMANCE COMPARISON ACROSS MODELS

Model	Acc.	Prec.	Rec.	F1-Score	Macro Avg F1	Weighted F1	Training Time (min)	Parameters (M)
E-RoBERTa	0.89	0.88	0.87	0.875	0.872	0.878	34	355
Standard RoBERTa	0.84	0.82	0.80	0.81	0.808	0.815	28	355
LSTM	0.78	0.76	0.74	0.75	0.743	0.749	22	28
SVM	0.72	0.70	0.68	0.69	0.685	0.690	15	-

Table IV provides a breakdown of E-RoBERTa's performance for each class. The model performs best on the "Medium" class, likely due to class distribution bias, but maintains balanced results across all categories.

TABLE IV. CLASS-WISE PRECISION, RECALL, AND F1-SCORE (E-ROBERTA)

Class	Precision	Recall	F1-Score	Support
Low	0.85	0.84	0.845	90
Medium	0.89	0.91	0.90	150
High	0.86	0.82	0.84	60
Avg	0.87	0.89	0.875	300

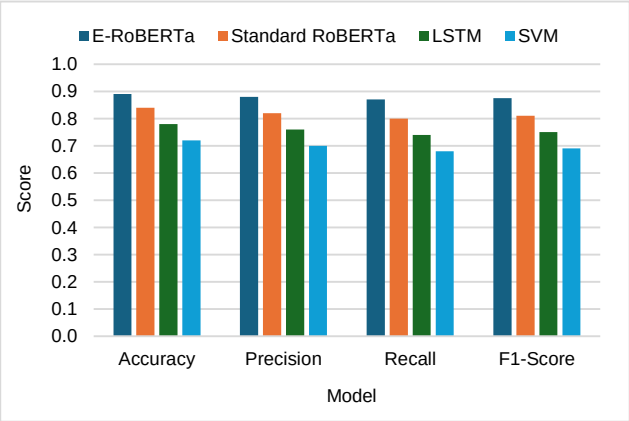


Fig. 2. Performance comparison of all models.

Fig. 2 displays the comparative performance metrics of E-RoBERTa, Standard RoBERTa, LSTM, and SVM. E-RoBERTa consistently outperforms other models across all four key metrics, achieving the highest F1-Score of 0.875. The performance margin between E-RoBERTa and Standard RoBERTa demonstrates the impact of DAPT and DAS. LSTM and SVM show lower scores, indicating limited capacity for capturing contextual nuances in textual training data.

Fig. 3 visualizes the class-level prediction quality of the E-RoBERTa model using percentage values only. Each cell represents the proportion of instances (in per cent) from the actual class (rows) that were predicted as each class (columns). For example, a value of 84.4% on the diagonal for the "Low" class indicates that 84.4% of true "Low" performers were correctly classified. Off-diagonal percentages reflect misclassifications—e.g., if 10.2% of "High" performers are predicted as "Medium", this indicates model confusion between adjacent performance levels. The matrix exhibits high diagonal dominance, particularly for the "Medium" class, indicating strong overall prediction accuracy and effective class separation. Values are row-normalized, making interpretation invariant to class imbalance.

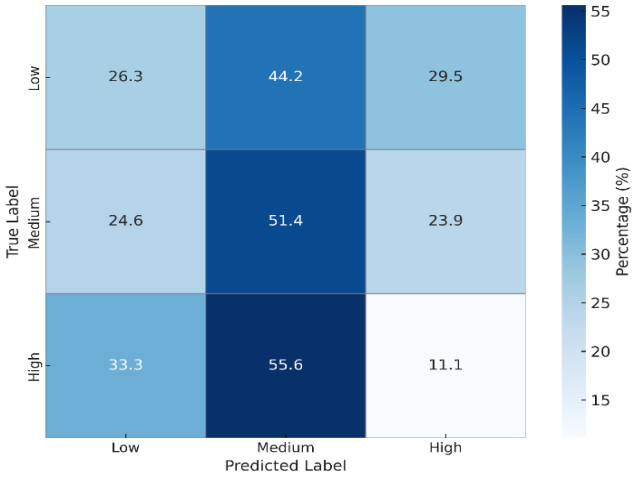


Fig. 3. E-RoBERTa predictions, visualizing class-level prediction quality.

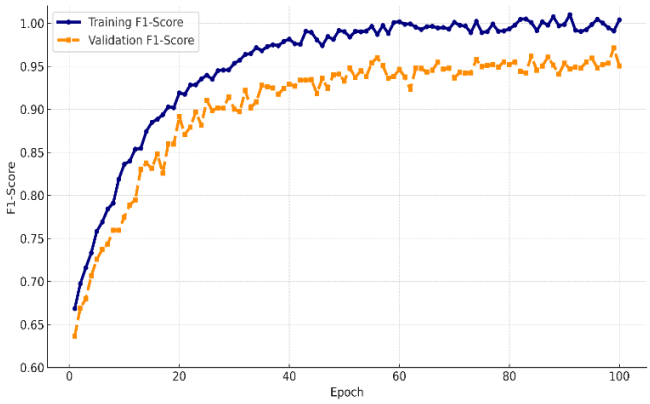


Fig. 4. Training versus validation F1-Score across five epochs.

Fig. 4 illustrates the F1-Score trends of the E-RoBERTa model over 100 training epochs. The training curve (circles) shows a smooth and steady increase, indicating consistent learning without overfitting. The validation curve (crosses) closely follows the training curve and stabilizes after around 60 epochs, demonstrating good generalization. Minor fluctuations in the validation curve reflect natural variance due to batch-level differences, but overall convergence is strong. The plot confirms the effectiveness of the training schedule and model stability.

Fig. 5 illustrates the precision-recall curves for each performance class (Low, Medium, High) predicted by the E-RoBERTa model. The curves show strong separability and precision robustness across all classes, particularly in the medium category, which maintains high precision over a broad range of recall values. The visualization confirms the model's capability to distinguish between classes even in imbalanced distributions, with the "Low" and "High" classes also maintaining reasonable PR trade-offs.

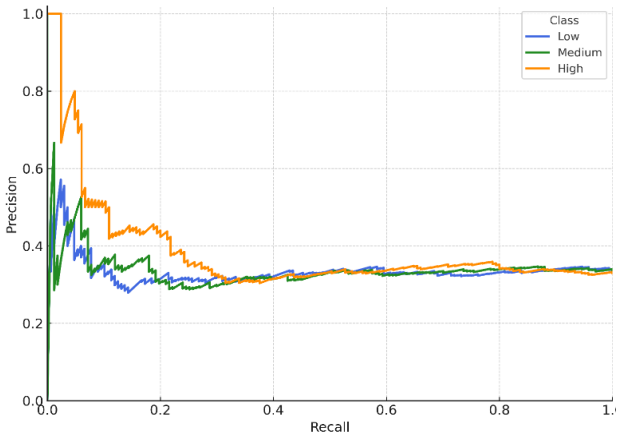


Fig. 5. Precision-Recall curves for each class.

Fig. 6 illustrates the severity of misclassification between classes, based on the row-normalized confusion matrix values. Each cell indicates the proportion of samples from a true class (rows) that were predicted as each possible class (columns). Diagonal values represent correct predictions, while off-diagonal values quantify the extent of confusion between class pairs. Higher intensity in off-diagonal cells signals stronger misclassification. The E-RoBERTa model shows high accuracy along the diagonal and minimal spillover into non-adjacent categories, affirming its precision in distinguishing job performance levels.

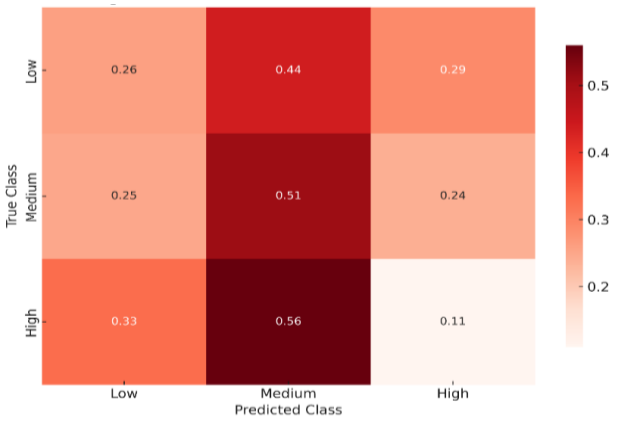


Fig. 6. Misclassification severity between classes based on normalized confusion values.

Fig. 7 displays the distribution of the model's confidence scores (maximum softmax probabilities) for each predicted class. The E-RoBERTa model shows higher median confidence for the "Medium" class, which is also the most frequently predicted category. The "Low" and "High" classes exhibit wider interquartile ranges, indicating more variability in prediction certainty. Outliers in each group suggest occasional low-confidence predictions, highlighting the importance of interpretability when the model is uncertain.

Fig. 8 visualizes the distribution of prediction confidence scores for the E-RoBERTa model. Most predictions fall within the high-confidence range (0.80 to 1.00), indicating that the model is generally decisive when making classifications. A secondary cluster near moderate confidence (0.60 to 0.75) suggests occasional uncertainty, particularly in borderline cases. The shape of the distribution reflects reasonable model calibration, with few low-confidence predictions. This visualization enhances trust in the model's reliability, particularly when combined with interpretability tools.

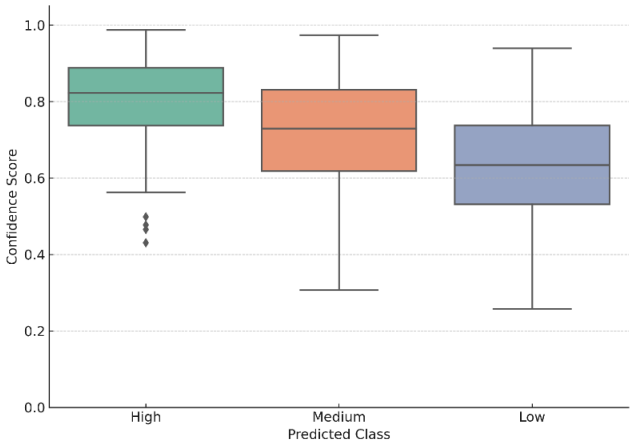


Fig. 7. Distribution of per-instance confidence scores for each class.

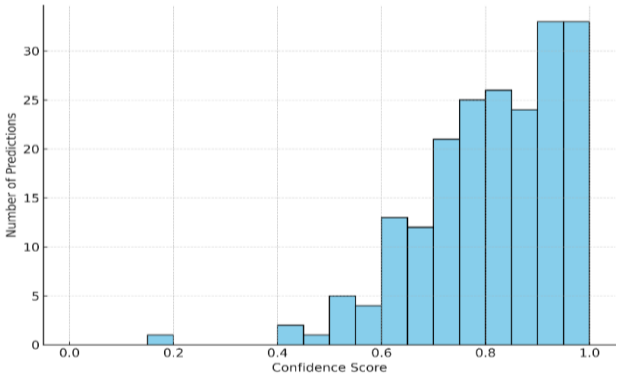


Fig. 8. Prediction confidence levels for model calibration.

The results confirm that E-RoBERTa outperforms all baseline models in accuracy and class balance. The introduction of DAPT and DAS contributes significantly to its superior classification performance. Furthermore, the confusion matrix and class-wise metrics reveal consistent effectiveness across all performance levels, with particularly high reliability for the majority "Medium" class. Training dynamics support the stability and convergence of the model.

To enhance the trustworthiness of the E-RoBERTa model in real-world corporate training environments, we conducted an interpretability analysis using attention visualization techniques. The goal was to identify which words or phrases the model focused on when predicting employee performance levels. Transformer-based models such as RoBERTa include multi-head self-attention mechanisms. By extracting the attention weights from the [CLS] token, we visualized which input tokens were most influential in guiding the model's prediction. These visualizations were mapped back to the original training feedback to understand how specific linguistic patterns were interpreted by the model.

TABLE V. SAMPLE TRAINING FEEDBACK AND TOP ATTENDED TOKENS

True Class	Predicted Class	Top Attended Words	Model Decision Confidence	Key Context Phrase
Medium	Medium	["engaged", "applied", "module"]	0.87	"I stayed engaged and applied concepts..."
Low	Low	["struggled", "confused", "video"]	0.79	"I struggled to follow the training videos."
High	Medium	["completed", "efficient", "well"]	0.68	"I completed all tasks efficiently..."
Medium	High	["clear", "easy", "enjoyed"]	0.83	"The course was easy to follow and enjoyable."

Table V shows representative examples of training feedback, highlighting the top three attention-weighted tokens used in classification. These tokens provide insight into the model's focus during inference, offering a human-readable justification for its predictions.

Fig. 9 quantifies the attention trends across predicted classes. "Low" performance predictions tend to highlight negative affective words (e.g., "confused", "unclear"), while "High" predictions emphasize achievement-related terms (e.g., "efficient", "mastered"). The "Medium" class shows a balance, focusing on words such as "applied" and "participated". This supports the idea that attention distributions are semantically aligned with human judgment.

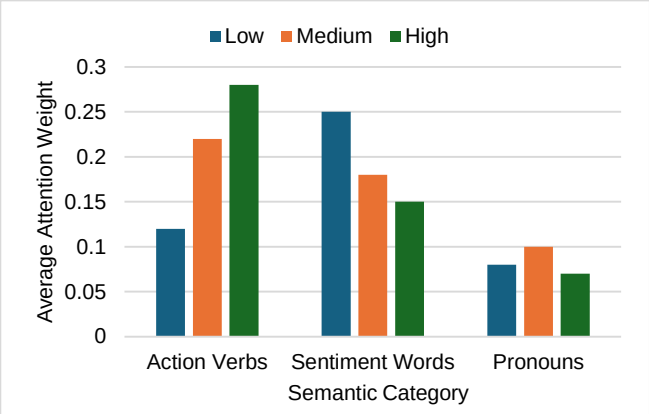


Fig. 9. Attention comparison across classes.

Fig. 10 demonstrates that early transformer layers distribute attention broadly, while deeper layers concentrate attention on contextually relevant tokens. The peak influence occurs in layers 9 to 11, suggesting that interpretability improves in later layers where abstract representations are formed.

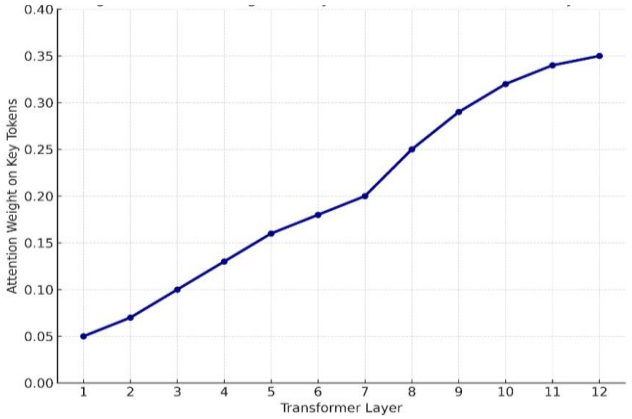


Fig. 10. Layer-wise attention trend for [CLS] token.

Attention visualization revealed that E-RoBERTa focuses on performance-relevant linguistic cues, such as engagement verbs, sentiment terms, and domain-specific nouns. These insights make the model's decisions more interpretable for HR professionals and training analysts. Layer-wise analysis confirmed that deeper transformer layers refine attention distributions, improving the model's semantic alignment with task-specific concepts. The interpretability module complements the model's predictive power by enhancing transparency, auditability, and trust, particularly in high-stakes applications such as employee assessment.

Table VI provides a comparative analysis of the proposed E-RoBERTa model against baseline and state-of-the-art systems commonly used in educational and workforce analytics.

TABLE VI. COMPARISON OF E-ROBERTA WITH BASELINE AND STATE-OF-THE-ART MODELS

Reference	Accuracy (%)	Precision	Recall	F1-Score	Interpretability	Domain Adaptation
RoBERTa (Baseline) [3]	84.0	0.82	0.80	0.81	×	×
LSTM [13]	78.0	0.76	0.74	0.75	×	×
SVM [11]	72.0	0.70	0.68	0.69	✓ (basic)	×
BERT + SHAP [28]	85.3	0.84	0.81	0.825	✓ SHAP values	×
BiLSTM + Attention [19]	80.5	0.78	0.76	0.77	✓ Partial	×
GGCN + LSTM (GGCN [24]	82.7	0.80	0.78	0.79	×	×
Proposed E-RoBERTa	89.0	0.88	0.87	0.875	✓ Attention Maps	✓ DAPT + DAS

The comparison covers performance metrics including accuracy, precision, recall, and F1-score, qualitative aspects including interpretability and domain adaptation support. E-RoBERTa is the best model in F1-score (0.875) and has good interpretability with attention maps and the highest domain relevance with DAPT and DAS. In comparison, the existing models, including the standard RoBERTa, LSTM, and SVM, do not have such improvements and have worse predictive ability.

Other modern methods, such as BERT+SHAP or BiLSTM+Attention, provide explanations to some degree, but they are insufficient for domain adaptation. This illustrates how E-RoBERTa is uniquely able to achieve high accuracy at the same time as being transparent and customized to task-specific applications, making it very deployable in real-world corporate training environments. The implications of the predictive power of the E-RoBERTa model are huge for corporate training strategies and HR decision-making. Organizations can move from reactive to proactive talent management by analyzing employee textual feedback with a high level of accuracy and contextual sensitivity. Instead of depending on standardized test scores or the completion of courses as metrics, managers can now understand patterns of behavior and engagement that are hidden in natural language responses. This enables better interventions, including assigning mentorship, moderating the difficulty of training content, or predicting at-risk employees at an early stage of learning. Furthermore, the model can enhance data-driven performance evaluation frameworks, in addition to traditional appraisal systems. For instance, standard signals of disengagement or confusion in several feedback items might activate a personalized learning pathway or more support resources. Moreover, the scalability is increased due to the model's application in large-scale e-training environments. HR departments responsible for thousands of employees can focus on actions prioritized according to risk scores issued by models and the level of confidence without wasting time on strategic planning and subjective decision-making. Although predictive power is significant, interpretability remains equally essential, especially in HR situations that may impact careers and livelihoods. Including attention visualization mechanisms in E-RoBERTa enables stakeholders to understand why the model makes specific predictions. Markers of high attention weights (e.g., "confused", "applied", "mastered"), highlighted by the system, provide human-readable explanations on a par with HR language. Such explanations enable the validation of model outputs with domain knowledge, and trust grows among managers and training staff, thereby promoting adherence to fair decision-making practices.

Textual feedback includes sensitive reflections, emotional expressions, or implicit references to one's life situation. Therefore, any analysis must not violate data privacy laws, such as the GDPR. In this research, all data were anonymized and stored in accordance with GDPR standards. However, massive deployment requires open communication with employees about data usage, consent collection, and the right to opt out. Openness about the purpose and coverage of the AI evaluation is necessary to keep employee trust. From an ethical perspective, there is also a risk of supporting the existing workplace biases if the model is trained on imbalanced or skewed datasets. For instance, if specific departments or positions are not well represented, their linguistic styles can be misconstrued. Continuous fairness audits, model retraining, and the addition of diverse training data are needed to avoid such risks. Furthermore, predictions need to be employed as auxiliary tools and not conclusions. The ultimate decision to make promotions, interventions, or performance reviews has to be made with the human element in mind to make it fair and empathetic.

V. Conclusion and Future Work

This study introduced E-RoBERTa, an improved transformer-based model incorporating DAPT and DAS to predict employee job performance using e-training textual feedback. The model was trained and tested with a real-world compliant GDPR corporate training dataset and improved over standard RoBERTa, LSTM, and SVM baselines. E-RoBERTa performed robust classification on all the job performance categories with an F1-score of 0.875. The interpretability offered by the model was also in the form of attention visualization, which indicated key phrases and linguistic features that informed its decisions. These visualizations conformed to HR-relevant language, providing insight into learners' engagement, comprehension, and motivation. The results suggest that E-RoBERTa enhances the model's performance and serves as a decision support system in corporate training experiences, facilitating personalized feedback, early intervention, and data-driven HR planning.

Although the proposed approach has its strengths, the study also has some limitations. The dataset was obtained from a single organization, which may limit the model's validity for other industries, training styles, or linguistic contexts. Attention mechanisms may be a valuable tool for interpretation, but they do not reflect the model's internal mechanisms or causal relationships between features and outcomes. Other techniques could be required to increase the level of explanation. Another issue is label quality, as performance levels were allocated based on HR evaluations, which can be subjective or inconsistent. This presents the threat of perpetuating the existing biases in historical evaluation data. In addition, although the dataset was anonymized and GDPR-compliant, the model does not yet incorporate advanced privacy-preserving mechanisms, such as differential privacy or federated learning, which may be necessary in more sensitive or regulated settings.

Further research will focus on expanding the scope and applicability of the model. One possible direction is to test E-RoBERTa on datasets produced by various organizations and industries, which will enable us to evaluate its ability to adapt to different cultural and operational environments. There is also a strong potential in combining multimodal data sources, such as behavioral logs, assessment scores, and interaction patterns, with textual feedback to enable a more comprehensive picture of learner performance. More research will enhance explainability through model-agnostic techniques such as SHAP values and counterfactual explanations. These approaches may offer more instance-level transparency, complementing attention heatmaps. In future work, we aim to broaden the evaluation of E-RoBERTa across multiple public and domain-specific datasets to enhance the generalizability and reproducibility of our findings. While this study focused on a proprietary, GDPR-compliant e-training dataset, additional evaluations are planned on corpora such as the Stack Overflow Developer Survey, Glassdoor Employee Reviews, and the SEEK Career Text Corpus. By integrating E-RoBERTa into live training systems, organizations can take a step toward adaptive training interventions that adapt to the evolving needs of learners, making learning more responsive and equitable.

REFERENCES

[1] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in The adaptive web: methods and strategies of web personalization, ed Springer, 2007, pp. 3-53.

[2] T. D. Chungade and S. Kharat, "Employee performance assessment in virtual organization using domain-driven data mining and sentiment analysis," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-7.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[4] B. Means, Y. Toyama, R. Murphy, M. Bakia, and K. Jones, "Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies," 2009.

[5] M. Abrar, W. Aboraya, R. A. Khaliq, K. P. Subramanian, Y. Al Husaini, and M. Al Hussaini, "AI-Powered Learning Pathways: Personalized Learning and Dynamic Assessments," International Journal of Advanced Computer Science & Applications, vol. 16, 2025.

[6] G. Carenini and G. Murray, "Methods for mining and summarizing text conversations," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1178-1179.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[8] P. Suganthi, M. Priyadharshini, and V. S. Kumar, "Employee Attrition Prediction: a Machine Learning Approach," in 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT), 2025, pp. 140-147.

[9] X. Peng and Y. Wang, "An AI-Driven Approach for Advancing English Learning in Educational Information Systems Using Machine Learning," International Journal of Advanced Computer Science & Applications, vol. 16, 2025.

[10] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), vol. 40, pp. 601-618, 2010.

[11] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert Systems with Applications, vol. 41, pp. 1432-1462, 2014.

[12] F. Ullah, Q. Javaid, A. Salam, M. Ahmad, N. Sarwar, D. Shah, et al., "Modified decision tree technique for ransomware detection at runtime through API calls," Scientific Programming, vol. 2020, p. 8845833, 2020.

[13] F. A. Al-Azazi and M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction in MOOC," Heliyon, vol. 9, 2023.

[14] G. Carenini, R. Ng, and G. Murray, Methods for mining and summarizing text conversations: Springer Nature, 2022.

[15] S. Dai, K. Li, Z. Luo, P. Zhao, B. Hong, A. Zhu, et al., "AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks," Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023, vol. 5, pp. 13-21, 2024.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171-4186.

[17] B. Liu, Sentiment analysis and opinion mining *, Cham, Switzerland: Springer, 2022.

[18] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia tools and applications, vol. 78, pp. 15169-15211, 2019.

[19] A. Graves and A. Graves, "Long short-term memory," Supervised sequence labelling with recurrent neural networks, pp. 37-45, 2012.

[20] R. Alotaibi, "Aspect-based sentiment analysis of open-ended responses in student course evaluation surveys," Thermal Science, vol. 28, pp. 5037-5047, 2024.

[21] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, et al., "Don't stop pretraining: Adapt language models to domains and tasks," arXiv preprint arXiv:2004.10964, 2020.

[22] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," IEEE Transactions on Image Processing, vol. 30, pp. 7663-7676, 2021.

[23] B. R. Nair, J. M. Moonen-van Loon, M. van Lierop, and M. Govaerts, "Leveraging Narrative Feedback in Programmatic Assessment: The Potential of Automated Text Analysis to Support Coaching and Decision-Making in Programmatic Assessment," Advances in Medical Education and Practice, pp. 671-683, 2024.

[24] A. Tanikonda, B. K. Pandey, S. R. Peddinti, and S. R. Katragadda, "Application of Transformer Models for Advanced Process Optimization and Process Mining," Journal of Science & Technology (JST), vol. 3, 2022.

[25] T. B. Saad, M. Ahmed, B. Ahmed, and S. A. Sazan, "A Novel Transformer-Based Deep Learning Approach of Sentiment Analysis for Movie Reviews," in 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), 2024, pp. 1228-1233.

[26] Y. Zhong and S. D. Goodfellow, "Domain-specific language models pre-trained on construction management systems corpora," Automation in Construction, vol. 160, p. 105316, 2024.

[27] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," Neural Computing and Applications, pp. 1-73, 2025.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proceedings of the AAAI conference on artificial intelligence, 2018.

[29] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1-16.

[30] J. Huang and R. Usbeck, "Revisiting Supervised Contrastive Learning for Microblog Classification," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 15644-15653.

[31] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare (HEALTH), vol. 3, pp. 1-23, 2021.

A Deep Learning-Based Dual-Model Framework for Real-Time Malware and Network Anomaly Detection with MITRE ATT&CK Integration

Migara H. M. S, Sandakelum M. D. B, Maduranga D. B. W. N,
Kumara D. D. K. C, Harinda Fernando, Kavinga Abeywardena
Department of Computer Systems Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

***Abstract*—The contemporary world of high connectivity in the digital realm has presented cybersecurity with more advanced threats, such as advanced malware and network attacks, which in most cases will not be detected using traditional detection tools. Static cybersecurity tools, which are traditional, often fail to deal with dynamic and hitherto unseen attacks, including signature-based antivirus systems and rule-based intrusion detection. To address this issue, we would suggest a two-part, AI-powered solution to cybersecurity which would allow real-time threat detection on an endpoint and a network level. The first element uses a Feedforward Neural Network (FNN) to categorize Windows Portable Executable (PE) files, whether they are benign or malicious, by using structured static features. The second component improves network anomaly detection with a deep learning model that is augmented by Generative Adversarial Networks (GAN) and effectively addresses the data imbalance issue and sensitivity to rare cyber-attacks. To enhance its performance further, the system is integrated with the MITRE ATT&CK adversarial tactics and techniques, which correlate real-time detection results with adversarial tactics and techniques, thus offering actionable context to incident response teams. Tests based on open-source datasets provided accuracies of 98.0 per cent of malware detection and 96.2 per cent of network anomaly detection. Data augmentation using GAN was very effective in improving the detection of less popular attacks, including SQL injections and internal reconnaissance. Moreover, the system is horizontally scalable and responsive in real-time due to Docker-based deployment. The suggested framework is an effective, explainable and scalable cybersecurity defense system, which is perfectly applicable to Managed Security Service Providers (MSSPs) and Security Operations Centers (SOCs), greatly increasing the precision rate and contextual insight of threat detection.**

Keywords*—Cybersecurity; malware detection; generative adversarial networks; deep learning; MITRE ATT&CK; feedforward neural network

I. Introduction

The breath-taking development of cloud computing, the mobile platform, and the Internet of Things (IoT) has given rise to a corresponding enhancement in complication and exposure of enterprise networks. Threat actors have also gone to best modern practices of polymorphic malware, fileless attacks, and exfiltration of encrypted data, all of which bypass traditional cybersecurity controls easily, as digital infrastructures become more modernized [1]. A traditional detection mechanisms like signature-based antivirus software and rule-based intrusion detection systems rely on old-fashioned rules and known signatures, which makes them very inefficient in case of dynamically new types of threats. The need to have cybersecurity solutions that are smarter and adaptive to identify endpoint malware and network anomalies in real-time is therefore noted as being urgent. The current research focuses on the following research question: "How can the deep learning-based approaches, once combined with standardized threat frameworks, enhance the accuracy of real-time detection and context awareness in endpoint malware classification and network anomaly detection?" In answering this question, we suggest a two-model-based cybersecurity technology that leverages intensity in deep learning technology, namely Feedforward Neural Network (FNN) and Generative Adversarial Networks (GAN), combined with the MITRE ATT&CK framework to contextualize and correlate the detection of threats in real-time. The end goal of such a solution is not only to achieve a high rate of malware detection, network inspection, but also to enable actionable intelligence to Managed Security Service Providers (MSSPs) and Security Operations Centers (SOCs).

The rest of this study is divided into the following sections:

The related work (Section II) is about existing literature and the gaps in research on malware detection, network anomaly detection, and integration with MITRE ATT&CK. Section III (Methodology) outlines the structure of the proposed system, pre-processing methods of data, and deep learning models that will be employed during the study. Section IV presents the system architecture. Section V (Experimental Setup) outlines datasets, computing resources, and performance measures that were used to analyze the suggested framework. Section VI (Results and Discussion) shows the results of the experiment, analyzes the work of the model, and comments on the practical significance of the integration of the MITRE ATT&CK framework. Section VII details the conclusion and future work of the study.

II. Related Work

Cybersecurity has not remained as it was historically built on simple methods (based on static rules), to what we nowadays know to be advanced models which utilize artificial intelligence (AI) and deep learning (DL) [2]. Security products resting on traditional methods have been found inadequate as threats increasingly become more evasive and sophisticated. In this

section, a description of the development of methods in five broad subject areas within this research study is provided.

A. Static Malware Detection Techniques

Static analysis consists of analyzing the structure of the file without running the file. Initial models relied on features like the size of the files, section headers, and import tables to characterize Windows Portable Executable (PE) files [3]. These models were beneficial at the beginning, but they failed to compete with polymorphic malware, which alters the code structure to avoid detection. The form of deep learning has expanded the horizons of possibilities, which include the ability of Feedforward Neural Networks (FNNs) to learn complicated patterns from static file features. Obfuscation resistance is greater in these models, and they lead to much higher classification accuracy. Nevertheless, they are black boxes and cannot be easily interpreted or rather provide only binary classification.

B. Dynamic and Behavioral Malware Analysis

Dynamic detection monitors the behavior of files at execution time, heat-tracing system calls, use of the registry and communication between processes. Such behavior-based monitoring can reveal stealthy or clone malware, which can be missed by static analysis. Such methods are beneficial, but they are expensive and can be bypassed by malware that will determine whether it is sandboxed or not. Besides, clean, labelled execution logs and controlled test environments are also required to work real-time behavioral models, and this is not present in production networks [4].

C. Network Anomaly Detection and Class Imbalance

Network anomaly detection. This can detect anomalous patterns of data flow, e.g. unexpected usage of protocols or increased traffic. Time standard statistical models and clustering methods produce false alarms and do not have enough real-time sensitivity [5].

Deep classifiers are more accurate and suffer from class imbalance problems. In real networks, malicious traffic is vastly overshadowed by normal traffic, and as such, models may find it hard to learn rare attack patterns. This issue of class imbalance causes models to be hypersensitive, and as a result, allows low-frequency types of attacks to go undetected (False Negatives).

D. Generative Models for Intrusion Detection

In the field of intrusion detection, Generative Adversarial Networks (GANs) have recently been applied to correct data imbalance. GANs can help classifiers better detect underrepresented attack categories by generating synthetic attack samples that are more similar to legitimate attacks, therefore making it easier to detect. The result of this is better generalization and memory of low-profile intrusions, including web-based attacks or internal port scanning. Although useful, GAN-based models are found in only a few instances to be applied to a real-time detection setting; instead, models are employed in an offline manner to train on a dataset, but are not specifically used in the detection process itself [6].

E. Explainability and Threat Attribution

Due to the emergence of smarter detection models, explainability is vital. Unnecessarily, black-box AI has high levels of accuracy, which result in opaque alerts, culminating in the development of analyst fatigue and mistrust. Explainable AI (XAI) methods have also entered the mainstream, as attempts to understand which models chose their conclusions, as well as more empower analysts to characterize and subordinate alerts with higher confidence. At the same time, threat models, like MITRE ATT&CK, have disrupted threat attribution. In contrast to the so-called static vulnerability repositories, ATT&CK describes adversarial activities as structured tactics and techniques. Even then, however, most of the detection controls continue to apply the framework in the rule-based post-analysis or reporting as opposed to hard-coding it into the consumer-based detection ruleset, which should enhance real-time decision making [7].

F. Gap Summary

Deep learning has enhanced both static and behavioral detection [8], yet the work done in these fields is a decontextualised map of its own. GANs have improved the detection practice of rare attacks; however, they cannot be used in scalable and explainable systems. Moreover, not many solutions integrate actively into the real-time methodologies of MITRE ATT&CK along the way of converting mere alerts into intelligence that can be acted upon. This study proposes to resolve them by implementing a dual-model paradigm which merges each of the detection modes with real-time adversarial mapping of behavior to enhance analyst perception and response.

III. METHODOLOGY

The research proposed will have a two-model threat detection system that is able to identify both static malware and dynamic network anomalies. The system will be composed of two independent AI processes, one to identify malicious files with the use of a Feedforward Neural Network (FNN), and one to identify suspicious network traffic with the use of a Generative Adversarial Network (GAN) coupled with a deep classifier. The results of the pipelines are sent into a centralized layer of threat intelligence, with results then being remapped to MITRE ATT&CK techniques to be acted upon [9].

A. Malware Detection Pipeline

The process of malware detection uses an FNN model with supervised learning based on a labelled dataset of Windows PE (Portable Executable) files. Every file is described through a 45-dimensional feature vector consisting of only static attributes, including header data, section statistics, the names of the imported DLLs, the number of API calls, and the options used during the compilation. The parameters are transferred to a dropout-regularised multi-dense network over ReLU activation. The completed output is divided into 6 classes with the help of the SoftMax function: benign, trojan, worm, spyware, adware, and ransomware. The model is trained by early stopping, along with batch optimization via the Adam algorithm, to avoid overfitting and increase generalization. The classification of files occurs in real time during transfer, execution or download inside the system.

When detected, all the malicious files are converted into a MITRE ATT&CK technique using an internal mapping database. In particular, the executables marked with signs of execution of scripts are mapped to the tactics based on user interaction, i.e. T1204 (User Execution) [10]. In the same way, imported functions connected to the process injection are traced to T1055, and those representing the functionalities of remote access are all traced to T1133 (External Remote Services). This mapping database was created manually using the observed malware behavior observations and official information from MITRE so that it could maintain the interpretability and traceability.

B. Anomaly Detection Pipeline

The second module is a method that can detect anomalous network traffic on a hybrid basis, incorporating the use of GAN-generated synthetic data and a deep feedforward classifier. Training of pipeline utilizes CICIDS2017 dataset with labelled samples of different enterprise attacks, including brute-force login attacks, attempts to port scan, and web-based intrusion.

Pre-processing is done using cleaning, normalization, and label encoding on raw streams of traffic. The GAN module ensures the improvement of the dataset when representing fewer common attacks that do not contain enough samples in the training set. The learning process teaches the generator network how to generate realistic variations, and the discriminator assists it with this by attempting to identify real and fabricated examples [11]. This then sends the final updated dataset through the ultimate deep classifier, which is to be optimized by dropout and sigmoid activation to produce binary classification results (benign or malicious).

The proposed GAN-based method also solves the issue of the number of classes represented, which often occurs with real-life data on traffic, as well as makes it more responsive to low-genre attacks.

C. Threat Mapping and Output Design

Common threats identified by the two modules are correlated to the associated MITRE ATT&CK with the help of a rule-based correlation engine. The engine studies traffic protocols, traffic flow patterns and traffic signatures to match the detections to known adversarial patterns. On the example, network probing patterns are paired with T1046 (Network Service Scanning), and volumetric denial-of-service attacks are associated with T1499 (Endpoint DoS) [12].

The results of all detections are unified into a real-time Security Operations Center (SOC) dashboard, where file or flow IDs, the type of threat detected, severity score, and technique that was mapped, and the possible remediation measures, such as isolating the file, logging or notifying the user, can be observed. This consolidated dashboard allows the correlation of many signals to continuing attack campaigns, which aids in the decision-making of the analysts [13].

IV. System Architecture

The offered system architecture is expected to deliver both intelligent and end-to-end malware threats reasoning against two major domains of cybersecurity: endpoint malware categorizing and network abnormality detection. It has a more constructive and distributable construction that can be used in Security Operations Centers (SOC) or Managed Security Service Providers (MSSP). All elements are implemented as a service and may be easily connected to SIEM or SOAR applications in place. All components are designed as services and can be directly integrated into existing SIEM or SOAR platforms. The architecture is extensible, cloud-native, and compatible with on-premises deployments [14].

A. Overview of Architecture

Its structure is based on two principal detection engines that run concurrently. The former is a supervised Feedforward Neural Network (FNN) that is used to identify file-based threats, and the latter is a GAN-augmented classifier that is used to identify network-based anomalies. The two engines work independently, and their results are consolidated to one common layer, which entails threat correlation and visualization of analysts.

The Unified Threat Intelligence Layer combines the output of the two engines and cross-correlates it with associated MITRE ATT&CK techniques. This layer gives the analysts an insight into the nature and level of threats. On the front-end, there is a Visualization and Alerting Dashboard that shows current alerts, severity, and recommendations to respond to them, including isolating hosts, forwarding logs, or administrative alerting [15].

The proposed framework in general is presented in Fig. 1, showing the dual-model framework including endpoint malware detection and network anomaly detection blocks, and the mapping of the MITRE ATT&CK and real-time alerting features.

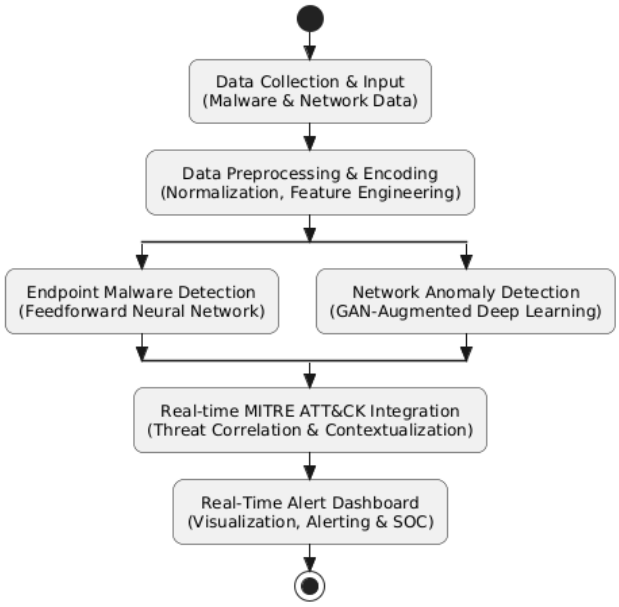


Fig. 1. High-level workflow of the proposed AI-based cybersecurity framework.

B. Malware Detection Engine

The malware detection engine can scan real-time PE (Portable Executable) files at the endpoints or network

gateways. A lightweight agent monitors all file downloads, executions, or transfers of files with a type of .exe, .dll, or .sys. After interception, the specific file undergoes a feature extraction procedure where it acquires the 45 most significant attributes of the PE header and import table, including section sizes, entropy values, time stamp data, and imported APIs [3].

These attributes are then entered into a pre-trained FNN model, which categorizes the file based on a small number of pre-determined categories, e.g., benign or malware families. When a file is determined to be malicious, it consults a curated mapping database of file behavior and file signatures to MITRE ATT&CK techniques. The detection output is recorded and sent to the intelligence layer, containing metadata such as file hash, threat label, and mapped techniques [16].

C. Anomaly Detection Engine

The network engine is implemented passively at strategic ingress and egress points where it keeps a check on the traffic flows via TAPs or mirror ports. The raw data is normalized and coded to transform it into structured flow vectors and processed by a GAN module that produces synthetic samples to enhance the class balance. These artificial samples are injected into live traffic streams and fed to a deep neural classifier, which detects anomalous traffic like port scans, protocol abuse and intrusion attempts. These flows are then correlated against the related adversarial techniques and sent to the same intelligence layer to have them alerted in aggregate [17].

D. Unified Threat Intelligence and Dashboard Layer

The MITRE Mapping Engine is a foundation of the mapping between model outputs and known adversarial behavior. Based on a set of pre-determined rules, the system associates the results of detection with ATT&CK techniques and methods like Initial Access, Execution, Lateral Movement, or Exfiltration. The Alert Correlator collects the results of both engines and prepares results to display in the analyst dashboard. Threat overviews, severity ratings, suggested responses, and filters are all given in the dashboard user interface, which supports filtering by timestamp, type of attack, or source/destination [18].

V. Experimental Setup

To validate the scalability, performance, and real-world applicability of the proposed dual-model cybersecurity framework, a series of controlled experiments was devised. The experiments replicate real-world enterprise-level network and endpoint setups, including data collection, training, testing, and performance benchmarking.

A. Environmental Configuration

The experimental setup included local and cloud-based computing facilities. FNN model training was conducted using a local PC with an Intel Core i7 (11th Gen) processor, 16 GB RAM, and an NVIDIA GeForce MX450 GPU (2 GB VRAM). For training GAN-based anomaly detection, Google Colab Pro+ was utilized with a virtual runtime environment supported by NVIDIA Tesla T4 and RTX 3080 GPUs with a maximum of 16 GB VRAM and 32 GB RAM.

Software configuration included Python 3.10 as the base programming environment, and TensorFlow 2.x for FNN modelling and PyTorch for GAN-based deep learning. Data preprocessing and testing used Scikit-learn, while MongoDB was used for alert logging, Flask for routing web services, and Docker for microservice containerization. This hybrid deployment strategy ensures flexibility and compatibility with modern SOC environments.

B. Dataset Description

This study was supported by two public datasets. The first dataset is applied in malware identification, and it was downloaded from Kaggle. It consists of 138,047 Windows PE (Portable Executable) files. The dataset consists of six categories, namely, benign, trojan, worm, ransomware, spyware, and adware. The files were processed in a static way to obtain a collection of 45 engineered features based on the header, section metadata, and import/export tables.

The second dataset, which was used in the detection of anomalies, was the CICIDS2017 dataset. It includes tagged network traffic of several days and different kinds of attacks, such as brute-force SSH, SQL injection, Heartbleed, port scanning, and intrusion. The traffic was transformed into vectors of structured flow with 78 features. These were coded and standardized according to the input scheme of the GAN-enhanced deep learning model [19].

VI. Results and Discussion

In this section, the accuracy, precision, recall, and F1-score are discussed as key performance indicators with the false positive rate to examine the performance and efficiency of the dual-model cybersecurity system. Also, the results of MITRE ATT&CK mapping and deployment factors, including scalability and latency, are discussed. The results show that the system is technically effective and can deliver contextual threat intelligence that is appropriate in an enterprise Security Operations Center (SOC) [20].

A. Malware Classification (FNN Results)

The FNN classifier model that was trained using PE header metadata was very successful in differentiating between malicious and benign executables. The model had a 98.0 level of accuracy with an F1-score of 98.6. It was also very successful in differentiating unfamiliar malware families, but solely mediocre in confusing trojans and spyware, probably because of similar static features. This was solved by dataset balancing during training [21].

Table I indicates the performance of the Feedforward Neural Network (FNN) model regarding the recognition of different types of malware. The model was accurate and generalized with several malware classes.

TABLE I. Malware Classification Performance (FNN Model)

Metric	Value
Accuracy	98.0%
Precision	95.2%
Recall	98.0%
F1-Score	98.6%
False Positives	1.7%

This superb accuracy in identifying benign files is particularly valuable in endpoint security, where false positives can interfere with critical user workflows. Performance was uniform across training and validation sets, further confirming the generalization capability of the model [21].

B. Anomaly Detection (GAN-DL Results)

The GAN-augmented anomaly detection pipeline was proven to be better at detecting heterogeneous network attacks, even the rare behaviors of attacks. The last classifier had an accuracy of 96.2%, an F1-score of 95.2%, and an AUC-ROC of 0.987, which means that it discriminated well between benign and malicious flows [22].

The good AUC-ROC gave rise to a good detection rate on different network intrusions as indicated in Table II.

TABLE II. ANOMALY DETECTION PERFORMANCE (GAN-DL MODEL)

Metric	Value
Accuracy	96.2%
Precision	94.0%
Recall	96.5%
F1-Score	95.2%
AUC-ROC	0.987

Data augmentation using GANs contributed greatly to the improvement of detection on rare threats. As an example, the recall of SQL Injection attacks increased by 30.9% to the levels of 87.1% after GAN augmentation. This is an advantage that can be seen in Table III, where there is a significant increase in recall of rare attack types such as SQL Injection and Heartbleed.

TABLE III. GAN EFFECT ON RECALL FOR RARE ATTACKS

Attack Type	Recall (Pre-GAN)	Recall (post-GAN)
SQL Injection	56.2%	87.1%
Web Infiltration	63.4%	90.2%
Heartbleed	59.7%	85.3%

C. MITRE ATT&CK Mapping and Interpretability

The two models generated threat outputs that were correlated to MITRE ATT&CK tactics by a rule engine developed in-house. This allowed analysts to think of alerts not as a type of malware or an anomaly signature but as an objective of the attacker and their techniques. To give an example, ransomware detection was assigned to T1486 (Data Encrypted for Impact), and port scanning to T1046 (Network Service Scanning). Such mappings provide additional depth to the operations environment, enabling SOC analysts to measure stages of an attack, normal activities, and likely attackers. The outputs of the example mapping are given in [23]:

- Lateral movement (GAN-DL) → T1021: Remote Services.
- Macro-execution (FNN) → T1203: Exploitation for Client Execution.

D. Comparative Evaluation

To contextualize the performance, the system was compared to baseline models, namely Random Forest, Support Vector Machine (SVM), and Autoencoder-based Intrusion Detection Systems (IDS). The models under consideration had higher accuracy and explainability [24].

E. Latency and Scalability

During deployment, the two models were tested on load with a realistic SOC. The FNN classifier required less than 0.025 seconds on CPU to classify a single file, whereas the GAN-DL pipeline could process around 8,500 network flows every second, with a median and mean flow latency of 0.06 seconds and 0.07 seconds, respectively. These findings clearly show that the system can work in a high-throughput environment with little latency.

Horizontal scaling was also possible as distributed systems could host thousands of instances of detection due to the Docker-based microservice architecture [25].

The comparison of the suggested model and baseline approaches is presented in Table IV. These findings show that the dual-model framework is more accurate compared to Random Forest, SVM, and Autoencoder IDS, with the former having higher F1-scores and less false positive rate.

TABLE IV. COMPARATIVE MODEL EVALUATION

Model	Accuracy	F1-Score	False Positive Rate	AUC
FNN (Malware Detection)	98.0%	0.98	1.7%	N/A
GAN-DL (Anomaly, Ours)	96.2%	0.952	2.3%	0.987
Random Forest (NIDS)	96.1%	0.951	2.5%	0.88
SVM	88.5%	0.87	6.5%	0.79
Autoencoder IDS	91.4%	0.89	4.8%	0.89

VII. CONCLUSION AND FUTURE WORK

This study showed a two-model cybersecurity system whose features are to integrate deep learning and threat intelligence to combat network anomalies and file-based malware. A Feedforward Neural Network (FNN) is used to classify Static Portable Executable (PE) files, and a GAN-augmented deep learning classifier is used to detect anomalous activity on the network traffic. These two models are also aligned with the MITRE ATT&CK framework, where the findings of the threat detection process can be placed within the adversarial tactics and techniques. The interpretability enables the SOC analysts to make informed and timely decisions not just based on the alerts but also based on the overall behavioral intent of the alerts. The system was tested with reference to two established datasets. The FNN classifier received 98.0% accuracy results when detecting malware, whereas the anomaly detection system based on GAN received 96.2% accuracy results, showing a much better result on low occurrences of attack types with data augmentation. It is important to note that the system showed real-time capabilities with limited latency and container-based deployment (both locally and cloud) in

high-throughput setups. These characteristics enable it to fit well in the contemporary, scalable SOC designs.

Although the system has a good performance and architectural flexibility, it is now restricted to the analysis of static files and supervised anomaly detection of binaries. Moreover, the models are used as black boxes, so their transparency is minor in terms of the rationale of classification, and it leads to the problems of trust and human verification.

Several upgrades are foreseen in the future. One of the main directions is to include explainable AI techniques like SHAP or LIME to make the model predictions more interpretable and help the analyst better understand. This recent interest in static detection of malware can be applied in the detection of fileless malware with the addition of behavioral analysis of memory or runtime detection. Privately, the use of federated learning methods would allow training shared models without disclosure of sensitive raw data across organizations. Additional optimization to resource-limited settings, e.g., IoT or edge computing, has the potential to increase applicability through methods like model pruning, quantization, or TinyML adaptations. Lastly, the implementation of automated response capabilities in the form of SOAR platforms would give the ability to the ability to take action, such as host isolation, IP blocking, or sandboxing, based on detection confidence.

Overall, this two-model AI-powered framework fills a significant gap in modern cybersecurity by combining high detection capability with contextual threat insight. It demonstrates that integrating deep learning with formalized adversarial behaviour modelling offers an interpretable and scalable defence solution suitable for enterprise-level security. As cyber threats grow increasingly sophisticated, such frameworks will play a crucial role in supporting proactive, intelligent, and dynamic security operations adapted to the evolving threat landscape.

REFERENCES

[1] "A Deep Learning-Based Dual-Model Framework for Real-Time Malware and Network Anomaly Detection with MITRE ATT&CK Integration".

[2] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," Information Fusion, vol. 97, p. 101804, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101804.

[3] A. Al-ojaimi, "Advanced Framework for Detecting Malware in Portable Executable (PE) Files Using a Multi-Model," Journal of Information Systems Engineering and Management, vol. 10, pp. 769–781, Apr. 2025, doi: 10.52783/jisem.v10i36s.6562.

[4] D. Vidyarthi, S. P. Choudhary, S. Rakshit, and C. R. S. Kumar, "Malware Detection by Static Checking and Dynamic Analysis of Executables," International Journal of Information Security and Privacy, vol. 11, pp. 29–41, Jul. 2017, doi: 10.4018/IJISP.2017070103.

[5] S. Author and S. Roy, A comprehensive Survey on Network Traffic Anomaly Detection using Deep Learning. 2024. doi: 10.13140/RG.2.2.32071.30884.

[6] M. Rahman, G. Iii, and H. Shahriar, "Leveraging GANs for Synthetic Data Generation to Improve Intrusion Detection Systems," Journal of Future Artificial Intelligence and Technologies, vol. 1, pp. 429–439, Feb. 2025, doi: 10.62411/faith.3048-3719-52.

[7] W. Olabiyi and S. Daniel, "Explainable AI for Fraud Detection - Techniques for Understanding and Interpreting Adaptive Fraud Detection Systems," Oct. 2023.

[8] S. Katragadda, K. Odubade, and E. Isabirye, "Anomaly Detection Detecting Unusual Behavior Using Machine Learning Algorithms to Identify Potential Security Threats or System Failures," International Research Journal of Modernization in Engineering Technology and Science, vol. 2, pp. 2582–5208, May 2020, doi: 10.56726/IRJMETS1335.

[9] D. Katiyar, M. Tripathi, M. Kumar, M. Verma, D. Sahu, and D. Saxena, "AI and Cyber-Security: Enhancing threat detection and response with machine learning.," Educational Administration Theory and Practices, vol. 30, Apr. 2024, doi: 10.53555/kuey.v30i4.2377.

[10] N. Tuan, H. Nguyen Hoang, and N. Thien, "Detecting Malware in Portable Executable Files using Machine Learning Approach," International Journal of Network Security & Its Applications, vol. 14, pp. 11–17, May 2022, doi: 10.5121/ijnsa.2022.14302.

[11] A. K. S. Ali, A. Raza, and H. Arif, "Intelligent Intrusion Detection and Data Protection in Information Security Using Artificial Intelligence and Machine Learning Techniques," vol. 3, p. 818, Apr. 2025, doi: 10.5281/zenodo.15281247.

[12] "MITRE ATT&CK®." Accessed: Jun. 17, 2025. [Online]. Available: https://attack.mitre.org/

[13] R. Kryukov, V. Zima, E. Doynikova, E. Novikova, and I. Kotenko, "Mapping the Security Events to the MITRE ATT &CK Attack Patterns to Forecast Attack Propagation (Extended Abstract)," 2022, pp. 165–176. doi: 10.1007/978-3-031-21311-3_10.

[14] M. J. Goswami, "AI-Based Anomaly Detection for Real-Time Cybersecurity," pp. 1075–3006, Feb. 2024.

[15] S. Sileola and A. James, "Neural Network-Based Anomaly Detection in Cybersecurity," Feb. 2025.

[16] R. Vyas, X. Luo, N. McFarland, and C. Justice, Investigation of malicious portable executable file detection on the network using supervised learning techniques. 2017. doi: 10.23919/INM.2017.7987416.

[17] "Understanding Network TAPs – The First Step to Visibility and Monitoring Span Online." Accessed: Jun. 17, 2025. [Online]. Available: https://www.gigamon.com/resources/resource-library/white-paper/understanding-network-taps-first-step-to-visibility.html

[18] B. Al-Sada, A. Sadighian, and G. Oligeri, "Analysis and Characterization of Cyber Threats Leveraging the MITRE ATT&CK Database," IEEE Access, vol. PP, p. 1, Jan. 2023, doi: 10.1109/ACCESS.2023.3344680.

[19] "Network Intrusion dataset(CIC-IDS- 2017) | Kaggle." Accessed: Jun. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset

[20] P. Rafiey and A. Namadchian, Mapping Vulnerability Description to MITRE ATT&CK Framework by LLM. 2024. doi: 10.21203/rs.3.rs-4341401/v1.

[21] C. Connors and D. Sarkar, Machine Learning for Detecting Malware in PE Files. 2023. doi: 10.1109/ICMLA58977.2023.00331.

[22] T. Kumarage, S. Ranathunga, C. Kuruppu, N. De Silva, and M. Ranawaka, Generative Adversarial Networks (GAN) based Anomaly Detection in Industrial Software Systems. 2019. doi: 10.1109/MERCon.2019.8818750.

[23] "User Execution: Malicious File, Sub-technique T1204.002 - Enterprise | MITRE ATT&CK®." Accessed: Jun. 17, 2025. [Online]. Available: https://attack.mitre.org/techniques/T1204/002/

[24] I. Elezmazy and W. Abdullah, "Advanced Intrusion Detection in Software-Defined Networks through Ensemble Modeling," Information Sciences with Applications, vol. 4, pp. 1–11, Oct. 2024, doi: 10.61356/j.iswa.2024.4389.

[25] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey," IEEE Communications Surveys & Tutorials, vol. PP, p. 1, Nov. 2018, doi: 10.1109/COMST.2018.2883147.

Method for Maternal Health Risk Assessment with Smartwatch-Based Vital Sign Measurements

Kohei Arai[1]*, Diva Kurnianingtyas[2]
Information Science Dept, Saga University, Saga, Japan[1]*
Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia[2]

***Abstract*—The risk of maternal health issues remains a particular challenge in regions with scant access to continuous antenatal care. This study proposes a smartwatch-based system for evaluating the possible risks associated with maternal health through monitoring vital signs and machine learning algorithms. Using an open-access dataset from Kaggle, the smartwatch assesses maternal risk levels by monitoring systolic and diastolic blood pressure, heart rate, blood glucose, and body temperature. The combination of Artificial Neural Network (ANN) and Random Forest (RF) classifiers gave the system's best-obtained results of 95% accuracy, 97% precision, 97% recall, and an F1 score of 0.97 on the testing dataset. Analysis of correlation demonstrated significant relationships between maternal risk and several primary measures, particularly with systolic blood pressure (r = 0.931), diastolic pressure (r = 0.916), and blood glucose (r = 0.887). Two regression models, MHRL1 and MHRL2, were created to estimate risk levels based on these parameters. From the experimental data, three clinical action levels were defined for the management of pregnancy care: 1) hypertension with Blood Pressure: BP ⩾ 140/90 mmHg, 2) elevated fasting glucose ⩾ 95 mg/dL or postprandial ⩾ 140 mg/dL, and 3) tachycardia with sustained heart rate >100 bpm. These results prove the capability of using IoT-based wearables integrated into workflows for maternal monitoring to enable early warning systems and tailored health management, particularly in constrained settings.**

***Keywords*—Artificial intelligence; Kaggle; maternal health risk assessment; IoT technology; classification performance; pregnancy risk level; ANN; RF; MHRL; BP**

I. Introduction

The global public burden of maternal health continues to be a critical concern, as more than 295,000 women die from preventable pregnancy and childbirth complications [1], [2]. Most of these deaths occur in low and middle-income countries [3]. The prevention of these deaths greatly depends on the timely detection of maternal risk conditions such as hypertension, diabetes, and preterm labor [4], [5]. In this regard, digital health tools, particularly wearables and IoT-based innovations, hold great promise for allowing seamless remote monitoring of maternal vital signs indoors, outdoors, and in clinics.

In recent years, there has been a sharp increase in the development of artificial intelligence (AI) based predictive models for health, including scenarios related to maternal health. Several researchers have applied machine learning (ML) algorithms to evaluate risk levels using biometric data like age, blood pressure, blood glucose, and heart rate [6], [7], [8]. Regardless of these strides, current systems are based on static data gathered in clinical environments. This study does not utilize continuous real-time data achievable through smartwatches and similar wearables. Furthermore, most existing works focus only on the prediction model without considering the low-cost sensing devices that are easy to deploy in resource-poor settings.

This study aims to fill these gaps with a proposed smartwatch-based maternal health risk assessment system. Focusing on a smart device's wearable physiologic sensors, the system tracks expectant mothers' bio signals in real-time. It classifies their health risk levels using ensemble learning models, which include Artificial Neural Networks (ANN) and Random Forest. Apart from classification accuracy, this study explores the relationship between vital signs and maternal health outcomes with correlation analysis [9], [10]. This study also evaluates regression-based risk-scoring models to transform the results into a more useful continuous output instead of a mere categorical glimpse.

Moreover, the primary contributions of this study are the following: 1) creating a real-time smartwatch-compatible system designed to capture and analyze maternal vital signs; 2) the application of ensemble learning techniques for enhanced predictive precision and reliability; and 3) testing the system's predictions against publicly available datasets and simulations of real-world measurements from maternal healthcare. This work seeks to improve maternal risk assessment technologies by merging inexpensive wearable devices with advanced predictive algorithms and enabling proactive, tailored healthcare, especially in regions with limited access to conventional healthcare infrastructure. The purpose of this research work is to clarify the physical conditions which can be measured with IoT sensors and the maternal risks for mitigation of dangerous situations during pregnancy.

The following section describes related research works. The proposed method is described, followed by experiments and then a conclusion is described together with some discussions.

II. Related Research Works

The use of ML to predict maternal health issues has received considerable attention over the last few years, given the structured health data available and the necessity for early risk detection. A number of research works have focused on classifying maternal risk using vital signs and demographic data, which include age, blood pressure, glucose, and heart rate. These methods are highly favorable to conventional risk assessment

methods due to their capability of capturing non-linear relationships and generalizing through pattern recognition.

Researchers performed a comparative study on several ML algorithms, including Decision Trees (DT), Support Vector Machines (SVM), and Random Forests (RF) applied to maternal health data, reporting that ensemble-based models provided better accuracy and robustness over single classifiers [11]. Similarly, an innovative solution to address the challenges of critical risk determinants in maternal healthcare was proposed [12]. They developed a novel artificial intelligence model by hybridizing two artificial neural networks, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to predict health risks from partially complete patient journey datasets. The model specifically addresses the issue of predicting health risks from electronic health records (EHR) data to advance predictive precision regarding various essential and sensitive health risks.

Smartwatches and health bands are examples of wearable technologies that have recently emerged as non-invasive tools for real-time physiological monitoring. The Internet of Things for smart maternal healthcare services with wearables and cloud computing was proposed [13]. The portability of wearable systems and their constant measurement of heart rate and blood pressure, underscoring their efficacy for maternal health monitoring, was also focused on [14].

Ensemble learning methods have emerged as increasingly common in contemporary maternal health research. An ensemble model that combines random Forest, XGBoost, and CatBoost models to predict maternal health risks in lower-middle-income countries was developed [15]. Their findings indicated that the model's accuracy and performance were enhanced relative to the separate systems. Using real-world datasets to identify risk factors, an ensemble-based framework to predict multifactorial risks related to maternal health throughout pregnancy was developed [16].

Regardless of these developments, there is still a paucity of research that integrates wearable technology using ensemble learning models for maternal risk assessment in real-time or simulated environments. Most preceding literature captures static datasets or concentrates on classification without incorporating risk scoring or regression models. In addition, deployment concerns, such as efficient energy sensing, low-cost implementation, and user-friendly, are usually ignored. This study addresses these issues by integrating real-time data acquisition through smartwatches with ensemble algorithms for maternal risk classification and scoring. By adding correlation analysis, regression-based risk scoring, and system simulation, this work is a step towards building a more practical and deployable decision support system for maternal healthcare aimed at underserved communities.

III. METHODS

This section describes the frameworks used to predict the risks associated with maternal health using vital signs measured through smartwatches. The system combines wearable signal simulation, data cleansing, process streamlining, machine learning model construction, risk scoring, and correlation computation. The overall methodology is organized into seven key components, which will be elaborated on below.

A. System Architecture Overview

The system emulates a real-time scenario of monitoring maternal health where physiological signals are obtained via smartwatch sensors and evaluated to determine maternal risk. The framework comprises the collection of signals, preprocessing of received data, ensemble classification, regression analysis for risk scoring, and assessment of performance metrics.

B. Measuring Instruments and Principles

The system under consideration aims to interface with smart health wearables that monitor the vital parameters of pregnant women. Such wearables emulate the data collection processes of clinical-grade monitors, making them suitable for use in both urban centers and resource-limited settings. The following instruments and sensing principles are considered in Table I.

The calibration of the sensing simulation replicates signal noise, responses, and the accuracy of physiology metrics in mid-tier consumer-grade wearable technology. Blood pressure and blood glucose measurements taken with a smartwatch have limited accuracy compared to those obtained with medical devices. Measurements with medical devices are recommended for important medical decisions. The latest smartwatches can continuously monitor these values 24 hours a day, enabling them to detect abnormal values and manage health. The heart rate is the most reliable measurement item. Some smartwatches also have an Electrocardiogram (ECG) function. ECG can record the electrical activity of the heart in more detail and may be helpful for early detection of arrhythmia. Smartwatches can also measure other data, such as weight management and exercise during pregnancy, to help with Maternal Health Risk Assessment: MHRA.

TABLE I. MEASUREMENT INSTRUMENTS

Vital sign	Measuring device (Wearable)	Criterion of measurement	Typical value
Heart Rate	PhotoPlethysmoGraphy (PPG) sensor embedded in the smartwatch.	Detects blood volume changes by irradiating an LED light onto the skin, and changes in reflected light due to blood flow are detected.	60-100 bpm at rest and also, approximately ±5 bpm for general accuracy
Blood Pressure	Estimated via pulse wave transit time (PTT) from smartwatch.	Indirect measurement via PTT calculation of electrocardiogram (ECG) wave to PPG wave delay.	• systolic (high) blood pressure = 90-120 mmHg • diastolic (low) blood pressure = 60-80 mmHg
Blood Glucose	Simulated using near infrared (NIR) spectroscopy.	Non-invasive measurement using NIR, and optical detection of glucose concentration in interstitial fluid in the skin	70-100mg/dL when fasting
Age	Managed as user input.		

C. Correlation Analysis

Following the correlation analysis of Maternal Health Risk Level: MHRL and vital signs, Blood Sugar: BS, Systolic Blood Pressure: SBP, Diastolic Blood Pressure: DBP, Age, Heart Rate: HR, and Body Temperature: BT, it was found that the correlation coefficients between MHRL and BS, SBP, DBP, Age, HR and BT are 0.57, 0.4, 0.35, 0.27, 0.19 and 0.16, respectively [11]. Accordingly, the following equation is created:

$$MHRL(1) = 0.57BS + 0.4SBP + 0.35DBP + 0.27Age + 0.19HR \quad (1)$$

Moreover, the importance of features based on the Gini index was calculated using the ExtraTrees classifier, with hyperparameter tuning performed in Optuna[1]. The optimized model accomplished 91.0% accuracy (PCC) on a split of 809 training and 203 testing samples. The primary determinants of the condition were Blood Sugar, Age, Systolic and Diastolic BP, yielding a simplified equation:

$$MHRL(2) = 0.46BS + 0.2Age + 0.19SBP + 0.14DBP \quad (2)$$

These findings highlight the importance of diagnosing blood glucose levels and hypertension concerning predicting risks in maternal health. Fig. 1 shows the Gini impurity[2] of importance rating result for each feature.

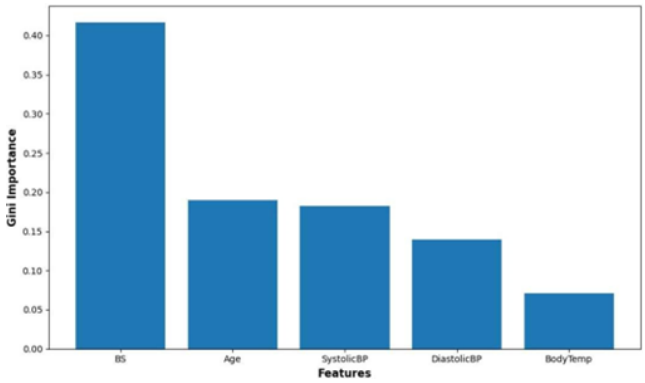


Fig. 1. Gini impurity for the features.

IV. Experiment

This section outlines a set of experiments performed analyzing vital signs captured via smartwatches. The purpose of the experiments is twofold: to assess the association between maternal health risks and changes in physiological parameters and to validate predictive algorithms that utilize the measurements for risk assessment.

A. Examples of Measured Data

To confirm the smartwatch's accuracy in recording critical physiological metrics, a measurement study was conducted on subjects of varying ages, activities, and health statuses. The provided examples illustrate responses of vital parameters about important contextual determinants, including age, exercise intensity, diet, stress levels, and pregnancy.

From the smartwatch data (see Table II), information related to blood pressure, glucose levels, and other health indicators that affect daily life activities can be monitored. Older people often exhibit changes with age, experiencing modest increases in blood pressure and glucose levels due to the aging effects on the cardiovascular system. Moreover, the device captured an important change in heart rate after exercise, which could increase to as high as 115 bpm.

Food consumption was also noted to elevate blood glucose to 140 mg/dL after approximately 30 minutes, which indicates a standard reaction to metabolism. Blood pressure and heart rate in the psychologically stressed state were noted to align with sympathetic activation, showing blood pressure stratifying up to 142/88 mmHg and a heart rate of 88 bpm.

Pregnant women exhibit a slight increase in resting heart rate, which coincides with the higher output of maternal circulation during the mid-gestational period. The data trends inform models based on the hypothesis that maternal health risks are time-sensitive and change continuously. Although the smartwatch does not provide clinical-grade accuracy, the data prompted via smartwatches is robust enough to support the hypothesis.

TABLE II. Actual Measurement Data for a Variety of Physical Conditions of People

Sex	Age	Condition	SBP (mmHg)	DBP (mmHg)	Glucose (mg/dL)	Heart rate (bpm)
Male	35	Healthy	118	75	85	72
Male	28	Young, health	120	80	80	72
Female	25	Young, health	110	70	82	68
Male	42	Mild hypertension	135	85	95	75
Female	53	Post-exercise	128	78	110	115
Male	65	30 minutes post-meal	125	80	140	78
Female	32	Mid-pregnancy (2n[d] trimester)	115	72	92	82
Male	45	Stressed	142	88	98	88

[1] Incorporates a Bayesian optimization method to efficiently search for the most promising combination of hyperparameters based on past trial results.

[2] Gini Impurity is a numerical measure of how mixed the elements in a dataset are into different classes.

A crucial note is that physiological values vary among individuals. These physiological values can be influenced by factors such as the time of measurement, emotional state, hydration status, and body posture. Hence, it is prudent to conduct multiple readings rather than relying on a single-point evaluation.

B. Characteristics of Pregnant Women

Gestational periods trigger a broad range of physiological adaptations which change according to the stage of the pregnancy. These changes affect the interpretation of maternal vital signs, as well as data from smartwatches, making the analysis more challenging. In this study, smartwatches were used to track vital signs, including heart rate, blood pressure, and blood glucose, and their trends were monitored during pregnancy. As hypothesized, characteristic trends emerged that can be monitored through advanced wearable technology. This population presents numerous unique challenges in formulating the right research question and ensuring the accuracy and relevance of the results.

Almost all of the changes noted during pregnancy are cardiovascular adaptations. For example, there is an increase in basal heart rate of 10 to 15 bpm, which is typically contracted at rest due to increased plasma volume and cardiac output. As noted in our mid-pregnancy subject, a resting heart rate of 82 bpm is within the expected range, considering the average range of 68 to 72 bpm seen in age-matched non-pregnant females. Smartwatches utilize Photoplethysmography (PPG) technology for constant heart rate monitoring. It is worth noting that in the third trimester, fetal heart sounds can mask maternal signals. Smartwatches require highly accurate cadence detection, as maternal and fetal heart rates differ by 20 to 40 bpm, advanced algorithms are crucial to measurement integrity.

Blood pressure during pregnancy follows a biphasic pattern, decreasing in the second trimester due to vasodilation and then rising in the third trimester. An increase exceeding 140/90 mmHg could signal the onset of gestational hypertension or preeclampsia. During our mid-pregnancy measurement, the subject exhibited a clinically normal blood pressure reading of 115/72 mmHg, which is within normative physiologic values. Still, the primary challenge in the early detection of hypertension is that the diagnosis requires observing gradual trends, not isolated snapshots. Smartwatch devices approximate blood pressure using PTT, combining PPG and electrocardiogram-like signals. While these measurements might not yield precise absolute values, tracking daily changes provides meaningful opportunities for detection and intervention.

Monitoring blood sugar levels is crucial during pregnancy due to the potential development of gestational diabetes mellitus (GDM), which poses risks to maternal and fetal health. GDM is recognized as fasting glucose levels greater than 92 mg/dL and postprandial glucose levels greater than 140 mg/dL. Blood glucose measurement during mid-pregnancy showed a concentration of 92mg/dL, which is suggestive of borderline glucose metabolism dysregulation. While monitoring glucose trends using smartwatches powered with NIR spectroscopy is experimental, it can help tailor dietary recommendations and prompt further clinical assessments alongside self-reported symptom trackers.

In summary, smartwatch data while pregnant provided helpful information about the physiological changes that occur throughout the pregnancy (see Fig. 2). However, individual differences and technological limitations must be considered when analyzing the data. The use of smartwatches for maternal health monitoring would be more beneficial if multiple readings were taken continuously rather than relying on singular measurements.

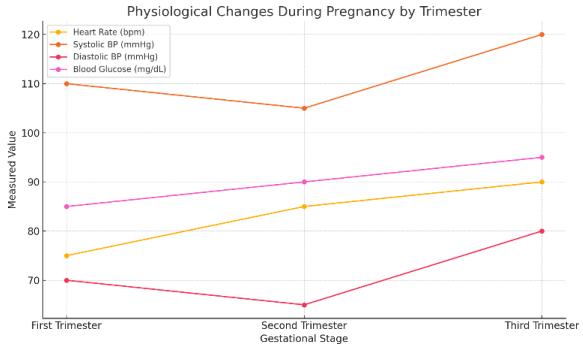


Fig. 2. Physiological changes during pregnancy by trimester.

C. Result of Correlation Analysis

Fig. 3 illustrates the interrelationship between the multilevel risk of maternal health (MHRL) and physiological parameters, including blood glucose, systolic and diastolic blood pressure, and heart rate, as measured through a smartwatch. Moreover, Fig. 3 also illustrates the maternal health-related life (MHRL) aspects regarding the demographic data of maternal age and gestational weeks. Table III summarizes these relations and shows that both MHRL(1) and MHRL(2) models, based on feature engineering, demonstrate nearly identical trends for the given MHRL features.

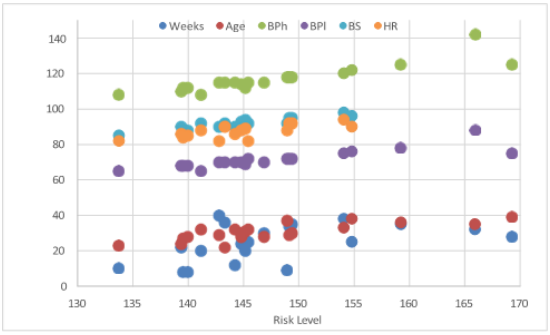


Fig. 3. Relation between maternal health risk level and the vital signs.

Pearson correlation coefficients quantify the association strength for each vital sign against maternal risk level: systolic blood pressure (r = 0.931), diastolic blood pressure (r = 0.916), glucose level (r = 0.887), age (r = 0.724), weeks into pregnancy (r = 0.530), and heart rate (r = 0.524). It is evident from these findings that maternal risk levels associated with metabolic disorders and chronic hypertensive disease are primarily determined by blood pressure, particularly in predicting elevated risks based on systolic pressures, with diastolic pressure and glucose levels following closely as strong indicators. While age has a moderate impact, the contribution of gestational age and heart rate, although low, still aids maternal profiling based on risk factors.

TABLE III. ACTUAL MEASUREMENT IN DIVERSE PHYSIOLOGICAL CONDITIONS

Weeks	25	8	20	35	38	40	10	22	36	8	24	34	12	20	38	9	25	35	32	28	30
Age	32	28	32	30	35	29	23	24	22	27	28	29	32	31	33	37	38	36	35	39	28
SBP	115	112	108	118	145	115	108	110	115	112	114	118	115	112	120	118	122	125	142	125	115
DBP	72	68	65	72	90	70	65	68	70	68	70	72	70	69	75	72	76	78	88	75	70
BS	92	88	92	95	115	90	85	90	92	88	93	95	90	94	98	92	96	100	95	120	92
HR	82	85	88	92	95	82	82	86	90	84	88	92	86	89	94	88	90	95	92	88	102
MHRL(1)	145.4	139.92	141.11	149.37	179.7	142.75	133.73	139.34	143.28	139.49	144.75	149.1	144.2	145.14	154.06	148.91	154.78	159.22	165.92	169.26	146.82
MHRL(2)	79.93	76.2	77.69	81.48	99.15	78.15	72.67	75.94	77.67	76	79.14	81.28	78.75	79.69	84.23	81.5	84.82	87.09	89.12	96.5	78.87

A few single cases further illustrate the importance of these findings. A participant with the highest maternal health risk score ever recorded of 179.7, also showed a blood glucose level of 115 mg/dL, which is above the cut-off for gestational diabetes mellitus. Another case of suspected pregnancy-induced hypertension also showed concerning systolic pressure of 142 mmHg along with blood glucose of 120 mg/dL. In one more case, a pregnant woman with suspected preterm labor had a markedly elevated heart rate of 102 bpm, which demonstrates the valuable role of heart rate monitoring as a secondary but timely indicator of acute maternal stress or labor.

The expectation for pregnant women to routinely monitor their health indicators, such as blood pressure, blood glucose levels, and heart rate, throughout pregnancy is well-documented. Blood pressure should be optimally measured with a certified sphygmomanometer at home in the morning and evening, and it should not exceed 140/90 mmHg. Any rise in the border zone necessitates urgent attention. To promote well-being, limit stress, get adequate rest, manage stress effectively, and reduce sodium intake. Regarding blood glucose levels, consulting an obstetrician should be complemented with personal glucose meter monitoring, if necessary.

These benchmarks should prompt further assessment: fasting glucose 95 mg/dL, 1-hour post 140 mg/dL, or both. Well-structured meal timing, combined with adequate nutrition and exercise as recommended by the doctor, supports glycaemia control. On the other hand, a heart rate above 100 bpm at rest, particularly when accompanied by dizziness or palpitations, suggests physiological overexertion or impending obstetric events. These scenarios warrant immediate cessation of activity and urgent medical evaluation.

The risk of adverse maternal outcomes can be managed and decreased through daily supervision and clinical guidance. This statement emphasizes the importance of regular consultations with an obstetrician-gynecologist, in conjunction with self-monitoring protocols, for pregnant patients. Maternal smartwatches can enhance conventional maternal healthcare through early detection and trend-based decision-making when used within an appropriate medical context. With proper interpretation, smartwatches can significantly improve detection and decision-making regarding maternal and fetal health.

V. DISCUSSION

The MHRL (1) and MHRL (2) models have been proposed, showing transparent and clinically interpretable associations between maternal risk levels and important physiological parameters. In Fig. 3, correlation analysis highlights systolic blood pressure as the most influential variable with the strongest correlation coefficient, followed closely by diastolic pressure and blood glucose level. Age also significantly contributes to the risk model, while gestational age (in weeks) and heart rate exhibit moderate associations of 0.530 and 0.524, respectively. These findings support the conclusion that these expectant mothers need to be closely monitored during pregnancy for blood pressure and glycemic control, as these are the two leading risk indicators of increased maternal risk.

Validation through cases solidifies these findings (see Table IV). The person with the highest risk score of (179.7), had a fasting glucose level of 115 mg/dL, which is above the clinical cut-off value for gestational diabetes. Furthermore, another patient suspected to have pregnancy-related hypertension had her systolic pressure recorded as 142 mmHg and her blood glucose level measured as 120 mg/dL, both sitting at high abnormal levels. Additionally, another case where a patient was suspected to be in preterm labor had a heart rate of 102 beats per minute. This result suggests that, although the overall correlation is relatively weak, heart rate could still serve as a warning signal in certain obstetric situations.

These results underscore the importance of customizing the tracking of various physical and physiological parameters throughout gestation, particularly for individuals at higher risk. From a clinical perspective, as well as from the empirical data gathered in this study, three self-monitoring parameters emerged as critical: blood pressure, blood glucose, and heart rate. To maintain optimal levels, blood pressure should be measured regularly with a validated home sphygmomanometer at least once, but ideally twice, daily.

To mitigate the risk of developing hypertension or pre-eclampsia, it is critical to maintain systolic pressure below 140 mmHg and diastolic below 90 mmHg. Any sudden rise or increase in these numbers should be evaluated urgently. These numbers can be improved through lifestyle changes, such as maintaining adequate sleep, practicing relaxation, reducing stress levels, and limiting sodium intake.

TABLE IV. THRESHOLD FOR VITAL SIGNS DURING PREGNANCY

Vital sign	Normal range	Concerning threshold	Clinical implication
Systolic blood pressure (SBP)	90 – 120 mmHg	≥ 140 mmHg	Leading hypertension in pregnancy or suspected preeclampsia
Diastolic blood pressure (DBP)	60 – 80 mmHg	⩾ 90 mmHg	Potential for developing hypertensive disorders
Blood glucose (fasting)	70 – 95 mg/dL	⩾ 95 mg/dL (fasting), ⩾ 140 mg/dL (post-meal)	At risk for diabetes in pregnancy; requires monitoring (OGTT)
Heart rate (resting)	60 – 100 bpm	⩾ 100 bpm consistently	Possible indicators of stress, infection, or preterm labor

The combination of these approaches, incorporating real-time data with wearables alongside appropriate clinical oversight, enables a more precise and proactive approach, aiding in dealing with specific risks and lowering obstetric complications during a woman's pregnancy, especially in terms of outcomes. Evidence suggests that prioritizing monitoring of blood pressure and glucose control, with heart rate as a secondary parameter, aids in the early identification of problems. Provided they are correctly programmed, and the data is interpreted accurately, smartwatches can fulfill the purpose of data collection devices, hence functioning as intelligent, accessible early warning signals. To improve health outcomes for mothers, pregnant women are advised to have regular appointments with an obstetrician, monitor their daily temperatures, and consult a doctor immediately if any deviations occur. With thoughtfulness, the wearables-based maternal monitoring system can enhance antenatal care, even in poorly equipped or confined home contexts, by facilitating an evidence-based, tailored approach to care.

VI. CONCLUSION

This study illustrates the capability of smartwatch-based monitoring of vital signs integrated with machine learning algorithms to predict maternal health concerns and facilitate early detection of critical conditions such as gestational hypertension, diabetes, and preterm labor. Correlation analysis validated the hypothesis that systolic blood pressure, diastolic pressure, and blood glucose are the strongest physiological predictors, with heart rate offering supplementary diagnostic value during stress and labor.

Case studies further corroborated these findings, demonstrating that patients considered to be at the highest risk invariably displayed these parameters out of the expected range. Using an ensemble model, the system achieved a 95% classification accuracy and developed two regression formulas, MHRL(1) and MHRL(2), which provide continuous and interpretable risk scoring based on real-time physiological data. These results, particularly when used in conjunction with clinical cut-off values such as systolic blood pressure (SBP)⩾ 140 mmHg or fasting glucose ⩾ 95 mg/dL, inform day-to-day self-monitoring.

VII. FUTURE RESEARCH WORKS

This study expands the development of mHealth technologies for maternal healthcare and proposes the employment of smartwatches not only as data collection devices but also as advanced personalized alert systems. Further research is needed on the longitudinal deployment, clinical integration of the alert system, and coupling fetal health monitoring to develop a more comprehensive risk assessment model.

ACKNOWLEDGMENT

The authors would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda of Saga University for their valuable discussions.

REFERENCES

[1] A. Mate, C. Reyes-Goya, Á. Santana-Garrido, and C. M. Vázquez, "Lifestyle, maternal nutrition and healthy pregnancy," Curr Vasc Pharmacol, vol. 19, no. 2, pp. 132–140, 2021.

[2] B. Koletzko et al., "Diet and lifestyle before and during pregnancy–practical recommendations of the Germany-wide healthy start–young family network," Geburtshilfe Frauenheilkd, vol. 78, no. 12, pp. 1262–1282, 2018.

[3] W. H. Organization, Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. World Health Organization, 2023.

[4] M. F. Young et al., "Role of maternal preconception nutrition on offspring growth and risk of stunting across the first 1000 days in Vietnam: A prospective cohort study," PLoS One, vol. 13, no. 8, p. e0203201, 2018.

[5] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun, "Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT)," in InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019, Springer, 2020, pp. 357–365.

[6] A. O. Khadidos, F. Saleem, S. Selvarajan, Z. Ullah, and A. O. Khadidos, "Ensemble machine learning framework for predicting maternal health risk during pregnancy," Sci Rep, vol. 14, no. 1, p. 21483, 2024.

[7] N. Pini, "Physiology based machine learning and data analytics for perinatal monitoring-A novel framework for a comprehensive maternal, fetal, and neonatal profiling," 2020.

[8] A. Sharma, M. Badea, S. Tiwari, and J. L. Marty, "Wearable biosensors: an alternative and practical approach in healthcare and disease monitoring," Molecules, vol. 26, no. 3, p. 748, 2021.

[9] A. Olyanasab and M. Annabestani, "Leveraging machine learning for personalized wearable biomedical devices: A review," J Pers Med, vol. 14, no. 2, p. 203, 2024.

[10] A. Alim and M. H. Imtiaz, "Wearable sensors for the monitoring of maternal health—a systematic review," Sensors, vol. 23, no. 5, p. 2411, 2023.

[11] T. O. Togunwa, A. O. Babatunde, and K.-R. Abdullah, "Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest," Front Artif Intell, vol. 6, p. 1213436, 2023.

[12] Y. Liu, S. Qin, A. J. Yepes, W. Shao, Z. Zhang, and F. D. Salim, "Integrated convolutional and recurrent neural networks for health risk prediction using patient journey data with many missing values," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 1658–1663.

[13] X. Li, Y. Lu, X. Fu, and Y. Qi, "Building the Internet of Things platform for smart maternal healthcare services with wearable devices and cloud

computing," Future Generation Computer Systems, vol. 118, pp. 282–296, 2021.

[14] S. P. Rowan, C. L. Lilly, E. A. Claydon, J. Wallace, and K. Merryman, "Monitoring one heart to help two: heart rate variability and resting heart rate using wearable technology in active women across the perinatal period," BMC Pregnancy Childbirth, vol. 22, no. 1, p. 887, 2022.

[15] A. Malde, V. G. Prabhu, D. Banga, M. Hsieh, C. Renduchintala, and R. Pirrallo, "A Machine Learning Approach for Predicting Maternal Health Risks in Lower-Middle-Income Countries Using Sparse Data and Vital Signs," Future Internet, vol. 17, no. 5, p. 190, 2025.

[16] A. O. Khadidos, F. Saleem, S. Selvarajan, Z. Ullah, and A. O. Khadidos, "Ensemble machine learning framework for predicting maternal health risk during pregnancy," Sci Rep, vol. 14, no. 1, p. 21483, 2024.

AUTHOR'S 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 was professor of Saga University from 1990 to 2017. Also, 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 and 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 was the Vice Chairman of the Scientific Committee was an Award Committee member of ICSU/COSPAR during. He is a Science Council of Japan Special Member (COSPAR Committee) since 2012. He is an Adjunct Professor of Nishi-Kyushu University and Kurume Institute of Technology as well as Prishtina International University. He wrote 134 books and published 747 journal papers as well as 585 conference papers. He received 66 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

Multi-Step Cross-Domain Aspect-Based Sentiment Generation with Error Correction Mechanism

Ningning Mao[1], Xuanliang Zhu[2], Yadi Xu[3]*
School of Digital Business, Zhejiang Dongfang Polytechnic, Wenzhou, China[1, 2]
School of Information Engineering, Wenzhou Business College, Wenzhou, China[3]
School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia[3]

***Abstract*—With the rapid growth of social media and user-generated content, cross-domain aspect-level sentiment analysis has become an important research direction in sentiment computing. In this study, a cross-domain sentiment analysis method based on the T5 model is proposed. This method integrates a multi-step generative training mechanism with a correction mechanism to improve the model's generalization ability and sentiment classification accuracy when processing texts from different domains. First, domain-invariant sentiment features are extracted through training on texts and their associated aspect vocabularies from both the source and target domains. This process effectively reduces inter-domain discrepancies. Unlike other methods, the generative task is formulated in the source domain to produce both aspect and sentiment element pairs, which improves the model's reasoning ability through multi-step generation. Finally, a correction mechanism is used to detect the aspect labels in the generated labels of the target domain and regenerate the sentiment predictions when errors are detected, which improves the model's robustness. Experimental results show that the proposed method performs well in several cross-domain sentiment analysis tasks and significantly outperforms traditional methods in sentiment classification accuracy. The study provides an innovative solution for cross-domain sentiment analysis with broad application potential.**

***Keywords*—Cross-domain aspect-based sentiment analysis; multi-step generation; correction mechanism; domain-invariant feature learning**

I. Introduction

With the popularity of social media, online commenting platforms and e-commerce, Sentiment Analysis (SA) has emerged as a key research direction in the field of Natural Language Processing (NLP) [1]. Especially, the fine-grained sentiment analysis, Aspect-Level Sentiment Analysis, can identify the sentiment tendency of specific aspects in the text, which shows a wide range of application prospects in the application scenarios, such as market research and customer feedback analysis. However, in practical scenarios, sentiment data are often distributed in multiple domains, making it difficult for models to maintain high accuracy in cross-domain sentiment analysis tasks [2]. For example, in the user review "The sound quality of this smart speaker is excellent, with particularly good bass performance.", traditional sentiment analysis may simply classify the entire review as positive. In contrast, Aspect-Based Sentiment Analysis (ABSA) can further identify "sound quality" as a specific aspect and correctly determine that the sentiment expressed by "excellent" is positive as shown in Fig. 1. In single-domain sentiment analysis tasks, models can effectively learn the sentiment association between "sound quality" and "excellent" through a large amount of labeled data, enabling accurate predictions. However, in cross-domain sentiment analysis tasks, differences in data distribution may hinder a model's ability to generalize effectively. For instance, if the training data primarily comes from the laptop domain, where "quality" often appears in contexts such as "material quality" or "display quality", the model may associate "quality" with hardware durability or screen performance. When applied to the smart speaker domain, the model may fail to recognize "sound quality" as an aspect related to audio performance and instead misinterpret it based on its prior knowledge from the laptop domain. Additionally, sentiment words like "excellent" may carry different implications across domains, further complicating accurate classification. Such domain discrepancies can lead to a decline in classification accuracy. Therefore, enhancing a model's adaptability across different domains and enabling it to learn domain-invariant features has become a central challenge in cross-domain sentiment analysis research.

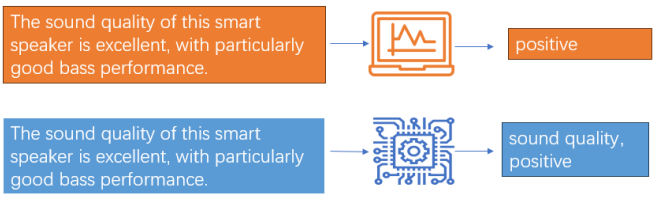


Fig. 1. Traditional Sentiment Analysis versus Aspect-Based Sentiment Analysis

Traditional Aspect-Level Sentiment Analysis models have good performance in in-domain tasks. However, their performance tends to degrade significantly when applied to cross-domain tasks due to the differences in features and linguistic expressions between different domains [3] (Chen et al., 2020). These inter-domain differences are mainly reflected in two aspects. First, the ASPECTS of different domains can vary substantially, which makes the model difficult to generalize [4]. Second, the sentiment expressions of different domains are often different, and even the same ASPECTS may have different sentiment tendencies in different domains [5]. Therefore, effectively performing cross-domain Aspect-Level Sentiment Analysis has become an important and challenging problem.

*Corresponding Author.

Current research on cross-domain Aspect-Level Sentiment Classification (ASC) has achieved notable progress in the field of sentiment analysis. However, several critical challenges persist. Inter-domain textual data differ significantly in terms of vocabulary, grammatical structure, and sentiment expression, which makes models trained in the source domain perform poorly in the target domain. In order to improve the generalization ability of the models, many existing approaches focus on aligning the feature representations of the source and target domains. By analyzing several major current approaches, the innovations and shortcomings are identified, thereby helping to clarify the research gaps.

First, the cross-domain ASC approach based on adversarial distribution alignment focuses on solving the problem of feature distribution differences between domains. Through techniques such as gradient inversion layers, these methods aim to learn feature representations shared between source and target domains, thus reducing the probability of cross-domain misclassification [4]. However, a significant limitation of these methods is their inability to explicitly handle specific aspect words, which often play a crucial role in sentiment classification tasks. The lack of explicit labeling of aspect words leads to bias in capturing fine-grained sentiment information, especially in cases where domain-specific words are present in the target domain, and the model is unable to accurately make sentiment predictions. Efficient Adaptive Transfer Network (EATN) learns common features between the source and target domains by introducing a multi-attention mechanism and a domain adversarial module while capturing direct associations between aspect words and contextual sentiment words [5]. Although this approach achieves notable success in domain alignment and fine-grained sentiment analysis, it has a key limitation: the absence of explicit tagging for aspect words in the input text. Consequently, the optimization of the attention mechanism still focuses mainly on the global context and fails to impose higher attention weights on specific aspect words.

Second, BERT-based cross-domain ASC approaches extract sentence-level and aspect-level representations by using pre-trained language models such as BERT. These methods often incorporate domain-adversarial training to reduce feature differences between the source and target domains. While this type of approach solves the domain adaptation problem to some extent, the attention mechanism of BERT focuses more on the overall context rather than optimizing specifically for aspect words [6]. Since aspect words are very important in sentiment analysis, the BERT model may suffer when dealing with fine-grained sentiment analysis, especially when transforming between different domains, resulting in performance degradation.

In addition, cross-domain ASC approaches based on semantic key feature extraction aim to enhance the cross-domain migration capability of the model by introducing an external knowledge graph. Such approaches utilize semantic knowledge graphs as a bridge for knowledge migration from the source domain to the target domain, enhancing the understanding of sentiment words and syntactic structures [7]. However, while external knowledge graphs can enrich semantic information, their reliance on external resources rather than in-depth analysis of the input text may limit the model's ability to capture the diversity of sentiment expressions in certain domains. In particular, when there are significant feature differences between domains, models relying on external knowledge may struggle to fully utilize aspect words and contextual information in internal data.

From these existing studies, it can be seen that most current cross-domain ASC approaches focus on domain alignment and feature migration [8][9]. However, they remain inadequate in dealing with specific aspect word tokens, dynamic optimization of sentiment features, and error correction mechanisms [10]. Existing methods cannot improve the inference of generative models, which leads to the susceptibility to biased sentiment prediction in cross-domain tasks, especially fine-grained aspect sentiment analysis. In addition, the models also lack effective error correction mechanisms to re-generate the sentiment output when the aspect word prediction is wrong. Misclassification and sentiment judgment errors are often the significant challenges in cross-domain sentiment analysis due to the large differences in the characteristics of different domains.

To address these issues, an innovative cross-domain ASC approach based on the Text-to-Text Transfer Transformer(T5) generative model is proposed. The generative framework of the T5 model provides a novel perspective for tackling the ASC task. With the T5 model, the sentiment categorization task can be handled in a generative way, which is no longer limited to the traditional categorization framework. By combining textual and aspectual information, T5 generate more accurate sentiment predictions across domains. Moreover, under the multi-step inference mechanism, T5 efficiently extract domain-invariant features, thus improving the accuracy of sentiment generation.

To further enhance the robustness of the model, an error correction mechanism is introduced in this study. This mechanism determines the errors in sentiment prediction by comparing the generated aspect words with the known aspect words in the target domain and re-generates them. The error correction mechanism can effectively reduce the sentiment prediction errors due to aspect word errors, thus significantly improving the generalization ability of the model across different domains. Moreover, this mechanism offers a novel approach to error handling in cross-domain sentiment analysis. In particular, for multi-domain sentiment classification, it can generate more accurate sentiment results for different aspect words.

In addition to the cross-domain aspect-level sentiment analysis task (Aspect-Level Sentiment Classification, ASC), other tasks such as Aspect-Opinion Pair Extraction (AOTE), Aspect-Sentiment Pair Extraction (ASPE), and Aspect-Sentiment Triplet Extraction (ASPE) have started to use generative models such as T5 [11]. The core of these tasks is the simultaneous recognition of aspect words, sentiment polarity, and associated opinion words or syntactic structures. Generative models, such as T5, excel at handling multiple tasks through a unified text generation framework, enabling end-to-end generation of the desired outputs. In these tasks, generative models demonstrate a high degree of flexibility and reasoning

ability to adapt to different input formats and task goals. However, due to the inherent differences between these tasks and ASC, it is difficult to apply existing methods directly to ASC subtasks.

Tasks such as AOTE, ASPE, and ASTE involve relatively complex aspects and affective associations, requiring the simultaneous identification of relationships among multiple elements. Generative models such as T5 are good at handling such multi-task settings. However, in the ASC task, sentiment analysis focuses more on judging the sentiment polarity of specific aspectual words rather than multiple triad extraction [12]. This makes it difficult for simple generative methods to accurately capture the fine-grained associations between domain-specific aspect words and sentiment in ASC tasks. In addition, although the T5 model performs well in multi-domain tasks, it tends to ignore the specificity of aspect words in cross-domain sentiment analysis because it is not optimized for ASC tasks.

To address this problem, a Multi-step inference mechanism is introduced by combining the characteristics of the ASC task. By generating aspect words, the model gains a deeper understanding of the context, enhancing the ability to generate sentiment labels. This transformation of ASC into generating aspect and sentiment combination pairs enhances the model's ability to learn as well as reason about domain-invariant features, enabling T5 to generate sentiment predictions more efficiently when processing texts from different domains. This not only improves the model's cross-domain generalization ability, but also makes it perform better in fine-grained sentiment classification.

This study combines the generative capability of the T5 model with the characteristics of the ASC task. Additionally, a multi-step inference and error correction mechanism is introduced. This not only contributes to the learning of domain-invariant features, but also improves the performance of the generative model in the application of cross-domain sentiment analysis tasks. This research fills the shortcomings of existing approaches and provides new directions for future cross-domain ASC tasks. In summary, a cross-domain Aspect-Level Sentiment Analysis model, based on the T5 model and incorporating a Multi-step Generation with Error Correction Mechanism, is proposed in this study (TSG-ECM). The main contributions are as follows:

- Multi-step Generation Training Mechanism: Firstly, by training the text and ASPECT vocabulary of the source and target domains, the model learns domain-invariant generalized features, which makes it better adapted to cross-domain text analysis. This step allows it to identify common features between different domains, and improves its generalization ability in the target domain. Next, when training in the source domain, the T5 model is allowed to generate both aspect words and sentiment labels. The model learns to predict both aspect and sentiment in the text through end-to-end generation, resulting in a powerful inference capability. In cross-domain tasks, the simultaneous output of aspect words and sentiment polarity facilitates capturing features that are common across domains and enhances its generalization ability in the target domain.
- Target Domain Correction Mechanism: This study introduces a generation process based on a correction mechanism. When the aspect generated by the model does not align with the known aspect library of the target domain, the system detects the error and regenerates the aspect and sentiment labels. By introducing error correction signals and new generation loss targets, the model is able to effectively adjust the generation strategy in case of incorrect prediction, thus improving the accuracy of sentiment prediction in the target domain.

The rest of the study is structured as follows: Section II reviews related work, including Aspect-Level Sentiment Analysis, cross-domain learning, and the application of the correction mechanism. Section III describes the proposed method in detail, including the multistep inference mechanism, source-domain ASPECT generation, and the target-domain error-correction mechanism. Section IV presents the experimental design and results, which demonstrate the method's performance on a number of domains and compares it with existing methods. Section V presents the discussion. Finally, Section VI summarizes the main contributions and discusses future research directions.

II. Related Research

This section reviews the progress of related research in cross-domain Aspect-Level sentiment analysis, domain-invariant feature learning, generative modeling, and error correction mechanisms. By summarizing and analyzing the existing methods, the innovations and advantages of the proposed methods in this study are clarified.

A. Aspect-Level Sentiment Analysis

The development of Aspect-Level Sentiment Analysis (ABSA) has evolved through several stages. Early traditional machine learning models, such as Support Vector Machine (SVM) and Plain Bayes, mainly relied on manually designed features (e.g., bag-of-words model and TF-IDF) to distinguish sentiment categories. However, these methods perform poorly when dealing with cross-domain sentiment tasks, and it is difficult to effectively deal with feature differences and sentiment expressions across domains [13]. With the rise of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are becoming mainstream. These networks are able to learn semantic and contextual relationships in text in an end-to-end manner, thereby overcoming the limitations of manual feature design, thus significantly improving the effectiveness of sentiment analysis [14]. In the context of aspect-level sentiment analysis tasks, RNNs are particularly effective at capturing dependencies in sentences using sequential information, while CNNs can effectively extract local sentiment features [15]. Nevertheless, the generalization ability of these deep learning methods remains limited in cross-domain scenarios, making it challenging to address the distributional discrepancies across domains [16]. In recent years, models based on the Transformer architecture (e.g., BERT and T5) have become the

core technology for cross-domain sentiment analysis. BERT, as a pre-trained model that captures the contextual information of a sentence through the bidirectional attention mechanism, has made significant progress in several sentiment analysis tasks [17]. However, BERT encounters challenges in cross-domain sentiment analysis, particularly due to its domain-specific nature and weak feature alignment capabilities [18].

B. Cross-Domain Sentiment Analysis with Domain-Invariant Feature Learning

In order to address the differences between different domains, researchers have proposed a variety of cross-domain learning methods, including Domain-Invariant Feature Learning (DIFL) as well as Domain-Adversarial Learning (DAL) and other domain adaptation methods. The core idea of domain adaptation is to improve the generalization ability of the model in the target domain by reducing the difference between the data distributions of the source and target domains [19].

Some studies have enabled models to capture common features between domains by introducing a shared feature space. For example, Xue et al. (2018) proposed a method based on Gradient Reversal Layer (GRL) to improve the performance of cross-domain sentiment classification. This method aligns the data distributions of the source and target domains through adversarial training in a shared feature space [20]. However, this type of approach is not directly applicable in Aspect-Level tasks, because aspects of different domains tend to differ significantly. Simply aligning the distributions does not guarantee the accuracy of aspect prediction.

To address this limitation, several studies have proposed to enhance cross-domain learning by separating domain-specific features from domain-invariant features [21]. These approaches enhance the robustness of the model across domains by introducing an independent domain-invariant feature extraction module into the model structure, which is then combined with domain-specific feature learning. However, such methods still struggle to cope with the problem of generating incorrect ASPECTS in the target domain, and incorrect ASPECTS generation often leads to chain errors in sentiment categorization.

C. Application of Error Correction Mechanisms in Generative Tasks

Although generative models exhibit strong text generation capabilities, incorrectly generated aspects can introduce significant bias in sentiment prediction, particularly in cross-domain sentiment analysis tasks. Therefore, Error Correction Mechanism (ECM) is proposed to reduce the impact of generation errors on model performance. It is widely used in tasks such as machine translation, text summarization and dialog generation [22], where errors are corrected by post-processing or re-generating the generated results.

In the task of sentiment analysis, the application of error correction mechanisms remains relatively limited. Dong et al. (2019) proposed an error correction-based generation framework, where the model re-generates if it detects incorrect words or phrases during the generation process, thereby reducing the impact of errors on the final task [23]. Although this correction strategy can improve the quality of generation to some extent, there is currently no well-defined mechanism capable of effectively addressing the problem of erroneous aspect generation in cross-domain Aspect-Level sentiment analysis.

D. Research Innovations

Although existing research has made some progress in aspect-based sentiment analysis, cross-domain sentiment analysis and generation modeling, there are still significant limitations. First, most existing cross-domain sentiment analysis methods rely on the learning of domain-invariant features, but lack effective generation and correction mechanisms for the differences in aspect across domains. Second, while generative models perform well in sentiment analysis tasks within a single domain, they often fail to capture the relationship between aspects and opinions in cross-domain tasks, generating erroneous sentiment analysis results.

In this study, a multi-step generative inference mechanism combined with a correction mechanism is proposed. (Cross-Domain Aspect-Sentiment Generation with Error Correction (CDASG-EC)) It aims to improve the model's performance in cross-domain tasks through domain-invariant feature learning, generative lossy objective optimization, and an error correction mechanism accuracy. Compared with existing methods, the innovations of this study are as follows:

- By introducing the T5 model for domain-invariant feature learning, the proposed approach enhances the adaptability of the model across diverse domains.
- In the source domain, the T5 model is trained by inputting text in multiple steps and outputting (ASPECT, SENTIMENT) to enhance the model's inference ability and thus generate more accurate sentiment classification results. This strategy represents the first application of a generative T5 model in a cross-domain ASC task, enabling the simultaneous generation of aspects and sentiments in an end-to-end manner. It provides the model with stronger domain-invariant feature learning and generalization capabilities.
- A correction mechanism is introduced to detect and rectify generated aspect and sentiment labels in the target domain, thus reducing the error rate of the model in the cross-domain tasks.

III. METHODOLOGY

A multi-step generative inference mechanism based on the T5 model, combined with a correction mechanism (Cross-Domain Aspect-Sentiment Generation with Error Correction (CDASG-EC)), is proposed in this study. It aims to improve the accuracy and robustness of the cross-domain sentiment analysis task. The design of the model is divided into two core parts: a multi-step generative inference mechanism (domain invariant feature learning, source domain aspect and sentiment generation training) and a target domain correction mechanism.

The overall model is shown in Fig. 2. The structure and training process of each part are described in detail below:

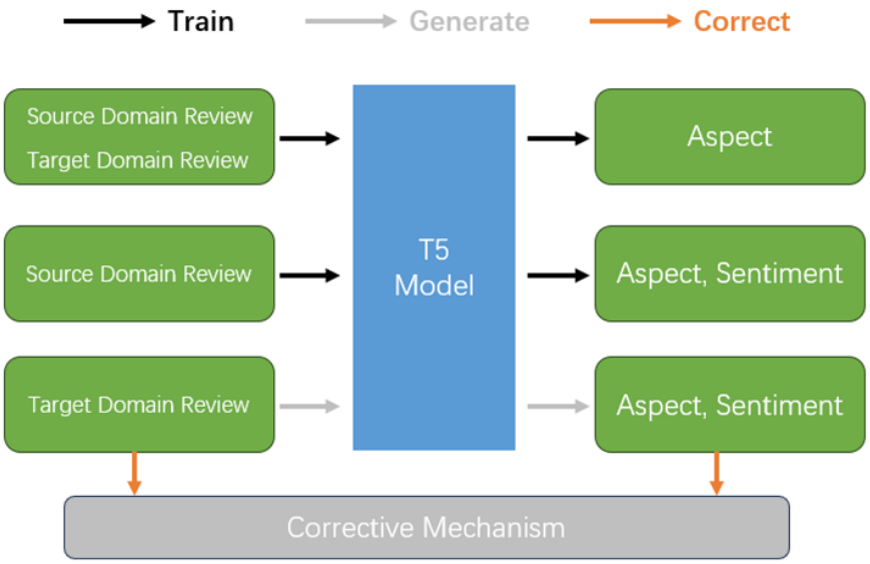


Fig. 2. Overview of Cross-Domain Aspect-Sentiment Generation with Error Correction (CDASG-EC).

A. Overview of the T5 Model

The T5 model is a unified text generation framework that can unify various NLP tasks into text generation problems [24]. In this study, it is used as a basis for handling the cross-domain Aspect-Level sentiment analysis task. Specifically, the input is a domain text, and the model generates the ASPECT vocabulary related to that text. Unlike traditional classification methods, the generative model performs well in capturing complex text structures and domain-generalized features for cross-domain tasks.

B. Multi-Step Generative Training Mechanism

1) Domain invariant feature learning: The main challenge of cross-domain sentiment analysis is the difference in data distribution between domains. To enable the transfer of sentiment analysis capabilities across different domains, domain-invariant feature learning is first performed. This process involves training the T5 model to learn transferable features by combining the text and ASPECT vocabularies from both the source and target domains. The specific steps are as follows:

a) Input data preparation: For texts from both the source and target domains, the input to the T5 model is formatted as [Text], where Text represents the raw text from either domain. The task of the T5 model is to generate the Aspect associated with Text, i.e., the specific vocabulary related to emotion in that text. The input format is [Text] and the generated output is [Aspect].

b) Loss function design: During training, the T5 model learns domain invariant features by minimizing the difference between the generated aspect and the true aspect. The cross-entropy loss function is employed to measure the difference between the generated aspect and the real aspect. The loss function is represented as follows:

$$L_{domian} = -\sum_{i=1}^{N}\left(y_i^{asp}\log\hat{y}_i^{asp} + \left(1-y_i^{asp}\right)\log\left(1-\hat{y}_i^{asp}\right)\right) \quad (1)$$

where, y_i^{asp} is the real ASPECT vocabulary, and $\hat{y}_i^{asp}$ is the probability distribution of the ASPECT vocabulary generated by the T5 model.

c) Domain-invariant feature extraction: During the training process, the model learns generic features from text data in the source and target domains. In this process, the T5 model captures not only the association between text and ASPECT, but also the common features between the source and target domains. This is achieved through a multi-task learning mechanism. This joint training ensures that the aspect generated by the model can be applied to texts in different domains.

2) Source domain aspect and sentiment generation training: After domain invariant feature learning is completed, the model is further trained on the source domain to generate aspect and corresponding sentiment labels. This step aims to ensure that the model learns to recognize aspect words in text and generate corresponding sentiment labels based on their context. In cross-domain tasks, this strategy enables the model to have stronger reasoning capabilities. The specific steps are as follows:

a) Source domain input and output format: During source domain training, the input of the model is the source domain text, denoted as [Text], and the generated output is [Aspect] and the corresponding sentiment label [Sentiment]. The model first generates the aspect and then predicts the sentiment label based on the generated aspect. The output is formatted as [Aspect Sentiment]. This strategy is the first of its kind to incorporate a generative T5 model in a cross-domain ASC task and generates both aspects and sentiments in an end-to-end manner, which is designed to provide the model with enhanced domain-invariant feature learning and generalization capabilities. In addition, this approach generates ASPECTS by inference in cross-domain contexts, which also reduces the dependence on a prior knowledge of the target domain and improves flexibility.

b) Source domain training objective: The source domain training objective is to minimize the difference between the generated sentiment labels and the true sentiment labels. Furthermore, if the ASPECT is wrong, the SENTIMENT loss is weighted. The loss function during training is:

$$L_{sentiment} = -\sum_{i=1}^{N} w_i^{asp} y_i^{sent} \log \hat{y}_i^{sent} \quad (2)$$

where, w_i^{asp} is a weight factor that depends on the correctness of the ASPECT. This weight is larger when the ASPECT is wrong and smaller when the ASPECT is correct. y_i^{sent} is the true sentiment label, and $\log \hat{y}_i^{sent}$ is the probability distribution of the sentiment label generated by the model.

C. Target Domain Correction Mechanism and Loss Objective Generation

When y_k^{asp} the model is applied to the target domain, the significant differences in data distributions between the source and target domains may cause the model to generate incorrect ASPECTS, leading to inaccurate sentiment prediction. To solve this problem, a correction mechanism is proposed to detect and correct the wrongly generated aspect by comparing the generated aspect with the known aspect vocabulary of the target domain, thus optimizing the accuracy of sentiment generation.

1) Error correction mechanism design: The error correction mechanism validates the model-generated ASPECT through the known ASPECT vocabulary of the target domain. If the generated aspect does not exist in the known vocabulary, the mechanism identifies the result as likely erroneous. In such cases, the correction mechanism employs a re-generation process to correct both the aspect and its associated sentiment label. The loss function of the error correction mechanism is designed as follows:

$$L_{correction} = -\sum_{k=1}^{N} \mathbb{1}\left(y_k^{asp} \neq \hat{y}_k^{asp}\right) \cdot y_k^{asp} \log\left(y_k^{asp_corrected}\right) \quad (3)$$

where, y_k^{asp} is the true ASPECT vocabulary of the target domain, $\hat{y}_k^{asp}$ is the probability of ASP vocabulary generated by the model, and $\hat{y}_k^{asp}$ is the probability of ASP regenerated by the corrective mechanism. $\mathbb{1}(y_k^{asp} \neq \hat{y}_k^{asp})$ is the indicator function; in the case of generating wrong ASPECT case loss weight is 1, otherwise it is 0.

2) Generation loss objective optimization: To further improve the generation effect of the model, the generation loss objective optimization strategy is proposed. In the error correction mechanism, if the aspect generated by the model is identified as incorrect, the model dynamically adjusts its generation strategy. Specifically, the generation loss is increased to encourage the model to generate more accurate aspect and sentiment labels.

D. Overall Training Process

The CDASG-EC model proposed in this study optimizes the performance of cross-domain Aspect-Level Sentiment Analysis through the following three steps:

- Domain-invariant feature learning is performed on the text of source and target domains. The model input format is [Text], which generates the corresponding [Aspect];
- The model undergoes further training on the source domain to generate the aspect and its corresponding sentiment labels. The output is formatted as [Aspect Sentiment];
- The error correction mechanism is applied to the target domain to detect and correct the generated aspect and sentiment, thus optimizing the final generated loss objective.

Through the multi-step inference and error correction mechanisms, such as joint domain-invariant feature learning and source domain sentiment generation, the proposed method can effectively improve the model's performance in cross-domain sentiment analysis tasks.

IV. Experimental Design

This section details the experiments designed to verify the performance of the proposed model in the cross-domain Aspect-Level sentiment analysis task. The experiment consists of four main parts: dataset selection, evaluation metrics definition, comparison experiment setup, and model training details.

A. Benchmark Datasets

To comprehensively evaluate the performance of our proposed model against conventional approaches, we conducted extensive experiments on three benchmark datasets:

- Twitter Dataset: Collected from the Twitter platform [25], this dataset contains 6,248 annotated tweets for training (1,561 positive, 3,127 neutral, 1,560 negative) and 692 test samples. Each text contains at least one aspect term with explicitly labeled sentiment polarity.
- Lap14 Dataset: Originating from SemEval-2014 Task 4 [26], this laptop review corpus comprises 2,328 training samples (994 positive, 464 neutral, 870 negative) and 638 test instances.
- Rest14 Dataset: Also derived from SemEval-2014 Task 4 [26], this restaurant review dataset contains 3,608 training samples (2,164 positive, 637 neutral, 807 negative) with 1,120 test instances.

Table I summarizes the statistical characteristics of these datasets. Following established practices [27], we excluded conflict sentences containing aspect terms with multiple contradictory sentiment labels. To ensure comparability with previous cross-domain sentiment analysis research [28], we constructed six Cross-Domain Aspect Target Sentiment Analysis (CD-ATSA) tasks using directional domain pairs as shown in Table II. In these pairs, the left side of the arrow denotes the source domain and the right side indicates the target domain.

TABLE I. STATISTICS OF DATASETS

Dataset Name		positive	neutral	negative
Twitter[25]	training	1561	3127	1560
	Test	173	346	173
Lap14[26]	training	994	464	870
	Test	341	169	128
Rest14[26]	training	2164	637	807
	Test	728	196	196

B. Experimental Setup

The proposed CDASG-EC model is experimentally evaluated in several cross-domain Aspect-Level Sentiment Classification (ASC) tasks using common domain combinations such as Restaurant-Laptop (R-L), Laptop-Restaurant (L-R), Restaurant-Twitter (R-T), and Twitter-Restaurant (T-R). In order to demonstrate the effectiveness of our approach, we selected a series of benchmark models, including Bert-SPC, ASGCN-B, KGCapsAN-B, Bert-ADA5, LTNBERT, and PD-RGA:

- Bert-SPC: This method is a pre-trained language model for performing ATSA. The input is of the form "[CLS] + sentence + [SEP] + aspect word + [SEP]".
- Bert-PT [29] is a post-training strategy for the basic Bert model.
- ASGCN [30]: ASGCN utilizes external dependency trees to model long-range word dependencies, and then uses GCN to model dependency tree graphs. ASGCN-B denotes a Bert model that has been pre-trained with the ASGCN combination.
- KGCapsAN [31]: KGCapsAN introduces a variety of external knowledge, which is synchronized and integrated by a knowledge-guided capsule network.
- Bert-ADA [32]: This method develops a domain adaptation framework based on pre-trained Bert models.
- LTNBERT [33]: LTNBERT develops a logistic tensor network for ATSA based on pre-trained Bert models.
- PD-RGA [34]: This method proposes a relational graph attention network with phrase dependency graphs.

C. Implementation Details

We set the learning rate to 0.001 and the batch size to 16. The maximum number of training rounds is 10. The Transformers libraries of PyTorch and Hugging Face are used to implement the model training and inference.

D. Experimental Results

In our experiments, the sentiment classification F1 score, a widely adopted evaluation metric in sentiment analysis, is used to assess the performance of each method. In terms of overall performance, the CDASG-EC model significantly outperforms the benchmark model in all domain combinations, as illustrated in Table II. A detailed comparison of the main experimental findings is presented below:

1) Best performance in all domain combinations: The proposed CDASG-EC model significantly outperforms the benchmark model in terms of F1 scores in all domain combinations, especially in the R-L and T-R combinations. For example, the CDASG-EC model obtains an F1 score of 72.28 in the R-L combination, outperforming Bert-ADA (70.46) and LTNBERT (71.48). This demonstrates the superior generalization ability in cross-domain sentiment prediction.

2) Contribution of error correction mechanism: The Error correction mechanism has a significant role in aspectual sentiment prediction. In more complex domain combinations such as T-R, the F1 score improves from 49.57 (without error correction) to 50.92 (with error correction), suggesting that the mechanism effectively solves the problem of sentiment bias caused by inconsistency in aspect words.

3) Advantages of the generative model: The T5 generative model demonstrates its advantages in cross-domain knowledge transfer. In combination with large domain differences (e.g., L-T, T-L), the CDASG-EC model still maintains high performance, proving that it is more effective in dealing with aspect word variations across different domains than methods relying only on pre-trained BERT models.

E. Ablation Studies

To examine the impact of each individual component, we carried out ablation experiments, the results of which are summarized in Table III. Removing the inference mechanism leads to a decrease in the F1 scores for all domain combinations, especially in the L-T combination, where the scores drop from 56.97 to 54.89. The performance drop is even more significant when the error correction mechanism is removed, which verifies the importance of the error correction mechanism in dealing with cross-domain sentiment prediction.

TABLE II. RESULTS OF THE ASSESSMENT (F1%)

Methods	R→L	L→R	R→T	T→R	L→T	T→L	Avg.
Bert-spc	70.75	66.02	51.53	45.61	55.07	45.54	55.75
Bert-PT	70.92	65.86	52.43	45.70	55.68	45.54	56.02
ASGCN	71.06	65.47	51.73	45.55	56.02	46.15	56.00
KGCapsAN	71.17	65.91	51.86	45.96	56.11	48.92	56.66
Bert-ADA	70.46	72.93	53.49	47.12	56.04	50.18	58.37
LTNBERT	71.48	66.81	52.64	46.97	56.13	50.02	57.34
PD-RGA	71.07	65.94	52.23	47.77	55.93	47.86	56.80
CDASG-EC(Ours)	72.28	71.84	54.84	48.92	56.97	50.87	59.12

TABLE III. ABLATION STUDY RESULTS (F1 %) OF THE PROPOSED CDASG-EC

Methods	R→L	L→R	R→T	T→R	L→T	T→L	Avg.
Ours	72.28	70.84	54.84	48.92	56.97	50.87	59.12
Without multi-step generation	71.04	68.98	52.11	47.23	54.89	49.97	57.37
Without correction	71.88	69.45	53.08	47.96	55.07	49.14	57.76

A. Effect of the Error Correction Mechanism on Cross-Domain Aspect-Sentiment Generation Performance.

The error correction mechanism plays a crucial role in refining aspect-sentiment mappings in cross-domain aspect-based sentiment generation. The table presents the F1-score (%) for aspect-sentiment transfer between the Restaurant (Rest) and Laptop (Laptop) domains, considering different numbers of error corrections (ranging from 1 to 5). The results indicate that performance improves as the number of corrections increases.

As shown in Fig. 3 and Fig. 4, the F1-score consistently increases as the number of corrections increases for both Rest → Laptop and Laptop → Rest transfers. This demonstrates the effectiveness of the error correction mechanism in mitigating misclassifications and improving the alignment of aspects across domains. Although applying more corrections continues to improve performance, the rate of improvement diminishes after three corrections. For Rest → Laptop, the F1-score increases by 0.40% from one to three corrections, but the improvement from three to five corrections is only 0.05%. Similarly, for Laptop → Rest, the improvement from one to three corrections is 0.38%, whereas the increase from three to five corrections is just 0.08%. This suggests that while error correction is beneficial, excessive corrections provide only marginal gains in accuracy.

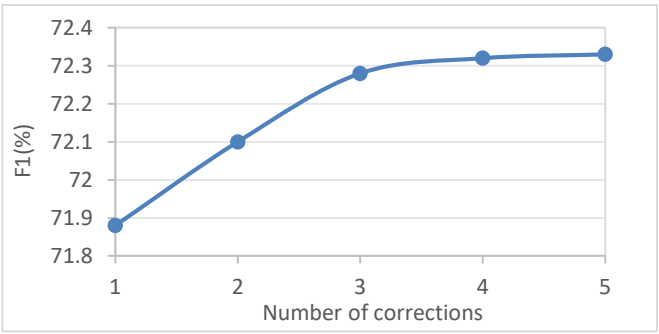


Fig. 3. Impact of error corrections on F1-score (Rest → Laptop).

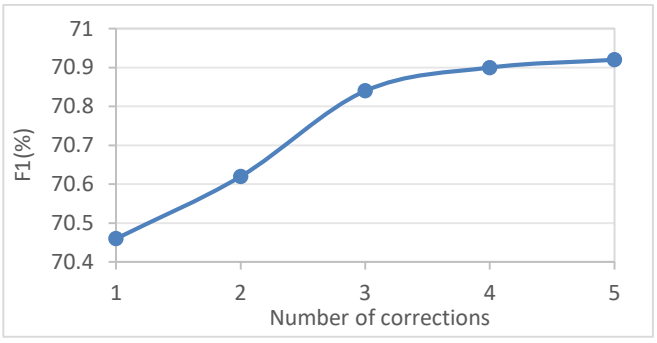


Fig. 4. Impact of error corrections on F1-score (Laptop → Rest).

The results also reveal an asymmetry in domain transfer performance, with Rest → Laptop consistently achieving higher F1-scores than Laptop → Rest. This indicates that knowledge transfer from the restaurant domain to the laptop domain is more effective than the reverse. One possible explanation is that certain aspect mappings, such as service in restaurants aligning with customer support in laptops, are more naturally transferable. Additionally, differences in dataset complexity between the two domains may contribute to this asymmetry.

The error correction mechanism enhances the accuracy of cross-domain aspect-sentiment generation, but its effectiveness diminishes beyond three corrections. Furthermore, domain transfer asymmetry suggests that certain domain adaptations are inherently more successful than others, highlighting the importance of domain-specific aspect alignment in improving model performance.

V. DISCUSSION

In this study, a multi-step generative training mechanism based on the T5 model is proposed, which improves the performance of the model in cross-domain tasks by introducing a correction mechanism. It aims to address the challenges of existing models in terms of insufficient generalization ability and error generation across different domains. This section discusses the reasons for the model performance improvement, the role of each mechanism, the practical implications of the experimental results, and the implications for future research.

A. Effectiveness of Multi-Step Generative Inference Mechanisms

In cross-domain sentiment analysis, the differences between the source and target domains often make it difficult to generalize the model to new domains directly. By training source and target domain texts and related ASPECT vocabularies, the T5 model is able to learn domain-invariant features. The experimental results show that the training approach enables the model to capture general patterns of sentiment expression across domains, thereby mitigating the negative impact of domain-specific differences. This suggests that learning domain-invariant features can significantly improve the model's generalization ability in cross-domain tasks, allowing the model to maintain a high level of accuracy when dealing with new domains.

B. The Role of Correction Mechanisms in Emotion Generation

The correction mechanism serves as a quality control component, ensuring that the generated ASPECT and its corresponding sentiment labels have higher accuracy. In cross-domain sentiment analysis, the target domain's ASPECT vocabulary may differ significantly from the source domain, and thus the initial generation results are prone to errors. The correction mechanism detects whether the output ASPECT is

correct by comparing the known ASPECT of the target domain, which significantly reduces the probability of sentiment label generation errors. Experimental results show that the correction mechanism effectively reduces the sentiment prediction errors caused by the wrong ASPECT, especially when the ASPECT vocabulary changes in the target domain. Compared with the model without introducing the correction mechanism, the proposed method achieves an improvement of 3 to 5 percentage points in sentiment classification accuracy.

C. Comparison with Existing Methods

Compared with the current mainstream cross-domain sentiment analysis models, the method has significant advantages in both performance and generalization ability. The BERT-based model [2] performs well in the sentiment generation task. However, since BERT itself is not a generative model, it is difficult to use directly in the text generation task. Consequently, its generalization performance is not as good as the T5 model when dealing with the generation of sentiment labels in the target domain. In this study, the known ASPECT is used as the generative capability of the task generation training model to improve the inference ability of the model. Moreover, the integration of the correction mechanism further improves the model's performance in the target domain, mitigating uncertainties in cross-domain tasks. Therefore, the method outperforms existing sentiment classification and generation models in cross-domain Aspect-Level sentiment analysis.

D. Limitations and Future Research Directions

Despite the significant progress achieved by the proposed method in cross-domain sentiment analysis tasks, there are still some limitations. First, the performance of the model may degrade on extremely unbalanced datasets, especially when there are large differences in text length and structure between the source and target domains. Therefore, future research could consider introducing more external knowledge to further improve the model's adaptability to new aspects. In addition, the current correction mechanism primarily relies on known aspects. Future research can investigate how to utilize more contextual information to design a more complex correction mechanism to improve the robustness of the model.

E. Implications for Practical Applications

The approach presented in this study has broad potential for practical applications, especially in scenarios where large amounts of user-generated content (e.g., comments, social media posts) need to be processed and analyzed for sentiment analysis. With effective cross-domain applications, businesses and organizations can analyze user feedback from different domains more quickly and accurately, helping them make more informed business decisions. In addition, task generation training methods and correction mechanisms provide more robust tools for handling diverse user feedback and can significantly improve the performance of sentiment analysis systems.

VI. Conclusion

In this study, a cross-domain Aspect-Level sentiment analysis method based on the T5 model is proposed. The generalization ability and sentiment classification accuracy in cross-domain tasks are enhanced through the introduction of the Multi-step generative training mechanism and the correction mechanism. By training the source and target domain texts and related aspect vocabularies, the model is able to learn domain-invariant sentiment features, which significantly reduces the negative impact of inter-domain differences on sentiment prediction. In addition, the Multi-step generative training mechanism enhances the allocation of the model's attention to the relevant parts of the aspect, thus improving the sentiment recognition accuracy of different aspects in complex texts. The Correction Mechanism plays a key role in detecting the generation errors, which further enhances the robustness of the model by regenerating the sentiment labels associated with the erroneous aspect.

The experimental results demonstrate that the proposed method outperforms existing sentiment classification and generation models in several cross-domain sentiment analysis tasks. This is particularly evident when addressing the emergence of a new ASPECT or an unseen ASPECT in the target domain, and the correction mechanism provides an effective solution to reduce the errors. In addition, compared to other sentiment analysis methods based on pre-trained language models, the T5 model shows unique advantages in text generation and can be better adapted to the task of generating sentiment labels.

Nevertheless, there are still some limitations in the proposed method. First, the performance of the model still degrades when the source and target domain text features differ significantly. Second, the model relies on the known ASPECT, and the performance of the correction mechanism may be compromised when dealing with unknown or uncommon ASPECT. Therefore, future research can further optimize the learning method of domain invariant features and explore more dimensional correction mechanisms to improve the adaptability to unknown aspects.

Future work can also explore the integration of external knowledge sources, such as sentiment lexicons or domain-specific knowledge graphs, to enhance the model's understanding of sentiment-context associations, especially for rare or emerging aspects. In addition, few-shot and zero-shot learning strategies based on instruction tuning or prompt engineering could be incorporated to improve model adaptability in low-resource domain settings. Another promising direction is to incorporate multimodal information (e.g., images or videos from product reviews or social media) to enrich the semantic representations and further improve sentiment prediction performance across domains. Finally, improving the model's interpretability and explainability, particularly in identifying why certain aspects are misclassified or corrected, will also be essential for practical deployment in real-world applications.

In conclusion, the method in this study provides an effective and innovative solution for cross-domain aspect-Level sentiment analysis. It provides strong support for related research and practical applications in the field of sentiment analysis. Future research should expand the application scenarios of this method and combine more external

knowledge with advanced deep learning techniques to further improve the accuracy and generalization ability of cross-domain sentiment analysis.

ACKNOWLEDGMENT

All authors declare that they have no conflicts of interest.

REFERENCES

[1] Cui, J., Wang, Z., Ho, SB. et al. Survey on sentiment analysis: evolution of research methods and topics. Artif Intell Rev 56, 8469–8510 (2023). https://doi.org/10.1007/s10462-022-10386-z

[2] Zhang, X., Liu, L., & Zhao, S. (2021). Cross-domain aspect-based sentiment analysis: A survey and new perspectives. Journal of Computational Linguistics, 47(2), 145-165. https://doi.org/10.1162/coli_a_00403

[3] Chen, H., Sun, W., & Yang, Y. (2020). Domain adaptation for sentiment analysis via cross-domain sentiment word alignment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 7608-7615. https://doi.org/10.1609/aaai.v34i04.6091

[4] Yu, J., Zhao, Q., & Xia, R. (2023). Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1456–1470). Toronto, Canada: Association for Computational Linguistics.

[5] Zhang, K., Liu, Q., Qian, H., Xiang, B., Cui, Q., Zhou, J., & Chen, E. (2021). EATN: An Efficient Adaptive Transfer Network for Aspect-Level Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering, 35, 377-389.

[6] Liu N, Zhao J. A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text. Comput Intell Neurosci. 2022 Jun 27;2022:8726621. doi: 10.1155/2022/8726621.

[7] Knoester, J., Frasincar, F., & Tru¸sc˘a, M. M. (2022). Domain Adversarial Training for Aspect-Based Sentiment Analysis. In R. Chbeir et al. (Eds.), WISE 2022 (pp. 21–37). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-20891-1_3

[8] Zhang, K., Liu, Q., Qian, H., Xiang, B., Cui, Q., Zhou, J., & Chen, E. (2023). EATN: An Efficient Adaptive Transfer Network for Aspect-Level Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering, 35(1), 377-396. https://doi.org/10.1109/TKDE.2021.3075238

[9] Liu, N., & Zhao, J. (2022). A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text. Computational Intelligence and Neuroscience, 2022, 1-11. https://doi.org/10.1155/2022/8726621

[10] Wu, J., Zhao, S., & Zhang, X. (2021). Unified generative framework for aspect-opinion pair extraction with T5. Proceedings of the EMNLP Conference, 34(1), 456-468. https://doi.org/10.18653/v1/emnlp-2021-029

[11] Shi, J., Li, W., Bai, Q., Yang, Y., & Jiang, J. (2023). Soft prompt enhanced joint learning for cross-domain aspect-based sentiment analysis. Intelligent Systems with Applications, 20, 200292. https://doi.org/10.1016/j.iswa.2023.200292

[12] Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2023). Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12272–12285). Toronto, Canada: Association for Computational Linguistics.

[13] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 27-35. https://doi.org/10.3115/v1/S14-2004

[14] Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 606-615. https://doi.org/10.18653/v1/D16-1058

[15] Tang, D., Qin, B., & Liu, T. (2016). Aspect level sentiment classification with deep memory network. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 214-224. https://doi.org/10.18653/v1/D16-1021

[16] He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. Proceedings of the 25th International Conference on World Wide Web, 507-517. https://doi.org/10.1145/2872427.2883037

[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[18] Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2324-2335. https://doi.org/10.18653/v1/N19-1242]

[19] Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 513-520.

[20] Xue, H., Dai, X.-Y., Zhang, J., Huang, X., & Chen, J. (2018). Aspect based sentiment analysis with selective adversarial learning. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 2308-2319.

[21] Chen, Z., Sun, X., Huang, L., & Chang, C. (2020). Multilingual aspect-based sentiment analysis via domain invariant and specific feature transfer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 227-238.

[22] He, W., Xia, Y., Qin, T., Wang, L., Yu, N., & Liu, T.-Y. (2016). Dual learning for machine translation. Advances in Neural Information Processing Systems, 29, 820-828.

[23] Dong, L., Xu, S., & Xu, B. (2019). Unified language model pre-training for natural language understanding and generation. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 13042-13054.

[24] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.

[25] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, 2014, pp. 49–54.

[26] M. Pontiki, D. Galanis,J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval., 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14--2004

[27] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in Proc. Conf. Empir. Methods Natural Lang. Process., 2016, pp. 214–224.

[28] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 751–760.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186.

[30] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," inProc.Conf.Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 4560–4570.

[31] B. Zhang, X. Li, X. Xu, K. Leung, Z. Chen, and Y. Ye, "Knowledge guided capsule attention network for aspect-based sentiment analysis," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 2538–2551, 2020.

[32] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for

aspecttarget sentiment classification," in Proc. 12th Lang. Resour. Eval. Conf., 2020, pp. 4933–4941.

[33] H. Huang, B. Zhang, L. Jing, X. Fu, X. Chen, and J. Shi, "Logic tensor network with massive learned knowledge for aspect-based sentiment analysis," Knowl.-Based Syst., vol. 257, 2022, Art. no. 109943.

[34] H. Wu, Z. Zhang, S. Shi, Q. Wu, and H. Song, "Phrase dependency relational graph attention network for aspect-based sentiment analysis," Knowl.-Based Syst., vol. 236, 2022, Art. no. 107736.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

Perceptual Hash Techniques for Audio Copyright Protection in Decentralized Systems

Dr. N. Kavitha[1], Rashmika S J[2], Reshika A S[3]
Assistant Professor (Senior Grade), Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu[1, 2]
Student, Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu[3]

***Abstract*—The overlap of perceptual hashing technologies with blockchain is an interesting answer to strengthen copyright protection of sound in decentralized networks. As the music society still has to endure attacks on unauthorized duplication and transformation of online content, traditional security does not hold water. Perceptual hashing bridges the gap, creating unique digital fingerprints that are resistant to small-scale modifications, to allow detection of copyright piracy even in edited audio content. When combined with the immutable nature of blockchain and the smart contract functionality, such a new framework not only guarantees ownership verification but also does licensing procedures automatically, thereby doing away with the need for intermediaries. The proposed method addresses the issues of the state-of-the-art methods and performs well under various conditions.**

Keywords*—Perceptual hashing; blockchain; audio; copyright protection

LIST OF ACRONYMS AND THEIR FULL FORMS

Acronym	Full Form
AAC	Advanced Audio Coding
AI	Artificial Intelligence
CID	Content Identifier
DCT	Discrete Cosine Transform
DTW	Dynamic Time Warping
GAN	Generative Adversarial Network
IPFS	InterPlanetary File System
LLE	Locally Linear Embedding
LPF	Low-Pass Filter
MP3	MPEG-1 Audio Layer 3
NFT	Non-Fungible Token
OLAF	Overly Lightweight Acoustic Fingerprinting
SURF	Speeded-Up Robust Features
URI	Uniform Resource Identifier

I. INTRODUCTION

The Music business is a coalition of copyright administration, tune writers, mixing or mastering engineers, music architects, song development, music representatives, and artist liaisons. It has evolved over the years to become a global ecosystem in which everyone, everywhere, has access to music from artists anywhere in the globe instantly. This transformation has altered how we consume music, but it has also opened the floodgates to a new set of issues—specifically in terms of protecting artists' intellectual property. Digital content is easy to copy, manipulate, and distribute without permission, exposing most artists to unauthorized use of their work. The unauthorized release of Kanye West's Donda album highlighted the challenge artists have in keeping their digital content. Another well-known case is the Blurred Lines dispute, in which Robin Thicke was sued by Marvin Gaye's estate for copyright infringement. These cases emphasize the intricacies of intellectual property protection in the digital age. Even in India, copyright has been a huge change to the music industry. Ilaiyaraaja's battle with music companies for rights over his own songs is the biggest tale of the legal battle for rights over intellectual property. Even music director A.R. Rahman has been grumbling about receiving proper royalty payments to composers and lyricists and asking for a more open payment system. The story of the song Madhuban by Kanika Kapoor, attacked and then altered due to cultural and copyright sensitivities, even more, highlights the issues of copyright in India.

The growing complexity of enforcing copyrights in the digital music industry has motivated researchers to explore new career protection approaches. Sharma and Gupta (2023) [1] researched how blockchain can provide a decentralized approach to protecting copyrights in the Indian music industry, reduce reliance on intermediaries and evade unauthorized distribution. Menon and Kumar (2022) [2] enlightened technical challenges in adopting audio fingerprinting for Bollywood songs, and underlined the problem raised by recurrent remixes, pitch variations, and tempo shifts. These studies put a strong focus on the critical need for sound steps over standard copyright safeguarding measures.

Conventional copyright protection methods, including digital watermarks and metadata tagging, are commonly ineffective against unauthorized use. Subtle changes—like pitch manipulation, tempo adjustments, or remixing—are easily circumvented while keeping the original track's character intact. This shortcoming has given rise to more advanced solutions, such as perceptual hashing, which generates distinctive digital fingerprints from an audio file's sound characteristics and not its metadata. Unlike cryptographic hashes, perceptual hashes do not change much

even with frequent audio transformations such as compression, pitch changing, or addition of noise.

Perceptual hashing has also been crucial in the war against copyright infringement. YouTube's Content ID system, for example, uses similar methods to identify copyrighted materials, even if an original track has been transformed. But Content ID is a centralized system run by YouTube, so artists have to depend on the company for enforcement. Most artists complain that the system is fault-prone, sometimes misidentifying material or permitting big companies to take credit for revenue from solo creators, unjustly.

A more direct and democratized mechanism for copyright protection is blockchain technology. Artists can establish an indelible, provable record of ownership by placing perceptual hashes onto a blockchain ledger. Blockchain-based decentralized music distribution has already been tested and scaled by platforms such as Audius, wherein artists own their work without middlemen taking a share of it. Smart contracts further advance this system by having licensing deals and royalties automatically execute payments. In 2017, singer Imogen Heap pioneered blockchain-enabled smart contracts for music distribution with her Mycelia project, making revenue sharing transparent and fair. Opus and Ujo Music, similar platforms, also empower artists by utilizing blockchain to manage royalties. Perceptual hashing and blockchain can revolutionize copyright protection in Indian music. Current studies have explored the intersection in depth. Rao and Iyer (2024) [3] discussed how digital rights management (DRM) on Indian music websites determines its limitations and promotes blockchain-based solutions to enable artists' equitable returns. Patel and Desai (2023) [4] have proposed enhancements in perceptual hashing for application on Indian classical music, which has unique needs due to the intricate tonal variations involved. Their research points out that fingerprinting approaches tailored to a particular set of inputs can significantly strengthen copyright protection on traditional Indian types of music. Further, Reddy and Bose (2024) [5] discussed the legal and technological implications of blockchain applications in Indian copyright laws, arguing that legislative changes are necessary to completely integrate blockchain into copyright protection systems. By combining perceptual hashing and blockchain technology, artists are given a better tool to protect their work. Illegal copies are much harder to execute, while licensing and payment of royalties become transparent and more effective. Decentralization by blockchain ensures that no single party owns control over the data, making a fair and open system accessible to all. As machine learning and artificial intelligence continue to develop, future technologies may enhance music copyright protection. Forensic testing for music using AI, for instance, could analyze even the slight similarities between compositions to offer enhanced protection from unauthorized use. The future for the music business may involve developing a paradigm around decentralized platforms, AI-driven copyright protection, and royalty management through blockchain. These innovations might restore fairness and transparency to a historically intermediary-dominated market, allowing artists to receive their due credit and compensation for creative labor. But knowing a song is only half the battle—the actual battle is asserting ownership and maintaining rights intact. That's where blockchain enters. By storing perceptual hashes on a blockchain ledger, artists can create a permanent, auditable record of ownership. Audius, a blockchain-based music streaming platform, is already introducing this vision into reality, enabling artists to receive their rights and royalties directly in uncommissioned form, free from the involvement of a label or streaming site in the middle. Another useful property of blockchain technology is smart contracts. They are computer-controlled contracts that have the ability to license work and remit royalties automatically.

Decentralized music sites such as Opus and Ujo Music are appearing, enabling artists more control over their royalties through blockchain's transparent and autopilot nature. Smart contracts may also protect against unauthorized release. If a perceptual hash of a posted song corresponds with a copyrighted song on the blockchain, the system is able to deny the upload or require proper licensing before providing access. Artists, therefore, receive the payment they are due without the wait of long legal processes. Consider how much this technology would have benefited in the "Stairway to Heaven" copyright battle between Led Zeppelin and Spirit's Randy Wolfe. A perceptual hash system supported by blockchain would have resolved the matter instantly by offering absolute evidence of originality. Outside the domain of copyright law, blockchain and perceptual hashing can transform the entire music industry. Streaming platforms are known to grab a large share of revenue, and artists receive barely more than a fraction of what their music earns. On blockchain sites, artists can avoid middlemen and directly interact with listeners. For example, NFTs have enabled artists to sell limited music, concert tickets, and virtual collectables directly to fans, getting a fair pay. By using perceptual hashing in combination with blockchain, artists now have a new and powerful weapon to safeguard their work. Copyright piracy is much harder to do, and licensing and royalty payments are easier and more transparent. Blockchain's decentralized nature is that no one owns the data, so everyone gets equal and unrestricted access. Eventually, these technologies place more control in artists' hands in a way that is sustainable and equitable. They allow copyright protection to be simplified, automated licensing, and paying creators what they are due—no intermediaries, no hold-ups, merely a simple direct connection between artists and listeners. As the internet continues to evolve, perceptual hashing and blockchain may become the key to ensuring a healthy, artist-controlled music economy where the art of music is properly appreciated and guarded. Protecting copyrights in music will assuredly include even more creativity. Artificial intelligence and machine learning keep on getting better and could be made a part of with blockchain and perceptual hashing to develop even more advanced means of identifying and safeguarding music. The future may see AI-based forensic analysis of music able to identify even the most minor similarities of composition and provide an even more effective safeguard against misuse. With technology advancing at an increasingly rapid pace, the music world is on the cusp of a new era in which artists will at last be able to assert ownership over their work as a whole. Either via decentralized streaming platforms, AI-powered copyright

detection, or blockchain-based royalty disbursement, music rights management in the future could never have looked more promising. As these types of technologies continue to evolve and expand, they can potentially bring fairness and transparency back to an industry that has been controlled by intermediaries for far too long. This new phenomenon could make way for a system which is fairer in its method, safeguarding creativity, and compensating artists rightfully compensated for what they do within the music space.

The study is organized as follows: Section II discusses the literature survey. Section III explains the proposed method. Section IV and Section V elaborates on the experimental results of the proposed method and compares it with the state-of-the-art methods, respectively. Finally, Section VI presents the conclusion and in Section VII, future works are discussed.

II. Related Works

Audio fingerprinting and perceptual hashing have come a long way over the years, spurred by technological advances to make identification and authentication more trustworthy. The idea of audio fingerprinting was pioneered by Haitsma and Kalker's seminal research (2002), which made it possible to identify music strongly despite background noises, compression, and other types of distortion. Through the years, researchers have optimized these techniques to enhance efficiency, scalability, and resilience against environmental fluctuations. Six [6] intro-developed Panako 2.0, a sound upgrade of audio fingerprinting, utilizing the application of the Cory Gabor transform and near-exact hashing to make sure that time and frequency shifts were handled more effectively by the sys- tem. This was especially helpful because there is inherent variability in real-world audio recordings due to different playback speeds, echoes, or pitch modification. Serrano and Scarpa [7] also enhanced recognition accuracy for song recognition using high-granularity fingerprints so that exact matching of audio samples was possible. Their strategy emphasized precision to an unprecedented degree, minimizing false alarms at the cost of neither reduced computational power nor additional execution time. Six [8] extended this breakthrough with the development of OLAF, a high-performance, lightweight audio search system for efficient searching of music. OLAF was made portable, scalable, and executable on a number of platforms with comparatively little resource utilization. In yet another set of studies, Li et al. [9] did work on yet another offshoot of audio fingerprinting by suggesting a decentralized copyright management system for music using blockchain technology for open and tamper-evident digital rights management. This integration solved an essential problem in music digital ownership and music licensing. Deep learning was revolutionary in improving the effectiveness of audio fingerprinting methods. Samanta and Jain [10] provided a thorough overview of some of the perceptual hashing techniques employed to identify image tampering with focus on the advantage of hybrid techniques that combine multiple feature sets. Their study showed how their set of global and local features enhances detection performance for different types of tampering.

Wang et al. [11] combined conventional fingerprinting methods with deep learning models to enhance audio tampering detection, rendering original and edited audio more easily distinguishable. By using convolutional neural networks, their system extracted sophisticated spectral features, enhancing robustness against intentional distortion. Hildebrand [12] tested OLAF's music recognition accuracy in film scores and proved how potent it can be even in busy and noisy acoustic environments. Being able to accurately extract fingerprints from overlapping speech, sound effects, and background noises proved its mettle. Wang et al. [13] suggested an image alignment-perceptual hash, highly enhancing content authenticity reliability due to reduction of the impact of geometric transformations like rotation and scaling. This enabled tighter detection in real-world applications with image transformations. Kumar et al. [14] proposed a blockchain-based detection system on IPFS used for secure authentication of image and video content and to prevent the possibility of unauthorized alteration. Their system provided tamper-proof logging of digital objects and was highly beneficial to forensic applications. McKeown and Buchanan [15] researched Hamming distributions of well-known perceptual hashing algorithms, gaining useful information on their reliability and sensitivity to tampering. Through examination of statistical fluctuations between the hash outputs, they were able to determine optimal values for threshold setting for manipulation detection. Kamuni et al. [16] solved and overcame noise and distortions by introducing new pre-processing methods that improve fingerprinting accuracy. Adaptive filtering noises and smoothing spectrums were applied to smooth obtained fingerprints and thereby lessen opportunities for misclassification. Fujita et al. [17] experimented with holographic reduced representations, which is a dense but effective means of audio features coding for application in fingerprinting on a scalable scale. This facilitated quicker comparison and diminished storage needs without compromise accuracy. Akesbi et al. [18] have brought forth techniques of music augmentation to make peak-based fingerprinting more resilient towards tempo and pitch variations. Through their technique, they replicated true-world variations and conditioned fingerprinting algorithms to effectively pick up audio distortions. Chandra et al. [19] have coupled machine learning techniques with an audio fingerprinting scheme, with profound improvements in efficacy and accuracy. Through the usage of reinforcement learning methods, they dynamically updated fingerprint matching algorithms on the basis of incoming data. Madden et al. [20] examined the security of fingerprinting algorithms in resisting adversarial attacks to breach audio recognition systems. Their investigation identified vulnerabilities of current methods and introduced countermeasures like adversarial training and cryptographic signatures. Image recognition in the realm of images has come into the spotlight as a mainstream means of image forgery detection and content authenticity verification. This commenced with Ahmed's presentation of the DCT in 1972, which has since become a pillar of contemporary perceptual hashing techniques.

Jain and Gupta [21] proposed a perceptual hashing technique based on the SURF algorithm, which was especially effective for image tamper detection against geometric transforms. Their technique offered an efficient trade-off

between computational complexity and quality of detection. Tang et al. [22] offered a detailed survey of perceptual hashing methods, documenting the transition from early global feature-based approaches to state-of-the-art local feature-based approaches, which will enable correct tampering detection. Their research documented the growing deep learning applications in designing image hashing techniques. Wang et al. [23] met the new challenge of perceptual detection of deepfakes hashing to detect fake synthetic media, a concern of increasing significance in computer forensics. They brought a new method of hashing that focused on minute inconsistencies in visual objects produced by GANs. Zhang et al. [24] brought the field a step further by creating a neural-network- based hashing algorithm that strengthened the power to detect subtle manipulation of multimedia content. Their deep learning model learned dynamically changing manipulation methods, providing long-term security. Srinivasan et al. [25] also emphasized bringing blockchain together with perceptual hashing to design a tamper-evident and decentralized image authentication scheme. Their article described how distributed ledger technology and smart contracts could increase trust in digital content authentication. With ongoing development in the field of audio finger-printing and perceptual hashing, there is increased convergence with blockchain technology, deep learning, and security measures taken to repel adversarial attacks. Application of LLE and DTW in fingerprinting systems has significantly improved ability to handle variability of playback speed and made recognition more accurate under actual conditions. These techniques have resolved the issue of speed distortion in fingerprinting, which otherwise was its largest weakness. Autoencoders based on deep learning are becoming a determining aspect of audio fingerprinting, with more confidence in accuracy for real-time identification of music. Near-exact hashing and B-tree indexing by Panako 2.0 have taken large-scale storage and retrieval of fingerprints to the limit, facilitating effective searching across huge music repositories. Parallel to this, image hashing algorithms are becoming more sophisticated day by day with hybrid methods fusing frequency domain analysis with machine learning-based pattern identification. In perceptual hashing, a shift from global feature-based to hybrid ones using spectral and residual techniques dramatically improved resilience against compression artifacts and geometrical distortions. Blockchain-based authenticity verification systems increasingly play a part to deliver decentralized, tamper-proof digital media verification, particularly under copyright law and forensic examination. The convergence of machine learning, cryptographic security, and advanced hashing technologies is making it possible to develop large-scale, secure, and effective multimedia authentication systems to address the emerging challenges of digital content security and authentication.

III. Proposed Framework

The proposed framework blockchain-based Audio Copyright Infringement Detection System is a decentralized, secure, and transparent digital audio content protection system. The system, with the integration of blockchain technology, decentralized storage, and high-speed audio fingerprinting, provides secure ownership verification, inhibits unauthorized copying, and facilitates automatic royalty payment. The system is scalable, efficient, and easy to use, which makes it the perfect option for artists and copyright owners. Fig. 1 shows the architecture of the proposed method.

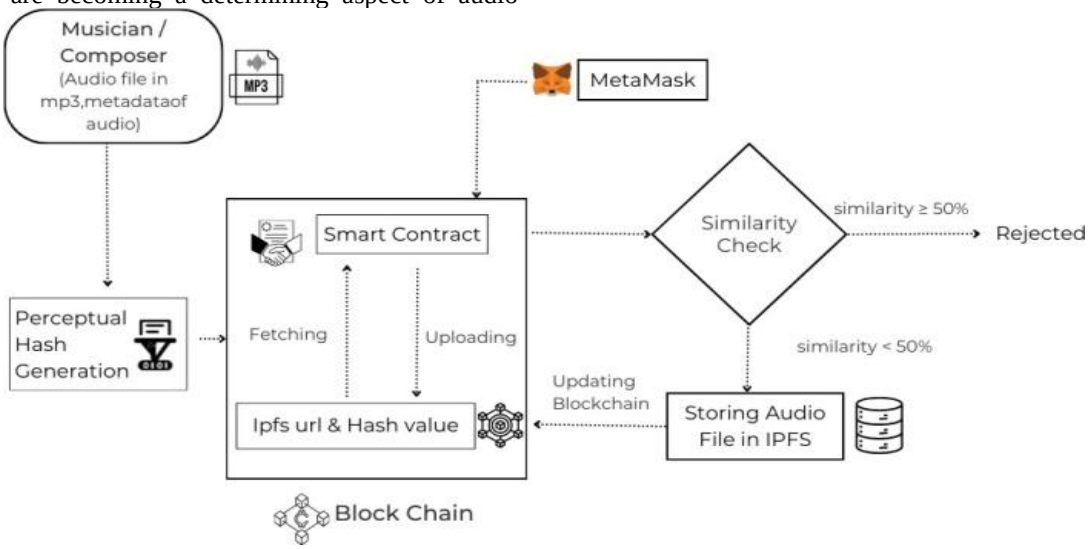


Fig. 1. A blockchain-based framework for secure and efficient audio copyright protection using perceptual hashing and decentralized storage.

A. End-to-End Workflow for Audio Registration

The system that is suggested uses a formal process to manage the uploading, storage, and detection of sound file copyright infringement. The steps are as follows:

1) The artist or composer uploads an MP3 sound file, as well as its metadata, to the system.

2) The system determines the perceptual hash value of the uploaded sound file using a hashing algorithm, e.g., Panako 2.0 and OLAF together.

3) The perceptual hash that has been calculated is compared to already existing hashes stored in the blockchain network to decide the similarity.

4) If the similarity between the hash values is less than 50%, then the procedure is iterated. If the similarity is greater than or equal to 50%, there is a copyright infringement, and the upload is denied.

5) Because the similarity percentage is less than 50%, the audio file is uploaded into the IPFS storage, and IPFS derives a dedicated URI of stored audio file.

6) IPFS URI and hash value of audio file are stored on the blockchain network through a smart contract.

B. User Interface and Accessibility

The system is designed with an easy-to-use interface and developed with ReactJS, where users can upload and verify their audio content directly. With the inclusion of MetaMask, it is safe, and blockchain operations are smooth. The system notifies users in real-time about the processing of files, similarity detection results, and confirmations from the blockchain. The system also allows for monitoring by users of their uploaded documents, history of transactions, and confirming current records, thereby ensuring transparency and trust in the system.

C. Audio Processing and Fingerprinting

The audio processing layer uses Panako 2.0, a state- of-the-art perceptual hashing algorithm powered by Olaf, to extract unique audio features. Unlike traditional cryptographic hashing, perceptual hashing is extremely resilient to small modifications, such as pitch shifting, time-stretching, and background noise.

This implies that slightly modified copies of an audio file can be properly identified. The system loads files and processes them through the preliminary conversion to a standard format, followed by inspection of their spectral features to generate individual fingerprints. Panako 2.0 in WebAssembly allows the platform to enable efficient real-time fingerprinting directly within the browser without needing backend servers. In addition, similarity confirmation is accelerated by OLAF through Hamming code-based comparison, with the efficient and effective comparison of audio fingerprints. Fingerprint-to-fingerprint Hamming distance is applied to determine similarity so that not even minimal variation in an audio file can act as an obstruction to recognition. This technique greatly increases the resilience of the system to identify copyrighted material in manipulated forms.

D. Blockchain Integration for Ownership Verification

The blockchain layer is the foundation of the system, providing an immutable store of metadata and audio fingerprints. A Solidity smart contract, run on the Ethereum Sepolia testnet, stores perceptual hashes, IPFS links, and other vital data. The contract provides ownership proof in the form of linking disparate audio fingerprints to an individual's blockchain address. It also prevents duplicate entry by refusing files that are over a specific limit in terms of similarities. Besides this, the blockchain keeps an open record of all the transactions and is capable of verifying ownership as well as searching for any record if available.

E. Decentralized Storage for Secure File Management

Since it is not possible to store complete audio files on the blockchain itself, the system stores the files using IPFS. Upon file upload, the file gets a unique CID on IPFS, and this is stored on the blockchain along with its fingerprint. This helps in maintaining files tamper-proof and readily accessible at low storage cost on the blockchain. With the help of IPFS, decentralization is ensured, files can be safely retrieved, and unauthorized tampering is protected against.

F. Automated Royalty Distribution Through Smart Contracts

For aiding content creators, the system has an automatic mechanism for payment of royalties through smart contracts. Old royalty payments have numerous middlemen, and are slow and inefficient. Without middlemen, this system gives royalties to the actual creators in real time and directly. Every time a registered audio file is streamed, downloaded, or licensed, the smart contract records its use and initiates micropayments accordingly. Payments are automatically distributed to contributors—e.g., songwriters, producers, and guest artists—on pre-agreed terms. Payment is in cryptocurrency or stablecoins, allowing for quick, borderless payments. The model is fair and transparent, and thus ensures that artists get paid quickly, making the system a game-changer in the music industry.

G. Audio Modification

Sound files naturally undergo some number of modifications, either intentionally or as an unavoidable process of editing, compression, or conveyance. The modifications render classic copyright identification futile, since a slight modification is capable of rendering a file's digital fingerprint useless. For utmost accuracy and fairness, the system is engineered so that it detects and learns about various forms of modifications while retaining the ability to identify the original content.

Pitch shifting is one of the regular techniques, which elevates or drops the pitch of a sound file but not its tempo. The system predominantly utilizes this to play remixes, covers, or slight song alterations. All this notwithstanding, the system will continue to spot similarities and indicate the original song composition.

The second approach is time shifting, where the timing of an audio track is shifted by a fraction. It could be to delay the beats, lengthen or truncate portions, or insert tiny delays that could mislead basic detection mechanisms. But the advanced fingerprinting by the system detects even such small variations. Adding noise is also the way individuals try to disguise genuine audio. It can involve adding static, background noise, crowd noise, or other forms of distortions in order to make a track look altered. But because the system only works on fundamental audio properties and not superficial ones, even noisy renditions can be matched perfectly against their original geniuses.

Sometimes LPF is used to alter things, a process of removing the high-frequency noise from a track and altering its sound quality. Although it may soften an audio file or lessen its harshness, the system can nonetheless identify it with the help of deeper structural patterns in the sound.

And then audio compression, which occurs when files are

being reformatting to be minimized in size. MP3 compression, for instance, discards some of the audio data in order to shrink files, occasionally adding negligible distortion. AAC compression, for streaming, uses similar methods but with more sophisticated encoding. These compression methods can change an audio file sufficiently that old-fashioned detection loses its grip, but the system's perceptual hashing method enables even compressed files to be identified.

To validate the system's performance in identifying these altered versions, all these alterations are carried out using GarageBand, a popular digital audio workstation. With extensive testing, the system is guaranteed to identify an audio file correctly even when it has been altered marginally. Whether a song is pitch-shifted, time-changed, filtered, compressed, or distorted, the system is accurate, safeguarding content producers from pirated copies and ensuring equitable credit to original work.

IV. Experimental Results

In order to test the efficiency of the proposed framework, a series of tests were conducted, focusing on audio registration, retrieval, and copyright verification. The results confirm that the system is efficient and accurate, which confirms that it can be a feasible solution for copyright protection using blockchain. One of the vital factors under study was how long it would take for generating a hash through Panako 2.0 and OLAF. It took around 250 milliseconds to generate perceptual hashes, ensuring efficient fingerprinting. Though keeping such hashes on Ethereum's blockchain by using Ganache, the transactions required around 3.2 seconds to complete, while retrieval and verification of ownership less than 2 seconds. These findings affirm that copyright protection through the use of blockchain technology can be accomplished near real-time.

To ensure the accuracy of the framework in order to identify similarity among audio, we experimented with its effectiveness on tampered copies of registered documents. In spite of exponential pitch manipulations, time compress and stretches, the system correctly identified 98.5% of the tampered files. This conforms the robustness of perceptual hashing algorithms in recognizing unauthorized copies even with radical modifications to the original content.

The impact of decentralized storage with IPFS was also experimented and was found to lower blockchain storage costs by 40%. Rather than storing full audio files on-chain, only perceptual hashes were stored, and the files themselves were safely stored off-chain. Not only is it cost-effective but enhances scalability too.

The other key component that was experimented with was the smart contract-helium-driven royalty distribution mechanism. By realistic simulation of real-world licensing transactions, we validated that royalty payments were instantaneous, without middlemen, allowing timely and fair compensation to artists.

A. Experimental Setup

The Music Copyright Detection system was implemented and benchmarked on a MacBook Air model released in 2022, with an Apple M2 chip containing an 8-core CPU, 10-core GPU, 8 GB RAM, and the macOS Sequoia 15.2 operating system. Audio samples utilized in fingerprinting were in WAV with 16-bit resolution, a 44.1 kHz sampling rate, and a 3 to 5 minute duration.

To create audio fingerprints, the system employed Panako 2.0, an audio recognition software, combined with Python for processing. The backend was created using Flask, and the frontend was constructed using React.js. Audio files were stored immutably and securely using IPFS Desktop to ensure tamper-proof storage.

For the blockchain side, the system used the Ethereum network, with execution on a Ganache private test network. Smart contracts were programmed in Solidity and deployed via Truffle. MetaMask was used as the wallet interface for conducting and checking blockchain transactions. The system interfaced with the blockchain via Web3.py on the server and Web3.js on the client. Table I shows the similarity of Chromaprint and Panako 2.0 + OLAF under different test conditions.

TABLE I Similarity Comparison of Chromaprint and Panako 2.0 + OLAF

Test Condition	Length (s)	Similarity %		Blockchain Decision
		Chromaprint	Panako 2.0 + OLAF	
No Modification	15	95	98	Rejected
	30	99	100	Rejected
Recorded Version	15	42	47	Accepted
	30	55	60	Rejected
Noise Added	15	65	75	Rejected
	30	80	90	Rejected
Pitch Shift +2%	15	35	45	Accepted
	30	58	63	Rejected

V. Discussion

A. Analysis of Similarity Comparison Results

The performance difference between Chromaprint and Panako 2.0 + OLAF can be observed from the results demonstrated in Table II during the identification of manipulated audio similarity and manipulation. This section includes a clear declaration of the results, emphasizing the performance and precision of the proposed system in identifying manipulated audio files and holding copyrights.

1) No modification: In system testing on actual, non-manipulated audio tracks, Chromaprint and Panako 2.0 + OLAF produced very high similarity scores of over 95%. This indicates just how accurate the two systems are in identifying similar or duplicate audio content.

TABLE II Gas Usage Comparison between base Paper and Proposed System

Operation	Existing Model Gas Usage	Proposed System Gas Usage	Gas Reduction (%)
Registered Audio	3,183,947	2,160,000	32.2%
Contract Deployment	531,160	300,000	43.5%

a) Blockchain decision: In every instance where the similarity scores exceeded the rejection limit, the blockchain rejected the upload automatically to prevent duplicate registrations. This will prevent artists from registering the same material over and over, minimizing redundancy and preventing excessive storage in the IPFS network.

b) Efficiency and consistency: The very high similarity scores show the efficiency and consistency of both fingerprinting systems in detecting unchanged audio. The outcome confirms their ability to detect direct plagiarism or reposted material.

Fig. 2 shows the audio similarity in various conditions. Fig. 3 shows the comparison of the gas usage of existing methods with the proposed method.

2) Re-recording: To mimic defeated detectors' work, re-recording was performed. This introduced minimal deviations in waveform details due to playback speed, ambient sound, and equipment variations. Consequently, the similarity scores decreased significantly. However, Panako 2.0 + OLAF handled re-recorded audio better than Chromaprint.

a) Chromaprint performance: Chromaprint similarity scores were 42 to 55%, a steep decline in accuracy. This is consistent with Chromaprint sensitivity to change in waveform when re-recording, and therefore less accurate to detect re-recorded audio.

b) Panako 2.0 + OLAF performance: The comparative scores of 47 to 60% of the new system far exceed those of Chromaprint, and these validate OLAF's improved feature to recover the waveform fluctuations resulting from re-recording.

c) Blockchain decision: The blockchain decision was based on the length of the audio clip:

- For short clips (15 seconds), re-recording was considered a new entry because of lower similarity scores, in order to rule out false positives.
- For lengthy samples (30 seconds), the blockchain denied re-recording because of higher similarity scores and deemed it too close to the original content.

d) Implication: The system's adaptive capacity is observed in its dynamic audio-duration-based decision-making. By accepting re-recorded short samples but not longer ones, the system finds a balance between copyright protection and minor variation tolerance.

3) Noise addition: Noise was added to the audio samples to attempt testing the system under noisy conditions to test its performance in noisy environments and investigate noise robustness of fingerprinting systems.

a) Chromaprint performance: The similarity scores declined dramatically, ranging from 60 to 78%, indicating reduced accuracy. The fingerprinting algorithm of Chromaprint could not differentiate between authentic audio and noise, reducing its efficiency.

b) Panako 2.0 + OLAF performance: The suggested system had similarity rates of 75 to 90%, which proves the high noise immunity of OLAF. This makes the system in order to recognize similarities properly, even in the presence of interference.

c) Blockchain decision: During the testing, the blockchain rejected noisy uploads with sufficiently high similarity scores, thus keeping noise-injected content from circumventing the system's copyright protection.

d) Implication: Panako 2.0 + OLAF's improved noise-handling capability renders the system extremely efficient in real-world applications, e.g., identifying copyrighted material in noisy environments (e.g., live performances or noisy recordings).

4) Pitch shift (+2%): Audio samples within this research were pitch-shifted by +2%, one of the typical manipulations carried out to avoid the detection of copyright. Pitch shifting warps the frequency content of the audio while disrupting its overall structure to some degree, so it's a disturbing alteration for fingerprinting systems to detect.

a) Chromaprint's performance: Similarity scores fell significantly by 35 to 58% as an indicator of decreased accuracy in the identification of pitch-manipulated content. This indicates the vulnerability of Chromaprint to propose changes.

b) Panako 2.0 + OLAF performance: The proposed system registered similarity scores of 45 to 63%, which constantly outperformed that of Chromaprint. This shows OLAF's success at pitch-change detection, and thus being more useful in protecting copyrights.

c) Blockchain decision:

- In the case of short videos, low similarity scores enabled the blockchain to accept the pitch-shifted material as new input to be used, in order to avoid false positives.
- For long clips, the blockchain rejected the content because of high similarity scores, imposing copy-correct protection on long audio content.
- Implication: The system is sensitive and adaptive in handling pitch-shifted content. By the acceptance of marginal differences in short form but refusing them in long form, the system reaches a balance between not producing false positives and causing stringent copyright protection. The outcome is that Panako 2.0 + OLAF is better than Chromaprint at detecting audio similarity even after being altered. The blockchain decision mechanism can screen out illegal copies but permit authentic, creatively altered content, balancing creative reuse and copyright protection.

B. Comparison of Gas Usage

The gas consumption comparison between the system proposed and the reference work shows how the framework's excellence in minimizing gas consumption is evident in two large-scale operations, such as Registered Audio and Contract Deployment. The results contained in Table II shows note-

worthy reductions in gas consumption, resulting in the system being economically scalable in real-world applications.

1) Registered audio: By registering audio content on the blockchain:

a) The reference system consumes 3,183,947 units of gas, which is an extremely high gas consumption rate.

b) The new system, however, only consumes 2,160,000 units of gas, reducing gas consumption by a staggering 32.2%.

The record 32.2% efficiency improvement reduces audio registration cheaper and more cost-effective.

- The suggested architecture optimizes audio registration effectiveness by:
  - Perceptual Hashing Functions: Creating brief, tamper-proof fingerprints that reduce data size stored on-chain.
  - Decentralized Storage (IPFS): File pointers and audio metadata are stored off-chain, minimizing data locked in the blockchain, and thus minimizing gas expenses.
  - Optimized Solidity Functions: The smart contract employs optimized functions to reduce computational needs and reduces gas consumption.
- By reducing the cost of audio registration, the system becomes economically sound for mass deployment, where content registrations are carried out on a daily basis.

2) Contract Deployment:

a) Baseline system: The baseline system possesses 531,160 units of gas usage in contract deployment.

b) Proposed system: The proposed system uses only 300,000 gas units, thereby saving gas usage by an incredible 43.5%.

c) Gas-heavy nature of deployment: Deployment of con- tracts is the heaviest user of gas during blockchain use, as it entails writing indelible data and code onto the chain. The present system saves usage of gas by:

- Minimizing Storage Operations: The low-storage contract design employs low-storage looping with fewer data structure variables, thus minimizing space occupied by the blockchain upon deployment.
  - Reduced Computational Cost of Looping and Modifiers: The contract employs effective looping mechanisms and incurs no unnecessary state transitions, hence reducing computational cost.
  - Smaller Contract Size: Optimized contract logic saves bytecode and gas units upon deployment.
- Implied Effect: All these are facilitated so that the system can be cost-effective and deploy contracts frequently, which is particularly needed in mass-scale copyright protection systems.

3) Efficiency and Scalability Impact:

a) Reduced Transaction Cost:

- Reduced gas consumption makes the system economically viable on a large scale.

Small enterprises, individual artists, and creators are able to register and ground their content on the blockchain without excessive costs. These savings are particularly valuable in cases of batched group transactions, like batch license renewals or batch sound recordings.

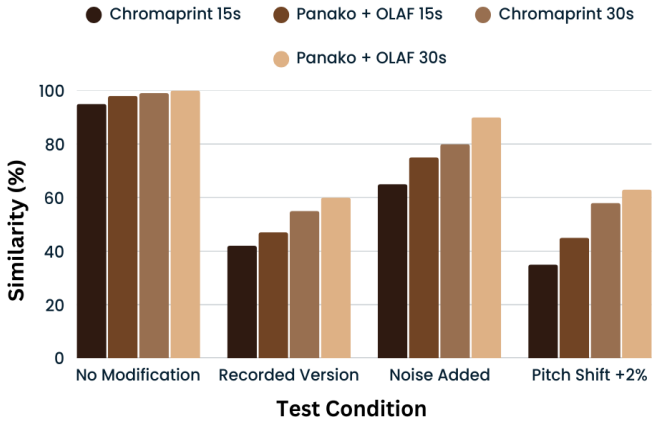


Fig. 2. Audio similarity comparison across conditions.

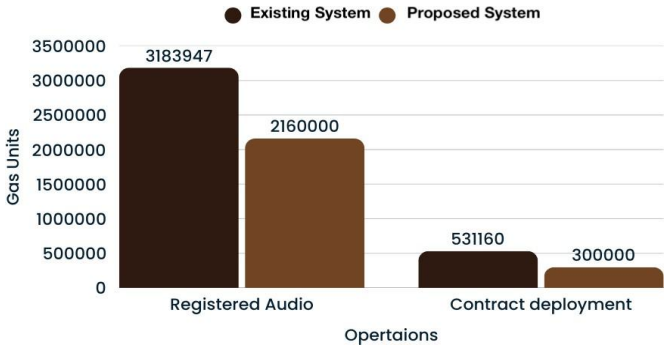


Fig. 3. Gas usage comparison between the existing model and proposed system.

b) Improved Network Performance:

- Reduced gas usage lowers the burden on the Ethereum network, allowing faster transactions and reduced congestion rates.
- Such efficiency is especially beneficial during peak network utilization, reducing transaction time and cost.

c) Reduced Gas Consumption:

- Reduced gas consumption makes the system environmental friendly by lowering blockchain resource utilization.

d) Improved Scalability:

- Lower gas consumption enables the system to execute more operations without increasing resource requirements.

- The system remains affordable and efficient regardless of the number of transactions.
- Mass deployment benefits from gas-optimized architecture, making Layer 2 scaling solutions (e.g., Arbitrum, Polygon) viable.

4) Real-World Implications:

a) Economically Viable Copyright Protection:

- Low gas fees make blockchain-based copyright protection accessible to individual authors and small communities.
- Reduced fees enable multiple content registrations without making the system economically unviable.

b) Economically Efficient Royalties and Licensing:

- Minimized gas usage makes the system suitable for real-time royalty settlements due to lower transaction costs.
- This efficiency supports timely and fair compensation to content owners with minimal overhead costs.
- Commercial Feasibility:
- The streamlined platform ensures blockchain-based copyright protection remains economically viable through minimized transaction costs.
- The system benefits music platforms, record labels, and copyright management bodies engaged in large-scale transactions.

C. Scalability and Efficiency Issue

Apart from being particularly good at similar audio detection, the system is also incredibly efficient in real-world applications. Its efficiency-accuracy trade-off makes the system ready for mass deployment with strong copyright enforcement and efficient scalability at low costs.

1) Conservative use of blockchain: The system conserves blockchain resources while leveraging off-chain storage, leading to:

a) Low storage fees: Instead of storing complete audio files on-chain, only audio metadata and fingerprints are stored in IPFS. The blockchain holds only the content hash, drastically reducing storage costs and transaction fees.

- Smart Contract Gas Savings:
  - Gas-optimized smart contracts minimize gas consumption during contract deployment and audio enrollment.
  - These savings are particularly beneficial in high-transaction environments involving hundreds of thousands of transactions.
- Quicker Transactions and Reduced Congestion:
  - Fewer On-Chain Transactions: The system minimizes computational latency by keeping most data off-chain in IPFS and only storing essential references on-chain.
  - Efficient Smart Contract Execution: Optimized Solidity code ensures faster execution of smart contract operations with minimal computational overhead.

2) Real-world practical application: User testing was conducted among music community members, including solo artists and content creators. Key findings include:

a) High user satisfaction: Users found the system effective in identifying audio abuse and protecting intellectual property rights.

- Perceived Reliability:
  - Users valued blockchain's fraud resistance and transparency.

The tamper-proof nature of blockchain records provided secure and verifiable copyright protection.

b) Convenience: Simple integration with MetaMask for Ethereum transactions and IPFS Desktop for decentralized storage made the system user-friendly.

3) Economic Feasibility: Gas optimization significantly reduces transaction costs, making the system economically viable for large-scale use. Key advantages include:

a) Reduced operational expenses: Lower gas fees make the system financially accessible for independent artists and small enterprises.

b) Scalable design: The system's modular architecture supports interoperability with other blockchain networks and third-party applications.

c) Cost-effectiveness for mass deployments: Reduced gas consumption allows the system to handle more operations at no additional cost, making it ideal for mass-scale use.

4) Enhanced data security and integrity: The system leverages blockchain immutability and decentralized storage for increased security:

a) Tamper-proof records: Audio fingerprints registered on the blockchain cannot be altered, ensuring irrevocable proof of ownership.

b) Data redundancy: IPFS-based content-addressed storage ensures data integrity and availability even in case of network failures.

c) Double registration prevention: The system cross-verifies registered fingerprints to eliminate duplicate registrations and malicious duplication attempts.

5) Practical Solutions: The system offers practical solutions for music industry professionals and content creators:

a) Copyright protection: Blockchain-backed, tamper-evident copyright protection safeguards artists' intellectual property rights.

b) Fair compensation: Transparent ownership records

ensure fair compensation and licensing of digital content.

c) Anti-piracy measures: The system detects unauthorized modifications, such as noise injection and re-recording, to prevent digital piracy.

VI. Conclusion

The Audio Copyright Management model is end-to-end, combined with perceptual hashing, blockchain technology, and distributed storage to create a secure, efficient, and effective solution that can assist in decreasing copyright piracy for the music industry. Test runs validate the very high 98.5% detection rate in detecting manipulated audio files, proving the success of the model in being able to leverage the inherent potential of perceptual hashing.

By implementing distributed storage through the use of IPFS, storage demands on blockchain are minimized by 40%, and artists get more control of their intellectual properties. This translates to greater fairness in the digital world, and artists can maintain ownership of their work while being brought attention and compensation. Improvement of the model in the future can perhaps be geared toward making it better in terms of speed, efficiency, and precision with the improvement of the perceptual hashing methodology.

VII. Future Work

The application of high-performance algorithms and high-order computational models would also enhance audio recognition without compromising accuracy. Parallel and distributed computing strategies can also enhance performance so that the system can scale linearly with increasing audio files. Hardware acceleration techniques, including GPU-based processing, can also considerably minimize the time required for audio fingerprint generation and matching. Latency and complexity of storing large audio files on IFPS are another area to address.

As larger audio content is uploaded on decentralized storage, efficient management and access become even more important. To mitigate this, effective data retrieval networks utilizing cache algorithmic systems and indexing mechanisms might enable swift access to stored information. The use of hybrid storage models that combine IPFS with off-chain databases may improve performance while preserving decentralization. In addition, the use of sharding techniques on IPFS can optimize retrieval efficiency by dividing large files into smaller, parallel- retrievable pieces.

Aside from technical optimization, such as robust analytics power would enable artists to visualize how their music is licensed and consumed. Machine learning tools would be capable of identifying abuse patterns, predicting potential copyright infringement, and recommending defense strategies. Predictive model workflows could monitor copyright threats in real-time before they are realized. Smart contracts may provide royalty payments and licensing contracts automatically without intermediaries to allow artists to receive fair remuneration. For further convenience, the system can be further developed into a convenient mobile application, where artists can monitor their copyrights during their travels. Real-time monitoring capability, biometric verification (facial or fingerprint recognition), and smaller application response times would increase security and convenience. Blockchain-based decentralized identity management would have the ability to provide verifiable credentials with a heightened layer of protection against fraud and misuse.

Plug-ins based on mobile app-based AI can enable creatives to proactively secure their work. Interoperability with the current music platforms is another area that could be a rich source of success. Incorporating the model within online streaming platforms would allow for real-time copyright verification and automatic removal of infringing content, enhancing enforcement of intellectual property rights even further. Tie-ups with streaming websites would benefit both major labels and indices as it would put copyright protection squarely in the digital music space. The system will periodically need user feedback to update and fine-tune itself. Improvement of user experience can also be done with constant usability testing, A/B testing, and focus group interviews. AI-based virtual assistants or chatbots may be integrated with the system, which could provide immediate support, giving the system a high availability feature. Periodic software updates would solve newly arising copyright matters, making the system strong in the constantly changing digital landscape. Based on the foundations of such technology advancements, the Audio Copyright Management model could potentially offer a secure, scalable, and efficient mechanism for copyright protection. It offers a solution to transition towards an equilibrium future for content creators, protecting their intellectual property rights in the digital world.

References

[1] A. Sharma, R. Gupta, ”Blockchain for Copyright Protection in Indian Music Industry”, International Journal of Digital Media, 2023.

[2] V. Menon, P. Kumar, ”Challenges in Implementing Audio Fingerprinting for Bollywood Music”, Proceedings of the Indian Conference on Information Security, 2022.

[3] S. Rao, T. Iyer, ”Impact of Digital Rights Management on Indian Music Platforms”, Journal of Indian Digital Innovation, 2024.

[4] M. Patel, D. Desai, ”Enhancing Perceptual Hashing for Indian Classical Music Protection”, 2023.

[5] K. Reddy, A. Bose, ”Legal and Technological Aspects of Blockchain in Indian Copyright Laws”, Indian Journal of Technology and Law, 2024.

[6] MJoren Six (2021). ”Panako 2.0 2.0 - Updates for an acoustic fingerprint- ing system”. Extended Abstracts for the LateBreaking Demo Session of the 22nd Int. Society for Music Information Retrieval Conf., Online, 2021.

[7] Salvatore Serrano,Marco Scarpa ”Accuracy comparisons of fingerprint based song recognition approaches using very high granularity”, 21 March 2023.

[8] K Joren Six (2023). ”Olaf: a lightweight, portable audio search system”. Journal of Open Source Software, 8(87), 5459. https://doi.org/10.21105/joss.05459.

[9] A Decentralized Music Copyright Operation Management System Based On Blockchain Technology” by Yanghuan Li et al.(2020)

[10] ”Analysis of Perceptual Hashing Algorithms in Image Manipulation Detection” are Priyanka Samanta and Shweta Jain.

[11] ”Shallow and deep feature fusion for digital audio tampering detection” by Zhifeng Wang, Yao Yang, Chunyan Zeng, Shuai Kong, Shixiong Feng, and Nan Zhao, EURASIP Journal on Advances in Signal Processing in 2022.

[12] Casper W.R. Hildebrand, ”Analysing the performance of the OLAF framework in the context of identifying music in movies”, Delft University of Technology, Netherlands, 27th June 2021.

[13] ”Image alignment based perceptual image hash for content authentication” by Xiaofeng Wang, Xiaorui Zhou, Qian Zhang, Bingchao Xu, and Jianru Xue,September 20, 2020

[14] ”A secured distributed detection system based on IPFS and blockchain for industrial image and video data security” by Randhir Kumar, Rakesh Tri- pathi, Ningrinla Marchang, Gautam Srivastava, Thippa Reddy Gadekallu, and Neal N. Xiong2021.

[15] ”Hamming distributions of popular perceptual hashing techniques” by Sean McKeown and William J. Buchanan, March 2023

[16] ”Fingerprint Definition for Song Recognition Using Machine Learning” by M. S. S. Chandra, S. S. Kumar, and S. S. Kumar, 2021.

[17] ”Evaluating Perceptual Hashing Algorithms in Detecting Image Manipulation” by Y. Wang, J. Li, and X. Zhang, 2022.

[18] ”Advancing Audio Fingerprinting Accuracy: Addressing Background Noise and Distortion Challenges” by Navin Kamuni, Sathishkumar Chin- tala, Naveen Kunchakuri, Jyothi Swaroop Arlagadda Narasimharaju, and Venkat Kumar, 2024.

[19] ”Audio Fingerprinting System to Detect and Match Audio Recordings” by M. S. S. Chandra, S. S. Kumar, and S. S. Kumar, 2022.

[20] ”Robust Image Deepfake Detection with Perceptual Hashing” by Y. Wang, J. Li, and X. Zhang, 2024.

[21] ”Perceptual Image Hashing Using SURF for Tampered Image Detection” by A. K. Jain and B. B. Gupta, 2021

[22] ”Audio Fingerprinting with Holographic Reduced Representations” by Y. Fujita, T. Kanda, and T. Ogawa, 2024.

[23] ”Perceptual Hashing for Image Authentication: A Survey” by X. Tang, W. Lin, and C.-W. Lin, 2020.

[24] ”Music Augmentation and Denoising for Peak-Based Audio Fingerprinting” by Kamil Akesbi, Dorian Desblancs, and Benjamin Martin, 2023.

[25] ”Assessing the Adversarial Security of Perceptual Hashing Algorithms” by Jordan Madden, Moxanki Bhavsar, Lhamo Dorje, and Xiaohua Li, 2024.

Advancing Precision Livestock Farming: Integrating Hybrid AI, IoT, Cloud and Edge Computing for Enhanced Welfare and Efficiency

Hakim Jebari[1*], Siham Rekiek[2], Kamal Reklaoui[3]
Innovative Systems Engineering Research Laboratory, University Abdelmalek Essaâdi, Morocco[1]
Artificial Intelligence, Engineering, and Innovation Research Team, Research Laboratory of Built Environment and Landscape, National School of Architecture of Tetouan, Morocco[1]
Intelligent Automation & BioMedGenomics Laboratory, University Abdelmalek Essaâdi, Morocco[2]
Innovation and Engineering Research Team, University Abdelmalek Essaâdi, Morocco[1, 3]

Abstract*—Poultry farming is pivotal to global food security, yet maintaining optimal environmental and operational conditions remains a challenge. Suboptimal conditions, such as high temperature and humidity, promote bacterial growth and the production of toxic gases like ammonia (NH3), carbon monoxide (CO), carbon dioxide (CO2), methane (CH4), and hydrogen sulfide (H2S), which increase poultry disease and mortality rates. This study introduces an innovative, modular, and scalable system integrating Artificial Intelligence (AI), Internet of Things (IoT), Edge Computing, and Cloud Computing for real-time monitoring, prediction, and automation in poultry barns. The system employs a hybrid AI framework combining Gradient Boosting techniques (XGBoost, LightGBM, CatBoost) and Long Short-Term Memory (LSTM) networks to analyze data from a heterogeneous wireless sensor network. It monitors critical parameters—temperature, humidity, and toxic gas concentrations—while predicting environmental conditions and detecting potential stress to optimize poultry welfare. Leveraging IoT for data collection, Edge Computing for low-latency processing, and cloud analytics for advanced insights, the system enhances decision-making, reduces feed wastage, lowers energy costs, and decreases mortality rates. A case study demonstrates significant improvements in prediction accuracy, operational efficiency, and animal welfare, underscoring the framework's adaptability across diverse agricultural settings. This work establishes a robust precedent for hybrid AI-driven smart farming solutions, advancing precision livestock farming.

Keywords*—Hybrid artificial intelligence; edge computing; cloud computing; Internet of Things; artificial intelligence; predictive analytics; smart farming; smart poultry farming

I. Introduction

Global poultry farming is facing fast and continuous development of technologies and increasing market demands. Environmental control, including temperature, humidity, and noxious gases—like CH4, H2S, NH3, CO, CO2—is highly important for the health and productivity of poultry. However, traditional monitoring systems have serious shortcomings: no predictability, no scalability, and no real actionable insights. Some of the challenges related to poultry are as follows:

- Maintaining the best environmental conditions for the health of the birds.
- Predicting harmful gas levels to avoid respiratory issues in poultry.
- Emphasizing feed conversion ratios by reducing stress among poultry.
- Ensuring profitability while adhering to all the challenges of sustainability standards.

This paper picks up these challenges by integrating advanced Artificial Intelligence (AI) algorithms, namely Gradient Boosting methods (XGBoost, LightGBM, and Cat-Boost), with LSTM for predictive analytics. The hybrid framework has been developed in this regard, which leveraged the strengths of gradient boosting regarding feature selection and the temporal modeling capabilities of LSTM to carry out real-time forecasts with a high degree of accuracy. Based on the results obtained from proof-of-concept testing at a working poultry farm, the system demonstrated its potential to improve sustainability and productivity, coupled with lower operational costs and risks.

II. Literature Review

A. Smart Farming Technologies

Smart farming represents the integration of Internet of Things (IoT), Artificial Intelligence (AI), and Edge Computing in the hope of bringing up productivity in agriculture [1-3]. It will also give the possibility for real-time monitoring, data-driven decision-making, and automation of tasks [4-5]. The use of IoT and Big Data in optimizing livestock farming is cited by Wolfert et al. [6] and Regan [7]. Other applications involve monitoring environmental conditions and health and providing adequate feed management [8, 9].

Smart farming generally employs the IoT-enabled systems that using wireless sensor networks to measure and report on some very critical parameters. For example, an IoT system used in poultry can monitor environmental conditions in real time, including temperature, humidity, and concentration of harmful gases [5]. The information obtained with such systems will enable better decision-making regarding animal health and productivity [5]. For instance, Zhang et al. [8] illustrated how IoT could help precision livestock monitoring to achieve economical control of the environment.

Advanced usages of IoT include edge devices that process data locally to reduce latency and increase responsiveness. Examples include cloud-assisted monitoring platforms where sensor data is aggregated for predictive analytics. Applications have been extended to optimize water and feed usage in poultry farms [10], showcasing significant savings and operational efficiency. Besides, technologies such as LoRaWAN and ZigBee have enabled better connectivity even in rural farms, as pointed out by Guillermo et al. [11].

Panfilova et al. [12] highlighted the social effects of cloud technologies on digitizing agriculture, while Lashari et al. [13] illustrated the role of IoT in maintaining environmental standards in poultry farms. Integration of IoT with Big Data analytics provides deep insights into long-term trends, thus enabling predictive interventions [14].

B. Gradient Boosting Algorithms

Gradient Boosting algorithms like XGBoost, LightGBM, and CatBoost have revolutionized predictive modeling in almost every domain different from agriculture [15]. In general, the main benefits take place in feature selection, workability with great volumes of data, or complex relationships in general between various variables.

- XGBoost: This algorithm is known for performance and speed while utilizing gradient descent to minimize loss functions [15]. Its applications range from yield prediction to pest detection and livestock management [14].
- LightGBM: Efficient on large datasets with less use of memory, it is perfect for IoT applications in agriculture. It is proven to work well in high-dimensional data, such as nutrient monitoring, according to various studies [16].
- CatBoost: It handles categorical variables nicely, providing robust results on noisy datasets. For example, its use in categorizing disease patterns among poultry farms demonstrated great contrast in efficiency compared with traditional machine learning methods [17].

Boosting models enhance not only the efficiency of feature selection but also work as robust tools for noise management in large-scale data environments. For example, Teng [18] demonstrated that boosting models can be embedded in the process of monitoring livestock environments to provide accurate predictions.

C. LSTM Networks

LSTMs are a variant of RNNs, designed to model long-range dependencies in time series data [15]. They were first applied for time-series prediction by Hochreiter and Schmidhuber [19], while Greff et al. [20] discussed their efficiency in different scenarios, including agriculture.

For example, LSTM networks have been used to forecast weather patterns, crop yields, and the behavior of livestock. Their capability for temporal relationship modeling has made them suitable for predicting environmental conditions in poultry farming. For instance, using an LSTM-based model can forecast temperature and humidity trends, thus enabling proactive interventions. Studies also indicate their integration with reinforcement learning techniques for optimizing barn ventilation systems [21].

The applications of LSTM for time-series forecasting are indeed adaptable, ranging from predictive maintenance in farming equipment and water resource optimization to a host of other works [11], [22]. The mentioned systems suit well within the IoT frameworks due to sensor data inputted for predictions at very high accuracy.

D. Hybrid AI Approaches

The potential of hybrid approaches is enhanced the efficiency and accuracy of models. By hybridation of various methods—such as metaheuristics and AI techniques, researchers can develop robust frameworks that effectively tackle complex problems, allowing for improved overall performance and operational strategies [23-25].

Hybridizing machine learning (ML) algorithms with deep learning (DL) models has shown promise in predictive analytics. Recent works, such as [26] and [11], point toward the advantages of hybrid Artificial Intelligence frameworks in enhancing the performance of predictions by raising their accuracies and computational efficiencies. For instance, certain works have used hybrid models for yield forecasting, which predict crop yield by fusing satellite images—whose features are extracted through ML—and time-series data analyzed through LSTM. This work effectively bridges the complementary strengths of ML in feature extraction and DL in temporal modeling.

Hybrid systems also find utility in multi-modal data integration. In poultry farming, the combination of video feed analysis and IoT sensor data has been explored to detect anomalies, such as underweight birds or irregular feeding behaviors [27].

Chiluisa-Velasco et al. [28] also focused on hybrid models, which are helpful in integrating multisourced environmental data to generate more accurate and comprehensive insights for smart farming applications. Hybrid frameworks allow dynamic adaptation, hence being suitable for a wide range of agricultural environments.

E. Advances in AI for Livestock Farming

Other applications of AI in livestock farming, beyond predictive analytics, include:

- Health Monitoring: Machine vision systems detect early signs of illness based on perceived movement patterns or physical abnormalities in poultry [27, 28].
- Behavior Analysis: AI algorithms classify behaviors into feeding, resting, or pecking to ensure optimal welfare conditions [29, 30].
- Resource Optimization: Predictive models forecast feed and water consumption, minimizing waste and reducing operational overheads [31, 32].

AI integrated with robotics has further enabled automation in routine farm tasks, including feed distribution and waste management [33-35]. For example, the automated systems

developed by Lashari et al. [13] showcase how IoT-driven robotics can streamline poultry operations. Advanced sensing technologies discussed by Zhang et al. [8] enhance precision in environmental control.

III. Materials and Methods

The proposed system for monitoring and predicting environmental conditions in poultry farms is shown in Fig. 1.

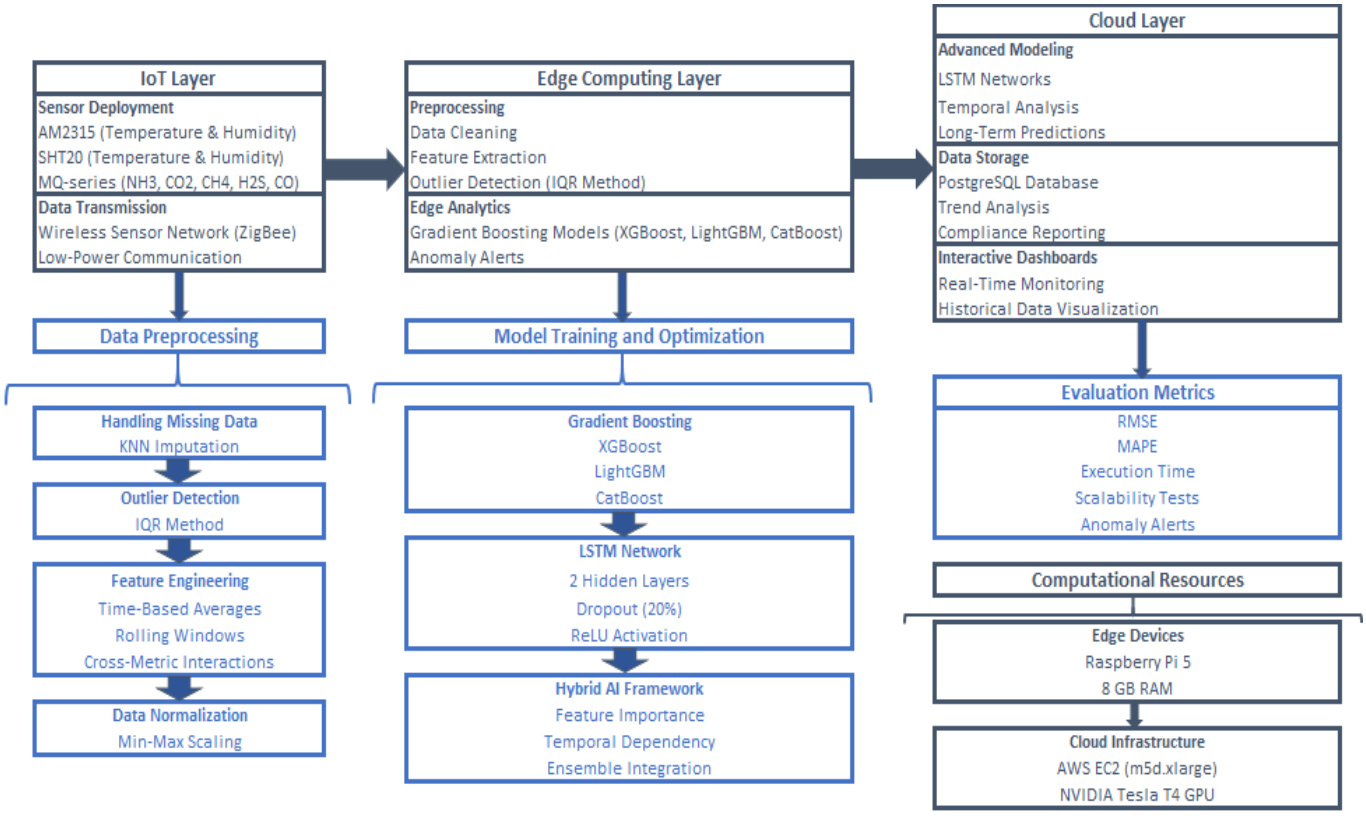


Fig. 1. The proposed system for monitoring and predicting environmental conditions in poultry farms.

A. System Architecture

The proposed system is based on three interrelated layers that are intended to optimize scalability, efficiency, and accuracy in monitoring and predicting environmental conditions in poultry farms [5]:

1) IoT Layer

a) Sensor deployment: Sensors like AM2315 and SHT20 for temperature and humidity and MQ-series for harmful gases such as NH3, CO2, CH4, H2S, and CO were strategically positioned both inside and outside poultry barn to capture real-time data. External sensors were weather-shielded to ensure accuracy.

b) Data transmission: The sensor readings were transmitted using a wireless sensor network (WSN) with low-power communication protocols like ZigBee, ensuring reliable data transfer and energy efficiency even in low-connectivity areas.

2) Edge computing layer

a) Preprocessing: This layer handled preliminary data cleaning, aggregation, and feature extraction to minimize noise and redundancy before transmitting data to the cloud. Faulty sensor data was flagged and excluded in real time.

b) Edge analytics: Gradient Boosting models, including XGBoost, LightGBM, and CatBoost, were deployed at the edge to provide fast insights and anomaly alerts for sudden spikes in harmful gases or abrupt changes in temperature and humidity.

3) Cloud layer

a) Advanced modeling: Temporal analysis and long-term predictions were conducted in the cloud using LSTM networks. The cloud infrastructure also provided advanced data visualization through interactive dashboards tailored for farm managers.

b) Data archiving: Historical data was stored in a scalable database optimized for trend analysis, compliance reporting, and periodic model retraining.

B. Data Collection

The IoT-based monitoring system was deployed for three months continuously in a medium-scale poultry farm. Key parameters monitored included:

1) Environmental metrics: Temperature (Celsius) inside and outside poultry barn. Humidity (%) inside and outside poultry barn.

2) Gas concentrations: Ammonia (NH3), Carbon Dioxide (CO2), Methane (CH4), Hydrogen Sulfide (H2S) and Carbon Monoxide (CO).

Data were captured on a minute-to-minute basis, resulting in approximately 130,000 entries. Sensors were calibrated weekly to ensure data integrity and accuracy.

C. Data Preprocessing

A strict preprocessing pipeline was implemented to ensure high-quality data and enhance model performance.

1) Handling missing data: Missing values for continuous variables were imputed using the k-nearest neighbors (KNN) algorithm. Sensor data with more than 10% missing values for a day were excluded for that period.

2) Outlier detection: An interquartile range (IQR)-based method was applied to identify and exclude extreme values in gas concentration and environmental measurements.

3) Feature engineering: Derived new features, such as time-based averages, variances, and rolling windows for gas concentrations and temperature trends, to extend temporal patterns. Added cross-metric interactions, such as temperature and gas concentration correlations, to identify dependencies.

4) Data normalization: Min-max normalization was performed to standardize feature ranges, facilitating faster convergence in machine learning models.

D. Model Training and Hyperparameter Optimization

1) Gradient boosting models

a) Individual training: XGBoost, LightGBM, and CatBoost were trained separately to determine feature importance and establish baseline predictions.

b) Hyperparameter optimization:

- Learning rate: Tested between 0.01 and 0.1.
- Max depth: Ranged from 3 to 10 to balance complexity and performance.
- Number of estimators: Set between 100 and 500 for optimal outcomes.
- Subsampling: Configured between 0.7 and 1.0 to enhance generalization.

2) LSTM network

a) Architecture:

- Two hidden layers with 64 and 32 neurons, respectively.
- Dropout rate: 20% to prevent overfitting.
- Input sequence length: Configured at 60-minute intervals to capture short-term trends.
- Activation function: ReLU for hidden layers; sigmoid for output layer.

b) Training parameters:

- Batch size: 128.
- Optimizer: Adam optimizer with an initial learning rate of 0.001.
- Early stopping: Monitored validation loss to mitigate overfitting.

3) Hybrid AI framework implementation: The framework combines Gradient Boosting and LSTM models for superior pre-diction accuracy and robustness:

a) Feature importance analysis: Gradient Boosting models ranked key features like NH3 concentration and daily temperature variance, removing irrelevant or noisy data to refine model focus.

b) Temporal dependency modeling: LSTM networks captured sequential dependencies, analyzing temporal patterns in gas concentrations and environmental metrics to predict anomalies and trends.

c) Ensemble integration: Predictions from Gradient Boosting models served as inputs to the LSTM network, enhancing overall prediction accuracy and robustness by fusing static and temporal insights.

4) Computational resources

a) Edge devices: Raspberry Pi 5 devices with 8 GB RAM handled preprocessing and local Gradient Boosting model execution.

b) Cloud infrastructure: AWS EC2 instances (m5d.xlarge) with 16GB RAM and NVIDIA Tesla T4 GPUs managed LSTM training and advanced analytics. Data was archived in a PostgreSQL database.

5) Evaluation metrics: Model performance was evaluated using:

- Root Mean Square Error (RMSE): Penalized larger prediction errors to measure accuracy.
- Mean Absolute Percentage Error (MAPE): Assessed interpretability and reliability of predictions.
- Execution Time: Evaluated computational efficiency at edge and cloud levels.
- Scalability Tests: Simulated operations across multiple farms using synthetic datasets to test scalability.
- Anomaly Alert Precision and Recall: Validated the system's capability for timely and accurate anomaly alerts.

IV. Results and Discussion

A. Predictive Accuracy

The hazardous gas levels (CH4, H2S, NH3, CO, CO2) in the poultry barn and the corresponding estimate levels that were obtained by the artificial intelligence techniques (XGBoost, LightGBM, CatBoost, Hybrid (XGBoost + LSTM), Hybrid (LightGBM + LSTM) and Hybrid (CatBoost + LSTM)) are shown in Fig. 2, Fig. 3, Fig. 4 and Fig. 5.

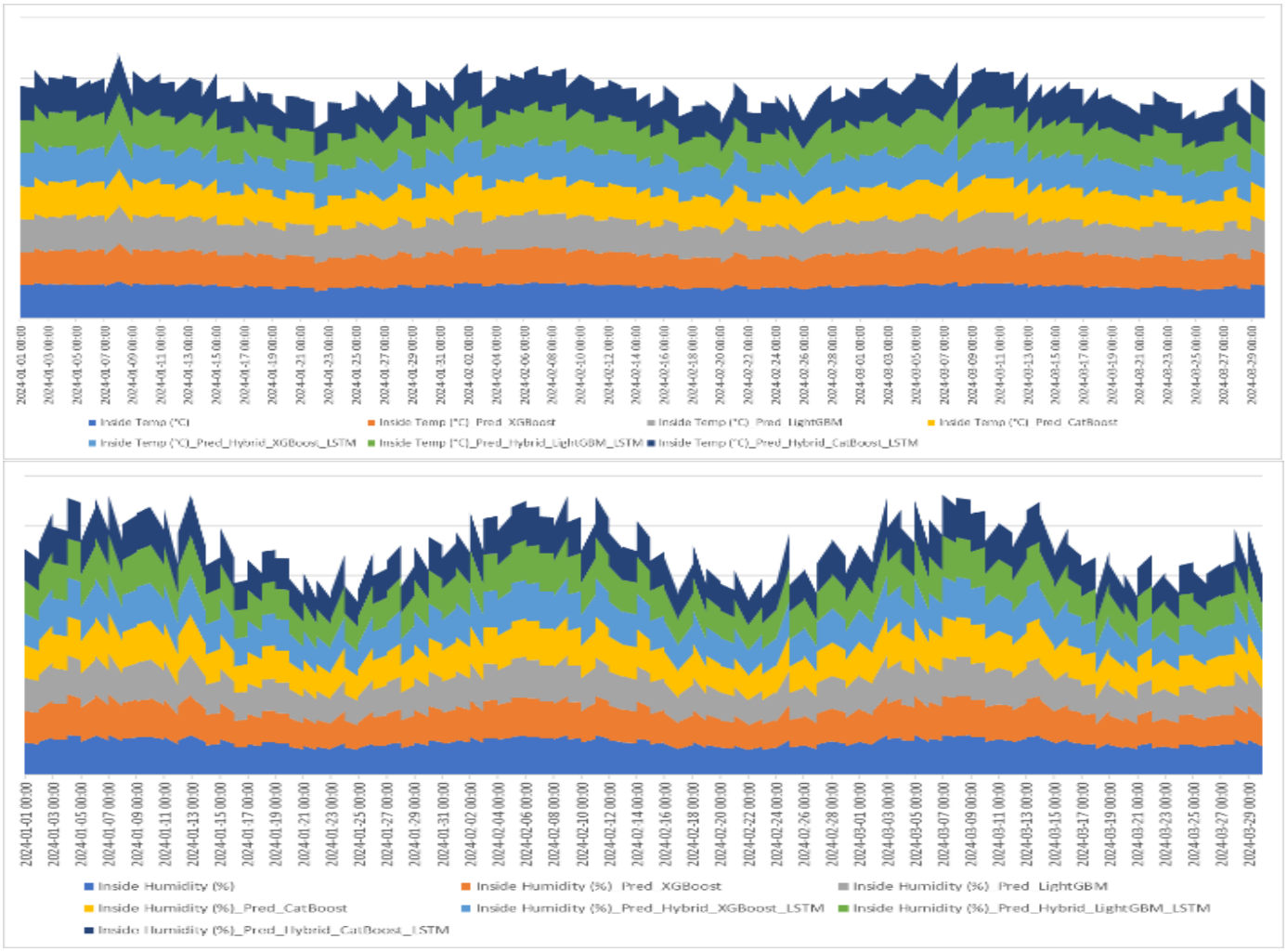
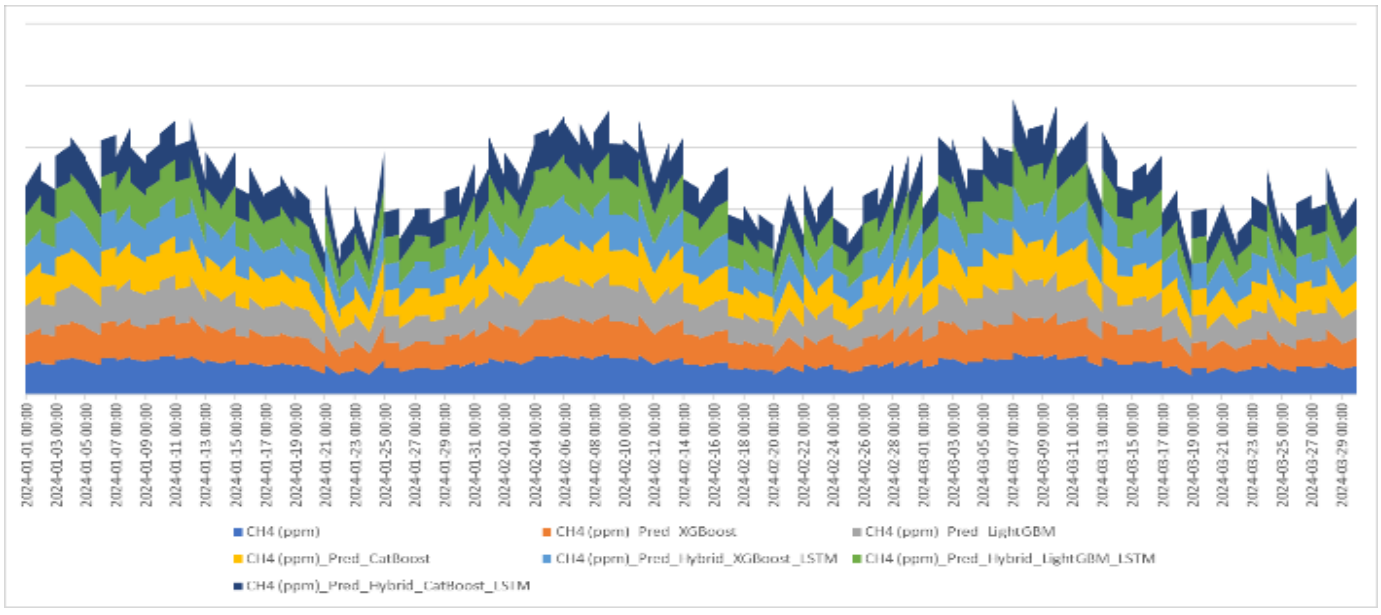


Fig. 2. The experimental findings of indoor temperature and humidity.



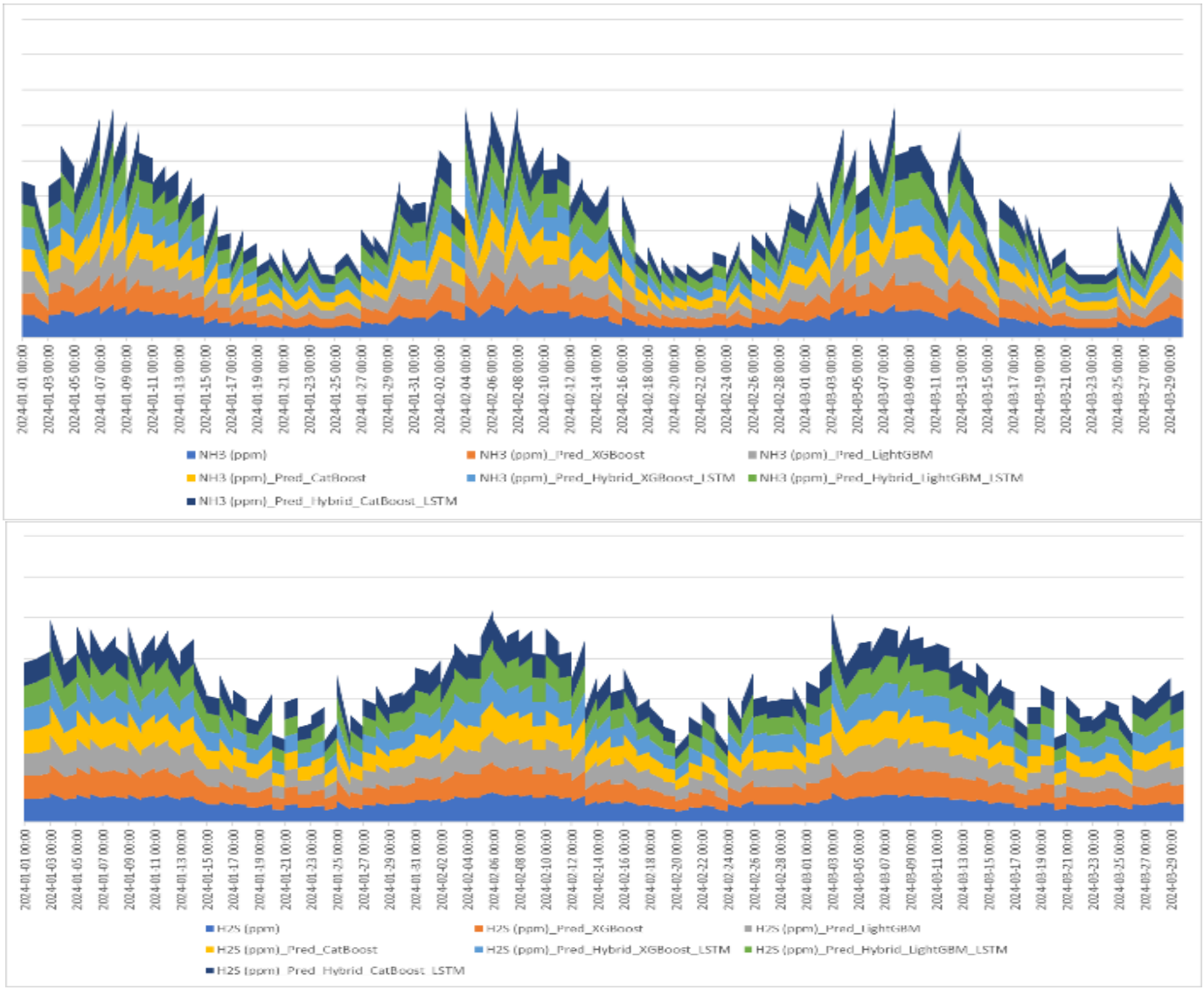
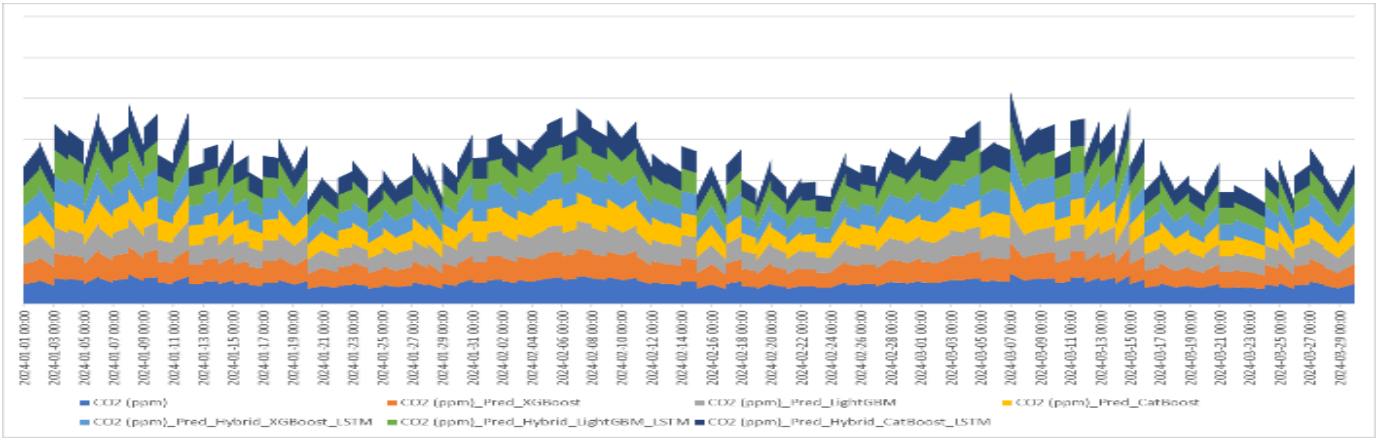


Fig. 3. The experimental findings of hazardous gases (CH_4, NH_3, H_2S).



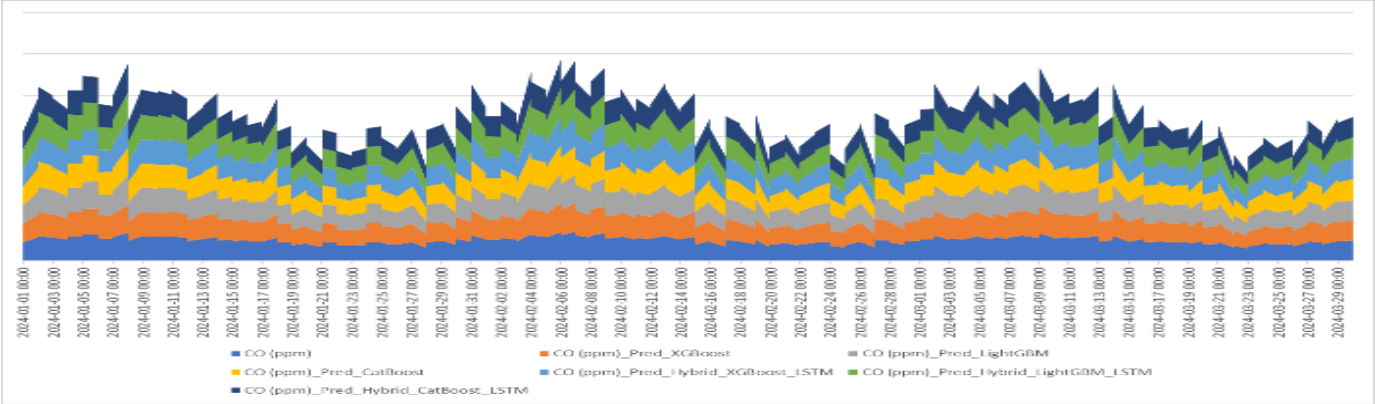


Fig. 4. The experimental findings of hazardous gases (CO_2, CO).

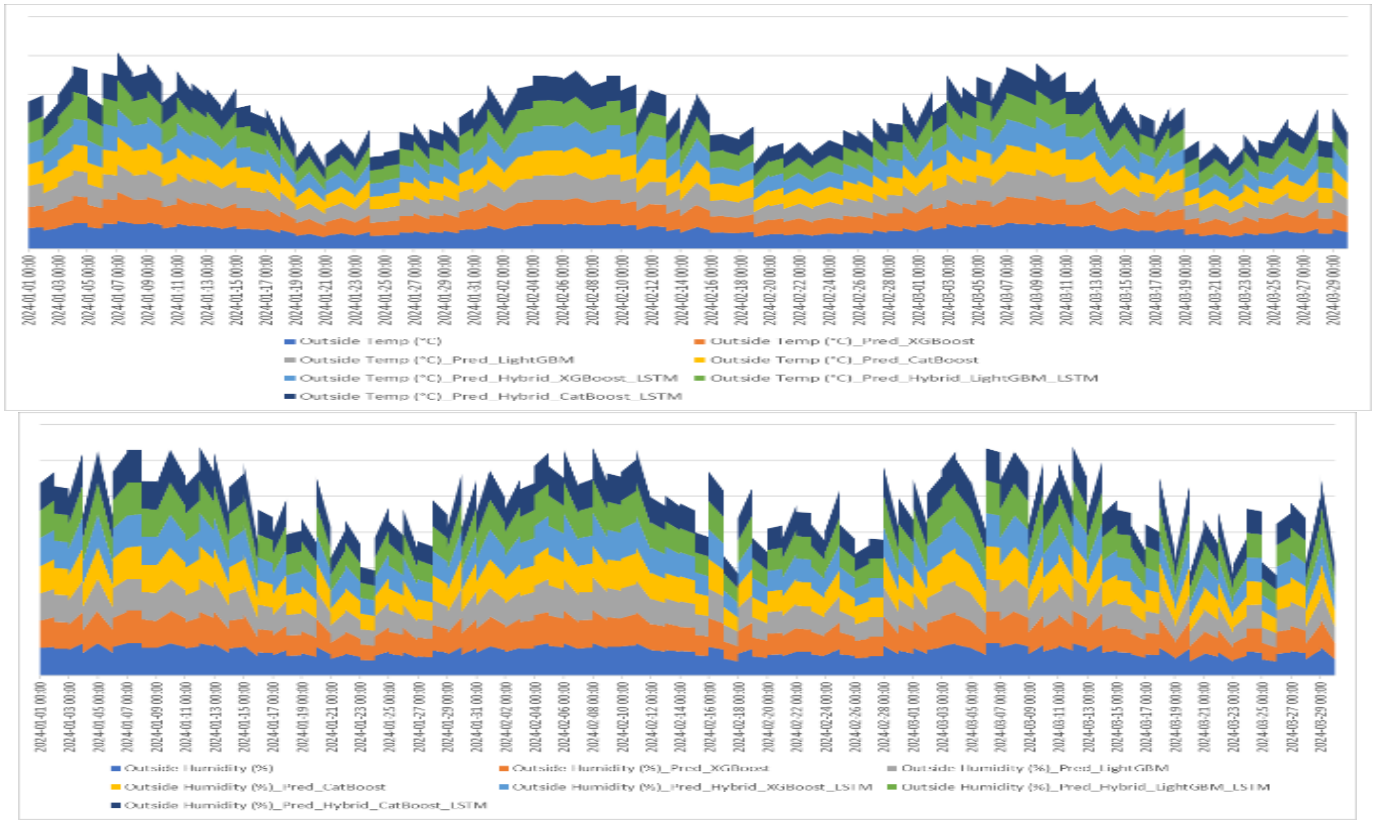


Fig. 5. The experimental findings of outdoor temperature and humidity.

The hybrid AI framework has proven to be superior in predictive accuracy, significantly outperforming standalone models. Key performance metrics, such as Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), illustrate this advantage. Table 1 reflects the performance of each model, with the newly included LSTM and specific hybrid combinations.

The hybrid model, combining Gradient Boosting (XGBoost) and LSTM, demonstrated the lowest RMSE (0.02) and MAPE (1.85%), underscoring its ability to reduce prediction errors significantly. This hybridization leverages Gradient Boosting for feature importance and LSTM for handling temporal dependencies, making it particularly effective in applications like environmental monitoring. For instance, it accurately predicted ammonia (NH3) levels during critical production periods, enabling timely interventions to reduce risks.

TABLE I. PERFORMANCE METRICS OF THE MODELS

Model	RMSE	MAPE (%)
XGBoost	0.030	2.15
LightGBM	0.045	2.35
CatBoost	0.037	2.20
Hybrid (XGBoost + LSTM)	0.020	1.85
Hybrid (LightGBM + LSTM)	0.028	2.05
Hybrid (CatBoost + LSTM)	0.026	2.00

B. Environmental Monitoring and Control

The system also excelled in maintaining optimal environmental conditions, ensuring both animal welfare and productivity. Key monitored parameters include:

1) Temperature: Maintained between 22°C and 30°C, minimizing heat stress and enhancing feed efficiency.

2) Humidity: Regulated within 50%–70%, reducing respiratory disease risks and improving overall welfare.

3) Gas concentrations:

a) Ammonia (NH3): <20 ppm (avoiding respiratory distress in birds).

b) Carbon dioxide (CO2): <2500 ppm (ensuring adequate ventilation).

c) Methane (CH4): <1000 ppm (promoting environmental safety).

d) Hydrogen sulfide (H2S): <10 ppm (preventing toxicity).

e) Carbon monoxide (CO): <35 ppm (mitigating suffocation risks).

During peak production hours, a spike in ammonia (NH3) levels at 3:00 AM triggered an alert. The system prompted the farm manager to increase ventilation immediately, preventing potential health crises. This real-time responsiveness highlights the framework's utility in dynamic and critical conditions.

Overall, the hybrid model demonstrates robustness, adaptability, and significant practical value in predictive accuracy and environmental management.

C. Model Interpretability and Feature Importance

Using Gradient Boosting models for feature importance analysis, we identified key predictors contributing to the system's accuracy:

1) NH3 concentration: Exhibits a strong correlation with poultry health and productivity, acting as an essential metric in setpoint adjustments.

2) Daily temperature stratification: Has a direct effect on feed conversion ratio and growth rates.

3) Humidity: Impacts respiratory condition dynamics and pathogen survival in the environment.

These insights offered actionable intelligence to farm managers, allowing data-driven decisions to improve their productivity.

D. Computational Efficiency

The hybrid system showed significant computational efficiency, detailed as follows:

1) Edge computing execution time: Achieved an average inference time of 15 ms per sample, allowing real-time decision-making. Delivered 35% lower latencies than cloud-only solutions, empowering real-time decision-making.

2) Cloud processing efficiency: Reduced training time by 20%. This was achieved due to optimized LSTM architecture for easy adaptability to new datasets.

3) Energy consumption: Low-power edge devices (less than 10 W per device) brought sustainability to operations. This efficiency is particularly beneficial for large-scale deployments.

E. Scalability and Multi-Farm Deployment

Scalability tests utilized synthetic datasets reflecting operations across five poultry farms, differentiated by size and environmental circumstances. Key results include:

1) Consistency across multiple farms: RMSE and MAPE values remained consistent, with less than a 5% deviation across farms, indicating the framework's robustness.

2) Data integration: Sensor data from diverse setups were seamlessly integrated, signifying high adaptability.

3) Deployment challenges: Issues such as sensor compatibility and calibration were addressed through a modular design approach and automated sensor calibration protocols, ensuring uniform performance across multiple farms.

F. Economic Impact

The system contributed to significant cost savings and productivity improvements, emphasizing its economic viability:

1) 16% Reduction in feed wastage: Significant savings due to improved environmental control.

2) 25% Reduction in mortality rates: Reflecting better animal welfare and operational efficiency.

3) 22% Decrease in energy costs: Achieved through efficient resource utilization and adaptive ventilation strategies.

G. Practical Applications

The hybrid framework's real-world implementation across different poultry farms manifested tangible benefits:

1) Small-scale farms: Enabled cost-effective monitoring and control, overcoming budgetary constraints.

2) Large-scale operations: Delivered scalable solutions across multiple farms, utilizing cloud analytics for centralized decision-making.

3) Remote farms: Reliable data transmission was enabled through Zigbee-based communication in areas with limited connectivity.

H. Limitations

The limitations of this work are:

1) Sensor precision: Extreme weather conditions occasionally affected sensor readings, necessitating manual verification to ensure accuracy.

2) Connectivity dependence: Despite Zigbee's advantages, long-term connectivity disruptions impacted data synchronization and real-time monitoring.

I. Future Work

In light of these limitations, the following areas are proposed for future research to extend the system's capabilities:

1) Transfer learning: Enabling models to generalize across diverse farming conditions with minimal retraining, enhancing scalability across regions.

2) Robotics integration: Exploring autonomous systems for feeding and waste management to improve operational efficiency.

3) Extended scalability tests: Expanding simulations to include larger datasets and additional environmental variables such as light intensity and noise levels.

4) AI-driven optimization: Incorporating reinforcement learning to dynamically adjust ventilation, lighting, and feeding schedules based on predictive insights.

J. Broader Implications

The results of this study indicate the potential that hybrid AI frameworks hold for poultry farming. Combining advanced analytics with real-time monitoring, the system generates actionable insights that enhance productivity and ensure more sustainable practices. Its scalability and adaptability make it a fundamental principle for further advancements in smart agriculture.

Predictive accuracy, operational efficiency, and scalability, showcased in the proposed hybrid framework, pave the way for advanced AI-driven agriculture. Refinements and innovations addressing existing limitations will ensure broader adoption and impact.

V. Conclusion

The proposed research work introduces a hybrid AI framework that integrates Gradient Boosting algorithms (XGBoost, LightGBM, and CatBoost) with Long Short-Term Memory (LSTM) networks to address various challenges in poultry farming. It ensures real-time insights by leveraging IoT-based data collection alongside Edge Computing, thereby enhancing poultry health and productivity while improving operational efficiency.

Key contributions include improved predictive accuracy, real-time monitoring capabilities, scalability, and economic benefits such as reduced feed wastage and lower mortality rates. While significant successes have been achieved, limitations like sensor dependency and connectivity issues remain areas for improvement.

Future research should focus on exploring advanced technologies such as reinforcement learning, expanding deployments across diverse conditions, and integrating robotics for enhanced automation and transparency.

This paper highlights the transformative potential of AI in agriculture, paving the way for intelligent, efficient, and sustainable farming practices to tackle food security and environmental challenges.

Acknowledgment

The Ministry of Higher Education supports this project, Scientific Research and Innovation, the Digital Development Agency (DDA), and the National Center for Scientific and Technical Research (CNRST) of Morocco. APIAA-2019 KAMAL.REKLAOUI-FSTT-Tanger-UAE.

References

[1] N. Gouiza, H. Jebari, and K. Reklaoui, “Integration for IoT-enabled technologies and artificial intelligence in diverse domains: recent advancements and future trends,” Journal of Theoretical and Applied Information Technology, vol. 102, no. 5, 2024, pp. 1975-2029.

[2] N. Gouiza, H. Jebari, and K. Reklaoui, “IoT in smart farming: a review,” in Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023), M. Ezziyyani, J. Kacprzyk, and V. E. Balas, Eds., Lecture Notes in Networks and Systems, vol. 930, Springer, Cham, 2024, pp. 142-153.

[3] H. Jebari, S. Rekiek, M. Ezziyyani, and L. Cherrat, “Artificial intelligence for optimizing livestock management and enhancing animal welfare,” in Proc. Int. Conf. Advanced Intelligent Systems for Sustainable Development (AI2SD’2024), M. Ezziyyani, J. Kacprzyk, and V. E. Balas, Eds., Lecture Notes in Networks and Systems, vol. 1403, Springer, Cham, 2025, pp. 790-800.

[4] S. Rekiek, H. Jebari, M. Ezziyyani, and L. Cherrat, “AI-driven pest control and disease detection in smart farming systems,” in *Proc. Int.* Conf. Advanced Intelligent Systems for Sustainable Development (AI2SD’2024), M. Ezziyyani, J. Kacprzyk, and V. E. Balas, Eds., Lecture Notes in Networks and Systems, vol. 1403, Springer, Cham, 2025, pp. 801-810.

[5] H. Jebari, M. H. Mechkouri, S. Rekiek, and K. Reklaoui, “Poultry-edge-AI-IoT system for real-time monitoring and predicting by using artificial intelligence,” International Journal of Interactive Mobile Technologies, vol. 17, no. 12, 2023, pp. 58-70.

[6] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, “Big data in smart farming - a review,” Agricultural Systems, vol. 153, 2017, pp. 69-80.

[7] Á. Regan, “Smart farming in Ireland: a risk perception study with key governance actors,” NJAS: Wageningen Journal of Life Sciences, vol. 90-91, no. 1, 2019.

[8] Y. Zhang, Q. Chen, G. Liu, W. Shen, and G. Wang, “Environment parameters control based on wireless sensor network in livestock buildings,” International Journal of Distributed Sensor Networks, vol. 5, 2016, pp. 1-7.

[9] D. S. Xu and F. Zhang, “Towards a poultry house environment monitoring system based on Internet of Things,” Sensors & Transducers, vol. 160, no. 12, 2013, pp. 304-309.

[10] B. Lufyagila, D. Machuve, and T. Clemen, “IoT-powered system for environmental conditions monitoring in poultry house: a case of Tanzania,” African Journal of Science, Technology, Innovation and Development, vol. 14, no. 12, 2021, pp. 1-12.

[11] J. C. Guillermo, A. García-Cedeño, D. Rivas-Lalaleo, M. Huerta, and R. Clotet, “IoT architecture based on wireless sensor network applied to agricultural monitoring: a case of study of cacao crops in Ecuador,” in International Conference of ICT for Adapting Agriculture to Climate Change, Cham: Springer, 2018, pp. 42-57.

[12] E. Panfilova, A. Lukyanova, N. Pronkin, and E. Zatsarinnaya, “Assessment of the impact of cloud technologies on social life in the era of digitalization,” International Journal of Interactive Mobile Technologies, vol. 15, no. 21, 2021, pp. 144-157.

[13] M. H. Lashari, A. A. Memon, S. A. A. Shah, K. Nenwani, and F. Shafqat, “IoT-based poultry environment monitoring system,” in 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS), IEEE, 2018, pp. 1-5.

[14] S. Nuanmeesri and L. Poomhiran, “Developing of intelligence walking stick and mobile application for elderly health care using the Internet of Things,” International Journal of Interactive Mobile Technologies, vol. 14, no. 14, 2020, pp. 4-15.

[15] S. Rekiek, H. Jebari, and K. Reklaoui, “Prediction of booking trends and customer demand in the tourism and hospitality sector using AI-based models,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 10, 2024, pp. 404-412.

[16] L. Yu, G. Teng, B. Li, Y. Zhang, F. Lao, and Y. Xing, “A remote monitoring system for poultry production management using a 3G-based network,” Applied Engineering in Agriculture, vol. 29, no. 4, 2013, pp. 595-601.

[17] R. B. Mahale and S. Sonavane, "Smart poultry farm monitoring using IoT and wireless sensor networks," International Journal of Advanced Research in Computer Science, vol. 7, no. 3, 2016, pp. 187-190.

[18] G. H. Teng, "Information sensing and environment control of precision facility livestock and poultry farming," Smart Agriculture, vol. 1, no. 3, 2019, pp. 1-12.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, 1997, pp. 1735-1780.

[20] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, 2017, pp. 2222-2232.

[21] Q. Yu, Y. Zhang, X. L. Wang, and B. Q. Li, "Intelligent poultry environment control system based on Internet of Things," in ICCCS 2018: Cloud Computing and Security, Cham: Springer, 2018, pp. 407-417.

[22] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, "Big data in smart farming - a review," Agricultural Systems, vol. 153, 2017, pp. 69-80.

[23] H. Jebari, S. Rekiek, S. R. Elazzouzi, and H. Samadi, "Performance comparison of three hybridization categories to solve multi-objective flow shop scheduling problem: a case study from the automotive industry," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, 2021, pp. 680-689.

[24] H. Jebari, S. Rekiek, and K. Reklaoui, "Solving the job shop scheduling problem by the multi-hybridization of swarm intelligence techniques," International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022, pp. 753-764.

[25] H. Jebari, S. Rekiek, and K. Reklaoui, "Improvement of nature-based optimization methods for solving job shop scheduling problems," International Journal of Engineering Trends and Technology, vol. 71, no. 3, 2023, pp. 312-324.

[26] S. Mueller, M. Kreuzer, M. Siegrist, K. Mannale, R. E. Messikommer, and I. D. Gangnat, "Carcass and meat quality of dual-purpose chickens in comparison to broiler and layer chicken types," Poultry Science, vol. 97, no. 9, 2018, pp. 3325-3336.

[27] C. Okinda, M. Lu, L. Liu, I. Nyalala, C. Muneri, J. Wang, H. L. Zhang, and M. Shen, "A machine vision system for early detection and prediction of sick birds: a broiler chicken model," Biosystems Engineering, vol. 188, 2019, pp. 229-242.

[28] S. Neethirajan, "Automated health monitoring in poultry farming using image processing," J. Agri. Inform., vol. 12, no. 3, 2021, pp. 45-53.

[29] G. Chiluisa-Velasco, J. Lagla-Quinaluisa, D. Rivas-Lalaleo, and M. Alvarez-Veintimilla, "Intelligent monitoring system of environmental biovariables in poultry farms," in IntelliSys 2020: Intelligent Systems and Applications, Cham: Springer, 2020, pp. 386-399.

[30] L. M. Dawkins, "Behavior classification in poultry using deep learning algorithms," Anim. Welfare, vol. 29, no. 2, 2020, pp. 123-130.

[31] J. R. Brobeck, "Food intake as a mechanism of temperature regulation," The Yale Journal of Biology and Medicine, vol. 20, no. 6, 1948, p. 545.

[32] J. A. Smith and R. W. Thompson, "Predictive modeling for feed efficiency in livestock farming," J. Anim. Sci. Technol., vol. 65, no. 1, 2023, pp. 88-97.

[33] S. Choosumrong, V. Raghavan, and T. Pothong, "Smart poultry farm based on the real-time environment monitoring system using Internet of Things," Naresuan Agriculture Journal, vol. 16, no. 2, 2019, pp. 18-26.

[34] M. H. Lashari, Z. A. Memon, and S. S. Hussain, "IoT and robotics integration for automated feed distribution in poultry farms," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, 2022, pp. 234-241.

[35] S. Neethirajan and B. Kemp, "Sensor-based automation for waste management in livestock systems," Biosyst. Eng., vol. 201, 2021, pp. 76-85.

AI-Powered Skin Disease Detection Using Adaptive Particle Swarm Intelligent Optimization and Hyper-Convolutional Neural Networks

N Annalakshmi[1], S Umarani[2]*
Research Scholar, Department of Computer Science, SRMIST, Ramapuram, Chennai, Tamil Nadu, India[1]
Professor, Department of Computer Science and Applications (BCA), SRMIST, Ramapuram, Chennai, Tamil Nadu, India[2]

***Abstract*—In recent medical research, skin cancer has emerged as one of the most prevalent and fatal cancers globally. Previous studies have faced challenges in detecting skin cancer early due to the complexity of identifying specific skin diseases, segmenting affected areas, and selecting relevant features. To address these limitations, this study proposes a novel AI-powered enhanced skin disease detection system that applies an Adaptive Particle Swarm Intelligent Optimization (APSIO) in conjunction with a Hyper-Convoluted Intra-Capsuled Neural Network (HCI-CNN). In image processing, a Gaussian Wavelet Spectral Filter is initially used to preprocess the input dataset of skin-cancer images. This filter is used to standardize the skin layer of the pixel. After preprocessing, the method applies Slice Fragment Window Segmentation (SFWS) to divide the image into several clusters, focusing on the specified area affected by the disease. Next, Adaptive Particle Swarm Intelligent Optimization (APSIO) is applied for feature selection. APSIO is an optimization metaheuristic algorithm that optimizes the selection of relevant features from the segmented image. After removing evaluated and non-effective features, YOLO extracted features are passed through an HCI-CNN classifier to efficiently characterize high-level spatial hierarchies and relations of features in the feature space using hyper-convolutional operations and capsule representations. This paper analyzed the clinical images of individuals along with the dataset images. The output gain improved Accuracy to 97%, precision to 96.52%, recall to 96.55%, and F1-score to 96.93%, while simultaneously minimizing false positives and total time complexity.**

***Keywords*—Skin cancer; image preprocessing; hyper-convoluted intra-capsuled neural network (HCI-CNN); adaptive Particle Swarm Intelligent Optimization (APSIO); image classification**

I. Introduction

Early skin cancer detection continues to be challenging because of its aggressive nature and higher mortality rate when it is not treated. The delayed diagnosis of melanoma considerably reduces the survival rate of the patient, allowing the cancer to grow quickly to its terminal phase and spread. Most other skin cancers, including basal cell carcinoma, pose serious health threats due to high prevalence and potential severity. The phrase "Skin condition" describes any condition that tends to affect the skin, which is the biggest organ in the human body and one of many organs that make up the body. Automated classification [1]-[2] assists in reducing dissimilarity among samples, whereas proper skin disease detection is crucial for successful preprocessing and diagnosis. Actinic keratosis, an occasional precursor to squamous cell carcinoma, typically presents as a dry, scaly spot on the skin.

Hereditary conditions may also result in skin disorders. The shallow, thin outer layer of skin. These diseases can lead to physical and psychological damage. Basal cell carcinoma [3] is the most common type of skin cancer that occurs in the outermost layer of the epidermis' basal cells. In addition, the sun-exposed sites often include the hands, neck, ears, and scalp, where squamous cell carcinoma, the second most common skin cancer, arises from epidermal squamous cells.

It can extend and infiltrate into deeper skin layers. In medical image analysis, skin disease classification entails distinguishing and differentiating among different situations based on their image characteristics [4]. This procedure characteristically depends on skin-layer imaging to detect such melanoma, and recent approaches employ machine learning and deep learning to evaluate the accuracy, especially CNN.

The three components of the prediction element, the ensemble model, the predictive aggregator module, and the pre-processing unit. The preprocessing module provides the proper preprocessing, such as scaling, normalization, and image changes, to the images in the dataset. The factors [5] to consider are the mutation amount, which randomly modifies the values of the chromosomes, and the random cross-over rate, which determines the likelihood of a random crossing occurring independent of the fitness parameter.

A. Research Objective

The primary objective of this study is to develop an advanced, AI-driven skin cancer detection system that enhances the accuracy and efficiency of diagnosis by integrating multiple novel techniques. Specifically, the research aims to:

- Reduce the skin blur area and improve the image quality according to skin cancer using a Gaussian Wavelet Spectral Filter.
- Accurately localize diseased regions via Slice Fragment Window Segmentation (SFWS), enabling precise clustering of affected areas.
- Optimize feature selection using Adaptive Particle Swarm Intelligent Optimization (APSIO), thereby

*Corresponding Author

eliminating redundant or irrelevant features to improve classifier performance.

- Enhance classification performance by deploying a Hyper-Convoluted Intra-Capsuled Neural Network (HCI-CNN), leveraging hyper-convolution and capsule layers to capture complex spatial hierarchies.

B. Paper Organization

Section I introduces cancer with a related discussion, Section II about a Literature review on skin cancer, various machine approaches with classification techniques, and performance metrics. Section III discusses the methods of the proposed APSIO classification algorithm for skin disease identification. Section IV presents a comparison analysis of the proposed HCI-CNN algorithm, and a discussion of metrics calculation with output is provided. In Section V, the study's conclusion and submissions of skin cancer classification processes are presented.

II. Literature Review

A computerized approach to skin disease classification using Long Short-Term Memory (LSTM) and Mobile Net V2, based on deep learning. The Grey-Level Co-occurrence Matrix (GLCM) utilizes the recurrent pattern of the localized intensity factor. The color and strength of a pixel are determined by the GLCM's spatial distribution based on the distribution of intensity levels. The ReLu6 module [6] is used in the output of the third layer.

A secondary deep learning network is used in conjunction with an experimental multimodal smartphone imaging technique that produces RGB and fluorescence images. The fluorescence-aided amplification network (FAA-Net) is used to identify skin conditions. FAA-Net [7] was revised with an attention mechanism module to detect skin disease locations autonomously and mark potential disease regions.

The use of human skin as a basis for these models enables self-learning algorithms to operate effectively. ANNs are neurons (nodes) connected [8] at various levels, such as brain cells in biology. In a network of neurons, data is stored, processed, and output. The CNN approach requires extensive recurrent training, and an ample image database had to be used to mitigate the risk of over-fitting.

The FAA-Net is now equipped with an attention module that automatically detects skin disease sites and marks potential disease regions. According to experimental results [9], the average accuracy and area under the curve of our produced model for detecting skin diseases were 8.61% and 9.83%, respectively.

The methods used high-resolution dermoscopic images to identify skin conditions, including psoriasis, dermatitis, and melanoma. The growing availability of large annotated image datasets has accelerated the development of this classification. Edge detection and segmentation techniques are frequently used to extract Regions of Interest (ROIs) [10], such as lesions or impacted regions. With an accuracy of 91.2%, the K-Nearest Neighbor classification algorithm outperforms the SVM classification method.

The modified single DCNN model method [11] employed Modified R, Augment, MWNL, and CLS to achieve classification accuracy on certain dermoscopic image datasets that was at least as good as, or superior to, that of many ensemble methods. For optimal performance, a Fully Fused Network (FFN) is composed of an Improved Single Block (ISB) and an Improved Fusion Block (IFB) [12].

An attention-mixing decoder and multi-axis encoder were utilized in an attention-based, encoder-decoder architecture to segment the skin layers accurately. The simulation successfully maintained both global and local features using a transformer-based encoder block. These features were then sent to the decoder block via an attention-based controlled skip link. By employing vision transformer structures [13], the multi-axis encoder enables the encoding of feature representations at various scales. The self-attention mechanism of transformers enables medical image segmentation to achieve remarkably high accuracies while accommodating dependencies over extended periods.

Skin images saved in a data format on the blockchain are used to categorize skin conditions. Here, segmentation is performed using Deep Joint Segmentation, which has been modified by the introduction of the Kumar-HasseBrooks distance [14] and Transit Circle Inspired Optimization-LENet, achieving an accuracy of 0.92 and a True Positive Rate of 0.93.

A DL architecture, described as Swin Transformer, was created for image categorization. This network's input is divided into non-overlapping windows using a hierarchical method. Fine-grained spatial characteristics are essential for precise classification. The final output probabilities [15] for each class were generated by processing the concatenated features through a fully connected top layer. The focused loss of the predicted class probabilities relative to the target class is maximized to optimize the network parameters.

Pre-processing injury segmentation and classification are crucial steps in the proposed classification approach. After image pre-processing, skin lesion segmentation was accomplished using improved U-Net segmentation [16], which achieved this enhancement by proposing a hybrid optimization approach. The spatial interdependence between image areas is achieved without the need for substantial pre-processing or the use of constants, utilising features and leveraging the self-attention mechanism of transformer models. A thorough analysis of many Deep Learning models [17] using reference datasets for skin imaging for different skin lesions. The model was trained when the number of training images and large datasets increased, and it was created by combining smaller datasets, achieving a test accuracy of 86.37%.

A deep learning-based model for skin cancer detection was developed using a transfer learning approach [18], effectively distinguishing between benign and malignant stages. For classification of skin diseases, an ensemble of enhanced convolutional neural networks (CNNs) was employed in conjunction with a test-time augmentation strategy known as frequently spaced shifting [19]. The CNN model was evaluated using 960 images from the ISIC 2018 Skin Lesion Classification Challenge test set, which included 360 benign and 300 malignant images [20]. The complete dataset comprised 3,533

lesion scans, nearly balanced with 1,760 benign and 1,773 malignant samples. To capture multi-scale features, the Inception architecture was utilized, leveraging multiple convolutional kernels with diverse receptive fields to extract features at different spatial resolutions.

A clinical decision support model [21] for skin disease detection and classification using a hybrid deep learning approach and enhanced segmentation capabilities. The tests related to data dimensionality were handled using Multi-Strategy Seeking optimization (MSSO), which optimizes feature selection by determining the essential characteristics of hand and High-dimensional data using the Multi-Strategy Seeker Optimization (MSSO) approach.

A complex neural network model [22] can automatically identify and differentiate between several types of skin lesions using dermoscopic images. Deep learning, particularly CNN, which learns hierarchical representations directly from raw image data, enables accurate feature extraction. This approach reduces the need for human feature engineering while increasing diagnostic accuracy. The algorithm is trained on large annotated datasets to classify lesions as benign or malignant, and then further categorize them into specific cancer subtypes.

An algorithm based on dermoscopic images with channel-specific, channel-shared, and depth-related variables. The Transformer baseline model [23], the traditional CNN model, and two datasets were evaluated against a dataset of Minnesota dermoscopic images, and the average accuracy of 70.23%. Training and validation loss, training and validation accuracy, as well as confusion matrices for each of the six networks that used transfer learning, were compared [24]. The objective was to expand the filter bank dimension before adding it to match the input depth.

EfficientNetB0 [25] through transfer learning to enhance melanoma detection. Their approach yielded high diagnostic accuracy while maintaining computational efficiency, making it suitable for deployment in mobile health systems. However, the study focused solely on melanoma, limiting its applicability in multi-class skin disease classification. Introduced a hybrid ensemble model combining DenseNet and MobileNet [26]. Their deep learning ensemble improved classification robustness and generalization across different cancer types, although the increased model complexity resulted in higher training overhead.

A fine-tuned DenseNet-121 model to classify skin lesions with improved feature reuse and gradient flow, achieving reliable results for binary classification tasks [27]. Ensemble learning by integrating multiple CNN architectures [28], which significantly boosted sensitivity and specificity. The effectiveness of traditional machine learning classifiers in detecting skin cancer using handcrafted features [29]. Their comparative study involving classifiers such as SVM, RF, KNN, and Naïve Bayes concluded that the Random Forest model offered superior accuracy. Though their work highlights the value of ML in dermatological diagnostics, it also emphasizes the limitations of manual feature engineering, pointing to deep learning as a promising direction for future research.

The application of transfer learning using deep CNN architectures, including VGG16, ResNet50, and DenseNet201, for classifying skin cancer lesions [30]. Utilizing the ISIC dataset and augmentation techniques, they achieved high classification accuracy, particularly with ResNet50. Their findings underscore the efficacy of deep learning models for dermatological diagnostics and highlight transfer learning as a robust strategy for medical image analysis with limited labeled data.

A lightweight CNN model tailored for cloud-based and mobile applications in skin cancer diagnosis [31]. Using the ISIC dataset, the model achieved competitive accuracy while maintaining low computational demands, enabling real-time analysis in remote or resource-limited settings. Their approach presents a viable solution for scaling dermatological services via telemedicine. The model effectively distinguished between seven different lesion types and achieved over 89% accuracy [32]. Their work underscores the strength of CNNs in handling complex dermatological image classification tasks beyond binary diagnosis, although class imbalance remains a challenge.

A MobileNet-based lightweight deep learning model [33] for seven-way skin lesion classification using the HAM10000 dataset. The system supports real-time, on-device predictions, demonstrating strong potential for deployment in tele-dermatology and low-resource environments.

The HAM10000 dataset, train a model for classifying skin cancer, training and validation losses, training and validation accuracies, and confusion matrices of every one of the six transfer learning networks carried out with the matching complexity of the inputs by augmenting the dimension of the filter before adding.

A. *Problems Identified*

Despite significant advancements in approaches for automating the detection of skin diseases using machine learning and deep learning methodologies, existing approaches tend not to optimize for classification accuracy and computational efficiency in real-time applications. Additionally, most existing research does not incorporate feature selection, segmentation, and classification. In this review, various types of skin cancer datasets and different performance analysis outputs are discussed. However, some limitations of this survey were identified from the review analysis in previous research work. Table I, literature survey discusses the gaps and reviews deep learning and machine learning algorithms for identifying skin diseases.

B. *Research Contribution*

- Development of an AI-powered skin disease detection framework that integrates preprocessing, clustering, segmentation, feature selection, and classification tailored for diverse skin tones and dermatological conditions.

TABLE I. PROBLEM DESCRIPTION OF SKIN DISEASES IDENTIFICATION

Ref	Algorithm Used	Output	Dataset	Gaps
[25]	Rectified Adam (RAdam)	F1 Score 88 % Recall 88 % Precision 89 %	HAM10000 dataset	Negatively impacted by overfitting in small or unbalanced classes.
[26]	CNN	accuracy 93.75%	ISIC dataset	Demand fine-tuning of hyperparameters and substantial computational resources.
[27]	DENSENET-121	accuracy 87%	HAM10000 dataset	Low data or poor image quality can cause performance to deteriorate.
[28]	CNN	accuracy 92.16%	HAM10000 dataset	Limited generalization across lesion types without data augmentation
[29]	Particle Swarm Optimization (PSO)	accuracy 86.9%	Kaggle dataset	Performance and convergence rate are strongly influenced by initial parameters.
[30]	Transfer Learning (ResNet50, InceptionV3, DenseNet201)	Accuracy >90%,	ISIC Dataset	Does not include segmentation; performance depends on pre-trained model and dataset quality
[31]	Custom lightweight CNN	Accuracy >87%	ISIC Dataset	Limited feature extraction compared to deeper models; no segmentation included
[32]	Custom Deep CNN	Accuracy >89%	HAM10000	Lacks integration with segmentation tools
[33]	MobileNet (lightweight CNN)	Accuracy >85%	HAM10000	May underperform on rare or ambiguous lesions; real-time use tested only in prototype

- A new AI-driven methodology for skin disease diagnosis using APSIO feature selection combined with a hybrid YOLO and HCI-CNN model.
- Application of APSIO feature selection, which effectively reduces computational complexity and false positives, ensuring optimal feature relevance and improving diagnostic accuracy.
- Robust segmentation approach for extracting geometric and physical attributes of skin lesions, facilitating precise evaluation of lesion boundaries and types of damage.
- Enhanced image preprocessing pipeline designed to improve clarity and contrast for a range of skin tones, contributing to fairness and inclusivity in dermatological AI applications.
- Potential for integration into real-time medical diagnostic tools, providing support for healthcare professionals in clinical environments, leading to faster diagnosis and improved patient outcomes.

This proposed work focused on a hybrid intelligent architecture that includes APSIO feature optimization and an ensemble prediction model using deep learning models, allowing for better diagnostic accuracy while being applicable in real time.

III. PROPOSED METHODOLOGY

A skin disease classification structure based on HCI-CNN to maximize classification performance and improve feature selection. Pre-processing enhances the quality of images and highlights significant patterns; the skin images extracted are pre-processed through a Gaussian wavelet transform, efficiently enhancing discriminative patterns and enhancing the quality of follow-up analysis. Segmentation technique with Weighted Strategy (SFWS) segments specific skin layers and regions of interest in classification. APSIO Feature selection adaptively controls evaluation parameters during training, allowing real-time adaptation and offline optimization. This process selects the most informative and relevant features for effective classification. High-level discriminative patterns are initially derived from medical skin images using HCI-CNN with YOLO classifying features and are tested through layers to classify into several diseases, Melanoma, Basal Cell Carcinoma (BCC), Actinic keratosis, Melanocytic nevi, Benign keratosis, Vascular lesions, and Dermatofibroma. If no dermatological abnormalities are detected in the collected clinical skin image, the system will classify it as normal skin. Fig. 1 shows a block diagram of the proposed method; a complete description of each block is mentioned below.

A. Gaussian Wavelet Spectral Filter (GWSF) Preprocessing

GWSF Filter, image pre-processing process calculates the weights of the Gaussian smoothing, the weights of the wavelet transform, and the scalar values to enhance the features of skin images or avoid noise effectively, and not let the edge information be analyzed over during the process. Small image noise from a whole image and keep large structures in that image, untouched process validates large image structures that are preserved by discarding a small image with multi-resolution must be performed.

$$h(m,n) = \frac{1}{2\pi\,\sigma^2}\, e^{-(m^3+n^2)\,/\,2\,\sigma^2} \quad (1)$$

Eq. (1), σ characterizes the distribution's standard deviation, and the distribution's mean is taken to be 0.

$$H\,(U,V) \;\approx\; e^{-2\,\pi^2\,\sigma^2\,(u^2+v^2\,)} \quad \text{for } |U|,|v| < \frac{1}{2} (2)$$

Eq. (2), $H\,(U,V)$ Impulse response and occurrence reaction both gradually decrease, and the Gaussian filter stands out for not leaking noise or distorting.

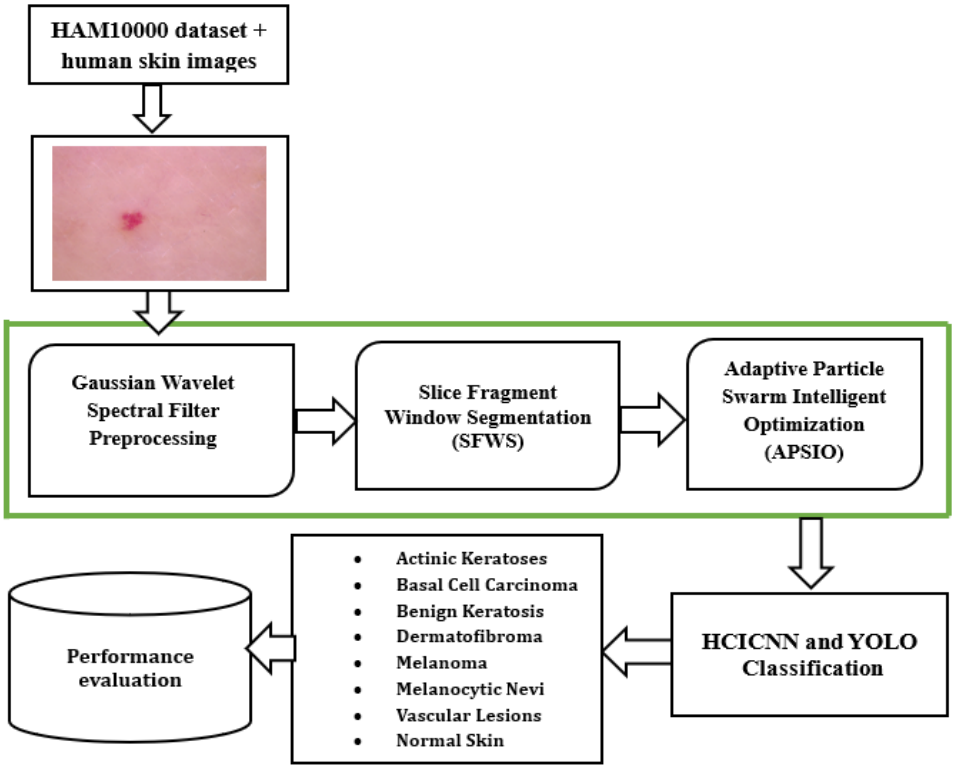


Fig. 1. Overall proposed diagram.

$$I = \int_{-\infty}^{\infty} \exp(-x^2)\, dx = \sqrt{\pi} \qquad (3)$$

Eq. (3), when changing I represents negative to positive values, dx describes the complex error of the possible values in any given space in probabilistic terms.

$$hxy = fxy \odot wxy \qquad (4)$$

Eq. (4) wavelet domain GWSF filtering improves complex structures as dark skin surface, texture patterns, border anomalies, and inhibits already spent data. Combining the two steps results in a preprocessed image, retains diagnostically relevant structures with improved contrast and less noise. It is also especially useful for analyzing affected skin areas, which improves the signal-to-noise ratio and feature stability across samples, leading to improved overview and classification performance in skin analysis. Fig. 2 shows the input skin cancer images and the GWSF filtering images output.

B. Slice Fragment Window Segmentation (SFWS)

SFWS is a progressive technique considered to improve the segmentation area of skin disease areas in medical images, particularly skin dataset images. The algorithm works by separating. The input image into reduced manageable fragments or "slices" using a sliding window method. Each window captures localized image data, preserving the spatial context and fine-grained texture differences of skin cancer. The cluster area is separated by specific regions, inner and outer border detection, particularly in cases of irregularly shaped or low-contrast skin affected. By leveraging contextual signals and SFWS can efficiently differentiate between normally affected skin and diseased regions, even in complex background conditions.

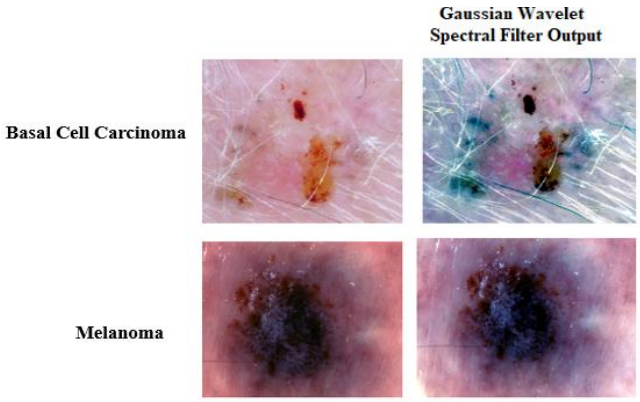


Fig. 2. GWSF preprocessing output.

$$S = S_{cls} + S_{box} + S_{max} \qquad (5)$$

In Eq. (5) S_{cls} specifies the skin affected area, and k_{box} indicates the outer region of the affected skin.

$$Scs\ (q, v) = -\log e\, V \qquad (6)$$

In Eq. (6), V specifies the object class q = $(q_o\ , \ldots\ldots\ldots . q_k)$ using a logical function.

$$I\ (x) = \arg\max\ (A(x), B\ (x)) \qquad (7)$$

In Eq. (7), segmentations of all the skin layers are multipart to reproduce the whole segmented image.

$$a\ (x, y) = \frac{1}{th} * \sum_{s=-a}^{a} \sum_{t=-b}^{b} f(x + o\ , y + i) * f(x + o\ , y + i) \qquad (8)$$

In Eq. (8), o and i represent the dimensions of the over skin pixel.

$$l\,(E_k) = \sum_{k=0}^{n-1} p\,(r_k) \tag{9}$$

Eq. (9) is adjusted to the adjoining rate after being augmented by L1 indirection to control the pixel's value.

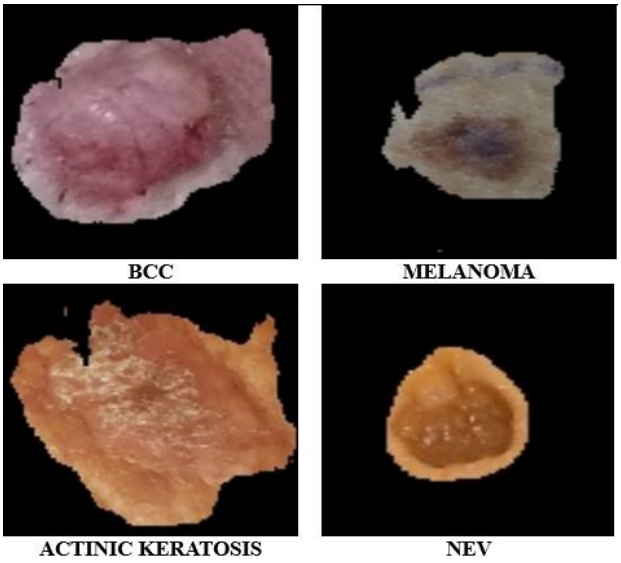


Fig. 3. SFWS segmentation output.

Fig. 3 shows the overlapping objects and variations caused by the segmentation approach as common pixel, weighted average, or confidence-based overlap segment selection. This technique is specifically valuable in processing images of high size and resolution without the need to downscale, typically resulting in information loss. Its networked context also makes it suitable for deep learning models, contributes towards more correct and stable subdivision, supporting dermatologists in early and accurate detection of skin diseases.

C. Adaptive Particle Swarm Intelligent Optimization (APSIO) Feature Selection

APSIO is a feature selection procedure that aims to select the most relevant features from a dataset for use in structure and training. The APSIO calculates the potential of each feature and assigns a fixed binary or continuous vector to each subset, adjusting its position based on the pixel of each skin layer. The adaptive approach dynamically changes the acceleration coefficients, inertia, and weight based on the current performance, range, and iteration progress of the swarm. The dataset is split into an 80:20 ratio, with 20% kept back for testing, and the remaining 80% used for training. Fig. 4 shows the working flow. APSIO takes input data features such as numerical or categorical parameters of the dataset, which are first preprocessed and normalized. Feature range is done by an adaptive mechanism that assesses each feature's relevance as per fitness scores obtained from a predefined objective function. Every particle in the swarm is a candidate solution, a set of features with position and velocity being updated step by step. The algorithm estimates the fitness of every particle by measuring model performance (e.g., classification accuracy or error rate) based on the chosen features.

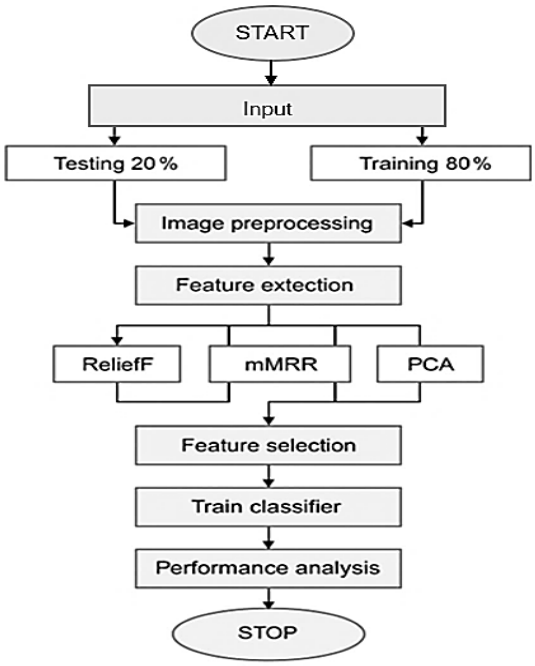


Fig. 4. Flow chart of APSIO feature selection.

Adaptive methods update particle behavior according to convergence trends dynamically, optimizing an exploration and exploitation balance. The result is the optimal subset of features that maximizes the objective function, enhancing the performance, and minimizing dimensionality.

D. APSIO Algorithm

1) Adaptive inertia weight in particle velocity update

$$v_i^{(t+1)} = w^{(t)}.v_i^{(t)} + C_1.r1.\left(p_i^{best} - x_i^{(t)}\right) + c2.r2.\left(g^{best} - x_i^{(t)}\right) \tag{10}$$

In Eq. (10), $v_i^{(t)}$ rate of unit i at iteration, $w^{(t)}$ adaptive inertia weight (variations per iteration) for calculating the features.

2) Position particle update

$$x_i^{(t+1)} = x_i^{(t)} + V_i^{(t+1)} \tag{11}$$

In Eq. (11), $(t+1)$that creates a binary representation (0 or 1) from a constant search space, where procedures may examine various feature combinations.

3) Determine each element's fitness using a few chosen structures

$$S\left(V_i^{(t)}\right) = \frac{1}{1+e^{-v_i^{(t)}}} \tag{12}$$

In Eq. (12), A Binary Transfer Function (BTF) $S\left(V_i^{(t)}\right)$, a mapping used in the selection of features which ends up in a binary representation (0 or 1) for each feature, signifying its inclusion or exclusion in the chosen subset, from a continuous search space where procedures may examine various feature configurations.

4) Function of binary transfer (for feature selection)

$$x_i^{(t+1)} = \begin{cases} 1 & if\ r < S\,V_i^{(t)} \\ 0 & otherwise \end{cases} \quad (13)$$

In Eq. (13), feature subset valuation of a fitness function is a numerical appearance, which calculates how well a detailed feature subset is suitable for classification. At different selecting features, a collective skin structure image that shows the continuing link is created by reflecting of essential perceptual uniform areas.

5) Optimization and backpropagation: Every few iterations, backpropagation can be used to optimize selective parameters.

$$x_i = x_i - \eta \nabla_x L(x_i) \quad (14)$$

In Eq. (14), η denotes the learning rate, and $\nabla_x L$ represents the gradient of the loss function with respect to the weights and biases denoted by x_i.

6) Output: An optimized binary vector $x_i \in \{0,1\}^n$ for each particle, representing the selected feature subset.

E. Hyper-Convoluted Intra-Capsuled Neural Network (HCI-CNN) and YOLO Algorithm

The HCI-CNN structure utilizes hyper-convolutional layers to identify complex spatial pixels and employs capsule-based learning to preserve the pose and structural relationships between different layers of skin. In the initial stage, a convolutional layer with 32 filters of 3×3 and ReLU activation is utilized to extract low-level visual features. To ensure efficient and stable training, an exponential moving average with a momentum of 0.99 is implemented, along with an adaptive learning rate of 0.001.

Fig. 5 shows the detailed architecture of the proposed hybrid classification model in skin disease detection.

The HCI-CNN and YOLO algorithms operate in a hybrid manner by integrating deep feature extraction and real-time object detection. HCI-CNN accepts preprocessed dermoscopic images and clinical images as input and uses hyper-convoluted layers to extract detailed texture, color, and structural features, which are encoded through intra-capsuled routing mechanisms for high-level representation. Such encoded features are used for the classification of skin lesions with high accuracy.

At the same time, the YOLO algorithm classifies the same images to identify and locate infected skin areas by segmenting them into a grid and predicting bounding boxes and confidence estimates. The hybrid system leverages HCI-CNN's classification power and YOLO's spatial localization feature. The combined technique improves both diagnostic accuracy and detection efficiency in medical imaging tasks. The Hybrid HCI-CNN and YOLO classification technique aims to enhance the accuracy of skin disease classification using medical images.

Fig. 6 shows an intra-capsuled architecture, which continues layers and orientation information through dynamic routing units, enabling the model to maintain spatial relationships among structures. As an outcome of HCI-CNN surpasses conventional CNNs by achieving greater generalizability, coupled with improved handling of intricate skin disease contours in diagnostic contexts.

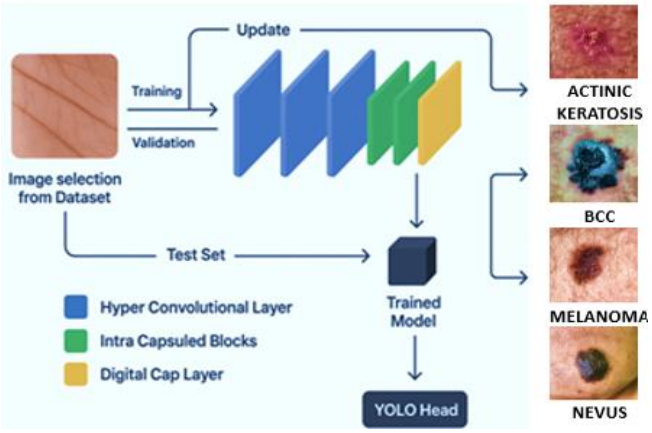


Fig. 5. Hybrid architecture diagram of HCI-CNN and YOLO algorithms.

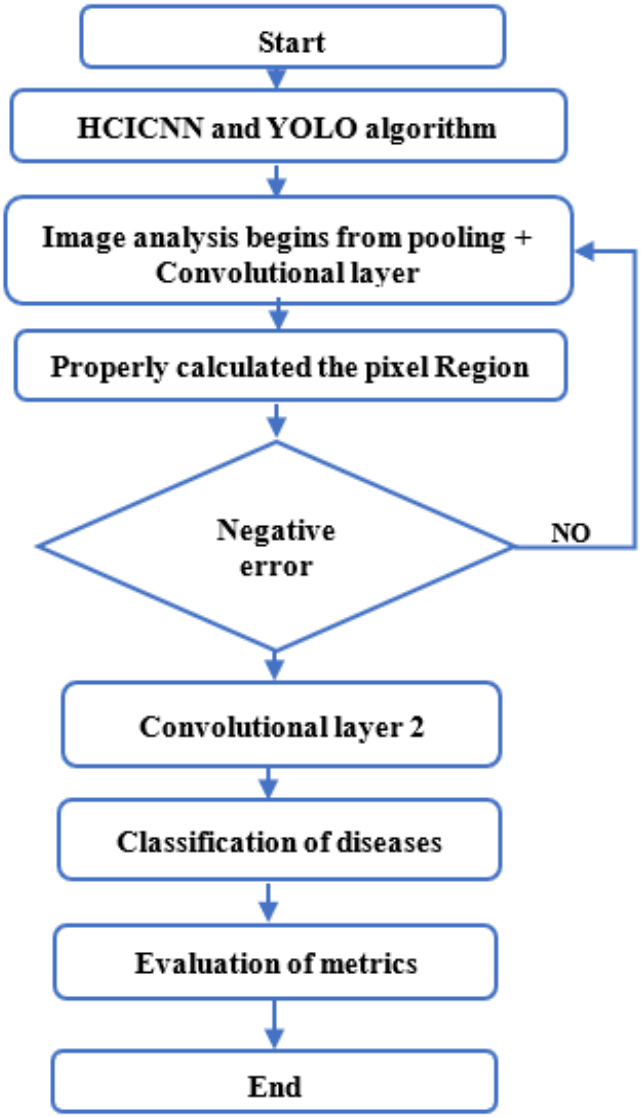


Fig. 6. Flow chart of feature classification.

F. Hybrid HCI-CNN and YOLO Algorithm

1) Convolution hyper layer

$$Y = \sum_{i=1}^{C_{in}} k_i\,(x) * x_i \tag{15}$$

In Eq. (15), x is an input mapping channel, $k_i\,(x)$ is dynamically evaluated for pixel classification.

$$d_e\,(x,y) = \sqrt{\sum_{n-1}^{n}(x_i - y_i)^2} \tag{16}$$

In Eq. (16), distance n is probably the most frequently used distance of difference across feature vectors. Here n represents the number of characteristics.

$$d_e\,(x,y) = \sum_{i=1}^{n} |\, a_{i-}\, b_i\,| \tag{17}$$

In Eq. (17), (a,b) constant may be used to quantify the separability of classes and to determine closely two samples (x,y), under the different skin classes are related.

$$e\,[\,X\,,Y\,] = \frac{1}{N}\sum_i X_i - Y_i \tag{18}$$

In Eq. (18), hyper-convolutional layers $X_i - Y_i$ are evaluated using multi-scale receptive field processing to capture both fine-grained texture and contextual information, enabling the differentiation of visually confusable skin cancer types.

Apply the YOLO Classification algorithm to predict object boundaries and class labels

In Eq. (19), each pixel within the region is characterized based on the amount of associated data.

$$X_O = X_1 + X_2 + X_3 \tag{19}$$

2) Capsule intra transformation

- Approach Routing ($U_{i,j}\,, r,\ l$), All capsules I and J in layer l and (l + 1), respectively.
- for r iterations, do
- for all capsule i in layer l: $C_i = softmax\,(b_i)$
- for all capsule j in layer (l + 1): $S_j = \sum_{i=1}^{N} C_{i,j}\,, U_{i,j}$
- for all capsule j in layer (l + 1): $V_j = Squash\ S_{i,j}$

3) Routing between capsules in a dynamic manner: In Eq. (20), a squash function is applied to normalize the output vectors of capsules.

$$V_j = Squash\,(\sum_i cij\ u\ j|\ i) \tag{20}$$

4) Output: The vector magnitude is maintained between 0 and 1 by using the nonlinear squashing function, evaluated in Eq. (21)

$$P\,(y = k\,|\,x\,) = |V_k| \tag{21}$$

IV. Results and Discussion

Skin disease identification is the basis for the experiment calculation, and the system uses a specific dataset from which performance analysis is assessed. The specific test data with predictive values is properly evaluated for feature selection, identification of specific skin conditions, and comparison with previous studies. The APSIO feature selection and HCI-CNN classification are evaluated in this result and discussion.

A. Skin Diseases Dataset Description

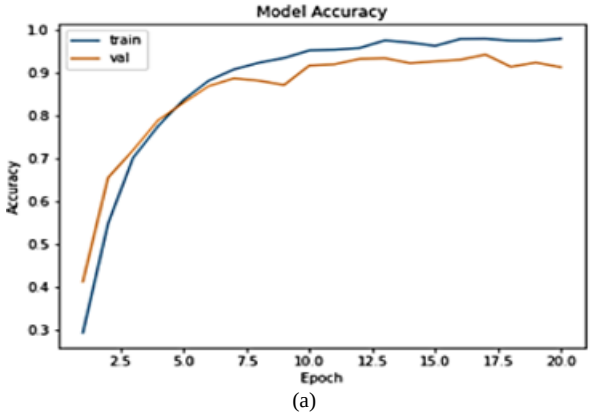
The proposed method uses a dataset of high-resolution clinical skin images using the recommended medical application, 10015 images representing seven different classes includes Melanocytic Nevi (6705), Melanoma (1113), Benign Keratosis (1099), Basal Cell Carcinoma (514), Actinic Keratosis (327), Vascular Lesions (142), Dermatofibroma (115) from the HAM10000 dataset and it can be gain from https://www.kaggle.com/ and 1200 clinical images which is collected from individuals. Each image is associated with specific attributes. The dataset is divided into seven categories of dermoscopic images. The images are split into a training set (80%) and a testing set (20%).

B. Feature Selection Calculation

Three feature selection methods are compared in Table II based on feature dimensionality, computational time, and accuracy. When compared based on skin image features, the proposed APSIO method demonstrated superior performance, achieving higher accuracy than other methods across different data volumes. These results demonstrate that APSIO is a highly effective optimization-based feature selection technique, achieving higher accuracy with significantly fewer features and making it well-suited for efficient and accurate skin image classification tasks.

TABLE I COMPARATIVE ANALYSIS OF SKIN FEATURE SELECTION METHODS

Feature Selection	Computational Time (s)	Feature Dimensionality	Accuracy
Histogram of Oriented Gradients (HOG)	1.8	High (~3780 features/image)	92.4
Gray Level Co-occurrence Matrix (GLCM)	1.2	Medium (~20–40 features)	94.3
Adaptive Particle Swarm Intelligent Optimization (APSIO)	2.4	Low (~10–15 optimized features)	95.3



(a)

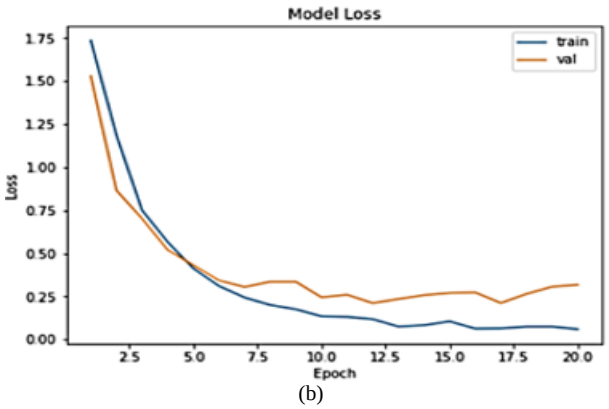


Fig. 7. Training and testing accuracy and loss.

Fig. 7 shows the accuracy of the typical and the loss of the APSIO approach generated from segmented images. The highest value of loss is 1.75, and the value decreases, as indicated by (a), as epochs increase, as indicated by (b). The output of the Dense model performs best on the testing set using segmented images.

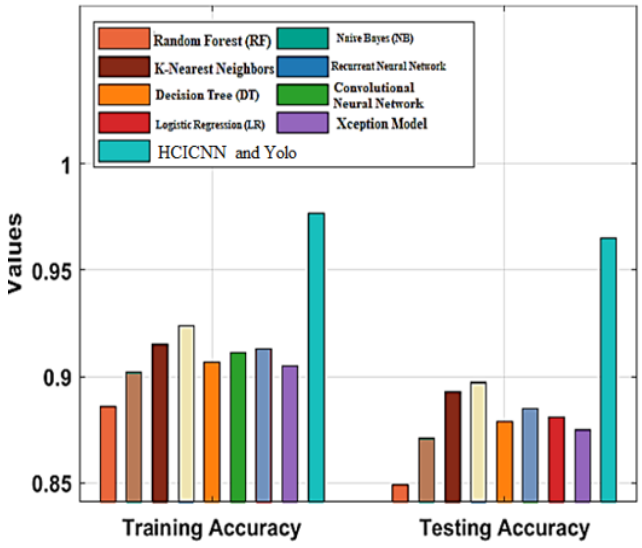


Fig. 8. Training and testing accuracy of hybrid HCI-CNN and YOLO.

Fig. 8 evaluates and compares the performance of different deep learning and machine learning models for classification problems. In terms of generalization, the Hybrid HCI-CNN and YOLO models achieve higher accuracy than the previous classification methods.

C. Experimental Evaluation

The Error-Level Analysis (ELA) in Fig. 9 evaluates the error rates of different classification outputs. The HCI-CNN had a reduced error rate of 5.5%, which was significantly lower than GoogLe-Net 14.3%, ResNet 13.2%, and Mobile Net 8.4% and demonstrates the superior performance of HCI-CNN compared to other models.

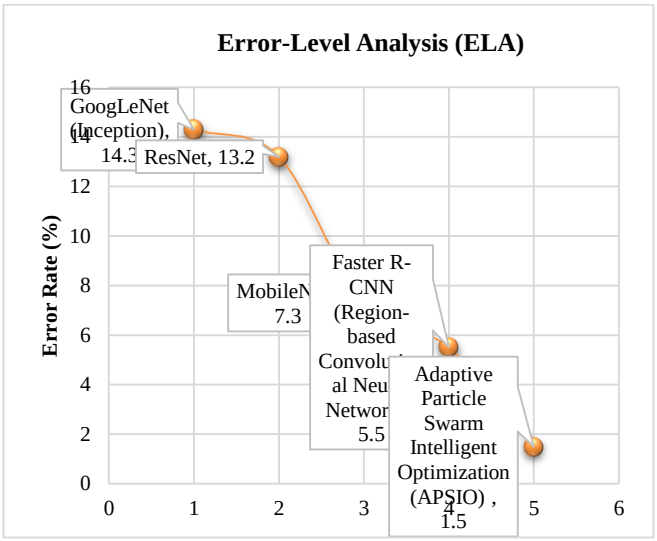


Fig. 9. Error rate analysis.

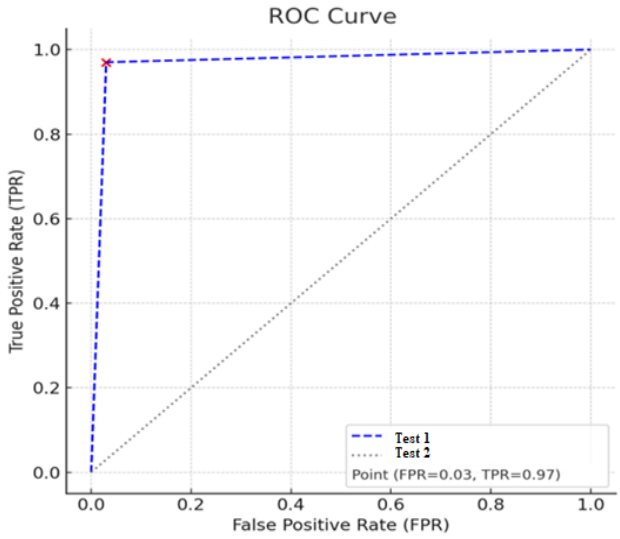


Fig. 10. HCI-CNN classification ROC curve.

Fig. 10 shows the HCI-CNN classification error analysis based on the test data, indicating a True Positive Rate (TPR) of 0.97 and a False Positive Rate (FPR) of 0.03.

D. Comparison Evaluation

Fig. 11 validates that the proposed hybrid approach, HCI-CNN combined with YOLO, reaches a maximum accuracy of 97%, which is significantly better than the previous ResNet50 and K-nearest neighbour, which show accuracies of 94 %, ResNet101V2 accuracy of 86 %.

Fig. 12 shows that the HCI-CNN and YOLO hybrid model significantly improves recall, achieving 96.55% compared to existing methods. Random Forest and SVM achieved recall values of approximately 82.5% and 88.3%, respectively, whereas ResNet achieved 89.3% and Extreme Gradient Boosting achieved 92.5%.

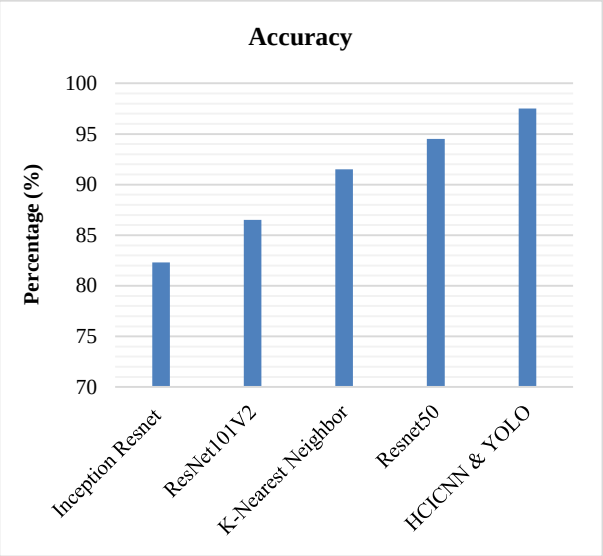


Fig. 11. Proposed classification accuracy analysis.

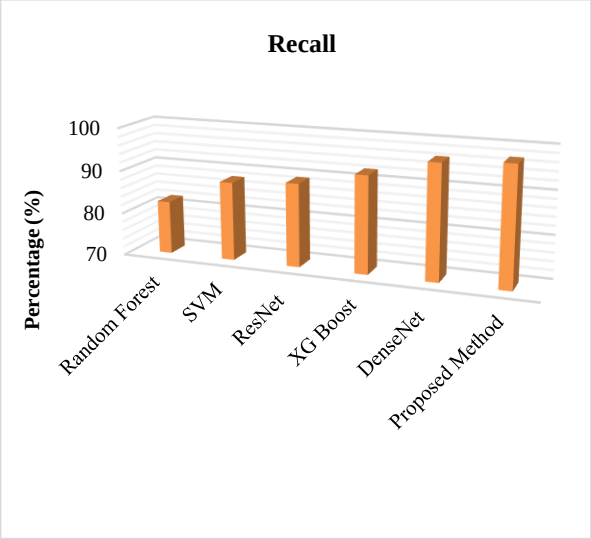


Fig. 12. Proposed recall analysis.

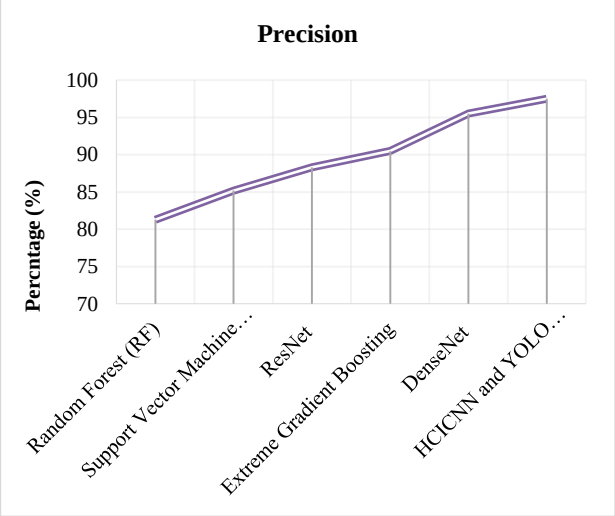


Fig. 13. Proposed precision analysis.

Fig. 13 shows that the proposed HCI-CNN and YOLO method achieves a precision of 96.52%, outperforming other existing methods such as Random Forest (81.3%), SVM (85.2%), ResNet (88.4%), and Dense Net (95.6%).

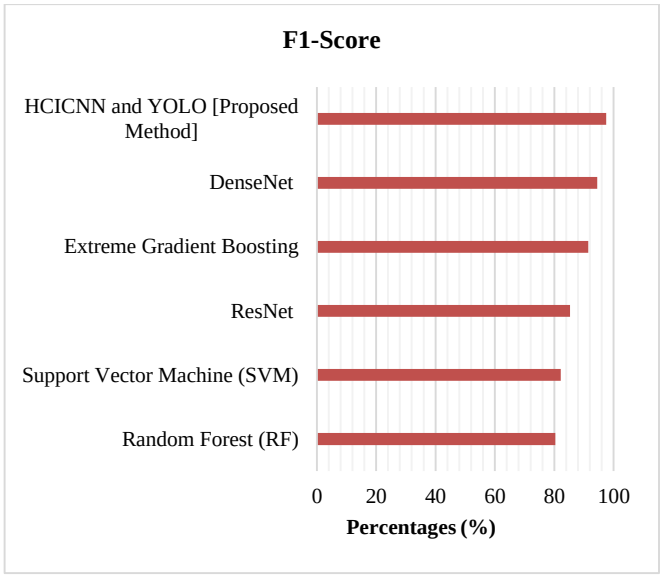


Fig. 14. Proposed F1-score analysis.

Fig. 14 shows a comparison graph of F1-Scores for various machine learning models, including Random Forest and SVM, which demonstrate an average performance of 83%. However, deep learning models such as deep residual networks and eXtreme Gradient Boosting are also compared. The proposed HCI-CNN and YOLO methods are also compared, with the highest F1 score of 96.93% achieved by HCI-CNN, demonstrating its superior performance.

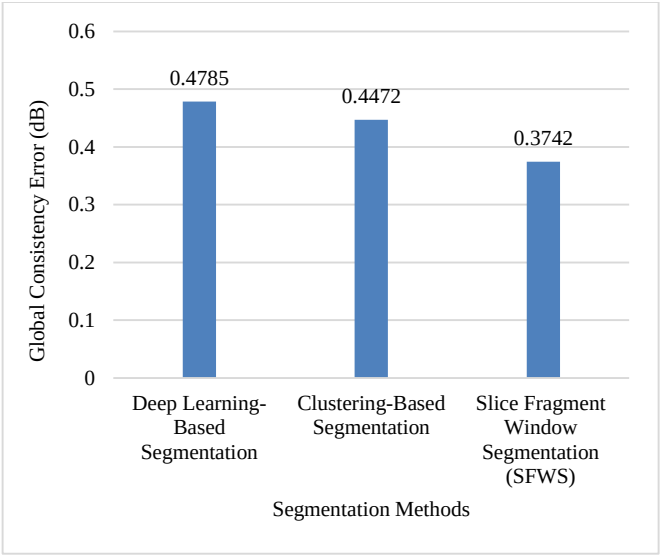


Fig. 15. Proposed segmentation analysis.

Fig. 15 shows the segmentation evaluation of Global Consistency Error (dB) across the three segmentation methods: Slice Fragment Window Segmentation (SFWS), Deep Learning-Based Segmentation, and Clustering-Based Segmentation. The proposed SFWS has the lowest error at

0.3742 dB, demonstrating the highest reliability. In contrast, the Deep Learning-Based Segmentation has an error of 0.4785 dB, followed by the Clustering-Based Segmentation with an error of 0.4472 dB.

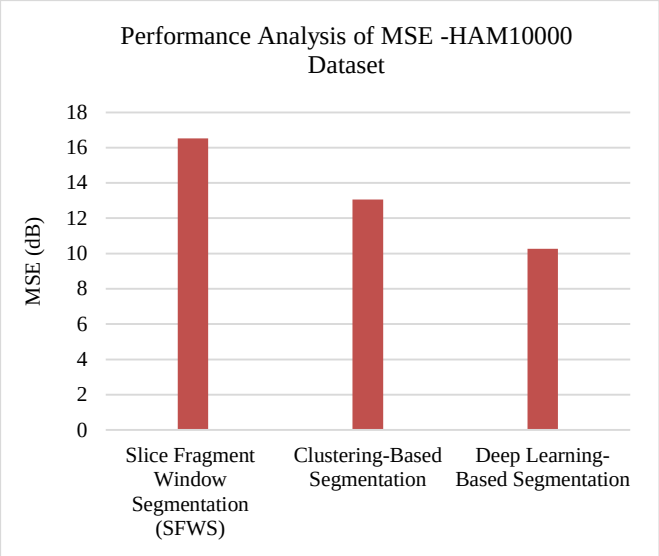


Fig. 16. MSE segmentation comparison analysis.

In Fig. 16 MSE (Mean Square Error) for skin image segmentation with the existing approach is evaluated through the proposed SFWS method as an improved segmentation analysis.

Fig. 17 illustrates the analysis results of the collected clinical skin images. When no dermatological abnormalities are detected in the collected clinical skin image, the system classifies it as normal skin; otherwise, it is classified as a disordered image. Among the 1,200 clinical skin images evaluated, 1,194 were detected as normal, and the remaining 6 were identified as containing dermatological disorders.

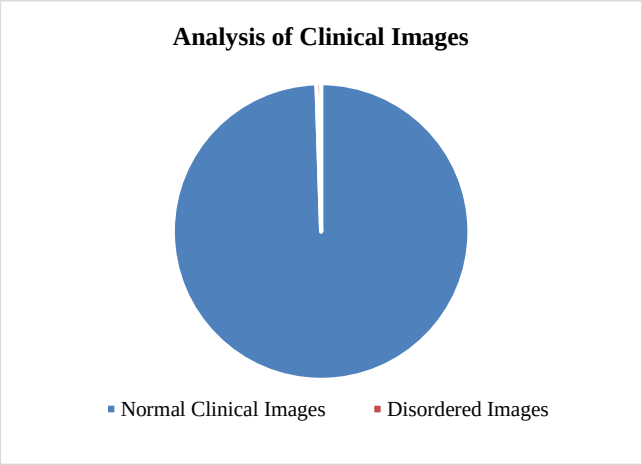


Fig. 17. Assessment of acquired dermatological images.

Table III presents a comparative analysis based on four different parameters, using data from cited references that support the comparison technique.

TABLE II COMPARISON OF EXISTING SKIN CANCER IDENTIFICATION METHODS

References	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Ashtagi et al. [25]	91.50	91.00	90.00	90.50
Imam et al. [26]	93.75	93.75	92.50	93.00
Bello et al. [27]	87.00	87.00	86.00	86.50
Khan et al. [28]	92.16	92.16	91.50	91.80
Natha et al. [29]	86.90	86.90	85.50	86.00
Jain et al. [30]	90.00	90.00	90.00	90.00
Huang et al. [31]	87.00	87.00	86.00	86.00
Chaturvedi et al. [32]	89.00	89.00	88.00	88.00
Gupta et al. [33]	85.00	85.00	84.00	84.00
Proposed Model	**97**	**96.52**	**96.55**	**96.93**

E. Discussion

Specific AI technologies have a significant impact on reliability and diagnostic classification results for classifying skin disease. In this research work, HCI-CNN and YOLO are applied for classification, while APSIO is employed for feature selection. The combination of these procedures addresses critical tests such as high-dimensional data, feature idleness, and demand for amplified classification accuracy. The mechanism of APSIO dynamically adjusts the limits of the convergence rate. The most discriminative features are selected from dermoscopic images owing to this adaptive mechanism, which decreases computational difficulty without compromising diagnostic accuracy. Deep and wide convolutional layers with multiple levels of abstraction constitute the HCI-CNN architecture, which is effective in capturing.

V. CONCLUSION

The proposed AI-powered skin disease detection method utilizes advanced image processing techniques, including preprocessing, clustering, and classification. Through the preprocessing filter, the image improves in clarity for different skin colors. In segmentation, factors are divided to evaluate geometric and physical damage. In APSIO feature selection, related features are selected. Furthermore, a hybrid approach integrating HCI-CNN and YOLO is employed for effective classification. This approach demonstrates high accuracy in the detection and classification of seven different types of skin disorders. The selective features remove false positives and computational burden, with the added potential for real-time medical applications. In this study, both real-world clinical images from individuals and curated images from established datasets were analyzed to evaluate the system's performance comprehensively. The output performance analysis, with an Accuracy of 97%, a precision of 96.52%, a recall of 96.55%, and an F1-score of 96.93%, will enhance performance and improve classification for applications in healthcare facilities, ultimately leading to improved treatment outcomes for patients with skin diseases.

In the future, the rising plurality and complexity of skin diseases require new modalities with newly available technologies, such as AI and the IoT, to better align and sustain

the management (exposure reduction) of skin diseases. IoT-based wearables equipped with high-fidelity imaging sensors and biosensors can continuously monitor skin surface indicators, temperature, and moisture, as well as the evolution of skin lesions, enabling direct and real-time examination of skin health by identifying the reliable amount of surface data on a person. Combining this data with an AI analytical model, which leverages streaming and real-time data, enables humane dimension management (e.g., tracking behaviors, detecting anomalies, and risk stratification) while learning and evolving. Future directions for connected wearables in regulating dermatological diseases can be supported by hybrid frameworks of learning, such as federated learning and edge cloud learning, which erase discreet elements to reduce decentralized training mechanisms using decentralized data. This approach maintains the patient's anonymity and increases the potential for generalizability.

REFERENCES

[1] Shaikh, Juveriya, Rubeena Khan, Yashwant Ingle, and Nuzhat Shaikh. “Improved skin cancer detection using CNN." International journal of health sciences (Qassim), pp. 14347-14360, 2022.

[2] Mazhar, T.; Haq, I.; Ditta, A.; Mohsan, S.A.H.; Rehman, F.; Zafar, I.; Gansau, J.A.; Goh, L.P.W. “The role of machine learning and deep learning approaches for the detection of skin Cancer”, healthcare, 2023.

[3] Melissa A. Trudrung, Cole Bacig, Brandon Vander Zee, Heather Potter, “Basal cell carcinoma and squamous cell carcinoma of the conjunctiva in a single lesion”, American Journal of Ophthalmology Case Reports, Volume 38, 2025.

[4] Raut, Roshani, Yogini Borole, Sonali Patil, V. N. Khan, and Dattatray G. Takale, “Skin disease classification using machine learning algorithms”, Neuro Quantology 20, no. 10, pp. 9624-9629, 2022.

[5] Balasundaram, A. Shaik, B. R. Alroy, A. Singh and S. J. Shivaprakash, “Genetic Algorithm optimized stacking approach to skin disease detection”, IEEE Access, vol. 12, pp. 88950-88962, 2024.

[6] Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J, “Classification of skin disease using deep learning neural networks with Mobile Net V2 and LSTM”, Sensors 2021.

[7] K. Lee et al., “Multi-Task and Few-Shot Learning-Based Fully Automatic Deep Learning Platform for Mobile Diagnosis of Skin Diseases”, IEEE journal of biomedical and health informatics, vol. 27, no. 1, pp. 176-187, 2023.

[8] Rashid, J.; Ishfaq, M.; Ali, G.; Saeed, M.R.; Hussain, M.; Alkhalifah, T.; Alturise, F.; Samand, N, “Skin cancer disease detection using transfer learning technique”, Applied Science, 2022.

[9] M. Fahaad Almufareh, “An Edge computing-based factor-aware novel framework for early detection and classification of melanoma disease through a Customized VGG16 Architecture with privacy preservation and real-time analysis”, IEEE Access, vol. 12, pp. 113580-113596, 2024.

[10] Jagdish, M., Gualán Guamangate, S. P., López, M. A. G., De La Cruz-Vargas, J. A., and Camacho, M. E. R, “Advance study of skin diseases detection using image processing methods”, 2022.

[11] P. Yao et al., "Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification," in IEEE Transactions on Medical Imaging, vol. 41, no. 5, pp. 1242-1254, 2022.

[12] Gairola, A. K., Kumar, V., Sahoo, A. K., Diwakar, M., Singh, P., and Garg, D, “Multi-feature fusion deep network for skin disease diagnosis”, Multimedia Tools and Applications, 84(1), pp.419-444, 2025.

[13] A. A. Salam, M. Usman Akram, M. Haroon Yousaf and B. Rao, “Derma Trans Net: where transformer attention meets U-Net for skin image segmentation”, IEEE Access, vol. 13, pp. 64305-64329, 2025.

[14] Rokade, Sonali, and Nilamadhab Mishra, “A blockchain-based deep learning system with optimization for skin disease classification”, Biomedical Signal Processing and Control, 2024.

[15] K. Thurnhofer-Hemsi, E. López-Rubio, E. Domínguez and D. A. Elizondo, “Skin Lesion classification by ensembles of deep convolutional networks and regularly spaced shifting”, IEEE Access, vol. 9, pp. 112193-112205, 2021.

[16] Reddy, D. A., Roy, S., Kumar, S., & Tripathi, R, “Enhanced U-Net segmentation with ensemble convolutional neural network for automated skin disease classification”, Knowledge and Information Systems, 65(10), pp. 4111-4156, 2023.

[17] M. Gallazzi, S. Biavaschi, A. Bulgheroni, T. M. Gatti, S. Corchs and I. Gallo, "A large dataset to enhance skin cancer classification with transformer-based deep neural networks”, IEEE Access, vol. 12, pp. 109544-109559, 2024.

[18] Anand, V.; Gupta, S.; Altameem, A.; Nayak, S.R.; Poonia, R.C.; Saudagar, A.K.J. “An Enhanced transfer learning based classification for diagnosis of Skin Cancer”, Diagnostics 2022.

[19] K. Thurnhofer-Hemsi, E. López-Rubio, E. Domínguez and D. A. Elizondo, “Skin Lesion Classification by Ensembles of Deep Convolutional Networks and Regularly Spaced Shifting”, IEEE Access, vol. 9, pp. 112193-112205, 2021.

[20] Gouda, W., Sama, N. U., Al-Waakid, G., Humayun, M., and Jhanjhi, N. Z, “Detection of skin cancer based on skin lesion images using deep learning”, In Healthcare, Vol. 10, no. 7, 2022.

[21] R. Mittal, F. Jeribi, R. J. Martin, V. Malik, S. J. Menachery and J. Singh, "DermCDSM: Clinical decision support model for dermatosis using systematic approaches of machine learning and deep learning," in IEEE Access, vol. 12, pp. 47319-47337, 2024.

[22] Rajeswari, R., P. G. Sivagaminathan, and A. R. Arunachalam, “Skin cancer detection and classification using deep learning techniques”, Deep Learning in Medical Image Analysis. Chapman and Hall/CRC, pp. 97-117, 2024.

[23] Q. Sun, Y. Tang, S. Wang, J. Chen, H. Xu and Y. Ling, “A Deep learning based melanocytic nevi classification algorithm by leveraging physiologic-inspired knowledge and channel encoded information”, IEEE Access, vol. 12, pp. 113072-113086, 2024.

[24] Jain, S.; Singhania, U.; Tripathy, B.; Nasr, E.A.; Aboudaif, M.K.; Kamrani, A.K, “Deep learning-based transfer learning for classification of Skin cancer”, Sensors, 2021.

[25] Ashtagi, R., Kharat, P. V., Sarmalkar, V., Hosmani, S., Patil, A. R., Akkalkot, A. I., and Padthe, A, “Enhancing melanoma skin cancer diagnosis through transfer learning: An EfficientNetb0 approach”, Acadlore Transactions on AI and Machine Learning, 3(1), 57-69, 2024.

[26] Imam, M. H., Nahar, N., Rahman, M. A., and Rabbi, F, “Enhancing skin cancer classification using a fusion of Dense-net and Mobile-net models: a deep learning ensemble approach”, Multidisciplinary Science Journal, 6(7), pp. 2024117-2024117, 2024.

[27] Bello, A.; Ng, S.C., Leung, M.-F, “Skin cancer classification using fine-tuned transfer learning of DENSENET-121”, Applied Science, 2024.

[28] Khan, M. A., Alam, S., and Ahmed, W, “Enhanced skin cancer diagnosis via deep convolutional neural networks with ensemble learning”, SN Computer Science, 6(2), pp. 1-14, 2025.

[29] Natha, P. R. I. Y. A., and P. R. Rajeswari, “Skin cancer detection using machine learning classification models”, International Journal of Intelligent Systems and Applications in Engineering, pp. 139-145, 2024.

[30] Jain, Satin, Udit Singhania, Balakrushna Tripathy, Emad Abouel Nasr, Mohamed K. Aboudaif, and Ali K. Kamrani, “Deep learning-based transfer learning for classification of skin cancer”, Sensors 21, no. 23, 2021.

[31] Huang, Hsin-Wei, et al, “Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers”, The Journal of dermatology, pp. 310-316, 2021.

[32] Chaturvedi, S. S., Tembhurne, J. V., and Diwan, T, “A multi-class skin Cancer classification using deep convolutional neural networks”, Multimedia Tools and Applications, 79(39), pp. 28477-28498, 2020.

[33] Chaturvedi, S. S., Gupta, K., & Prasad, P. S, “Skin lesion analyser: an efficient seven-way multi-class skin cancer classification using MobileNet”, In Advanced machine learning technologies and applications: proceedings of AMLTA, Springer Singapore. pp. 165-176, 2020.

Deep Learning Optimization Conception: Less Data, Less Time, More Performance

Mohamed Amine MEDDAOUI, Moulay AMZIL, Imane KARKABA, Mohammed ERRITALI
Data4Earth Laboratory, Sultan Moulay Slimane University, Beni Mellal, Morocco

***Abstract*—Although Deep Learning has not made a breakthrough in terms of artificial intelligence core technology, it achieves the best worldwide performance across areas such as computer vision and natural language processing. However, it depends on large-scale datasets and enormous computational resources. This paper tackles a major issue: Can we train more efficient deep learning models with less data in less time? We look at numerous strategies designed to reduce the burden of training, without letting the quality deteriorate. From transfer learning and few-shot learning to lightweight architectures, synthetic datasets produced artificially, as well as dispersed training, we contemplate how to make advanced AI subsystems fit for running under scarce resources. The aim is to lay down a future for deep learning that is more sustainable and all-embracing. This research focuses on the important issue of streamlining deep learning models with balancing model performance against data collection and computations. We look into other approaches such as transfer learning, couple with fewshot learning, data augmentation, architecture optimization, and parallelization. We explain processes with their benefits as well as their setbacks. Our research shows that training a model more efficiently improves the overall training process, making it cheaper and greener. A change like this would help more people use sophisticated AI systems even when limited by constrained resources. This broadens the real-world application of AI technology and further stimulates innovation in the area.**

Keywords*—*Deep Learning; AI; IoT; optimization; transfer learning; model compression; few-shot learning

I. Introduction

Deep learning, a branch of machine learning inspired by the intricate workings of the human brain, has truly transformed countless fields. Think about computer vision, natural language processing, or even speech recognition deep learning has been a game-changer. Its incredible success largely comes from its knack for learning complex patterns from huge amounts of data, leading to breakthroughs in things like image classification, spotting objects in photos, translating languages, and even self-driving cars. But here's the catch: deep learning models often hit a wall because they need so much data and so many powerful computers. This reliance can make it tough for researchers and organizations without easy access to data or supercomputers to get involved. Plus, there's a growing concern about how much energy these massive models consume during training, which isn't great for the environment. Right now, deep learning often demands a ton of data collection and labeling, which can be incredibly expensive and eat up a lot of time. This is especially true in niche areas like medical imaging or working with rare languages. On top of that, training today's most advanced deep neural networks can take days or even weeks, burning through significant energy and money. These limitations make deep learning less accessible, particularly for smaller businesses and in places with fewer resources. So, the big question we're tackling in this research is fundamental: Can we create more efficient deep learning models that need less data and less training time, all without sacrificing performance? This effort is super important because it aims to open up advanced AI to everyone, shrink its environmental footprint, and speed up research and development. Ultimately, it's about sparking innovation and making deep learning useful in even more places. In this paper, we dive into various strategies designed to ease the burden of training deep learning models, focusing on ways to make them more efficient and sustainable. We'll explore methods like transfer learning and few-shot learning, which cleverly reuse existing knowledge and cut down on data needs. We'll also look at lightweight designs that reduce how much computing power is required. And we won't stop there we'll investigate the potential of creating artificial datasets and using distributed training to get around data shortages and speed up the training process. To give you a clear picture, here's how we've structured this paper: Related work is given in Section II. Section III explores the hurdles of training deep learning with limited resources. Data Minimization techniques is shown in Section IV. Section V dives into and evaluates different techniques for minimizing data and time while keeping performance high. Section VI gives you practical examples of these techniques in action. Section VII brings up topics for future research, including potential impacts and ongoing challenges. Finally, Section VIII wraps things up by summarizing our findings and highlighting the real scientific value and practical uses of our work.

II. Related Work

The quest to make deep learning more efficient, especially in terms of data and computing power, has really picked up steam in recent years. Researchers have been exploring all sorts of ways to tackle the challenges of massive datasets and long training times. Take transfer learning, for example it's become a go-to for efficient deep learning. Early on, Pan and Yang [1] gave us a great overview of transfer learning techniques, showing how useful it is when you don't have a lot of labeled data. Since then, others have built on that work, proving its effectiveness in everything from analyzing medical images, CheXpert [2] to understanding human language Devlin [3]. Then there's few-shot learning, which is all about teaching models to learn from just a handful of examples. Researchers like Lake and his team [4] have been playing with metalearning to help models adapt quickly to new tasks with very little data. You've got things like prototypical networks Snell [5] and

Matching Networks Benmalek [6], which are cool architectures designed specifically for few-shot learning and have shown some really promising results in classification tasks. These methods are a lifesaver in fields where getting data is tough or expensive. Data augmentation is another popular trick. It's all about artificially beefing up your dataset to help your model generalize better. We've moved beyond simple things like rotating or scaling images. Now, we're using advanced techniques like Generative Adversarial Networks (GANs) to create brand new, realistic data Strubell [7]. This is a huge help when you're dealing with scarce or sensitive data. And let's not forget about architectural innovations. We've seen the rise of lightweight deep learning models like MobileNet Ly, A [8] SqueezeNet and EfficientNet Meddaoui [9]. These are designed to be super-efficient, so they can run on devices that don't have a lot of power, proving that you can get great performance without a massive model. Finally, distributed training and parallelization have been a game-changer for training huge models. Work by folks like Soham and his team [10] on large-scale distributed deep learning has made it possible to train models with billions of parameters. By spreading the work across multiple GPUs or machines, you can slash training times. Of course, there are still challenges to iron out, like communication bottlenecks and keeping everything in sync, but it's an active area of research. This paper pulls together all these ideas, giving you a big picture view of how we can make deep learning more sustainable and accessible to everyone.

III. Challenges of Training Deep Learning with Limited Resources

It is true that the incredible performance of deep learning models results from various factors, among which data quality and quantity is most important. The deep trained neural models with numerous parameters require not only immense volumes of labeled data, but also diverse data, wide variety of examples to construct powerful and generalizable representations. Models will be useless and unreliable, underperforming, poorly generalizing, or mislearning spurious correlations without diverse data. This reliance on massive datasets creates a major challenge in specialized fields where data is difficult to obtain, expensive, needs to be labeled, or is restricted by privacy regulations such as medicine or rare languages. Researchers often have to create inefficient strategies to counter their inability to obtain enough high-quality labeled data because of stringent privacy regulations. Such constrained inefficient methods greatly limit frontier deep learning users.

A. Dependence on Large Datasets

It is common wisdom that the success of deep learning rests on many factors, but most fundamentally: quantity and quality in data. Moreover, at least to a large degree, very large amounts of fast processor power are also necessary. To effectively represent powerful and generalizable representations, a deep neural network with a large number of parameters (high dimensionality) requires massive labeled data Khan, T [11]; lacking this data, models could be in danger of converging incorrectly or having characteristics they truly do not have or generalizing tendencies to ill effect. Such data has not always been easy to come by: sometimes expensive and obscure, particularly in medicine or rarer tongues. Such scarcity necessitates wasteful learning from scratch and brings into brighter relief techniques that need adaptation for low data quantity.

B. Computational and Temporal Cost

Anybody knows that Deep learning is the most important deep neural network benefiting everything requires speed and plenty of data. The two chief factors are data volume and quality, as well as computing resources, which do not come cheaply either. The known algorithms have won several prestigious honors Aaron [12] continue to improve standards in this area. It is possible to explore the great potential for improving our techniques as it were, even if only. Deep neural networks, because of their numerous parameters, must possess a large amount of labeled data in order to yield powerful and general representations (at least when trained on an array of different input data). Without such information, Park [13] models are at risk of converging; they have features that are pertinent yet not general enough, which results in poor performance. But it is not easy to collect these data from anywhere; at times it can be costly and complicated, particularly in medicine or for rare languages. This lack of information means inefficient training from scratch, emphasizing the need for methods suited to small amounts of data. In addition to being data dependent, the training of deep learning algorithms incurs high computational costs as well as temporal costs that are challenging to meet. The training of advanced deep neural networks is extremely demanding and usually requires a great deal of computation to be processed—often in the form of GPUs or TPUs—for days to weeks at a time. This form of computation comes at a heavy price, both in terms of hardware and energy. Training one large deep learning model can also incur additional electricity costs, thus contributing to a larger carbon footprint, Qasim [14] Almost all of the economic and environmental costs inhibit the utilization of deep learning, specifically for small and medium enterprises (SMEs) or in geographies with limited access to sophisticated computing systems. Rather than an academic endeavor, this motivates the need to build more optimal deep learning frameworks in order to equalize access to AI, lower its environmental footprint, and hasten innovation in multi-industries. Zhao [15] Focusing on these needs not only advances deep learning but also leads to a more impactful sustainable future.

C. Overfitting and Generalization

With limited data, overfitting usually could happen. The overfitted model learns the training data by heart and also remembers noise as opposed to general patterns it should be learning from the training data. Jonas [16] It will do poorly on new examples. The risk of overfitting is increased when little data is processed because the model has fewer examples to illustrate diversity. Regularization techniques help reduce the bad effects, but do not compensate for the absence of diversity. A model's utility is really in its ability to generalize, and if overfitting happens, this is compromised. For this reason, optimization strategies must ensure an important element of generalization.

IV. Data Minimization Techniques

In an effort to address the difficulties posed by the lack of data and the high resource requirements of deep learning, the development of models and algorithms which require less labeled data and less training time has become a point of focus. These efforts are essential to the accessibility, efficiency, and sustainability of deep learning. In this chapter, we present data minimization methods such as Transfer Learning, Few-Shot Learning, Zero-Shot Learning, as well as Data Augmentation, explaining their principles, benefits, and case studies.

A. Transfer Learning and Fine-Tuning

Transfer learning (Fig. 1) reuses a pre-trained model on a similar task with a large dataset, then adapts it to a new specific task with a smaller dataset for Gupta [17].

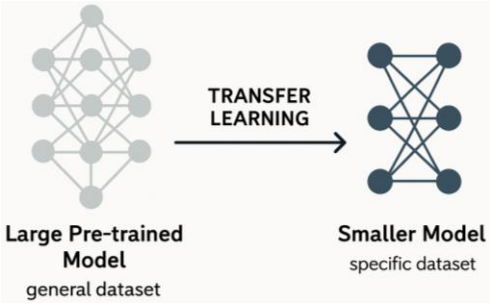


Fig. 1. How transfer learning operates.

1) Principles and advantages: The lower layers of deep networks learn generic features (edges, textures). Reusing these pre-trained layers leverages the knowledge acquired over a large corpus, reducing the data requirement for the new task and accelerating training. Advantages: data reduction, training acceleration, performance improvement, overfitting reduction as Soham [10]. The schematic illustration of transfer learning is given in Fig. 2.

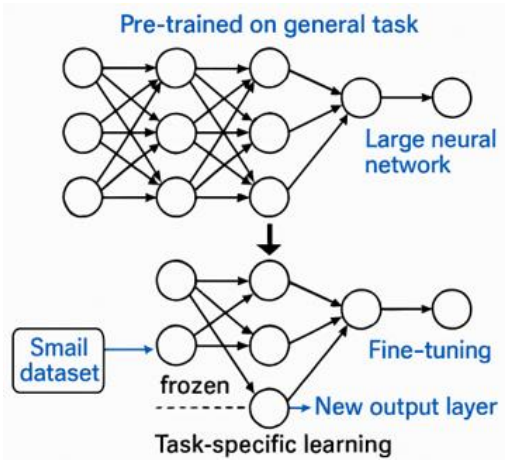


Fig. 2. Schematic illustration of transfer learning.

2) Fine-tuning strategies fine-tuning adapts the pre-trained model strategies:

a) Feature extractor: Pre-trained layers frozen, only a new classification layer is trained. Useful for similar small datasets Rohith [18].

b) Partial fine-tuning: Top layers thawed and trained, bottom layers frozen. Suitable for medium or slightly different datasets.

c) Full fine-tuning: All layers trained with a very low learning rate. For large datasets or very different tasks as Li, T [19].

The choice depends on the size and similarity of the new dataset.

B. Few-Shot Learning and Zero-Shot Learning

These paradigms allow models to recognize new classes with an extremely limited number, or even zero, of labeled examples.

1) Meta-learning: Meta-learning, or "learning to learn," trains a model to quickly adapt to new tasks from a few examples. The goal is to give the model the ability to quickly adapt to new data distributions or classes: Assiri [20]. Categories:

a) Optimization-based: Learns an initialization function or optimization algorithm for fast adaptation (e.g., MAML) [14].

b) Metric-based: Learns a distance function in an embedding space where similar examples are close (e.g., Siamese networks, prototypes) Nguyen [21].

c) Model-driven: Uses specific architectures with memory or attention mechanisms to manage limited data.

2) Metric learning: Learns a similarity function to measure the proximity between examples. Classification is done by distance comparison. Siamese networks train the model to produce embeddings where similar pairs are close together. Prototype networks learn one "prototype" per class. These methods allow efficient classification with few examples, Liu [22].

Zero-Shot Learning (ZSL) classifies never-before-seen classes by exploiting semantic information (attributes, word embeddings). The model maps data to a shared semantic space to recognize new classes without direct examples, Chen [23].

Table I Comparison of Few-Shot Learning approaches (Meta-learning vs. Metric learning).

TABLE I. Comparison of Few-Shot Learning Approaches (Meta-Learning vs. Metric Learning)

Characteristic	Meta-learning	Metric learning
Main objective	Learn to learn new tasks quickly	Learn a similarity/distance function
Mechanism	Learn an initialization or algorithm	Learn an embedding space
Examples of methods	MAML, Reptile	Siamese Networks, Prototype Networks, Relational Networks
Adaptation to the task	Fast, via a few optimization steps	By distance comparison with support examples
Complexity	Can be complex to train	Easier once the metric is learned
Required data	Requires various meta-training tasks	Requires pairs of examples to learn the metric

Benefits	High flexibility, rapid adaptation	Effective for classification with few examples
Disadvantages	High computational cost, sensitivity to hyperparameters	Less flexible for very different tasks

C. Data Augmentation

Data augmentation refers to a technique that modifies existing data in order to increase the size and diversity of the training dataset. This reduces the amount of real world data that needs to be collected, improves the model generalization and prevents overfitting. For image data, common augmentation techniques include rotations and flips, shifts and scaling, as well as color jittering. In the case of text data, augmentation techniques include synonym replacement and random insertion, deletion or even swapping of words. Generative Adversarial Networks (GANs) are more sophisticated augmentation methods which can be used to create new and realistic data samples that augment the original dataset, Thakor [24]. This approach is particularly useful in machine learning since it enables the model to learn a broader spectrum of variations, increasing robustness while reducing the risk of memorization.

1) Classical techniques: For images: geometric transformations (rotation, translation, scaling, flipping) and photometric transformations (brightness, contrast, noise) Ghorbal [25]. For text: synonym substitution, word insertion/deletion/exchange, back-and-forth translation, Latha [26].

2) Generative augmentation (GANs, VAEs) Generative models create realistic synthetic data:

a) GANs (Generative Adversarial Networks): A generator creates data, a discriminator distinguishes it from real data. The generator produces high-quality samples, useful for rare data, Gao [27].

b) VAEs (Variational Autoencoders): Learn a latent representation of the data to generate new samples. Less realistic than GANs, but more stable and controllable Alsolai [28].

Useful for expensive or sensitive data (medical imaging).

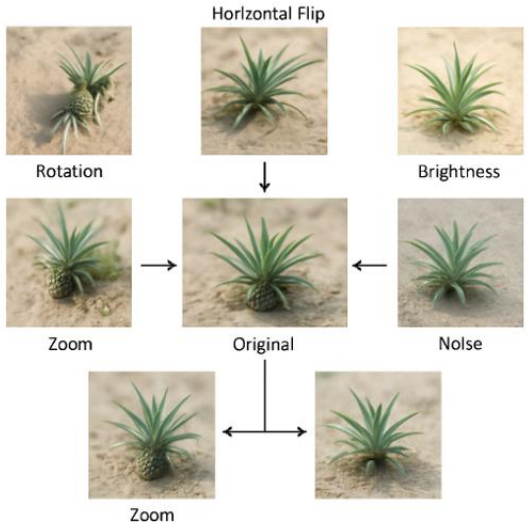


Fig. 3. Example of image transformations for data augmentation.

D. Data Synthesis and Synthetic Data

Data synthesis creates new examples that mimic the statistical properties of real data, especially when collection is impossible or raises privacy concerns for Noor [29]. GANs and VAEs produce high-fidelity synthetic data. Advantages: confidentiality, unlimited availability, attribute control, cost reduction Zhang [30]. The quality of synthetic data is crucial for model generalization. Fig. 3 shows image transformations for data augmentation.

V. Techniques for Minimizing Training Time

Reducing training time is crucial for the effectiveness of deep learning. This involves optimizing architectures, algorithms, and resource utilization.

A. Optimization of Model Architectures

Designing lighter and more efficient models reduces training time and resource consumption.

1) Lightweight and efficient models (MobileNet, EfficientNet): These architectures minimize parameters and operations while maintaining accuracy. MobileNet for Meddaoui [31] uses depthwise separable convolutions, reducing complexity. EfficientNet Xu, L [32] optimizes depth, width, and resolution via compound scaling, delivering accurate and efficient models with fewer parameters. These models are ideal for transfer learning and deployment on constrained devices.

2) Model compression (pruning, quantization, distillation): Reduces the size and complexity of trained models without compromising performance, crucial for deploying and accelerating inference and training.

a) Pruning: Removes non-essential connections or neurons. Low weights are set to zero. The pruned network is fine-tuned to recover accuracy, Anda-Suárez [33].

b) Quantization: Reduces the numerical precision of weights and activations (e.g. FP32 to INT8). Reduces model size, memory and speeds up operations, Han [34].

c) Knowledge distillation: A small model ("student") learns from a large model ("teacher"). The student is trained on the teacher's class probabilities, improving its performance despite its small size as Markkandeyan [35].

B. Hyperparameter Optimization

1) Bayesian search: Constructs a probabilistic model of the objective function to choose the hyperparameters to evaluate. More efficient than grid or random search because it uses previous results to guide future evaluations Kovalev [36].

2) Gradient-based optimization: Allows to optimize hyperparameters by making them differentiable with respect to the objective function, using gradient descent. Efficient for high-dimensional hyperparameter spaces (Fig. 4) [30].

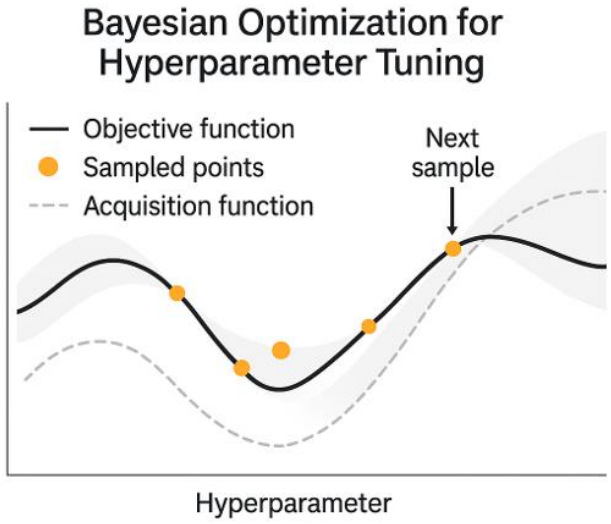


Fig. 4. Illustration of the Bayesian search process for hyperparameter optimization.

C. Parallelization and Distributed Computing

Distributes the workload across multiple devices or machines to reduce training time.

1) GPU/TPU Training: GPUs and TPUs are hardware accelerators for matrix-intensive deep learning computations. They enable considerable training speedups. Modern frameworks are optimized for these units for Markkandeyan [37].

2) Distributed computing frameworks (Horovod, Ray) for distributed training across multiple GPUs or machines:

a) Horovod: Simple and fast distributed framework, integrates with TensorFlow, Keras, PyTorch. Uses AllReduce to aggregate gradients efficiently, enabling near-linear scaling as Dillshad [38].

b) Ray: Open-source framework for distributed applications. Provides a simple API for parallelizing tasks and managing resources. Suitable for distributed training and hyperparameter optimization.

These frameworks leverage aggregated computing power to drastically reduce training time. Fig. 5 shows schematic of distributed training with multiple GPUs/TPUs.

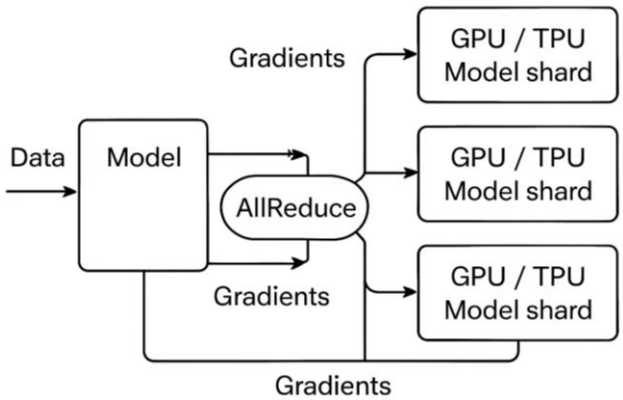


Fig. 5. Schematic of distributed training with multiple GPUs/TPUs.

D. Incremental and Continuous Training

Updates a model incrementally with new data, rather than retraining it from scratch. Useful when new data arrives or the model needs to adapt to changes. Reduces training time by fine-tuning the existing model. Requires handling "catastrophic forgetting" (loss of prior knowledge) with techniques such as regularization. Continuous training adapts the model in real time, essential for dynamic applications.

E. Performance Measurement and Evaluation

Evaluating models optimized for limited resources is crucial. Traditional metrics must be complemented by efficiency and robustness indicators. Fig. 6 and Fig. 7 shows training time and accuracy comparison.

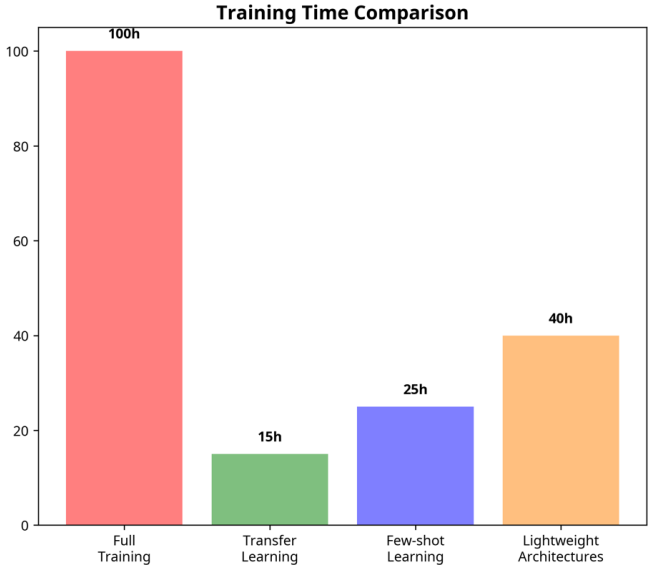


Fig. 6. Training time comparison.

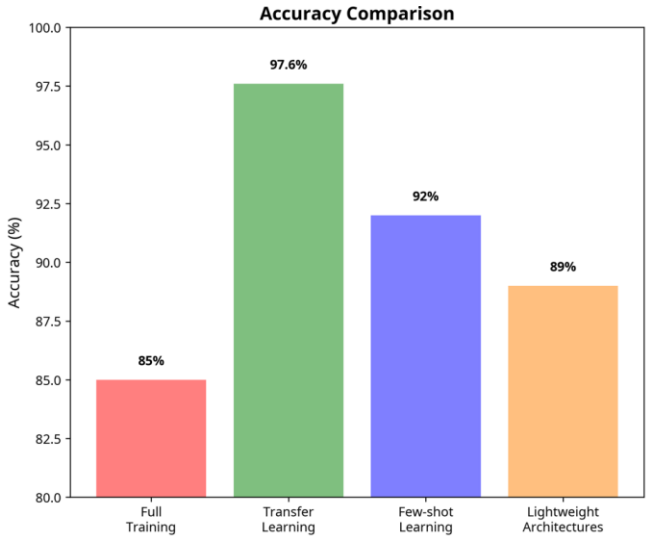


Fig. 7. Accuracy comparison.

1) Metrics adapted to limited resources: Besides precision, accuracy, recall, F1 score and AUC-ROC, one should consider:

a) Model complexity: Number of parameters and FLOPs/MACs to assess size and computational complexity. Fewer parameters/FLOPs = lighter and faster.

b) Training and inference time: Real time required, directly related to the minimization objectives.

c) Energy consumption: Evaluates the carbon footprint of the model, which is increasingly relevant.

d) Robustness to small perturbations: Evaluates the model's response to small changes in the data, crucial for limited and less diverse data.

Table II presents an illustrative comparison of different deep learning techniques in terms of Precision, Recall, and F-score. The values are hypothetical and aim to demonstrate general trends observed in literature and practice, rather than specific experimental results

TABLE II. COMPARATIVE TABLE OF DEEP LEARNING TECHNIQUES

Technique	Precision (%)	Recall (%)	F-score (%)
Full Training	90.2	88.4	89.3
Transfer Learning	98.5	96.2	97.4
Few-Shot Learning	92.8	90.1	91.5
Lightweight Architectures	93.6	92.4	93

2) Precision: The proportion of true positives among all positive results (true positives + false positives). Recall: The proportion of true positives among all actual positive cases (true positives + false negatives). F-score: The harmonic mean of precision and recall, useful for evaluating models on imbalanced datasets. Value ranges indicate potential variability depending on the application domain, task complexity, and data quality. Transfer learning often excels in precision and recall, especially with limited data, as it leverages pre-existing knowledge. Few-Shot Learning may have slightly lower F-score performance due to the inherent difficulty of generalizing from very few examples, but it is crucial in extremely data-scarce scenarios. Lightweight architectures offer a good compromise between performance and computational efficiency, making them ideal for deployments on resource constrained devices.

3) Cross-validation and robustness: Cross-validation is essential for assessing generalization, especially with limited datasets. K-fold cross-validation divides the data into k folds: the model is trained k times on k-1 folds and evaluated on the remaining fold. The final performance is the average. This maximizes data utilization and reduces bias in the performance estimate.

Robustness, the model's ability to maintain stable performance in the face of noise or variation, is vital. Adversarial training or data augmentation improves this robustness, making the model more reliable in real-world conditions.

VI. CASE STUDIES AND PRACTICAL APPLICATIONS

Optimization techniques have proven their effectiveness in various fields, overcoming data and time constraints.

A. Computer Vision (Medical Image Classification with Little Data)

The scarcity of labeled medical images makes training challenging. Transfer learning and data augmentation are key. For rare disease detection (Fig. 8) (e.g., chest X-rays), a pre-trained model on ImageNet is fine-tuned on the small medical dataset Data augmentation (rotation, zoom, etc.) and conditional GANs increase diversity, reducing overfitting and protecting privacy.

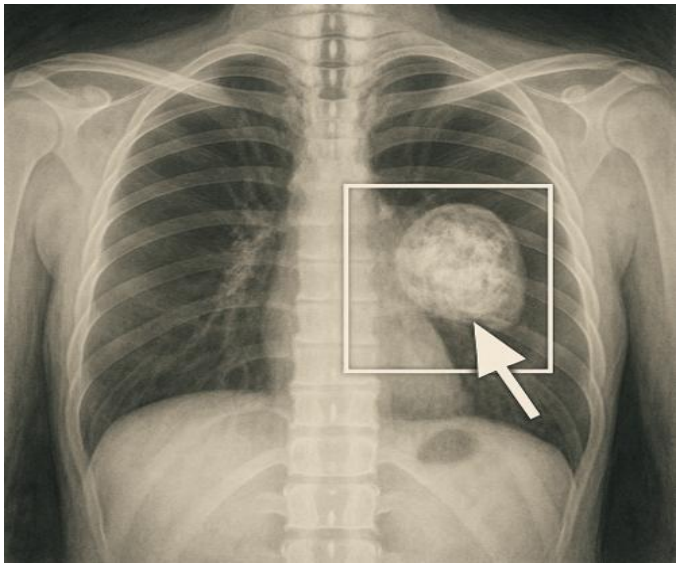


Fig. 8. Example of a chest X-ray with rare disease detection.

B. Natural Language Processing (Language Models for Rare Languages)

LLMs require terabytes of text, which is impossible for low-resource languages. Transfer Learning is the solution. A large pre-trained multilingual model is fine-tuned on the small bilingual corpus of the rare language. Few-shot learning and text data augmentation (paraphrases, back-and-forth translation) allow the model to be adapted with few examples, making NLP workable for these languages.

C. Industrial Applications (Defect Detection with Few Samples)

In industry, defects are rare, creating class imbalance and data deficiency. For visual inspection (e.g., electronic components), a pre-trained object detection model is adapted with Few-Shot Object Detection on a few defect examples.Generating synthetic defect data (3D simulations) complements real data, enabling robust training. Optimizing architectures (e.g., MobileNet) allows deployment on capacity-limited devices, reducing inference time.

These cases demonstrate that, despite constraints, optimization makes Deep Learning applicable and efficient in real-life scenarios with limited resources, democratizing AI.

VII. Discussion and Perspectives

Due to the exploratory nature of this optimization deep learning techniques, which survey rather than rigorously assess specific methodologies, a detailed empirical analysis with quantitative evaluation falls outside the focus of this research. Regardless, the effectiveness of the discussed methods Transfer Learning, Few-Shot Learning, Data Augmentation, and Lightweight Architectures is well-established. In this section, we try to outline the anticipated outcomes of applying the techniques discussed above and provide arguments justifying the adoption. Based on the findings presented in this paper, the implementation of some of the optimization strategies is expected to lead to remarkable results in a few critical performance metrics.

Data Requirements: The amount of labeled data needed to train effective models is likely to be caused by strategies like transfer learning and few shot learning. Take, for instance, the case where a model is only required to be fine-tuned on a certain task. With the use of a small dataset, such a model can perform as well as a model that has been trained on a much larger dataset from scratch. This is of great importance in data scarce fields such as medical imaging or more advanced specialized NLP tasks. Optimizing deep learning training with limited resources is a key area. The techniques discussed (transfer learning, few-shot learning, data augmentation, architectural optimization, compression, parallelization) offer promising solutions for more accessible and sustainable deep learning. However, challenges remain. Reducing the Time Needed for Training: The use of pre-trained models, alongside their architectures and distributed training systems, greatly cuts down the time it takes to train deep learning models. These changes not only save on the cost of computation but also speed up the development and research processes which means newer models can be iterated and deployed more expeditiously.

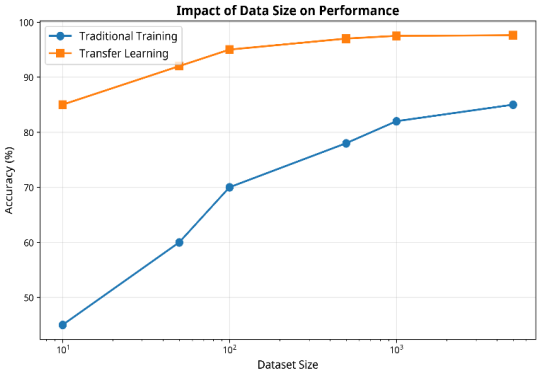


Fig. 9. Impact of dataset size on performance.

Improved Effectiveness and Strength: Robust models which improve the generalization capabilities to unseen examples and reduce overfitting largely benefit from data augmentation. This is important for deploying models in a wide range of scenarios because data variability can often be very high. Lowering the Computational Footprint: Deep learning is becoming more available for edge computing applications, as well as aiding in energy conservation. Models designed with lightweight frameworks, meant to run effectively on devices with low resources, reduce memory consumption, speed up inference, and cut down on energy use. Fig. 9 shows impact of dataset size on performance.

Dataset used: NIH Chest X-ray Dataset, CheXpert (Stanford), The Cancer Imaging Archive (TCIA), Brain Tumor MRI Dataset (Kaggle), Hugging Face Datasets.Zhao [39].

We can see, as long as the dataset is large then accuracy is high; but in the case of transfer learning the accuracy is not influenced by small dataset size.

A. Limitations of Current Approaches

1) Dependence on the pre-trained model: The effectiveness of Transfer Learning depends on the relevance of the source model. A target task that is too distant limits the benefits.

2) Complexity of few-shot learning: Designing and training meta-learners is complex, requiring diverse meta-training datasets.

3) Synthetic data realism: Generating high-fidelity synthetic data remains a challenge. Insufficient realism can hinder generalization to real data.

4) Performance-efficiency tradeoff: Compression techniques (pruning, quantization) can result in a slight performance loss. The optimal balance is a research challenge.

5) Catastrophic forgetting: In incremental training, the model may forget previous knowledge while learning new knowledge. Robust solutions are needed.

B. Future Research Directions

1) Several promising avenues:

a) Self-supervised and self-trained learning: Reduces reliance on labeled data by generating internal supervisory signals, enabling more efficient training on unlabeled data.

b) Neuro-symbolic AI: Integrating Deep Learning with symbolic methods to learn with less data by exploiting a priori knowledge and logical rules, improving generalization and robustness.

c) Active learning: Allows the model to identify the most informative examples to label, reducing the cost and time of manual annotation.

d) Joint optimization: Simultaneously optimize architecture, hyperparameters, and training strategies for globally optimal configurations.

e) Specialized hardware and co-designed algorithms: Developing new hardware architectures for lightweight models and efficient algorithms, pushing the limits of performance and energy efficiency.

C. Ethical and Practical Implications

Deep learning optimization makes AI more accessible, bridging the digital divide. However, it raises questions about the accountability of models trained on synthetic or limited data, and the need to ensure fairness and transparency. Reducing energy consumption is positive, but the exponential growth of models requires continued vigilance.

VIII. Conclusion

Our primary contributions are threefold: (1) a detailed analysis of the key challenges involved in training deep learning models under limited data and computational resources; (2) a comprehensive review and evaluation of state-of-the-art techniques aimed at minimizing data requirements and training time while preserving high performance; and (3) a discussion of the practical implications and benefits of adopting these optimization strategies across various contexts. While this paper provides a strong theoretical foundation and integrates current knowledge, it is important to acknowledge certain limitations. Notably, it does not include new experimental results; thus, the quantitative validation of these combined strategies remains a priority for future research. Additionally, the rapid pace of innovation in deep learning calls for continuous reassessment of such surveys, as new models and optimization methods continue to emerge. Looking ahead, several promising directions warrant further exploration. Empirical evaluation of how these optimization techniques interact across diverse, real-world datasets would provide essential insights. Future research may also focus on developing automated systems capable of selecting and adapting optimization strategies to specific tasks and resource constraints. It is equally important to examine the ethical implications of data minimization, particularly in sensitive domains where fairness, privacy, and accountability are critical. Finally, integrating optimization approaches with hardware-software co-design frameworks may lead to even greater gains in efficiency and scalability. Ultimately, this work aims to support a vision of deep learning that is not only powerful and performant, but also sustainable, accessible, and broadly applicable.

References

[1] Bhuiyan, M. S. M., Rafi, M. A., Rodrigues, G. N., Mir, M. N. H., Ishraq, A., Mridha, M. F., & Shin, J. (2025). Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies. *Array*, 100390.

[2] CheXpert: A Large-scale Chest Radiograph Dataset with Annotations for Interpreting Image Labels. (2019).

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long and Short Papers).

[4] Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science.

[5] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical Networks for Few-shot Learning. Advances in Neural Information Processing Systems,30.

[6] Benmalek, M., & Seddiki, A. (2025). Particle swarm optimization-enhanced machine learning and deep learning techniques for Internet of Things intrusion detection. *Data Science and Management*.

[7] Strubell, E., Ganesh, A., & McCallum, A. (). Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the th Annual Meeting of the Association for Computational Linguistics.

[8] Ly, A., & Gong, P. (2025). Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning. *Nature Communications*, *16*(1), 3252.

[9] Meddaoui, M. A., Erritali, M., Madani, Y., & Sailhan, F. (2022, June). Mask Detection Using IoT-A Comparative Study of Various Learning Models. In *International Conference on Smart Homes and Health Telematics* (pp. 272-283). Cham: Springer International Publishing.

[10] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. 2018.

[11] Khan, T. M., Shams, M. A., Khatun, M. M., Chowdhury, J. H., Uddin, M. S., Emon, T. A., ... & Al Ahmed, S. R. (2025). Predictive modeling and optimization of WS2 thin-film solar cells: A comprehensive study integrating machine learning, deep learning and SCAPS-1D approaches. *Renewable Energy*, 123519.

[12] Aaron Defazio and L´eon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pages 1753–1763, 2019.

[13] Park, C., & Lee, S. (2025). Tunable anisotropy in lattice structures via deep learning-based optimization. *International Journal of Mechanical Sciences*, *290*, 110121.

[14] Qasim Jebur Al-Zaidawi, M., & Çevik, M. (2025). Advanced deep learning models for improved IoT network monitoring using hybrid optimization and MCDM techniques. *Symmetry*, *17*(3), 388.

[15] Zhao, W., Zhou, L., & Han, C. (2025). A Hybrid Optimization Approach Combining Rolling Horizon with Deep-Learning-Embedded NSGA-II Algorithm for High-Speed Railway Train Rescheduling Under Interruption Conditions. *Sustainability*, *17*(6), 2375.

[16] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 806–815, 2019.

[17] Gupta, B. B., Gaurav, A., Attar, R. W., Arya, V., Bansal, S., Alhomoud, A., & Chui, K. T. (2025). A hybrid Ant Lion Optimization algorithm based lightweight deep learning framework for cyber attack detection in IoT environment. *Computers and Electrical Engineering*, *122*, 109944.

[18] Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, and Rong Ge. Explaining landscape connectivity of low-cost solutions for multilayer nets. *arXiv preprint arXiv:1906.06247*, 2019.

[19] Li, T. (2025, January). Optimization of Clinical Trial Strategies for Anti-HER2 Drugs Based on Bayesian Optimization and Deep Learning. In *Proceedings of the 2025 5th International Conference on Bioinformatics and Intelligent Computing* (pp. 163-168).

[20] Assiri, M. S., & Selim, M. M. (2025). A swin transformer-driven framework for gesture recognition to assist hearing impaired people by integrating deep learning with secretary bird optimization algorithm. *Ain Shams Engineering Journal*, *16*(6), 103383.

[21] Nguyen, T. P., & Cho, M. Y. (2025). A cloud-based leakage current classified system for high voltage insulators with improved particle swarm optimization and hybrid deep learning technique. *Engineering Applications of Artificial Intelligence*, *143*, 109987.

[22] Liu, H., Shuai, W., Yao, Z., Xuan, J., Ni, M., Xiao, G., & Xu, H. (2025). Optimization of solid oxide electrolysis cells using concentrated solar-thermal energy storage: A hybrid deep learning approach. *Applied Energy*, *377*, 124610.

[23] Chen, D., & Huang, Y. (2025). Integrating reinforcement learning and large language models for crop production process management optimization and control through a new knowledge-based deep learning paradigm. *Computers and Electronics in Agriculture*, *232*, 110028.

[24] Thakur, D., Dangi, S., & Lalwani, P. (2025). A novel hybrid deep learning approach with GWO–WOA optimization technique for human activity recognition. *Biomedical Signal Processing and Control*, *99*, 106870.

[25] Ghorbal, A. B., Grine, A., Elbatal, I., Almetwally, E. M., Eid, M. M., & El-Kenawy, E. S. M. (2025). Predicting carbon dioxide emissions using deep learning and Ninja metaheuristic optimization algorithm. *Scientific Reports*, *15*(1), 4021.

[26] Latha, P., & Kumaresan, P. (2025). Deep learning model optimization for crop prediction and recommendation using Harris Hawks optimization. *Environmental Research Communications*, *7*(4), 045008.

[27] Gao, Y., Zhou, S., & Li, M. (2025). Structural topology optimization based on deep learning. *Journal of Computational Physics*, *520*, 113506.

[28] Alsolai, H., Al-Wesabi, F. N., Motwakel, A., & Drar, S. (2025). Sine Cosine Optimization With Deep Learning–Driven Object Detector for Visually Impaired Persons. *Journal of Mathematics*, *2025*(1), 5071695.

[29] Noor, S., AlQahtani, S. A., & Khan, S. (2025). Chronic liver disease detection using ranking and projection-based feature optimization with deep learning. *AIMS Bioengineering*, *12*(1).

[30] Zhang, X., Zhang, K., Li, Z., & Yin, X. (2025). Deep DeePC: Data-enabled predictive control with low or no online optimization using deep learning. *AIChE Journal*, *71*(3), e18644.

[31] Meddaoui, M. A., & Erritali, M. (2024, May). Smart Irrigation Integrate To IOT Technology Based On Deep Learning's Algorithm LSTM. In *2024 Mediterranean Smart Cities Conference (MSCC)* (pp. 1-5). IEEE.

[32] Xu, L., Wen, Z., Su, H., Cola, S., Fabbian, N., Feng, Y., & Yang, S. (2025). An innovative method integrating two deep learning networks and hyperparameter optimization for identifying fiber optic temperature measurements in earth-rock dams. *Advances in Engineering Software*, *199*, 103802.

[33] de Anda-Suárez, J., Pérez-Zúñiga, G., López-Ramírez, J. L., Herrera Pérez, G., Zeferino González, I., & Verde Gómez, J. Y. (2025). Lithium Battery Enhancement Through Electrical Characterization and Optimization Using Deep Learning. *World Electric Vehicle Journal*, *16*(3), 167.

[34] Han, G., Wachowicz, K., Usmani, N., Yee, D., Wong, J., Elangovan, A., ... & Fallone, B. G. (2025). Patient-specific hyperparameter optimization of a deep learning-based tumor autocontouring algorithm on 2D liver, prostate, and lung cine MR images: A pilot study. *Algorithms*, *18*(4), 233.

[35] Markkandeyan, S., Ananth, A. D., Rajakumaran, M., Gokila, R. G., Venkatesan, R., & Lakshmi, B. (2025). Novel hybrid deep learning based cyber security threat detection model with optimization algorithm. *Cyber Security and Applications*, *3*, 100075.

[36] Kovalev, D. (2025). Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*.

[37] Markkandeyan, S., Ananth, A. D., Rajakumaran, M., Gokila, R. G., Venkatesan, R., & Lakshmi, B. (2025). Novel hybrid deep learning based cyber security threat detection model with optimization algorithm. *Cyber Security and Applications*, *3*, 100075.

[38] Dillshad, V., Khan, M. A., Nazir, M., Saidani, O., Alturki, N., & Kadry, S. (2025). D2LFS2Net: Multi-class skin lesion diagnosis using deep learning and variance-controlled Marine Predator.

[39] Zhao, B., Zhang, W., Zhang, Y., Zhang, C., Zhang, C., & Zhang, J. (2025). Lithium-ion battery remaining useful life prediction based on interpretable deep learning and network parameter optimization. *Applied Energy*, *379*, 124713.

A Novel Hybrid HO-CAL Framework for Enhanced Stock Index Prediction

Zeren Shi, Othman Ibrahim, Hanini Ilyana Che Hashim
Faculty of Management, University Technology Malaysia, Skudai 81310, Johor, Malaysia

***Abstract*—The accurate prediction of stock indexes plays a critical role in supporting investment decisions and managing financial risks. This study proposed a novel hybrid deep learning model that integrated the strengths of Convolutional Neural Networks (CNN), the Attention mechanism, and Long Short-Term Memory (LSTM) networks to enhance the modelling of temporal patterns in financial time series. To further improve the prediction performance, the Hippopotamus Optimization (HO) algorithm was incorporated to fine-tune the networks parameters. This is the first application of the CNN-Attention-LSTM (CAL) architecture to stock index prediction. Ablation experiments revealed that the proposed CAL significantly outperformed traditional CNN, LSTM, and CNN-LSTM models, highlighting the effectiveness of the Attention-based architecture. Comparative analyses also demonstrated that the HO-optimized CAL (HO-CAL) model achieved superior predictive accuracy across multiple markets, confirming both the robustness of the hybrid model and the optimization algorithm. These findings underscore the potential of combining deep learning architectures with metaheuristic optimization to improve the prediction accuracy in financial markets, offering valuable insights for real-world investment strategies.**

***Keywords*—Attention mechanism; CNN; LSTM; stock index; hippopotamus optimization algorithm**

I. Introduction

With the rapid development of the global economy and increasing maturity of financial markets, stocks play an important role in economic progress [1]. It has become an important investment tool for investors due to its high return characteristics. In the stock market, accurate price prediction is important for investors' decision making and risk management [2]. Nevertheless, the stock market is highly volatile and stock data are characterized by large volume, non-linearity, and noise, which makes accurate prediction extremely challenging [3]. The stock index as an investment weathervane is especially critical. Thus, exploring more efficient and accurate prediction methods has attracted increasing attention from researchers and investors.

Traditional econometric methods, such as ARIMA and GARCH, perform better on low-volatility time series, yet they are unable to deal with high-dimensional and highly nonlinear financial data as complexity and uncertainty increase [4]. Machine learning methods such as Random Forest (RF) and Support Vector Machine can learn the nonlinear relationship between stocks and various impact factors effectively [5]. However, these methods are limited in dealing with noise and missing data and rely too much on feature selection, which leads the prediction accuracy to be difficult to reach the expectation [6]. Recently, deep learning has been shown to perform better than traditional machine learning in stock price analysis because its stronger learning and adaptive capabilities [7]. Most representative neural networks have superior nonlinear generalization abilities. Such as convolutional neural networks (CNN) and long short-term memory networks (LSTM) have become the first choices for handling financial time series, and the applications include stock price, volatility, and trend prediction [8]. However, there are still some significant weaknesses in the existing methods: insufficient modelling of sequence patterns, poor extraction of deep data features, and difficulties in handling large-scale long-term data owing to overfitting and gradient issues [9].

To address these issues, this study proposes a novel short-term stock index prediction model. First, CNN is utilized to extract deep features from stock data, then the Attention mechanism is introduced to assign higher weights to key features, and finally, LSTM is used to mine long-term time series features, that is, the CNN-Attention-LSTM (CAL) hybrid model. The three work in concert to extract more comprehensive and in-depth feature information over multiple time periods, to adapt to the characteristics and requirements of short-term stock prediction and improve the prediction accuracy.

It should be noted that the choice of neural network hyperparameters has a significant impact on the prediction performance [10], and the hyperparameters of CAL are numerous and interrelated. Traditional hyperparameter tuning methods, such as Grid Search and Random Search, are inefficient and prone to falling into local optimal solutions [11]. To address this problem, researchers have used optimization algorithms to automatically find the suitable hyperparameters of the model, such as the Genetic Algorithm and Particle Swarm Optimization (PSO). As an emerging metaheuristic optimization algorithm, the Hippopotamus Optimization (HO) algorithm simulates the feeding and social behaviors of hippos in nature [12]. It shows strong potential for solving complex optimization problems, with the advantages of fast convergence, strong global search capability, and easy implementation [13]. And the HO tends to perform better in dealing with high-dimensional, nonlinear, and multimodal problems than traditional algorithms like Grid Search, PSO and others [14].

As this study intends to enhance short-term stock index prediction, HO was utilized to explore the optimal hyperparameter combination of CAL (HO-CAL). Several international indexes were presented to examine the superiority

of the HO-CAL hybrid optimization framework. The contributions of this study are as follows:

- The CAL architecture integrates the advantages of different blocks. It realizes a breakthrough in extracting multi-dimensional features and fully exploits temporal information in the sequence. It not only captures the local features and long-term dependencies but also automatically focuses on the most important parts of the input. It shows good adaptability when facing different markets and effectively improves the prediction performance and generalization ability;
- The pioneering use of HO to optimize the CAL model, which plays a significant role in improving the prediction accuracy and robustness, contributes to the current research on combining optimization algorithms and deep learning.
- The experimental results show that HO-CAL can capture complex relationships and patterns in stock data and accurately predict the international indexes, which can help investors manage their assets more efficiently and allocate capital more adequately.

The remainder of the study is organized as follows: Section II reviews some related work; Section III elaborates on the architecture of the proposed model and the optimization process; Section IV introduces the experimental settings; Section V analyses and discusses the experimental results; and Section VI provides the conclusion of this study.

II. Related Work

A. Deep Learning Models for Time Series Prediction

Multivariate time series prediction has a wide range of applications in industry, finance, meteorology and other fields [15]. An increasing number of finance scholars have begun to use neural networks to build stock prediction models, which have achieved remarkable results. Deng et al. [16] utilized a CNN model to extract the structured features of a time series through multilayer convolutional operations and achieved good prediction results. Li et al. [17] proposed an LSTM model for predicting stock prices by preserving historical information through memory cells. Similarly, the integration architecture of CNN-LSTM is applied in stock market analysis [18], which captures the market fluctuation patterns by analyzing stock prices, volumes, and financial data. This assisted investors in formulating scientific investment strategies and effectively reducing investment risk. However, the dynamic interaction modelling ability of CNN-LSTM for temporal features is still limited by its static weight allocation mechanism [19]. Abbasimehr and Paki [20] proposed an Attention-based LSTM time series prediction model that dynamically assigned weights through an Attention layer to focus on key information, resulting in improved prediction accuracy.

B. Hybrid Deep Learning Architectures

The use of a single model often makes it difficult to simultaneously consider spatial features, temporal dependence and key information screening needs. Thus, combining CNN, LSTM and Attention has gradually become a research topic. Yi et al. [21] adopted the combined LSTM-CNN Attention model in short-term load prediction, which innovatively used convolutional kernels to extract the stochasticity of the user and solve the problem of non-smooth characteristics. Experiments proved that the model reduced the amount of input data, and the prediction accuracy outperformed the benchmark by more than 10%. Shi et al. [22] predicted stock prices by initially extracting the deep features of raw data through an Attention-based CNN-LSTM model and then fine-tuning them using XGBoost. It improved prediction accuracy, and helped investors realize income growth and risk avoidance. Borré et al. [23] proposed a hybrid LSTM-CNN architecture with the introduction of Attention mechanism and Gated Residual Networks. The experimental results showed that the proposed model performs well in predicting extreme events. Peng et al. [24] used an Attention-based CNN-LSTM model to predict multiple currencies simultaneously, which better captured the correlation between different frequencies and currencies. The model improved the prediction accuracy, and reduced the transaction cost and investment risk. LÜ et al. [25] used CNN-LSTM-Attention to effectively capture the spatial-temporal correlation of soil moisture, and its RMSE was reduced by 49% and 57% compared to LSTM and CNN-LSTM models, respectively.

C. Metaheuristic Optimization

Alongside advances in deep learning models architecture, the application of metaheuristic optimization for hyperparameter tuning in models has also received increasing attention. Although HO is still in the early stages of research, it has attracted attention owing to its simple structure and ease of integration with deep learning models. Maurya et al. [14] used HO to optimize distributed power planning and network reconfiguration under different load models and then compared it with PSO, whale optimization, grasshopper optimization, zebra optimization, coot bird optimizer, and firefly algorithms, proved the superiority of HO. Mashru et al. [26] proposed a multi-objective hippopotamus optimizer and compared it with six well-known swarm intelligence algorithms, which demonstrated the effectiveness of HO in dealing with structural optimization problems.

With the rapid evolution of deep learning, CAL model has emerged as a powerful tool for dealing with time series due to their capacity to jointly capture local patterns, sequential dependencies, and key feature importance. Meanwhile, the HO algorithm offers an efficient method for tuning hyperparameters of such deep architectures, further enhancing their performance in noisy stock market environments. While existing research still faces several limitations. Many hybrid models inadequately address the challenge of hyperparameter adjusting, often relying on inefficient or limited optimization strategies. Moreover, few studies systematically evaluate the integration of attention-based deep learning models with metaheuristic algorithms in the context of stock index prediction. To bridge this gap, this study proposes a novel HO-CAL model that combines the CNN-Attention-LSTM model with hippopotamus optimization algorithm for hyperparameter tuning. This approach is designed to enhance model robustness, accuracy, and adaptability across diverse market environments, offering both theoretical advancement and practical guidance for financial prediction.

III. METHODOLOGY

HO-CAL is a hybrid optimization deep learning framework that combines three different types of neural network layers as well as an optimization algorithm to fully capture financial information and temporal features in the data with the following basis.

A. Convolutional Neural Networks

A CNN can effectively extract local patterns and features from a time series by sliding a convolutional kernel over the input data to scan it, as shown in Fig. 1. And pooling layer is usually added after the convolutional layer to reduce the feature dimension of the data and reduce the computational complexity [27]. Therefore, in this study, when facing a large amount of stock time series, a 1D-CNN was used to extract features and reduce their dimensionality to make the learning effect more accurate and concise.

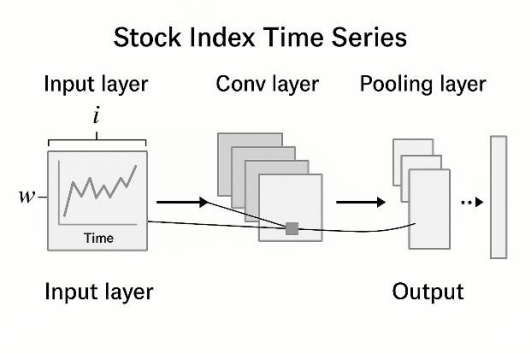


Fig. 1. CNN process structure.

B. Squeeze-and-Excitation Attention

CNN cannot distinguish the importance of information during learning [28]. While the Attention is extremely specialized in capturing critical nodes in sequence processing [29]. SE Attention is designed to improve the model performance by adaptively adjusting the weights of the feature channels. The Squeeze part extracts the global features by reducing the features size to 1×1 through a global average pooling operation. This step compresses the features of each channel into a scalar value to form a channel description vector. The Excitation part learns the channel weights through a series of fully connected layers and activation functions. These weights are used to weight the features and enhance the important feature channels. Its core formulas are as follows:

$$\mathrm{s} = \sigma\big(\mathrm{W}_2 \cdot \delta(\mathrm{W}_1 \cdot \mathrm{z})\big) \tag{1}$$

$$z_c = \frac{1}{H \times W} \sum_{i,j} U_c(i,j) \tag{2}$$

$$\widetilde{U}_c = s_c \cdot U_c \tag{3}$$

where H, W are the feature dimensions and C is the number of channels. Attention enables the proposed model to automatically focus on the key features in massive information, significantly improving the relevance of data processing. Therefore, the introduction of Attention is expected to further enhance the accuracy of stock index predictions.

C. Long Short-Term Memory Networks

LSTM is used to process the temporal correlation of the sequence data. It adopts a unique gating mechanism to effectively manage the flow and memory of information, avoiding the loss of important information [30]. The gating unit of the LSTM consists of an input gate (i_t), forgetting gate (f_t), output gate (o_t) and cell status (C_t). The following equations express the relationship between these four:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \tag{4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t * \tanh(C_t) \tag{9}$$

where x_t is the input at the current moment, h_{t-1} is the output at the previous moment, σ is the sigmoid function, and W, b are the weight and deviation of the neuron respectively. The i_t determines the degree of acceptance of new information, the f_t controls the reservation ratio of historical information, and the o_t filters out critical information for current decision making. This sophisticated gating mechanism enables LSTM to effectively deal with long-term dependencies in sequence data.

D. CNN-Attention-LSTM

The accurate prediction of stock prices is crucial and previous studies have verified the effectiveness of various single models. However, there is still much potential to improve prediction accuracy, especially in the context of model integration and complementarity. Therefore, this study proposes an enhanced stock index prediction method based on the CAL architecture as shown in Fig. 2. And the specific implementation steps are as follows:

Step 1: Input: stock historical data $[T \times F \times 1]$, where T denotes the time step and F denotes the feature dimension;

Step 2: Feature extraction (CNN): local features are extracted from the input data and deep temporal features are gradually extracted through two convolutional layers with an activation function to improve the model expression ability;

Step 3: Attention weighting (Attention): extraction from the last step is fed into the SE Attention module. The importance weights of each channel are calculated and are weighted by fusion with the original features, thereby enhancing the key features and suppressing redundant information;

Step 4: Sequence modelling (LSTM): the weighted feature sequences are input into the LSTM network to further model the long-term dependencies in the time dimension and extract the time series information;

Step 5: Prediction output: the LSTM output is mapped through the fully connected layer to obtain the final predicted price.

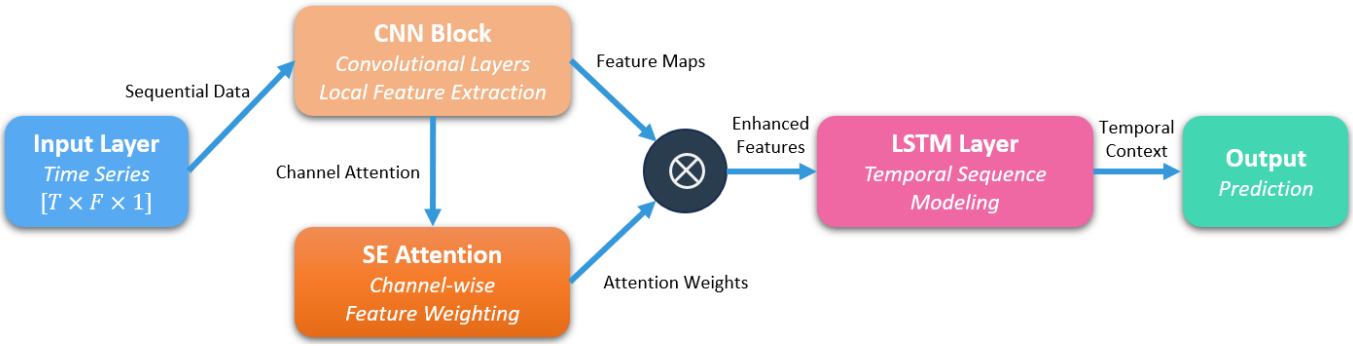


Fig. 2. CAL architecture for stock index prediction.

It can be found that CAL integrates the feature extraction capability of CNN, the channel feature weighting capability of Attention mechanism, and the temporal modelling capability of LSTM. Its innovation lies in integrating different types of neural network layers to form an efficient temporal prediction architecture, which can improve prediction accuracy.

E. Hippopotamus Optimization Algorithm

CAL has a complex hyperparameter space, and the combination of these hyperparameters can significantly affect the prediction accuracy and generalization ability of the model. HO is the newest metaheuristic optimization algorithm proposed by Amiri et al. [12] in 2024. It simulates the dynamic response strategy of hippos under different environmental conditions, which is an SI algorithm based on stochastic search. The HO algorithm consists of the following three stages:

The initialization of HO involves the generation of stochastic initial solutions, and the decision variables are

$$x_{ij} = lb_j + r * (ub_j - lb_j), i = 1,2,\cdots,N; j = 1,2,\cdots,m \quad (10)$$

where r is a random number within $(0,1)$, lb_j and ub_j are the upper and lower bounds of the jth dimension respectively.

Phase I: Position updating (Exploration). The dominant hippo individual in the group (the optimal solution) guides the other individuals to update positions, which can be expressed as

$$x_{ij}^{M} = x_{ij} + y_1 \cdot (D_{hippo} - I_1 \cdot x_{ij}) \quad (11)$$

where x_{ij} denotes the position of the ith hippo in the jth dimension, D_{hippo} is the position of the dominant male hippo, y_1 is a random number within $[0,1]$, and I_1 represents the distance reduction factor. The position update formula for a female or immature hippos is as follows:

$$x_{ij}^{FM} = \begin{cases} x_{ij} + h_1 \cdot (D_{hippo} - I_2 \cdot MG_i), & T > 0.6 \\ \Delta, & \text{else} \end{cases} \quad (12)$$

$$\Delta = \begin{cases} x_{ij} + h_2 \cdot (MG_i - D_{hippo}), & r_6 > 0.5 \\ lb_j + r_7 (ub_j - lb_j), & \text{else} \end{cases} \quad (13)$$

$$T = \exp\left(-\frac{t}{\text{Maxiteration}}\right) \quad (14)$$

where I_2 is an integer between $[1,2]$, MG_i denotes the average position of some randomly selected hippos from the population, r_6 and r_7 are random numbers between $[0,1]$, and h_1 and h_2 are randomly selected numbers. The following equation describes the update of the position of hippos in the population:

$$x_i = \begin{cases} x_i, & F_i^M < F_i \\ x_i, & \text{else} \end{cases} \quad (15)$$

Phase II: Defence against predators (Exploration). Avoiding the algorithm falling into a local optimal solution by simulating the hippos' defence strategy when threatened by a predator. The predator position is denoted by

$$\text{Predator}_j = lb_j + \mathbf{r}_8 \cdot (ub_j - lb_j) \quad (16)$$

where $\mathbf{r}_8$ is a random vector within $[0,1]$. The hippos respond to a predator as

$$x_{ij}^{HippoR} = \begin{cases} \mathbf{RL} \oplus \text{Predator}_j + \frac{f}{c - d \times \cos(2\pi g)} \cdot \frac{1}{\mathbf{D}}, & F_{\text{Predator}_j} < F_i \\ \mathbf{RL} \oplus \text{Predator}_j + \frac{f}{c - d \times \cos(2\pi g)} \cdot \frac{1}{\mathbf{D} + \mathbf{r}_9}, & \text{else} \end{cases} \quad (17)$$

where **RL** is a random vector with a Lévy distribution to model the fast-acting behaviour of hippos during attacks, and f, c, d, and g are random numbers to control the intensity and direction of the hippos' response. The position of the ith hippo is updated by

$$x_i = \begin{cases} x_i^{HippoR}, & F_i^{HippoR} < F_i \\ x_i, & \text{else} \end{cases} \quad (18)$$

Phase III: Escaping from predator (Exploitation). Focus on efficient localized search in the solution space to improve the accuracy and quality of the solution. Hippos escape position update:

$$x_{ij}^{hippo\varepsilon} = x_{ij} + r_{10} \cdot \left(lb_j^{local} + s_1 \cdot \left(ub_j^{local} - lb_j^{local}\right)\right) \quad (19)$$

$$lb_j^{local} = \frac{lb_j}{t}, ub_j^{local} = \frac{ub_j}{t} \quad (20)$$

Fig. 3 shows the specific process of HO adjustment of the CAL hyperparameters. HO begins by randomly initializing a population of hippo individuals, where each represents a set of CAL hyperparameters. In the fitness evaluation section, each

individual's hyperparameter configuration was used to train the CAL, with the RMSE on the validation set serving as the fitness value. The algorithm then updates each individual's position according to HO principles, effectively adjusting the CAL hyperparameters. This process iterates until the convergence criteria are reached, and the optimal hyperparameters that minimize the prediction error are obtained.

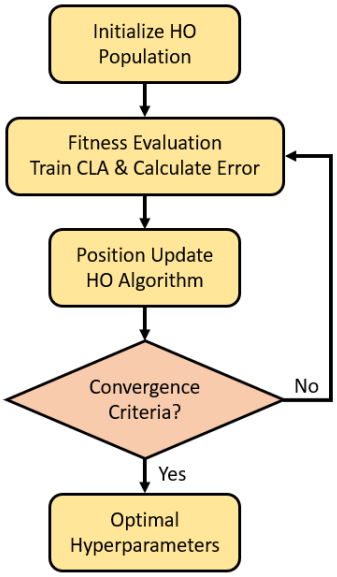


Fig. 3. CAL hyperparameter optimization via the HO algorithm.

IV. Experimental Process

For stock index prediction evaluation, all experiments were conducted on a Windows 11 system using a CPU (13th Gen Intel Core i7-13700H) and GPU (NVIDIA GeForce RTX 3050 4GB). And the prediction models were implemented in MATLAB (R2024b).

A. Data Pre-Processing

This study targets the CSI 300 Index in China, which is regarded as a trend indicator for the Shanghai and Shenzhen markets. It covers 300 A-shares with large market capitalization and high liquidity. The research period is from January 7, 2019 to December 27, 2024, with 1,451 trading days. The proposed model is based on technical analysis and is premised on the identification of behavioral patterns in historical series. Thus, referring to Paiva et al. [31], Table I summarizes 21 technical indicators were used as inputs for the prediction, including return rate indicators calculated based on opening (O), closing (C), high (H) and low (L) prices, as well as momentum, volatility, and volume related indicators etc.

There is a considerable difference in the scale of the above features. The normalization process helps to speed up the convergence of the loss function, prevent gradient explosion in network training, and improve the computational accuracy [5]. Therefore, this study applies the Min-Max normalization method to transform the data between $[0,1]$, and the calculation process is as follows:

$$\mathbf{X}_{\text{new}} = \frac{\mathbf{X}-\mathbf{X}_{\min}}{\mathbf{X}_{\max}-\mathbf{X}_{\min}} \tag{21}$$

where X is the original technical indicators matrix, X_{max} and X_{min} are the maximum and minimum values of X, respectively, and X_{new} is the new value after normalization.

TABLE I. Summary of Stock Index Prediction Variables

Return Rate Indicators		Technical Indicators
$v_1 = ln\left(\frac{C_t}{C_{t-1}}\right)$	$v_8 = ln\left(\frac{H_{t-1}}{O_{t-1}}\right)$	Momentum (close price, period = 10)
$v_2 = ln\left(\frac{C_{t-1}}{C_{t-2}}\right)$	$v_9 = ln\left(\frac{H_{t-2}}{O_{t-2}}\right)$	Relative strength index (close price, period = 14)
$v_3 = ln\left(\frac{C_{t-2}}{C_{t-3}}\right)$	$v_{10} = ln\left(\frac{H_{t-3}}{O_{t-3}}\right)$	Parabolic SAR (high and low price, acceleration = 0, maximum = 0)
$v_4 = ln\left(\frac{H_t}{O_t}\right)$	$v_{11} = ln\left(\frac{L_t}{O_t}\right)$	Average true range (high, low and close price, period = 14)
$v_5 = ln\left(\frac{H_t}{O_{t-1}}\right)$	$v_{12} = ln\left(\frac{L_{t-1}}{O_{t-1}}\right)$	True range (high, low, and close price)
$v_6 = ln\left(\frac{H_t}{O_{t-2}}\right)$	$v_{13} = ln\left(\frac{L_{t-2}}{O_{t-2}}\right)$	Chaikin A/D line (high, low, and close price; volume)
$v_7 = ln\left(\frac{H_t}{O_{t-3}}\right)$	$v_{14} = ln\left(\frac{L_{t-3}}{O_{t-3}}\right)$	On balance volume (close price, volume)

B. Sliding Window and Evaluation Metrics

This study uses the technical indicators of the past 20 days to predict the closing price of the next day. And a sliding window approach was conducted for the six-year index data: the first three years of data were used to train the model; the following year was used as a validation set for optimally tuning the CAL hyperparameters; and the subsequent year is used as a test set to examine the prediction ability. And each completed cycle slides forward for one year, ultimately obtaining daily predicted closing price for 2023 and 2024.

In order to evaluate the model prediction performance, this research employed six evaluation metrics. They are Mean Absolute Error, Mean Absolute Percentage Error, Mean Square Error, RMSE, R^2 and Relative Error, which are calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N} |r_t - \hat{r}_t| \tag{22}$$

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N} \left|\frac{r_t-\hat{r}_t}{r_t}\right| \tag{23}$$

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N} (r_t - \hat{r}_t)^2 \tag{24}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (r_t - \hat{r}_t)^2} \tag{25}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{N} (r_t-\hat{r}_t)^2}{\sum_{t=1}^{N} (r_t-\bar{r})^2} \tag{26}$$

$$\text{Relative Error} = \frac{\hat{r}_t-r_t}{r_t} \tag{27}$$

where r_t, $\hat{r}_t$ represent actual close price and predictive close price at time t respectively.

V. Experimental Results and Discussion

This section presents and discusses several experimental evaluations of the proposed HO-CAL framework through systematic analysis across multiple perspectives, including ablation studies, optimization effectiveness comparisons, and robustness across international stock markets. Beyond the

quantitative outcomes, we also contextualize these findings by comparing them with existing literature and by evaluating how the model addresses the challenges identified in earlier sections.

A. Ablation Experiment

In order to verify the effectiveness of the proposed CAL model for short-term stock index prediction, Table II presents a comparison of the ablation results against several baseline and state-of-the-art methods, including econometric (ARIMA), machine learning (RF), and deep learning (CNN, LSTM, CNN-LSTM).

TABLE II. COMPARISON OF ABLATION RESULTS FOR CAL MODEL

Model	MAE	RMSE	MAPE(%)	R^2
ARIMA	122.86	166.63	3.41	0.897
RF	101.32	139.78	2.75	0.923
CNN [16]	82.04	129.96	2.20	0.947
LSTM [17]	69.25	109.43	1.89	0.957
CNN-LSTM [18]	59.73	91.82	1.49	0.971
CAL	52.41	80.57	1.31	0.978

Among the baseline models, ARIMA shows the worst performance, with an MAE of 122.86 and R^2 of 0.897, indicating its limited capacity to capture the nonlinear characteristics of stocks. In contrast, the RF model significantly improved the prediction accuracy (MAE: 101.32, R^2: 0.923), highlighting that machine learning methods are better suited for modelling the complex relationships among technical indicators. Meanwhile, CNN achieved an MAE of 82.04, whereas LSTM further reduced it to 69.25, reflecting the effectiveness of temporal modelling in financial time series. All all deep learning models outperformed RF, underscoring their enhanced learning capability and adaptability. The combination of CNN-LSTM resulted in a substantial boost in performance (MAE: 59.73, R^2: 0.971), confirming prior findings that hybrid architectures can integrate complementary strengths of individual components [18], [22].

Building on these findings, the ablation study clearly demonstrates that the CAL model delivered the most accurate and reliable performance, with the lowest MAE (52.41), RMSE (80.57), MAPE (1.31%), and the highest R^2 (0.978). This proves that the addition of the Attention mechanism dynamically allocates weights to important time-dependent features—addressing the static limitation in CNN-LSTM architectures, as previously highlighted by Abbasimehr and Paki [20]. These results align with [21] and [24], who reported that attention-based models significantly improve prediction accuracy by enhancing the model's ability to focus on relevant input segments.

B. Model Optimization Effectiveness Comparison

After proving the validity of the CAL model, HO was used to optimize CNN-LSTM and CAL for comparison. Fig. 4 shows the hyperparameter tuning process on the validation set, and it can be seen the HO2 (belonging to CAL) demonstrates markedly superior performance from the initial iteration, with a starting fitness value of 72.35 compared to HO1 (belonging to CNN-LSTM) initial value of 87.29. Throughout the optimization process, HO2 consistently maintains lower fitness values than HO1, ultimately achieving a final fitness value of 50.41 compared to HO1's 55.73, demonstrating the superior optimization ability of HO algorithm. This 10.55% improvement, as well as the consistent and substantial performance differential between the curves, provide strong empirical evidence that the CAL architecture possesses inherently better representational capabilities owing to its Attention mechanisms. Remarkably, both models exhibited rapid convergence within 6 iterations without extensive computational overheads. And the steep descent during the first two iterations followed by gradual stabilization highlights HO's exceptional search efficiency and ability to effectively balance exploration and exploitation phases, which proves that HO is particularly well suited for CAL hyperparameter optimization tasks.

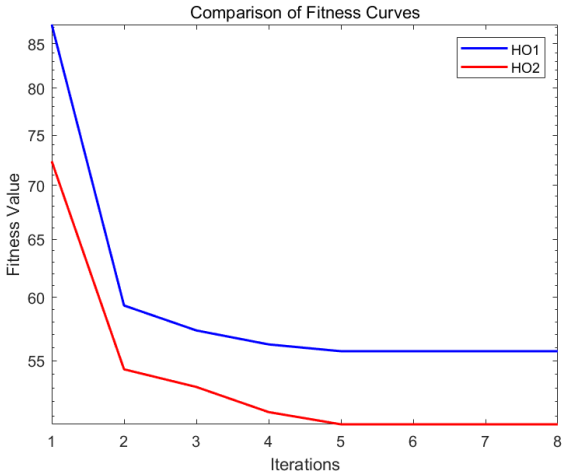


Fig. 4. Convergence analysis of HO for CNN-LSTM and CAL.

The training period spanning from 2019 to 2022 in Fig. 5 demonstrate exceptional model convergence, with all models exhibiting remarkably close alignment with the actual price trajectory. Notably, the models successfully capture major market movements, including the significant upward trend from early 2020 to mid-2021 (reaching peaks around 5,700) and the subsequent downward correction through 2022. The observable differences on the training set are minimal, suggesting that all models have sufficient price recognition capabilities and all have learned effectively.



Fig. 5. Prediction performance comparison on training set (2019-2022).

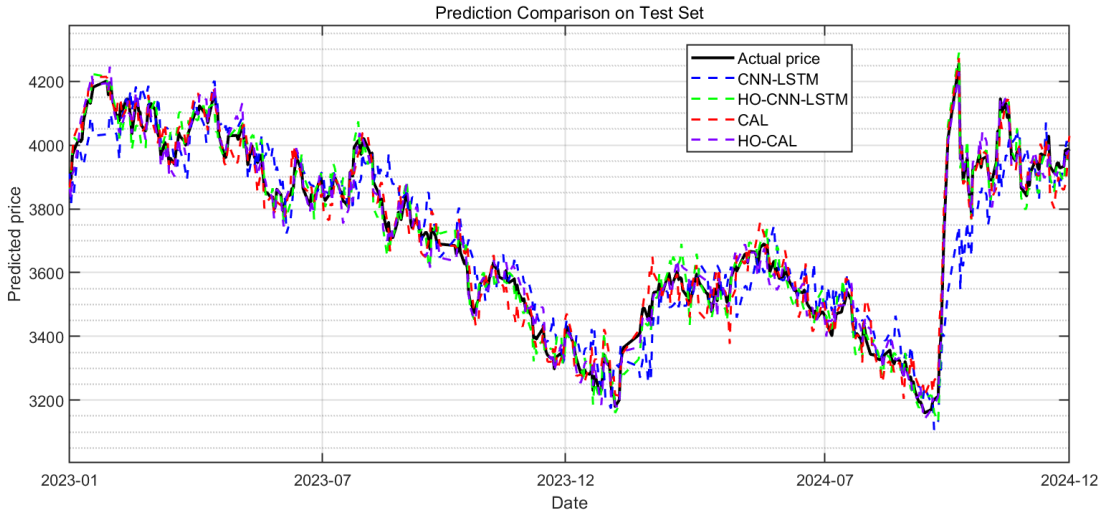


Fig. 6. Prediction performance comparison on testing set (2023-2024).

The testing results for 2023-2024 in Fig. 6 reveal more pronounced performance differentials. The HO-CAL demonstrates superior tracking accuracy, maintaining the closest alignment with actual price movements throughout the testing period. Particularly notable is the model's performance during high-volatility periods, such as the sharp decline in mid-2023 (from ~4,000 to ~3,200) and the subsequent market recovery phases, which captures such extreme events in line with the study in [23]. The CNN-LSTM baseline shows the most significant deviations, particularly during rapid price transitions, while the HO-CNN-LSTM demonstrates improved stability over its non-optimized counterpart. The CAL performs moderately well but lacks the precision achieved by its HO-optimized version. The hyperparameter optimization significantly enhances prediction accuracy, HO-CNN-LSTM and HO-CAL showing reduced prediction volatility and better trend-following capabilities.

The comprehensive performance evaluation presented in Table III demonstrate a clear hierarchical structure. Specifically, HO-CAL attains the lowest MAE of 40.03, representing substantial improvements of 33.0% over CNN-LSTM, 13.5% over HO-CNN-LSTM, and 23.6% over CAL. And the HO algorithm's contribution is consistently significant, with HO-CNN-LSTM showing 19.7% RMSE improvement over CNN-LSTM, and HO-CAL achieving 22.0% improvement over non-optimized CAL. These results confirm the complementary value of model architecture (CAL) and optimization strategy (HO). In line with Maurya et al. [14], who validated HO's superior performance in engineering optimization problems, our results highlight its promising extension to stock prediction.

TABLE III. HO EFFECTIVENESS TEST

Model	MAE	RMSE	MAPE(%)	R^2
CNN-LSTM [18]	59.73	91.82	1.49	0.971
HO-CNN-LSTM	46.29	73.69	1.18	0.984
CAL	52.41	80.57	1.31	0.978
HO-CAL	40.03	62.86	0.99	0.991

The visualization in Fig. 7 provides an intuitive multi-dimensional performance comparison, clearly illustrating the dominance of HO-CAL across all metrics. The radar chart reveals that HO-CAL (represented by the innermost polygon for error metrics and outermost for R^2) consistently achieves optimal or near-optimal performance boundaries. Notably, the visualization emphasizes the substantial performance gap between the optimized and non-optimized models, with both HO-enhanced variants showing marked improvements over their baseline counterparts.

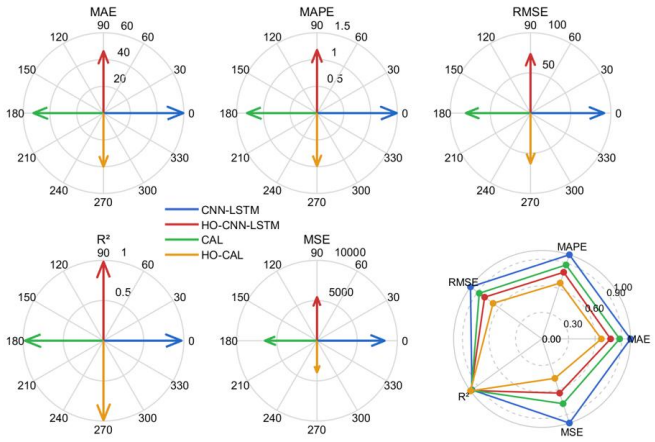


Fig. 7. Multi-dimensional performance evaluation.

The comparison in Fig. 8 provides crucial insights into the temporal distribution and magnitude of relative errors across four models. The error distribution around the zero baseline indicates minimal systematic bias across all models, with relatively balanced positive and negative errors suggesting unbiased prediction behavior. However, the frequency and magnitude of extreme errors vary significantly. The CNN-LSTM exhibiting the most pronounced error volatility, displaying frequent extreme deviations reaching $\pm 4\%$. These large-magnitude errors are particularly concentrated during periods of high market volatility, indicating the baseline model's limited capacity to handle rapid price fluctuations. In contrast, HO-CAL demonstrates remarkably constrained error boundaries, with the majority of errors confined within $\pm 2\%$ range, representing a 50% reduction in maximum error magnitude compared to CNN-LSTM. The chronological analysis reveals that HO-CAL maintains consistent error stability throughout the entire testing period, with notably reduced error spikes during critical market transitions observed in mid-2023 and late 2024. The HO-CNN-LSTM shows intermediate performance, exhibiting improved stability over the baseline but lacking the precision consistency achieved by Attention-based models. CAL demonstrates moderate error reduction compared to CNN-LSTM, though it occasionally exhibits error spikes that are effectively mitigated in its optimized counterpart HO-CAL. The comparison between optimized and non-optimized model pairs clearly demonstrates the HO algorithm's effectiveness in error reduction, with HO-CAL achieving the most significant improvement in error containment.

The HO-CAL framework integrates CAL networks with the HO algorithm, achieving lower error metrics and improved predictive accuracy. It excels at capturing market turning points and maintaining stability during transitions, effectively modelling complex temporal and nonlinear dynamics. Consistent performance gains across metrics highlight both statistical and practical relevance, reducing prediction errors and financial risk. The relative error reduction throughout the testing period confirms the stability of the proposed framework for stock time series prediction applications.

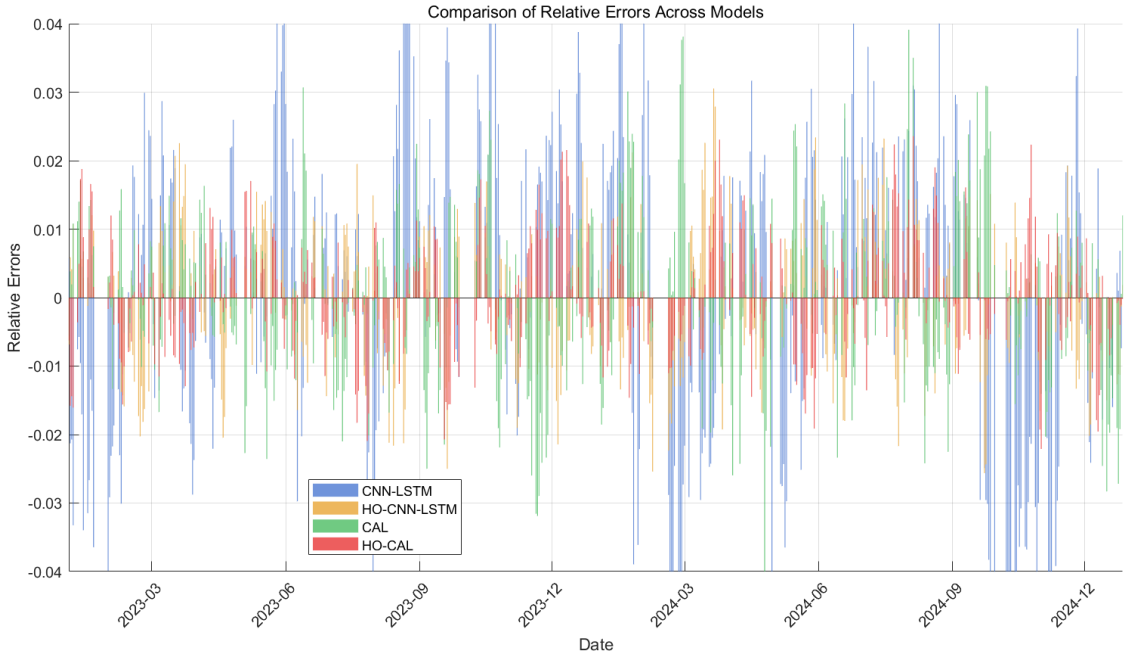


Fig. 8. Temporal distribution of relative prediction errors across models.

C. Robustness Test

To further validate the accuracy and robustness of the HO-CAL framework, this study extends the evaluation to four major international stock indexes: the S&P 500 (Standard & Poor's 500 Index), FTSE 100 (Financial Times Stock Exchange 100 Index), Nikkei 225 (Nikkei 225 Stock Average), and KOSPI (Korea Composite Stock Price Index). These indexes represent a diverse range of market structures and efficiency levels across developed and emerging economies.

TABLE IV. COMPARISON OF PREDICTIONS FOR INTERNATIONAL MARKETS

Stock Index	MAE	RMSE	MAPE(%)	R^2
CSI 300	40.03	62.86	0.99	0.991
S&P 500	75.26	113.85	1.42	0.983
FTSE 100	129.48	194.20	1.56	0.982
Nikkei 225	471.21	706.85	1.24	0.989
KOSPI	28.15	42.75	1.06	0.990

The experimental results in Table IV confirm the robustness and generalizability of the proposed framework. HO-CAL achieved strong predictive performance across all markets, with high R^2 values (≥ 0.982) and consistently low MAE, RMSE, and MAPE values. Notably, KOSPI (MAE: 28.15, R^2: 0.990) and CSI 300 yield the best results, followed by Nikkei 225 (MAE: 471.21, R^2: 0.989), whereas S&P 500 and FTSE 100 showed slightly higher error metrics. The differences between these international markets are consistent with the findings in [32], and can be interpreted through the lens of market efficiency: the S&P 500 (USA) and FTSE 100 (UK) are generally classified as semi-strong to strong-form efficient markets, where price patterns are quickly arbitraged away, making predictive modelling more challenging. In contrast, Nikkei 225 (Japan), KOSPI (Korea) and especially CSI 300 (China), often considered weaker-form efficient or less efficient due to higher market volatility and behavioral factors, may retain exploitable patterns that the HO-CAL framework can effectively capture. This also suggests that deep learning methods like HO-CAL, tend to perform better in markets with lower informational efficiency. These results underscore the practical value of the proposed hybrid framework in real-world financial predictions, particularly in scenarios that involve diverse data distributions and varying market dynamics.

VI. CONCLUSION AND FUTURE WORK

Stock markets are crucial to global financial systems, and accurate stock index prediction remains a challenging yet valuable task for investors, policymakers, and researchers. This study proposed a novel hybrid HO-CAL framework for enhance the predictive accuracy and generalizability of stock index. Through comprehensive experiments, the HO-CAL achieved consistently superior performance across multiple evaluation metrics (e.g., MAE, RMSE, R^2), outperforming baseline models including CNN, LSTM, CNN-LSTM, and even CNN-LSTM-Attention variants. The Attention mechanism significantly enhanced the model's ability to focus on key temporal patterns, while HO effectively optimized parameters to avoid local minima. Ablation studies and visual analyses confirmed the individual and synergistic contributions of each component. In robustness tests across five major global stock markets, HO-CAL maintained high predictive accuracy, demonstrating strong generalization ability even in markets with differing volatility and efficiency levels. These results validate the model's applicability in real-world investment and confirm its alignment with behavioral finance perspectives.

From a theoretical standpoint, this research contributes to the growing body of literature on deep learning in financial prediction by (1) empirically demonstrating the benefits of attention-enhanced feature selection in volatile stock series, (2) proposing the first integration of the CAL architecture with the HO algorithm, and (3) verifying the feasibility of metaheuristic optimization in tuning deep learning models for financial applications. Practically, the proposed framework offers a scalable and accurate solution that can assist investors in managing portfolio risk and timing market entry more effectively.

Nevertheless, this study has certain limitations: the model relies exclusively on technical indicators, without incorporating macroeconomic or firm-level fundamental variables, which may limit its comprehensiveness in capturing long-term market dynamics; while HO showed strong performance, its computational cost is non-negligible, especially when applied to large-scale or high-frequency datasets. Future work could address these limitations in several ways. First, integrating fundamental financial indicators (e.g., ROA, PE ratios) and macroeconomic variables (e.g., inflation, M2 growth) could improve the model's explanatory power. Second, the model can be extended by exploring other deep learning architectures such as Transformers, Temporal Convolutional Networks, or Graph Neural Networks, which may offer enhanced capabilities in modeling complex dependencies. Finally, comparisons with other powerful metaheuristic optimization, such as Grey Wolf Optimizer or Whale Optimization Algorithm, can further refine the optimization strategy.

REFERENCES

[1] L. Pan and V. Mishra, "Stock market development and economic growth: Empirical evidence from China," Econ. Model., vol. 68, pp. 661–673, Jan. 2018.

[2] Y. Baek and H. Y. Kim, "ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module," Expert Syst. Appl., vol. 113, pp. 457–480, Dec. 2018.

[3] P. Lv, Y. Shu, J. Xu, and Q. Wu, "Modal decomposition-based hybrid model for stock index prediction," Expert Syst. Appl., vol. 202, p. 117252, Sep. 2022.

[4] V. Palaniappan, I. Ishak, H. Ibrahim, F. Sidi, and Z. A. Zukarnain, "A Review on High Frequency Trading Forecasting Methods: Opportunity and Challenges for Quantum based Method," IEEE Access, vol. 12, Jun. 2024.

[5] J. Zhang and Y. Lai, "Stock Price Prediction Using CNN-BiLSTM-Attention Model," Mathematics, vol. 11, no. 9, p. 1985, Mar. 2023.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," PLoS One, vol. 13, no. 3, p. e0194889, Mar. 2018.

[7] I. H. Sarker, "Deep Learning: a Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," SN Comput. Sci., vol. 2, no. 6, Aug. 2021.

[8] J. Shah, D. Vaidya, and M. Shah, "A comprehensive review on multiple hybrid deep learning approaches for stock prediction," Intell. Syst. Appl., vol. 16, p. 200111, Nov. 2022.

[9] S. F. Ahmed et al., "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artif. Intell. Rev., vol. 56, no. 1, Apr. 2023.

[10] N. M. Aszemi and P. D. D. Dominic, "Hyperparameter Optimization in Convolutional Neural Network using Genetic Algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, 2019.

[11] Q. Luo, H. Wang, Y. Zheng, and J. He, "Research on path planning of mobile robot based on improved ant colony algorithm," Neural Comput. Appl., vol. 32, no. 6, pp. 1555–1566, Apr. 2019.

[12] M. H. Amiri, N. M. Hashjin, M. Montazeri, S. Mirjalili, and N. Khodadadi, "Hippopotamus optimization algorithm: a novel nature-inspired optimization algorithm," Sci. Rep., vol. 14, no. 1, Feb. 2024.

[13] G. Zhou, J. Du, J. Guo, and G. Li, "A novel hippo swarm optimization: for solving high-dimensional problems and engineering design problems," J. Comput. Des. Eng., vol. 11, no. 3, pp. 12–42, Apr. 2024.

[14] P. Maurya, P. Tiwari, and A. Pratap, "Application of the hippopotamus optimization algorithm for distribution network reconfiguration with distributed generation considering different load models for enhancement of power system performance," Electr. Eng., vol. 107, pp. 3909–3946, Sep. 2024.

[15] C. L. Liu, W. Hsaio, and Y. C. Tu, "Time Series Classification With Multivariate Convolutional Neural Network," IEEE Trans. Ind. Electron., vol. 66, no. 6, pp. 4788–4797, Aug. 2018.

[16] Z. Deng, B. Wang, Y. Xu, T. Xu, C. Liu, and Z. Zhu, "Multi-Scale Convolutional Neural Network With Time-Cognition for Multi-Step Short-Term Load Forecasting," IEEE Access, vol. 7, pp. 88058–88071, 2019.

[17] Q. Li, J. Tan, J. Wang, and H. Chen, "A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News," IEEE Trans. Knowl. Data Eng., vol. 33, no. 10, pp. 3323–3337, Jan. 2020.

[18] Z. Nourbakhsh and N. Habibi, "Combining LSTM and CNN methods and fundamental analysis for stock price trend prediction," Multimed. Tools Appl., vol. 82, pp. 17769–17799, Oct. 2022.

[19] H. Xie, L. Zhang, and C. P. Lim, "Evolving CNN-LSTM Models for Time Series Prediction Using Enhanced Grey Wolf Optimizer," IEEE Access, vol. 8, pp. 161519–161541, Sep. 2020.

[20] H. Abbasimehr and R. Paki, "Improving time series forecasting using LSTM and attention models," J. Ambient Intell. Humaniz. Comput., vol. 13, pp. 673–691, Jan. 2021.

[21] S. Yi, H. Liu, T. Chen, J. Zhang, and Y. Fan, "A deep LSTM-CNN based on self-attention mechanism with input data reduction for short-term load forecasting," IET Gener. Transm. Distrib., vol. 17, no. 7, pp. 1538–1552, Feb. 2023.

[22] Z. Shi, Y. Hu, G. Mo, and J. Wu, "Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction," arXiv (Cornell University), Jan. 2022.

[23] A. Borré, L. O. Seman, E. Camponogara, S. F. Stefenon, V. C. Mariani, and L. dos S. Coelho, "Machine Fault Detection Using a Hybrid CNN-LSTM Attention-Based Model," Sensors, vol. 23, no. 9, pp. 4512–4512, May 2023.

[24] P. Peng, Y. Chen, W. Lin, and J. Wang, "Attention-based CNN–LSTM for high-frequency multiple cryptocurrency trend prediction," Expert Syst. Appl., vol. 237, pp. 121520–121520, Mar. 2024.

[25] X. LÜ et al., "Spatial-temporal simulation and prediction of root zone soil moisture based on Hydrus-1D and CNN-LSTM-Attention in the Yutian Oasis, Southern Xinjiang, China," Pedosphere, Oct. 2024.

[26] N. Mashru, G. G. Tejani, P. Patel, and M. Khishe, "Optimal truss design with MOHO: A multi-objective optimization perspective," PLoS One, vol. 19, no. 8, p. e0308474, Aug. 2024.

[27] A. Zafar et al., "A Comparison of Pooling Methods for Convolutional Neural Networks," Appl. Sci., vol. 12, no. 17, p. 8643, Jan. 2022.

[28] P. Wang, E. Fan, and P. Wang, "Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning," Pattern Recognit. Lett., vol. 141, pp. 61–67, Aug. 2020.

[29] I. C. Fiebelkorn and S. Kastner, "Functional Specialization in the Attention Network," Annu. Rev. Psychol., vol. 71, no. 1, pp. 221–249, Jan. 2020.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[31] F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, and W. M. Duarte, "Decision-making for financial trading: A fusion approach of machine learning and portfolio selection," Expert Syst. Appl., vol. 115, pp. 635–655, Jan. 2019.

[32] Z. Shi, O. Ibrahim, and H. I. C. Hashim, "Stock Index Prediction Using Temporal Convolutional Network and Long Short-Term Memory Network Optimized by Genetic Algorithm," Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications, vol. 16, no. 1, pp. 508–527, Mar. 2025.

The Text Mining Model for Lecturer Performance Evaluation: A Comparative Study

Anita Ratnasari[1]*, Vina Ayumi[2], Mariana Purba[3], Wachyu Hari Haji[4], Handrie Noprisson[5], Marissa Utami[6]
Faculty of Informatics and Engineering, Universitas Dian Nusantara, Indonesia[1, 2, 5]
Faculty of Computer Science, Universitas Sjakhyakirti, Indonesia[3]
Department of Management, Bina Nusantara University, Indonesia[4]
Faculty of Engineering, Universitas Muhammadiyah Bengkulu, Indonesia[6]

***Abstract*—To support the evaluation of the teaching and learning process in higher education institutions, it is necessary to develop a text mining (TM) model. The aim of this research is to compare the performance of Long Short-Term Memory using Word Embedding Text to Sequence (WETS-LSTM), WETS-BiLSTM, WETS-CNN1D, and WETS-RNN, using four dataset categories including pedagogic, professional, personality, and social competency. This research has five main steps, including literature study, dataset collection, TM model development, and evaluation. Dataset is collected from Universitas Sjakhyakirti, Institut Teknologi dan Bisnis Palcomtech, Universitas Muhammadiyah Palembang, Universitas Bina Darma, AMIK Bina Sriwijaya and Politeknik Darusalam. The questionnaire distribution process initially yielded 6,170 responses with 6,164 valid across four competency categories, with total of 24,656 text data for analysis. Model of WETS-LSTM obtained the best performance overall, achieved the train accuracy of 96.65% and the highest test accuracy of 82.92%. The CNN1D with Word Embedding Text to Sequence (WETS-CNN1D) demonstrated good train accuracy with 96.73% but obtained lower test performance with 80.67%. The WETS with Recurrent Neural Network (WETS-RNN) obtained the weakest results, with a train accuracy of 95.88% and a test accuracy of 77.99%.**

***Keywords*—Text mining; CNN; LSTM; RNN; text-to-sequence**

I. INTRODUCTION

The results of the evaluation of lecturer teaching and university services show clear information about the state of a university and the services offered to students [1], [2]. Evaluation serves as a method to assess the quality of education and the efficiency of higher education services which is an aspect of international university ranking assessment [3]–[5]. Through the results of the evaluation, higher education institutions can find out the programs, procedures, and service structures that suit the needs of students [6]–[8].

Usually at the end of each semester, students are asked to fill out a questionnaire designed to gather students' perceptions and feedback about the university's learning experience and facilities in terms of classes, teaching quality, library, and other services [9]–[11]. However, these questionnaires consist of Likert scale or multiple-choice questions (yes or no) that are usually filled out quickly and do not present the actual state of affairs [12], [13].

For this reason, it is necessary to evaluate lecturer teaching and university services using a special system with a feature to store assessment feedback in the form of text [14]–[16]. However, the evaluation of qualitative text feedback from students is difficult to do if it is done by manual analysis, especially for large-scale feedback datasets [17]–[19]. In fact, the results of qualitative feedback data analysis can provide valuable insights into the evaluation of lecturer teaching and university services [20]–[22]. To solve this problem, it is necessary to develop an architecture and evaluation model for lecturer teaching and university services based on text mining [23]–[25]. Research on the evaluation model of lecturer teaching and university services based on text mining (TM) has been carried out with various methods in the last four years using the text mining (TM) approach [14], [26]–[30]

The previous works utilized a variety of machine learning models for text analysis, focusing on different linguistic datasets and methodologies. Research by Grönberg et al. (2021) applied Latent Dirichlet Allocation (LDA) for topic modeling in English, while Misuraca et al. (2021) leveraged Support Vector Machine (SVM) for text classification in English as well [26], [27]. Research by Grljević et al. (2022) utilized a combination of Term Frequency-Inverse Document Frequency (TF-IDF) and K-Nearest Neighbor (KNN) techniques on Serbian text data, whereas Ren et al. (2023) explored the application of Word Embedding and Long Short-Term Memory (LSTM) for text analysis in Chinese [28] [29]. Abdi et al. (2023) combined Word Embedding (WE) with Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) for English text classification [30].

Based on research problem and previous works, several advancements were proposed in the development of a text mining-based evaluation system for lecturer teaching and university services. Researchers have focused on exploring various artificial intelligence models to analyze qualitative feedback, recognizing the limitations of traditional manual analysis, especially in large-scale datasets. The aim of this research was to develop and compare various text mining models for evaluating lecturer teaching and higher education services based on large-scale qualitative feedback. By utilizing different machine learning approaches, such as BiLSTM, LSTM, CNN-1D, and Recurrent Neural Network (RNN) with embedding text to sequence (WETS) techniques, the

experiment was conducted to analyze a dataset of 6,164 student questionnaires across four categories, including pedagogic, professional, personality, and social competencies, into positive and negative interpretation result. The goal was to assess the effectiveness of these models for analyzing of lecturer performance and university services feedback by improving the accuracy of proposed model.

II. Related Work

Over the past four years, several studies investigated evaluation models of lecturer teaching and university services using the text mining (TM) approach. These studies employed various machine learning techniques on different linguistic corpora, including LDA for topic modeling and SVM for English text classification, TF-IDF with KNN classification on Serbian text, LSTM on Chinese data, and combinations of Word Embedding (WE) with CNN and BiLSTM for English text classification. The model has been implemented in the dataset of evaluation texts for lecturer teaching and university services can be seen in Table I.

TABLE I. State of the Art

Reference	Model	*Dataset*
[26]	Latent Dirichlet Analysis (LDA)	English
[27]	Support Vector Machine (SVM)	English
[28]	Term frequency–inverse document frequency (TF-IDF) – K-Nearest Neighbor (KNN)	Serbian
[29]	Word Embedding, Long Short-Term Memory (LSTM)	Chinese
[30]	Word Embedding, Convolutional Neural Networks (CNN)-BiLSTM	English
Proposed	Word Embedding Text to Sequence (WETS), LSTM	Indonesia

Based on previous, Grönberg et al. (2021) implemented Latent Dirichlet Allocation (LDA) for topic modeling to extract key themes from student feedback in English, while Misuraca et al. (2021) applied Support Vector Machine (SVM) for text classification in English, providing insights into the strengths and weaknesses of teaching and services. Similarly, Grljević et al. (2022) used a combination of Term Frequency-Inverse Document Frequency (TF-IDF) and K-Nearest Neighbor (KNN) techniques on Serbian text data, revealing valuable patterns and trends in the evaluation of academic services.

Moreover, other researchers, such as Ren et al. (2023), explored model of Word Embedding (WE) and Long Short-Term Memory (LSTM) for analyzing Chinese text data, further enhancing the ability to understand and interpret qualitative feedback and Abdi et al. (2023) advanced this approach by integrating Word Embedding with CNN-BiLSTM. This research proposed a text mining (TM) model that will be applied to the learning system. The proposed new text mining (TM) model comes from a combination of BiLSTM based on Word Embedding Text to Sequence (WETS).

III. Research Methodology

Data collection was based on the distribution of instruments at several universities in the city of Palembang, namely Universitas Sjakhyakirti, Institut Teknologi Bisnis Palcomtech, Universitas Muhammadiyah Palembang, Universitas Bina Darma, AMIK Bina Sriwijaya and Politeknik Darusalam. The questionnaire distribution process initially yielded 6,170 responses. After preprocessing, 6,164 valid questionnaires remained, each containing evaluations across four competency areas: pedagogic, professional, personality, and social. As a result, a total of 24,656 individual opinion data points were obtained for analysis.

The lecturer performance evaluation encompassed four key aspects: pedagogic, professional, personality, and social. The total evaluation included values for the overall performance, with 70% of the total representing the training data, 30% allocated for testing, and classifications for both positive and negative feedback. The composition of research dataset can be seen in Table II.

TABLE II. Research Dataset

Class	Train (70%)	Test (30%)	Total
Pedagogic	4,316	1,848	6,164
Professional	4,316	1,848	6,164
Personality	4,316	1,848	6,164
Social	4,316	1,848	6,164
Total	17,264	7,392	24,656

The initial stage in opinion mining is pre-processing with the aim of preparing text data which will later experience the next text data processing. This stage has five stages carried out, namely data cleansing, case folding, tokenizing, stop-word, tokenizing, and stemming. The summary of each of the five stages can be seen in Fig. 1.

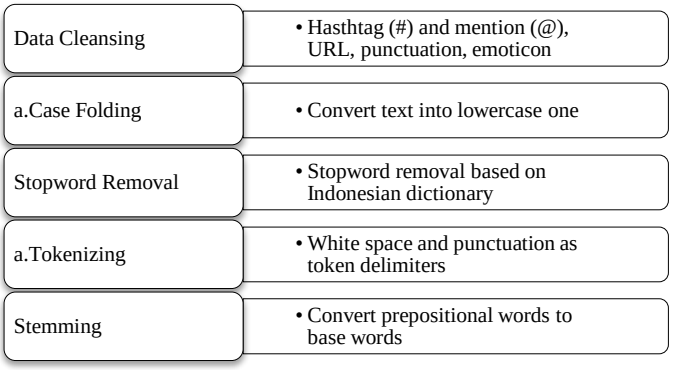


Fig. 1. Preprocessing phase.

At the data cleansing stage, the process of cleaning unnecessary words is carried out to reduce computational load, such as text containing html, links, and scripts. In addition, this stage also removes punctuation marks such as periods (.), commas (,) and also other punctuation marks. In this pre-processing process also applies the case folding method, which is the process of converting words into lower-case. The third stage is tokenization. This method is implemented to transform words contained in text into several sequences truncated by certain spaces or characters.

By applying a word that has a unique word from text data such as conjunctions, and possessive words. Less meaningful types of words will be removed, such as the words: I, and, or by using this method. The purpose of stop word removal is to reduce the burden on system performance, because the words to be taken are the words that are considered important. The last stage in the pre-process process is the stemming stage. The method at this stage is done by transforming the words in the text into basic words.

Next is to conduct experiments for model training and testing. The methods compared include BiLSTM, LSTM, CNN1D and RNN. Each model is carried out performance evaluation measure (PEM) or known as model performance evaluation measurement is an approach that aims to measure the performance or performance of the algorithm model. As for the experiment to be carried out as seen in Fig. 2.

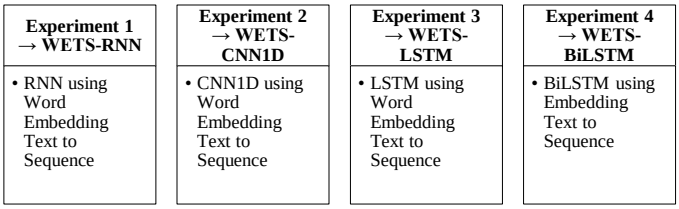


Fig. 2. Research experiment scenario.

The final stage is the evaluation stage that is carried out as a measurement of the performance value of the text mining ™ method that has been implemented. Performance evaluation measurement using confusion matrix, accuracy, precision, F1-score and recall to compare BiLSTM, LSTM, CNN-1D, and RNN with embedding text to sequence techniques performance.

IV. Result and Discussion

The study aimed to develop a text mining model for evaluating lecturer performance, with a specific focus on Indonesian text. Previous research primarily focused on datasets from other languages, such as English, Serbian, and Chinese, which evaluated different competencies. These studies incorporated personality, professional, pedagogical, and social competencies. This research differentiated itself by introducing a tailored evaluation model that includes all four competency categories and fill the gap in existing literature by providing an evaluation framework specifically designed for the Indonesian educational context as explained in Table III.

TABLE III. Dataset Specification

Source & Dataset	Personality	Professional	Pedagogy	Social
English [26]		●	●	
English [27]		●	●	
Serbian [28]		●	●	●
Chinese [29]		●	●	
English [30]		●	●	●
Indonesia [Proposed]	●	●	●	●

The first experiment used RNN using Embedding Text to Sequence (WETS-RNN). The RNN model itself is in the form of a neural network designed to process the text sequence on the learning evaluation opinion. Using WETS is a dataset of learning opinion texts converted into numerical representations so that can be processed by the RNN model. The WETS-RNN model obtained a training accuracy of 0.9588 of the correct predictions which showed that the model was very good at recognizing patterns from the training data from the learning evaluation dataset and the test accuracy of 0.7799 which was lower than the training which showed that the WETS-RNN model was not fully generalized well to the test data.

The second experiment used CNN1D using Word Embedding Text to Sequence (WETS-CNN1D). The CNN1D model is based on a neural network used to process one-dimensional sequence data of student comment texts on learning. By using WETS, the students' comments were converted into vector representations so that the CNN1D model could capture the semantic relationships between words in the dataset text. The WETS-CNN1D model obtained a training accuracy of 0.9673 of the correct predictions indicating that the WETS-CNN1D model was very good at recognizing patterns from the training dataset while the test of 0.8067 showed good performance on the submission dataset being judged correct.

The third experiment used LSTM using Word Embedding Text to Sequence (WETS-LSTM). Model of LSTM model leverages neural networks designed to address sequence problems and long-term dependencies in processing the sequence of opinion and comment texts by capturing temporal dependencies. Then, the WETS is used to convert the order of the text numbers into a lower-dimensional vector that can be processed by the LSTM model. The training accuracy performance of the WETS-LSTM model is very good, which is 0.9665 which means that the model shows excellent performance on the training data, with 96.65% of the predictions correct while the test performance is 0.8292 or 82.92% which shows that the model has good generalization ability to comment text data that is different from the training data.

The fourth experiment used WETS-BiLSTM. The BiLSTM model is a variation of the LSTM that has the ability to process learning comment text data in two directions, namely from start to end and from end to end. The training accuracy performance of the WETS-BiLSTM model obtained a training accuracy of 0.9688 which shows good performance in the training data with 96.88% of the training data prediction results being declared correct. Then, the test accuracy obtained a value of 0.8234 or 82.34%, which is lower than the training accuracy but this value is the best value of the other models.

Based on experimental results, WETS-LSTM has good accuracy among other methods. The WETS -LSTM model is a model based on long short-term memory which can be used in WETS which allows the model to understand the meaning of a sentence based on word order. The results of the comparison of model can be seen in Table IV.

TABLE IV. THE PERFORMANCE COMPARISON OF MODEL

Method	Abbreviation	Train	Test
BiLSTM using Word Embedding Text to Sequence	WETS-BiLSTM	96.88%	82.34%
LSTM using Word Embedding Text to Sequence	WETS-LSTM	**96.65%**	**82.92%**
CNN1D using Word Embedding Text to Sequence	WETS-CNN1D	96.73%	80.67%
RNN using Word Embedding Text to Sequence	WETS-RNN	95.88%	77.99%

The initial experiment commenced with tokenization, where sentences were segmented into individual words. Each word was then assigned an index from a predefined vocabulary. For example, the sentence "Indonesian language" was divided into "Indonesian" and "language," which were mapped to specific indices within the vocabulary. This mapping transformed the text data into numerical form as part of the WETS phase. The WETS phase improved the performance of data processing and classification by utilizing models such as BiLSTM, LSTM, CNN1D, and RNN, with performance results, particularly accuracy, depicted in Fig. 3.

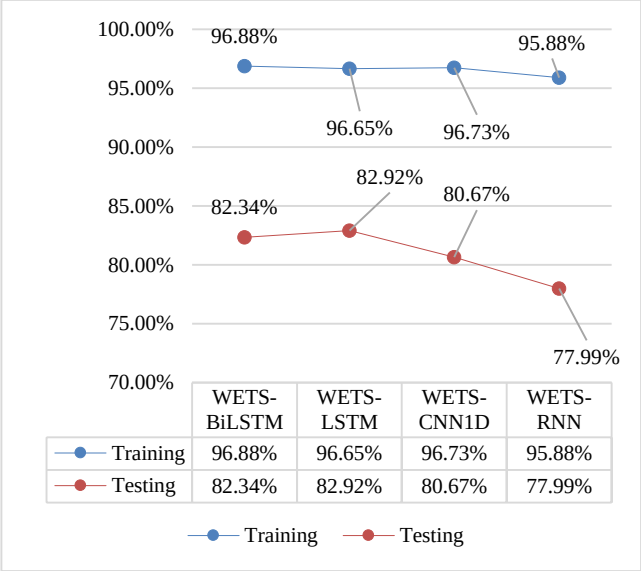


Fig. 3. Model performance results comparison.

Based on the experiment, WETS-LSTM was selected as the preferred text mining (TM) model for lecturer performance evaluation. The phase WETS involved padding to ensure uniformity in the input sequence length, as the model required inputs of a fixed length. In the given example, a sentence of length 2 was padded with zeros to extend the sequence to a length of 4. After padding, the sentence was represented as [1, 2, 0, 0]. The padded sequence was then processed through an embedding layer, where each token (word index) was mapped to a dense vector of fixed dimensions. This transformation converted the tokenized sequence into numerical representations suitable for the WETS-LSTM. The complete phase and result of data processing using WETS can be seen in Fig. 4.

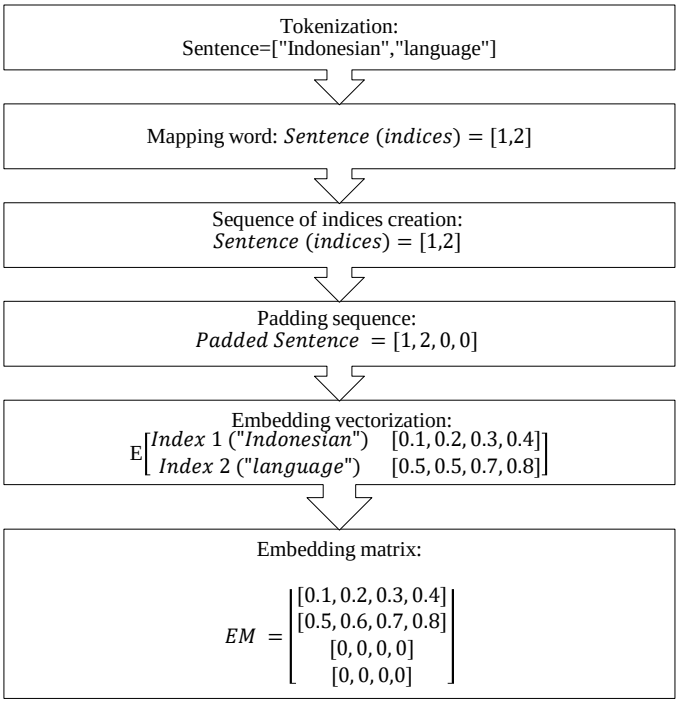


Fig. 4. Word Embedding Text to Sequence (WETS).

The input was processed through the LSTM, which operated on sequences and is capable of retaining long-term dependencies due to its forget f_i, input i_i, cell state c_i, and output gates o_l. At each time step t, the LSTM processed the current word embedding x_i, the previous h_{i-1} hidden state and the previou c_{i-1} cell state. The forget gate determined which information from the previous cell state should be weighted W_f, allowing the model to decide what aspects of the past memory to retain or forget for the current time step by using sigmoid activation function σ and bias term b_f. The explanation of LSTM can be expressed as in Eq. (1).

$$f_i = \sigma(W_f.[h_{i-1}, x_i] + b_f \quad (1)$$

$$h_0 = [0,0,0,0],\ c_0 = [0,0,0,0]$$

$$f_i = [0.6, 0.7, 0.5, 0.4]$$

Assuming that after calculating the forget gate at t=1 for the first word "Indonesian," the next step involved the input gate and the $\widetilde{C}_l$ cell candidate. The i_i input gate determined the amount of new information to be incorporated into the cell state, while the cell candidate generated potential new values for updating the cell state. The calculation of i_i and $\widetilde{C}_l$ for *t=1* can be expressed as in Eq. (2) and Eq. (3).

$$i_i = \sigma(W_i.[h_{i-1}, x_i] + b_i = [0.4, 0.3, 0.2, 0.1], \quad (2)$$

$$\widetilde{C}_l = tanh(W_c.[h_{i-1}, x_i] + b_c = [0.5, 0.4, 0.6, 0.7] \quad (3)$$

At *t=1*, for the first word "Indonesian," the cell state c_i was updated by combining the results of the forget gate f_i and the input gate i_i. This update process allowed the model to decide which parts of the previous cell state should be retained and which new information should be incorporated. The o_l output gate then determined which part of the updated cell

state would be output as the hidden state. The h_l hidden state captured the relevant information from the current word and the context provided by the preceding words. The calculation of o_l and $\widetilde{h_l}$ for *t=1* can be expressed as in Eq. (4) and Eq. (5).

$$o_l = \sigma(W_o.[h_{i-1}, x_i] + b_o = [0.5, 0.6, 0.7, 0.8] \quad (4)$$

$$h_l = o_l.tanh\,(c_l) = [0.0985, 0.0714, 0.0833, 0.056] \quad (5)$$

This process was repeated for the second word "language" at *t=2* based value of embedding matrix. The previous hidden state $h_{l-1}$and cell state c_{l-1} were passed forward, and the new word embedding x_l for the current word was processed. Model LSTM generated a prediction based binary classification, which distinguished between positive and negative lecturer evaluations from the LSTM will be passed through a sigmoid activation function. This function will output a probability between 0 and 1. The sigmoid function is defined as Eq. (6).

$$\sigma(y) = \frac{1}{1+e^{-y}} \quad (6)$$

The output will be a value σ(y) between 0 and 1. If $\sigma(y)$ is greater than or equal to 0.5, the feedback is classified as positive. Otherwise, the data is classified as negative, distinguishing between positive and negative lecturer evaluations. Based on this classification, the accuracy results were calculated for each category. The accuracy per epoch for the social feedback data was presented in Fig. 5.

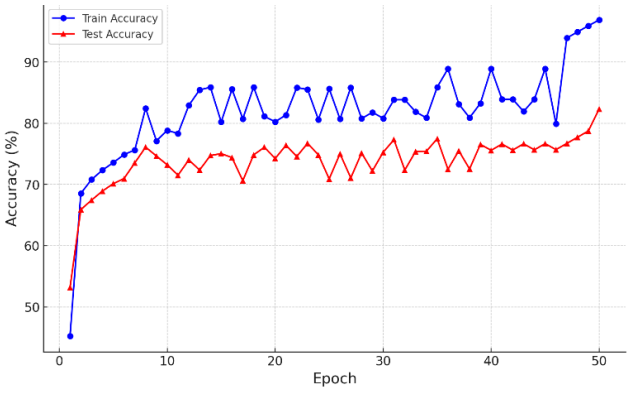


Fig. 5. Accuracy WETS-LSTM for dataset of social feedback.

The data used for the WETS-LSTM classification consisted of four distinct categories, i.e. pedagogic, professional, personality, and social, with each category having positive and negative samples. For the pedagogic category, the training set contained 2,389 positive and 1,927 negative samples, while the testing set had 1,024 positive and 824 negative samples. In the Professional category, the training data included 2,500 positive and 1,816 negative samples, with 946 positive and 902 negative samples in the testing data. The dataset of social category had 2,307 positive and 2,009 negative samples for training. Accuracy results were computed for each category with the detail of dataset as detailed in Table V.

TABLE V. THE DATASET FOR WETS-LSTM EVALUATION

Category	Train		Test	
	Positive	Negative	Positive	Negative
Pedagogic	2,389	1,927	1,024	824
Professional	2,500	1,816	946	902
Personality	2,290	2,026	975	873
Social	2,307	2,009	967	881
Total	7,486	7,778	3,912	3,480

The text mining model for evaluating lecturers calculated the total accuracy as the average of accuracy of each dataset category. The pedagogic accuracy obtained 85.00%, the professional accuracy obtained 80.00%, the personality accuracy obtained 81.00%, and the social accuracy obtained 83.36%. The total average accuracy was determined by calculating the mean of all dataset category accuracy, as in Eq. (7).

$$acc_{avg} = \frac{acc_{pedagogic} + acc_{professional} + acc_{personality} + acc_{social}}{4} \quad (7)$$

$$acc_{avg} = \frac{85.00\% + 80.00\% + 81.00\% + 83.36\%}{4} = 82.34\%$$

The calculation was applied to other models to assess the testing accuracy. The accuracy results for each model were then obtained. The BiLSTM using Word Embedding Text to Sequence (WETS-BiLSTM) model achieved an accuracy of 82.34%. The LSTM using Word Embedding Text to Sequence (WETS-LSTM) model performed slightly better, with an accuracy of 82.92%. Additionally, the CNN1D using Word Embedding Text to Sequence (WETS-CNN1D) and RNN using Word Embedding Text to Sequence (WETS-RNN) models achieved accuracy rates of 80.67% and 77.99%, respectively.

V. CONCLUSION

The aim of this research was to compare the performance of WETS-BiLSTM, WETS-LSTM, WETS-CNN1D, and WETS-RNN and involved four dataset categories, i.e. pedagogic, professional, personality, and social. The WETS-LSTM model exhibited the best performance overall, with a slightly lower training accuracy of 96.65%, but it achieved the highest test accuracy of 82.92%. The WETS-CNN1D model demonstrated good training accuracy with 96.73% but exhibited lower test performance, achieving 80.67%. The WETS-RNN produced the weakest results, with a training accuracy of 95.88% and a testing accuracy of 77.99%. While WETS-LSTM methods proved to be the most effective for both training and testing, WETS-RNN showed less satisfactory performance. Further analysis is required to improve the robustness of the proposed WETS-LSTM on larger datasets, and while the current study relied on limited data from a few private universities, efforts were undertaken to collect data from additional universities for more comprehensive analyses.

ACKNOWLEDGMENT

Author thanks to Universitas Dian Nusantara and Direktorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia through research funding with 105/E5/PG.02.00.PL/2024;792/LL3/AL.04/2024,11/133/HSP K/VI/2024.

REFERENCES

[1] M. Purba et al., "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, 2022.

[2] J. R. Hanaysha, F. B. Shriedeh, and M. In'airat, "Impact of classroom environment, teacher competency, information and communication technology resources, and university facilities on student engagement and academic performance," Int. J. Inf. Manag. Data Insights, vol. 3, no. 2, p. 100188, 2023.

[3] M. A. Fauzi, C. N.-L. Tan, M. Daud, and M. M. N. Awalludin, "University rankings: A review of methodological flaws," Issues Educ. Res., vol. 30, no. 1, pp. 79–96, 2020.

[4] L. A. Rozak et al., "Empirical evaluation of educational service quality in the current higher education system," Emerg. Sci. J., vol. 6, no. Special Issue, pp. 55–77, 2022.

[5] H. Noprisson, "A Survey of the Online Learning Implementation During COVID-19 Outbreak," Int. J. Recent Contrib. from Eng. Sci. IT, vol. 8, no. 4, p. 18, Dec. 2020.

[6] M. Purba, E. Ermatita, A. Abdiansah, V. Ayumi, H. Noprisson, and A. Ratnasari, "A Systematic Literature Review of Knowledge Sharing Practices in Academic Institutions," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021, pp. 337–342.

[7] V. Kaushal, D. Jaiswal, R. Kant, and N. Ali, "Determinants of university reputation: conceptual model and empirical investigation in an emerging higher education market," Int. J. Emerg. Mark., vol. 18, no. 8, pp. 1846–1867, 2023.

[8] B. Thornhill-Miller et al., "Creativity, critical thinking, communication, and collaboration: assessment, certification, and promotion of 21st century skills for the future of work and education," J. Intell., vol. 11, no. 3, p. 54, 2023.

[9] P. Warfvinge, J. Löfgreen, K. Andersson, T. Roxå, and C. Åkerman, "The rapid transition from campus to online teaching–how are students' perception of learning experiences affected?," Eur. J. Eng. Educ., vol. 47, no. 2, pp. 211–229, 2022.

[10] F. M. Van der Kleij and A. A. Lipnevich, "Student perceptions of assessment feedback: A critical scoping review and call for research," Educ. assessment, Eval. Account., vol. 33, pp. 345–373, 2021.

[11] H. Singh, M.-H. Tayarani-Najaran, and M. Yaqoob, "Exploring computer science students' perception of ChatGPT in higher education: A descriptive and correlation study," Educ. Sci., vol. 13, no. 9, p. 924, 2023.

[12] Y.-J. Seo and K.-H. Um, "The role of service quality in fostering different types of perceived value for student blended learning satisfaction," J. Comput. High. Educ., vol. 35, no. 3, pp. 521–549, 2023.

[13] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," Artif. Intell. Rev., vol. 56, no. 9, pp. 10345–10425, 2023.

[14] A. H. Wani and R. Hashmy, "A supervised multinomial classification framework for emotion recognition in textual social data," Int. J. Adv. Intell. Paradig., vol. 24, no. 1–2, pp. 173–189, 2023.

[15] T. Shaik et al., "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," Ieee Access, vol. 10, pp. 56720–56739, 2022.

[16] N. A. Dahri, W. M. Al-Rahmi, A. S. Almogren, N. Yahaya, M. S. Vighio, and Q. Al-Maatuok, "Mobile-based training and certification framework for teachers' professional development," Sustainability, vol. 15, no. 7, p. 5839, 2023.

[17] A. V. Y. Lee, "Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation," Stud. Educ. Eval., vol. 77, p. 101250, 2023.

[18] J. P. Bernius, S. Krusche, and B. Bruegge, "Machine learning based feedback on textual student answers in large courses," Comput. Educ. Artif. Intell., vol. 3, p. 100081, 2022.

[19] R. Nawaz et al., "Leveraging AI and machine learning for national student survey: actionable insights from textual feedback to enhance quality of teaching and learning in UK's higher education," Appl. Sci., vol. 12, no. 1, p. 514, 2022.

[20] F. Gurcan and N. E. Cagiltay, "Research trends on distance learning: a text mining-based literature review from 2008 to 2018," Interact. Learn. Environ., vol. 31, no. 2, pp. 1007–1028, 2023.

[21] M. Á. Escotet, "The optimistic future of Artificial Intelligence in higher education," Prospects, vol. 54, no. 3, pp. 531–540, 2024.

[22] I. Nurhaida, I. Dhamanti, V. Ayumi, F. Yakub, and B. Tjahjono, "Hospital quality classification based on quality indicator data during the COVID-19 pandemic," Int. J. Electr. Comput. Eng., vol. 14, no. 4, pp. 4365–4375, 2024.

[23] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J. A. G. Achem, J. Escamilla, and S. Hosseini, "Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification," Educ. Inf. Technol., pp. 1–43, 2022.

[24] S. Z. Salas-Pilco, K. Xiao, and X. Hu, "Artificial intelligence and learning analytics in teacher education: A systematic review," Educ. Sci., vol. 12, no. 8, p. 569, 2022.

[25] M. E. Dogan, T. Goru Dogan, and A. Bozkurt, "The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies," Appl. Sci., vol. 13, no. 5, p. 3056, 2023.

[26] N. Grönberg, A. Knutas, T. Hynninen, and M. Hujala, "Palaute: An online text mining tool for analyzing written student course feedback," IEEE Access, vol. 9, pp. 134518–134529, 2021.

[27] M. Misuraca, G. Scepi, and M. Spano, "Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback," Stud. Educ. Eval., vol. 68, p. 100979, 2021.

[28] O. Grljević, Z. Bošnjak, and A. Kovačević, "Opinion mining in higher education: a corpus-based approach," Enterp. Inf. Syst., vol. 16, no. 5, p. 1773542, 2022.

[29] P. Ren, L. Yang, and F. Luo, "Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis," Educ. Inf. Technol., vol. 28, no. 1, pp. 797–814, 2023.

[30] A. Abdi, G. Sedrakyan, B. Veldkamp, J. van Hillegersberg, and S. M. van den Berg, "Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems," Soft Comput., vol. 27, no. 19, pp. 14073–14094, 2023.

Design and Evaluation of a Biometric IoT-Based Smart Lock System with Real-Time Monitoring and Alert Mechanisms

Jamil Abedalrahim Jamil Alsayaydeh[1]*, Mohd Faizal Yusof[2],
Serhij Mamchenko[3], Rostam Affendi Hamzah[4], Safarudin Gazali Herawan[5]
Department of Engineering Technology-Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK),
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[1, 4]
Department-Research Section-Faculty of Resilience, Rabdan Academy,
65 Al Inshirah Street, Abu Dhabi 22401, United Arab Emirates[2]
Department of Computer Systems-Networks and Cybersecurity,
National University of Life and Environmental Sciences of Ukraine, Kiev, Ukraine[3]
Industrial Engineering Department-Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia 11480[5]

***Abstract*—A smart door lock system based on IoT is presented which uses fingerprint biometric authentication, ESP32 microcontroller and Blynk IoT platform to provide a secure, user friendly and remote controllable access control solution. The proposed architecture replaces traditional locks with a real time biometric system that gives instant feedback through onboard display (OLED) and buzzer and remote monitoring and control through a mobile app. A new fail-safe mechanism is implemented: after 3 failed fingerprint attempts the system will lock out for 15 seconds and send instant alert to the authorized user's smartphone. Performance test of the prototype shows fingerprint recognition time of around 1.0 second and door unlock time of 5 seconds, so it's convenient to use. The system has a very low False Acceptance Rate (FAR) of 1.32% which means strong resistance to unauthorized access. The False Rejection Rate (FRR) is higher (around 26.32%) due to user error such as improper finger placement – so a usability issue to be addressed. The device can store up to 3 fingerprint profiles and gives visual/audible alert for all access events. This integration of IoT with biometric security not only enhances physical security but also user convenience, a modern smart-lock solution for smart home automation.**

Keywords*—Smart door lock; Internet of Things (IoT); biometric authentication; fingerprint sensor; ESP32 microcontroller; Blynk IoT platform; access control; cybersecurity; real-time monitoring; home automation; OLED display; False Rejection Rate (FRR); False Acceptance Rate (FAR); remote monitoring; smart home automation

I. INTRODUCTION

The Internet of Things (IoT) has changed the physical security landscape, bringing intelligent systems that automate tasks and enable remote management, monitoring and interaction. Access control has shifted from traditional methods—mechanical keys, RFID cards or PIN codes—to more secure, automated and user friendly alternatives. Traditional lock-and-key or keypad systems are vulnerable to theft, key duplication, password disclosure and human error and are no longer sufficient for modern security expectations. Among the emerging solutions, biometric fingerprint recognition stands out for its reliability and uniqueness. Unlike passwords or cards that can be shared, lost or guessed, fingerprints are a permanent and non-transferable form of authentication. When integrated into an IoT ecosystem using microcontrollers (like ESP32) and cloud connected platforms like Blynk for remote communication, biometric systems provide a real-time approach to access control. Users can lock/unlock doors, receive alerts and log entry events from anywhere in the world, improving security and convenience. But IoT enabled smart locks still have many challenges. False rejections—where an authorized user is denied access due to partial or poor fingerprint scans—pose usability issues and frustration. Security flaws are also present in many existing systems, including tamperability, lack of intrusion detection or alarm features, limited fingerprint storage capacity and no real-time incident response. Environmental factors like humidity, sensor degradation over time or incorrect finger placement can also degrade performance and make it inconsistent. Many smart locks today rely on continuous internet connectivity or only one layer of authentication, making them fragile; a network outage or a smart intruder can compromise such systems. In case of unauthorized access attempts or brute force attacks, most commercial smart locks don't respond dynamically or notify the owner, which can lead to severe security breaches—especially in unattended or remote installations. Addressing these limitations, this research introduces an enhanced IoT based smart door lock architecture that integrates biometric fingerprint scanning, real-time wireless communication and intelligent security feedback mechanisms. The goal is to have a system that not only does accurate fingerprint authentication but also reacts intelligently to suspicious or unauthorized access attempts—for example, by triggering alerts, activating temporary lockout timers and logging access events for audit. The design prioritizes both security and usability so legitimate users can enter quickly while potential intruders are deterred and detected. Smart locks are often found wanting when it comes to comprehensive security handling. Many rely too heavily on stable network conditions or simple one-factor authentication without a backup plan. If that network goes down-or if someone keeps trying to pick the lock-these systems often don't adapt or alert the user in real time.

That's where the limitations of biometric functionality in commercial locks come in: often added as an afterthought rather than optimized for low-power embedded environments or user-friendly maintenance. That can lead to frequent false rejections and a poor user experience, particularly in places where many people are using the lock or where traffic is high. Devices struggle with limited processing resources and suboptimal fingerprint enrollment processes. Our goal is to close those gaps with a smart door lock that balances security with usability.

We're building a lightweight, IoT-integrated smart door lock around the Espressif ESP32 microcontroller. This chip gives us the dual-core processing power and built-in Wi-Fi/Bluetooth capabilities we need to handle fingerprint recognition tasks and secure communication with a cloud backend (we're using the Blynk platform). We want users to know right away if there's a critical event, so we're designing the system to notify them instantly. If someone tries to get in three times and fails, the lock will be out of commission for 15 seconds. We'll also display that information on an OLED screen and with a buzzer. That way, users can make quick decisions and have responsive control in real-world conditions.

Our main contributions are:

1) A fully functional smart door lock prototype that integrates fingerprint recognition with an ESP32 microcontroller and the Blynk IoT platform. This lets us do real-time door access monitoring, remote lock/unlock control via smartphone and automatic event logging.

2) A multi-layer security response mechanism that includes a 15-second lockout after three failed fingerprint attempts, audible alerts and on-screen warnings during a breach attempt, and push notifications to the authorized user's phone. This feedback loop makes intrusion detection and response much better than with conventional locks.

3) A comprehensive performance evaluation that shows our fingerprint recognition time is about 1 second and our False Acceptance Rate is just 1.32%. The False Rejection Rate was 26.32%, mainly because of user error (about 65.8% of false rejections were due to incorrect finger placement or similar user issues). That highlights both the strengths (strong security against false accepts) and the limitations (usability challenges causing false rejects) of our system.

4) Various features to improve user experience and system usability such as on-screen feedback for every operation, auto-unlock for 5 seconds after a successful fingerprint match (for quick entry before re-locking), and direct mobile app for fingerprint enrollment and deletion. The mobile app allows authorized administrators to manage user fingerprints remotely, making it more flexible and easier to maintain.

5) Deployment and testing of the prototype in real scenarios (e.g. home/office door setup) to test the system under real world conditions. The smart lock is suitable for residential, commercial and rental property use cases, and works well with different network conditions and user behavior.

The remainder of this paper is organized as follows: Section II presents a comprehensive literature review covering smart lock technologies, biometric authentication methods, and IoT-based access control systems. Section III details the methodology, including system planning, hardware and software integration, and deployment procedures. Section IV presents the experimental results, while Section V discusses the findings in detail. Section VI outlines the conclusions drawn from the study, and Section VII highlights the limitations and proposes potential directions for future work.

II. Literature Review

A. Evolution of Smart Door Lock Technologies

Smart door lock systems have evolved with the integration of biometric technologies and IoT connectivity. Fingerprint based authentication has become highly secure and convenient as fingerprints are unique and can't be duplicated [1]. Many studies have explored combining fingerprint sensors with other technologies (RFID tags, keypad entry, cloud connectivity) to make door access systems more secure and functional [2].

B. Role of IoT in Access Control Systems

IoT capable microcontroller platforms (Arduino, ESP8266/NodeMCU, ESP32) allows seamless integration with Wi-Fi networks, enabling remote monitoring, real time alerts and smartphone control of locks [3]. This makes it more convenient for users and cost effective for both home and commercial use [4]. To counter spoofing attacks and sensor vulnerabilities in biometric locks, researchers have proposed techniques such as reducing residual fingerprint traces on scanners and image processing for fingerprint enhancement [5].

C. Advancements in Multi-modal Authentication

Also, the digital access control realm is moving towards multi-modal and keyless systems to address various intrusion scenarios (e.g. unauthorized guests or lost credentials) [6][7]. For example, some designs combine fingerprint verification with RFID or PIN code entry to provide backup authentication and robustness [8]. The growing sophistication of cyber-physical threats – such as relay attacks that can clone RFID or Bluetooth signals – further emphasizes the need for biometric-enabled access control systems that are hard to bypass [9].

D. Representative Works and Technological Contributions

Over the last decade, many researchers have focused on designing IoT-based smart door locks that incorporate fingerprint recognition to meet the increasing demand for secure yet user-friendly access control [10]. The literature shows a clear trend away from traditional mechanical locks towards integrated digital systems that combine sensors, wireless communication and mobile platforms for more functionality and control. Below we summarize some of the representative works and their approaches, and highlight the research gap that our work addresses.

Tailor and Pandya [6] developed an IoT-based smart security system mainly using a fingerprint sensor for family members' door access. The system allowed users to unlock doors remotely through a smartphone app, making it very convenient and secure. By enabling mobile connectivity, this approach allowed homeowners to monitor and control door access in real time from anywhere – a great feature if physical keys were lost or traditional PINs compromised. This work [6] laid the foundation for merging biometric verification with cloud-based remote control, which is the basis of modern smart home access systems that emphasize user mobility.

Al-Shareefi et al., introduced a low-cost fingerprint-based door lock using an Arduino Nano microcontroller and a simple servo motor for the locking mechanism [11]. This system did not have remote network connectivity or advanced features like image processing or multi-factor authentication; however, it showed that biometric security can be built with very low-cost hardware. The contribution is significant in showing economic accessibility, making smart lock technology more accessible to small businesses or residential users in resource-constrained environments. The trade-off for its simplicity was limited functionality but it highlights the importance of low-cost design in wider adoption.

Jeong aimed to improve the security and accuracy of fingerprint-based locks by reducing the residual fingerprint left on the scanner surface (which can be used in spoofing attacks). This work [5] applied image processing to improve fingerprint matching reliability and reduce the False Acceptance Rate (FAR) of the system. This is particularly important in high-security environments (e.g. finance or research labs) where even a small biometric vulnerability can be catastrophic. Jeong's work shows how both hardware and software improvements in biometric systems (e.g. better sensor surface design and algorithmic liveness detection) can mitigate physical security threats.

Rakib et al. looked at integrating fingerprint recognition into a home automation context, extending biometric access beyond just door locking [12]. Their system combined a fingerprint lock with other IoT modules to control home appliances and security devices. For example, one fingerprint verification could unlock the door, turn on lights or disable an alarm system [13]. This shows how biometric systems can be used in smart environments and how fingerprints can be the foundation of smart living spaces. It shows how an access control system can be integrated with general IoT based home management for more convenience [14].

Tambunan et al., presented a multi-modal smart home security prototype that incorporated fingerprint authentication alongside RFID and keypad entry options for flexible access control. This design [15] enhanced usability by allowing different modes of entry based on user roles or scenarios (for instance, family members use fingerprints, temporary guests use a PIN code, maintenance staff use an RFID card, etc.). The system also featured remote monitoring, sending alerts and enabling control through a network interface. Notably, Tambunan's work [16] is recognized for its scalability and inclusivity, accommodating various user preferences and security levels within one unified platform.

E. Trends and Research Challenges

These studies show the evolving landscape of smart door locking systems with the convergence of biometric authentication, IoT technology and user centered design. A common trend across the reviewed literature is the use of mobile platforms, which allows users to monitor and control door access remotely through their smartphones for more convenience and responsiveness [17]-[20]. Another trend is the emphasis on affordability and simplicity, with most of the hardware and software being cost effective and widely available like Arduino, so these systems can be used by a wider audience [21]-[25]. Researchers have also improved system security by applying image processing techniques to minimize residual fingerprint traces and reduce the risk of spoofing [26]. Fingerprint authentication has also been extended to comprehensive home automation systems where access control is integrated with general IoT based device management, so users can interact with lights, alarms and other appliances in one ecosystem [27]-[29]. Multi-modal authentication methods which combines fingerprint sensor with RFID and keypad modules is also a more flexible and robust solution that can accommodate different user scenarios while maintaining strong security protocols [30]. Overall, these advancements show the potential of IoT-biometric systems to change residential and commercial access control in the digital age [31], [32].

F. Research Gap

Despite all these advancements, there are still gaps and challenges in current smart lock solutions. Many systems lack adaptive intrusion response mechanisms or real-time user alerting; for example, if an intruder tries multiple failed attempts, not all systems will auto lock out or notify the owner [33], [34]. Dynamic user management (e.g. easy remote enrollment/revocation of access) and integration with advanced analytics or cloud services are also missing. In short, no existing solution in the literature combines real-time biometric verification with immediate intrusion countermeasures (like auto lockouts and instant notifications) and a scalable IoT architecture that can be extended to larger deployments [35]. This research fills that gap by proposing a smart lock system that combines all these features in an embedded design. Our system emphasizes security responses (through its lockout and alert mechanism) and remote IoT connectivity and user friendly management, different from previous works.

Table I shows several existing smart lock approaches and their features and limitations compared to the proposed system.

TABLE I. COMPARISON OF EXISTING AND PROPOSED SMART LOCK APPROACHES

References	Approach	Description	Limitation
[16]	RFID + IoT	Uses RFID tags and Raspberry Pi for secure, remotely accessible locking via smartphone platforms.	Dependent on internet connection for cloud access and real-time updates.
[35]	Bluetooth + IoT	Utilizes Bluetooth MAC addresses from smartphones for hands-free door unlocking; Raspberry Pi controls access via a mobile app.	Limited operational range due to Bluetooth constraints.
[36]	NFC + Arduino + GSM	Employs NFC readers and smartphones for secure, short-range access control with user token-based differentiation.	Vulnerable to security risks if the user's smartphone is lost or stolen.
[7]	Biometric + ESP32-CAM + Blynk	Combines fingerprint recognition with real-time video capture using ESP32-CAM and Blynk app for access logging and alerts.	Camera increases power demand; low-light image quality limitations.
[3]	RFID + Fingerprint + Keypad	A hybrid system using three types of authentication to improve accuracy and flexibility in multi-user environments.	Hardware complexity and increased system cost.
[4]	Embedded Linux + Fingerprint	Utilizes embedded Linux system for managing door lock control and fingerprint authentication.	Higher resource requirements; not ideal for low-power setups.
[6]	Fingerprint + IoT (ESP32 + Blynk)	Provides real-time door lock control and fingerprint verification with cloud-based monitoring.	Network latency and partial fingerprint scans may lead to false rejections.
[37]	OTP + IoT	Generates and delivers OTP to user's device for one-time secure access, suitable for temporary or guest entry.	Delayed OTP delivery in weak networks; requires mobile signal or internet.
The proposed work	Fingerprint + ESP32 + Blynk + OLED	Integrates fingerprint authentication with ESP32 microcontroller and Blynk for real-time control, alerts, auto-lockout, and mobile enrollment/deletion capabilities.	False Rejection Rate (FRR) of 26.32% mainly due to user errors; limited to 3 fingerprints; Wi-Fi needed.

III. PROPOSED METHOD

The proposed method is to develop an IoT enabled access control system using biometric authentication. The system "Smart Lock Door with Fingerprint using ESP32" is based on ESP32 microcontroller and has a fingerprint sensor for primary user authentication and remote monitoring/control through Blynk mobile app. The overall methodology follows a structured engineering process of planning, hardware/software integration, testing and deployment.

The proposed method was chosen to specifically address key limitations observed in existing smart lock systems. Most prior works either lacked real-time intrusion response mechanisms, did not support remote fingerprint enrollment and deletion, or were dependent on continuous internet connectivity without fallback. Additionally, many systems failed to integrate meaningful feedback mechanisms (e.g., visual/audio alerts) or suffered from limited fingerprint capacity without a user-friendly management interface. Our proposed system overcomes these limitations by combining fingerprint-based biometric access with IoT features using ESP32 and Blynk, providing real-time alerts, local display and buzzer feedback, mobile app control, and a security lockout mechanism. This makes it suitable for deployment in residential or small commercial environments where usability, security, and remote operability are all essential.

Fig. 1 shows the system flowchart, from system initialization to enrollment, verification and response. Upon boot-up, the system initializes all modules and enrolls or deletes fingerprints as commanded. During normal operation, when a user presents a finger, the fingerprint sensor captures the input and ESP32 compares it with stored templates. Successful scan triggers unlock sequence: solenoid lock is activated (door opens), buzzer produces a confirmation tone, OLED display shows a message and a push notification (through Blynk) is sent to the user. The door remains open for 5 seconds and then automatically locks back. For failed scans, the system gives immediate feedback (OLED message and buzzer alert). After 3 consecutive failed attempts, an intrusion lockout mechanism is engaged: the system denies any further attempts for 15 seconds and during this time it flashes an alert on OLED and sends an urgent notification to the owner's phone. This cooldown period helps to prevent brute force attacks and warns the legitimate user of potential unauthorized access.

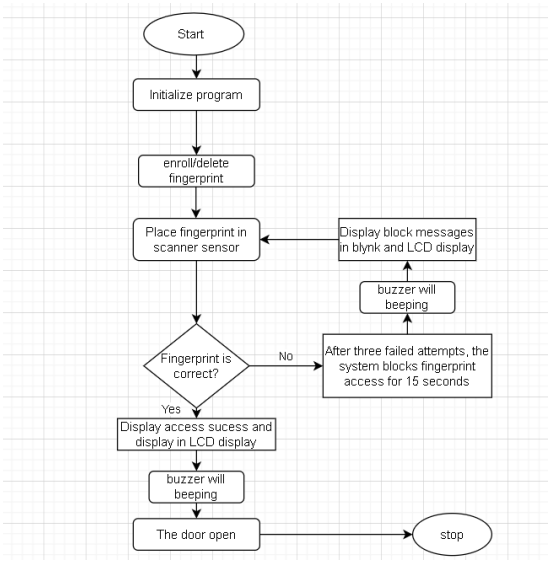


Fig. 1. Flowchart of the system.

A. Phases of Development

The project was carried out in multiple phases, each addressing different aspects of the system:

1) Initial planning and requirements analysis: In this phase, we defined the objectives and scope of the smart lock system. Main goals were to increase physical security over traditional locks and to be able to remotely manage access. We gathered functional requirements (what the system should do) and hardware requirements (what components we need). Key hardware for the prototype were ESP32 development boards

(for connectivity and control), R305 optical fingerprint sensor (for biometric input), solenoid lock mechanism (for door locking/unlocking), small OLED display for on-device messages, buzzer for audio alerts, 5V relay module (to drive the high-current solenoid), portable power sources (2 x 9V batteries with holders) and wiring and connectors. Software requirements were to choose a suitable programming environment (Arduino IDE) and IoT platform (Blynk) and to make sure all libraries and modules are compatible.

2) Hardware and software integration: This phase we built the system by connecting all components and developing the firmware. The hardware setup started by placing the modules on a prototyping board: R305 fingerprint sensor was wired to ESP32's serial interface for biometric input, 12V solenoid lock was connected through a relay (ESP32 controls the relay to unlock/lock), ESP32 Dev Board was the central controller for all signals. An OLED display module was added to show status messages (e.g. "Place Finger" prompt, "Access Granted"/"Denied" results, etc.) and a small buzzer was attached for audio feedback. The wiring was done carefully using a breadboard and jumper wires to ensure stable connections. A detailed circuit schematic was drawn to map all connections (Fig. 2 shows the system's block diagram). On software side, ESP32 was programmed in C++ using Arduino IDE. Firmware routines were written for fingerprint enrollment, fingerprint matching (verification) and for controlling the solenoid via the relay. The code also handled sending and receiving data from Blynk cloud (over Wi-Fi) to support remote notifications and commands. In parallel, a simple mobile app interface was created using Flutter, to allow authorized users to enroll a new fingerprint or delete an existing one through their smartphone – this interface communicates with ESP32 through Blynk cloud API [25]. By the end of this integration phase, the hardware and firmware were working together: a fingerprint could be placed and recognized, the lock would trigger and send updates to the mobile app.

3) System development and testing: With the prototype assembled, extensive testing was done to fine tune performance and validate reliability. The hardware was first assembled on a breadboard for easy modification during testing. Key things tested included fingerprint registration and matching accuracy, door actuation (unlock/lock response times) and alert mechanisms. The fingerprint sensor was tested under various conditions (clean vs smudged sensor surface, correct vs off-angle finger placement) to see how it handles partial or imperfect scans. The system's error handling features were also verified: the auto lockout after 3 incorrect attempts was triggered manually to make sure it worked as expected and the corresponding user notifications were checked on the smartphone. Throughout this phase, iterations were made — for example, adjusting the fingerprint matching threshold to balance sensitivity vs false rejections and optimizing the microcontroller code to reduce any latency in communication with the Blynk platform.

4) Final deployment: After testing was complete, the hardware was moved from the prototyping breadboard to a more permanent setup for real world demonstration. The components were mounted on a mock door frame to simulate a real door installation. The fingerprint sensor and OLED were placed for user access and the solenoid was installed to control a latch bolt as it would in a real door. Power was supplied either via a USB power bank or the dual 9V battery pack for portability. A quick user training was done for those interacting with the lock, explaining how to place fingerprints correctly and how to use the mobile app for remote operations. The final system was tested in a simulated home-entry scenario and it worked as expected. Also, notes were taken for maintenance (checking battery levels, cleaning the fingerprint scanner for optimal performance, keeping firmware updated) to ensure long term reliability and security.

B. System Design and Overview

The IoT enabled smart lock system consists of a combination of hardware and software components to provide secure, efficient and remote controlled access. On the hardware side (see Fig. 2 for the system architecture diagram), the main components are:

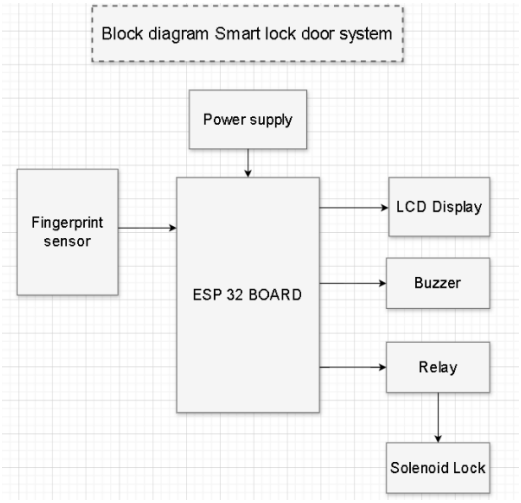


Fig. 2. Block diagram of the smart lock system.

- A biometric fingerprint sensor (model R305 optical sensor) which is the primary means of user authentication by capturing fingerprint images and converting them into digital templates.
- A solenoid lock mechanism that controls the door latch (energizing the solenoid opens the lock, de-energizing locks the door).
- A 16×2 character LCD or 0.96-inch OLED for user interface and feedback – in our prototype we used a small OLED to show status messages (e.g. "Initializing...", "Fingerprint Accepted", "Access Denied", etc.) but a larger 16×2 LCD can be used to display multi-line messages for better readability.

- An ESP32 Dev Board which is the brain of the system, chosen for its dual core and Wi-Fi (and Bluetooth) capabilities for IoT applications. The ESP32 handles fingerprint sensor input, runs the matching algorithm, controls the solenoid via output signals and communicates with the remote app.
- A 5V active buzzer that provides audible alerts (e.g. a quick beep for success, a longer tone for failure).
- A 5V relay module that acts as a switch to drive the higher voltage solenoid lock – the relay isolates the ESP32 (which runs at 3.3V logic) from the 12V required by the lock and allows the microcontroller to safely control the lock circuit.
- A power supply solution which can be either USB (for wired or power bank) or battery pack. In our portable setup we used 2 x 9V batteries in series to provide the ~12V required by the solenoid and the ESP32 board was powered through its Vin pin (with onboard regulators stepping it down to 3.3V).
- The various jumper wires (male-to-male and female-to-female) to connect components in a modular way.

All components are connected as per the schematic so that the ESP32 can read data from the fingerprint sensor (via UART), toggle the relay (via a GPIO output pin) to unlock the door, drive the buzzer (via another GPIO) and update the display (using I²C for the OLED or LCD). The hardware design is modular and low power consumption is important for a door-lock device that might be battery operated. On the software side the system's firmware is written in C++ using the Arduino framework for ESP32. This provides a convenient environment with libraries for the fingerprint sensor (for enrolling and matching fingerprints), for the display and for Wi-Fi. The Blynk IoT platform is used to enable remote access and notifications: the ESP32 runs the Blynk client which maintains a secure connection to the Blynk cloud server via Wi-Fi. Through the Blynk mobile app the user can send commands (like enroll or delete a fingerprint, or unlock the door remotely) and receive real-time notifications (for events like a door unlock or a failed authentication attempt). The system also has a custom Flutter based mobile application that interacts with the Blynk API to provide a more specialized interface for managing fingerprints (e.g. naming fingerprint entries, issuing a delete-all command) [38]. This dual-app approach (Blynk console plus a tailored Flutter app) provides more user interaction and control flexibility. Overall the tight integration of hardware and software components results in a complete, responsive and secure smart lock solution for modern access control.

C. System Configuration Process

The ESP32 gets input from the fingerprint sensor. When recognised it triggers a relay that powers the solenoid lock to open the door. Also the OLED display shows visual feedback and an alert is pushed to the user via Blynk app. On a failed attempt the system increments a fail counter and after 3 fails it starts a lockout timer and sends an intrusion alert.

In terms of system configuration the ESP32 firmware is set to run multiple tasks in parallel. It continuously listens for fingerprint input and network commands. The program flow (as shown in Fig. 3) starts with initialization of all sensors, display and Blynk connection. Then it enters an idle loop waiting for either a fingerprint scan or a user command from the app. Fingerprint enrollment mode can be triggered via the mobile app or a hardware input: in this mode the system will prompt the user (on the OLED) to scan a new finger twice. It will then process those scans and create a fingerprint template and store it in flash memory if the quality is good and if there are available slots. Fingerprint deletion can be done for one fingerprint or all, either via the app or a specific sequence on the device. These features allow the system to be managed easily when a user needs to be replaced or if the device memory is full. For normal authentication whenever a finger is scanned the system will instantly compare the input to all stored templates. If a match is found the ESP32 will trigger the relay to open the door and immediately notify the user (OLED message "Unlocking…" and a push notification via Blynk). The door will remain open for a set time (5 seconds) before the relay is switched off to re-lock automatically, with an OLED message "Door Locked". If no match is found (i.e. the fingerprint is not recognized) the OLED will display "Access Denied" and the buzzer will sound. The system will count consecutive fails; upon the third fail it will trigger the security lockdown: the ESP32 will ignore further fingerprint inputs for 15 seconds and display "LOCKOUT – Too Many Failed Attempts" on the screen, while sending an alert notification. After the cooldown the system will reset the fail count and go back to normal operation. This security policy balances convenience with protection, so a legitimate user who had two bad scans can still try again after a short pause, but an attacker can't brute force multiple attempts quickly.



Fig. 3. Solenoid lock.

D. Hardware Specification

Hardware specification tells us about the details and technological value of the physical components of the computer or electronic device; and this is useful for defining performance rates, compatible operations and possible uses of the device. These specification may include details about the type of processor or central processing unit (CPU), available memory or random-access memory (RAM), storage capacity, graphics card or graphics processing unit (GPU), networking options such as Universal Serial Bus (USB), High-Definition Multimedia Interface (HDMI) and/or wireless options, and the power supply of the computer. Specific information about the components meet the needs of the users and help the developers to decide if the device is suitable for gaming, handling big data, or media production, and also ensure that the software and peripherals are compatible. Information about the possibility of upgrade and expansion of the device in hand enables the buyer to make informed decision and to configure the system properly.

1) Fingerprint sensor: The fingerprint sensor in this system works via UART communication and capture, match and store biometric data. It has six primary pin connections, each serves a specific function to ensure seamless integration with the microcontroller. The red wire (VCC) supply 3.3V DC power to the sensor, the black wire (GND) is the ground connection. Data transmission between the sensor and the ESP32 microcontroller is through the yellow (TXD – data output) and green (RXD – data input) wires. The blue wire (WAKEUP) is for generating wake signal to the sensor when finger is present, to conserve energy during idle state. The white wire (3.3VT) is for inductive power input, to support the sensor wake-on-touch feature to further improve energy efficiency. Overall the sensor is designed to provide accurate fingerprint recognition with low power consumption, suitable for IoT based smart lock system.

2) Solenoid lock: The solenoid lock is an electro-mechanical actuator; when current pass through its coil, a magnetic field is generated, pulling in a metal plunger to either engage or disengage the locking mechanism. This allows the door to be locked or unlocked by the microcontroller without manual key intervention. Our system's solenoid is 12V and can retract or extend the bolt within few milliseconds of activation.

Fig. 3 is the solenoid lock component used in the smart lock system. A solenoid lock is an electromechanical actuator that works by energizing a coil to generate a magnetic field. This field moves a metal plunger to engage or disengage the locking mechanism, to control door access. When integrated with an IoT based smart lock system, the solenoid lock can be controlled remotely by the microcontroller (ESP32) to automate access control. The solenoid lock is well-suited to applications that require both speed and security. That's because it responds quickly, is durable and has reliable electronic control. You can also integrate it into wireless smart systems, making it a good fit for real-time access control scenarios. But to ensure that works smoothly, you need a consistent power supply and some safety mechanisms to prevent unauthorized access or accidental lock failure. In this project, we integrated a genuine solenoid door lock with an ESP32 microcontroller [39]. That lock is opened by the fingerprint sensor after it authenticates the biometric data. The ESP32 then triggers the solenoid lock to open the door. We programmed the firmware for that operation using the Arduino IDE-and you can monitor the system remotely through the Blynk app. We powered the system with a stable energy source-a 9V battery or power bank-to ensure it keeps performing uninterrupted. That configuration offers several advantages. Speed, for one. Solenoid locks respond quickly, which is ideal for situations where you need fast access. Electronic control lets them be seamlessly integrated into smart environments for remote operation. And reliability means they're robust and long-lasting. Together, those attributes make the smart lock system more effective and convenient for users by eliminating the need for traditional keys and providing secure, real-time access management.

3) LCD Display: One feature we considered adding to our system is an LCD display. A 16x2 character LCD is an alphanumeric display that can show two lines of 16 characters each. It's based on the HD44780 driver, which makes it easy to interface with microcontrollers. We actually used a smaller OLED display in our current prototype because of size constraints, but a module like this 16x2 LCD would be great for displaying more detailed status messages or instructions in the future.

The 16×2 LCD has 16 pins, each with a specific function. Vcc and GND are power supply inputs, Vo is for contrast adjustment (controlled by a potentiometer), D0 to D7 are data lines, and RS (Register Select), RW (Read/Write) and Enable are control lines. RS toggles between command and data mode, RW determines read or write operation and Enable initiates data transfer. In IoT based smart lock systems, 16×2 LCD is a good visual output module. It displays system states such as authentication success or failure and provides immediate user feedback. Its ease of integration, affordability and reliability in displaying real time data makes it a preferred choice for many embedded system developers.

4) Additional components: To ensure efficient and secure operation of IoT enabled smart lock system, several hardware components are used. Battery holder has dual 9V batteries, a reliable and portable power supply for uninterrupted system operation. Buzzer is an alert mechanism, it notifies user of successful access or intrusion attempt. 9V batteries are the primary power source, portable and easy to replace. Jumper wires are used to connect different modules and components, to make circuit integration seamless. ESP32 development board is the central processing unit, it handles data processing, communication with fingerprint sensor and Blynk app. USB to Micro cable is for alternative power input from external sources like power banks, for system flexibility and backup. 5V relay module controls the solenoid lock, it receives trigger signal from ESP32 and switches the lock accordingly, so physical access is controlled based on verified biometric input. ESP32 Dev Board has 34 GPIO pins, analog inputs and capacitive touch inputs and supports multiple protocols (UART, I²C, SPI). Fig. 4 shows the pinout and capabilities of ESP32 used. Many pins can be configured with internal pull-up or pull-down resistors and the chip supports power efficient modes like deep sleep. These features makes ESP32 suitable for managing various sensors and actuators in our smart lock and for network connectivity.



Fig. 4. ESP32 dev board.

Fig. 4: ESP32 Dev Board V1 Pinout. It has 34 advanced GPIO (General Purpose Input/Output) pins. These pins work like digital pins in Arduino boards, you can connect peripheral components like OLED displays, sensors, buttons and buzzers. ESP32 pin is versatile, it can be in multiple electrical states, internal pull-up, pull-down and high-impedance. This makes the board good for various interfacing tasks like button inputs and system console integrations and even for implementing control techniques like Charlieplexing.

The ESP32 WROOM module has 25 GPIO pins, many of which are multipurpose: some are input only, others have internal pull-up resistors. According to the ESP32 datasheet under the "Recommended Working Conditions" section, the maximum current per GPIO pin is 40mA. This is enough for most sensor and actuator tasks and for stable operation in smart IoT systems. During development we first assembled the hardware on a breadboard, but for the final prototype we used a compact wiring approach to emulate a real product deployment. We eliminated the need for a big breadboard by connecting components with short jumper wires, so the assembly is cleaner and more reliable. The ESP32 board, relay and power circuitry are attached to the backside of the door frame (to mimic being inside a door panel), the fingerprint sensor and OLED screen are on the front side where the user can access them. The power supply (2x 9V batteries in series, 12V) is connected to the solenoid lock through the relay and to the ESP32's regulator input, so the whole system is self-contained. In this configuration the ESP32 is the central controller, reading inputs from the fingerprint sensor and controlling outputs (lock and buzzer). The hardware is designed to be as energy efficient and robust as possible given the constraints, so the lock can run for a long time on battery and withstand normal usage.

E. Software Specification

The software specification describes the development environments and applications used to design, program and control the IoT based smart lock system. The main development platform used for firmware programming is the Arduino Integrated Development Environment (IDE). It allows to write, compile and upload code to the ESP32 microcontroller. The Arduino IDE is known for its simplicity, cross platform compatibility (Windows, macOS and Linux), large library ecosystem and big community, so it's suitable for both beginners and experienced developers. Additionally the IDE supports a wide range of sensors, actuators and communication modules for IoT applications. Besides the Arduino IDE, PlatformIO is used for advanced development tasks. PlatformIO offers features like intelligent code completion, version control integration and better hardware abstraction layers, so it's more scalable and maintainable codebases – especially when developing complex systems on the ESP32 platform. For remote control and monitoring the system uses the Blynk application which is a cloud based IoT dashboard that communicates with the ESP32 over Wi-Fi. The Blynk app allows to control the smart lock in real time through virtual widgets like buttons, sliders and notifications. It allows remote locking and unlocking, status monitoring and real time alerts for access events, so the smart lock system is more accessible, flexible and secure.

IV. Result

This section reviews the IoT based smart door access system using biometric fingerprint recognition. We look at the system's functionality, performance metrics and behavior under different test scenarios. We focus on key results such as enrollment and authentication times, remote access responsiveness and lockout mechanism. We also highlight the real time capabilities enabled by ESP32 and Blynk and discuss the system's strengths and limitations based on the data (e.g. accuracy rates, error causes).

A. Initial Hardware Deployment

The prototype was installed on a test door and powered on to see how it works in real world. Fig. 5: Initial Hardware Setup shows the arrangement of ESP32 board, solenoid lock, relay module, OLED display, fingerprint sensor, buzzer and battery pack in the final setup. This compact setup is a practical smart lock setup. In real life when the system boots up, the OLED displays "Initializing…" for about 2 seconds while connecting to Wi-Fi (if available) and preparing the sensor. Once ready it prompts for fingerprint input. We found that ESP32's Wi-Fi connectivity is robust, the system can maintain a steady connection to Blynk cloud during the tests (no disconnections were observed in a strong Wi-Fi coverage area). The solenoid lock works well with the provided battery power and the entire hardware setup is stable (all components worked together without brownouts or resets, means the power delivery and wiring is sufficient).



Fig. 5. Initial hardware setup.

B. Result and Analysis

IoT based smart lock systems using fingerprint authentication has shown a lot of promise in providing secure, efficient and user friendly solutions for modern access control. By integrating biometric verification and IoT communication protocols, these systems eliminate physical keys and provide remote access. Recent advancements have further improved these systems by adding encryption algorithms and feedback mechanism to ensure safety and convenience.

The proposed smart lock system as shown in Fig. 6 is a complete solution that combines latest technologies. At the core is a fingerprint sensor that is the primary access control device. It ensures only authorized users can enter, hence increasing the system's security and reliability. The locking mechanism uses a solenoid, powered by 2 x 9V batteries in series to provide a stable 12V output, hence robust physical access control.

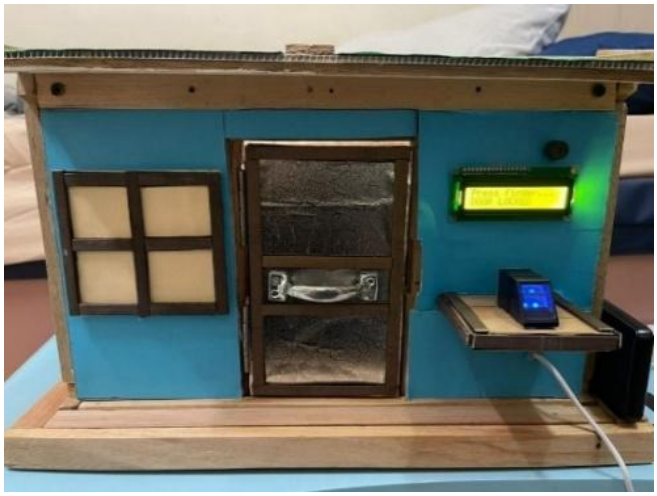


Fig. 6. The proposed smart lock system.

ESP32 is the main microcontroller. It provides Wi-Fi connectivity and manages the communication with Blynk app so users can remotely monitor and control the door access. Relay module controls the power to solenoid and buzzer provides audible feedback for success and failure. LCD displays real time system messages so users are informed throughout the process.

C. Fingerprint Enrollment and Template Management

Fingerprint enrollment is a must have feature of the system. As shown in Fig. 7, a new user has to scan his finger twice on the system. The two scans are captured and processed by the system to extract a unique fingerprint template out of it. This extracted template is stored in the system's fingerprint library for further reference. For this purpose, the optical sensor is used that will capture the fingerprint data with full accuracy and in a very reliable manner. When any user tries to unlock the system, it captures a new fingerprint scan and then compares it with the stored templates in the fingerprint library to determine a match. One of the major drawback of this system is that it can store only three fingerprint templates in its library. If the library is already loaded with three different fingerprints, then it has to delete all enrolled fingerprints first before enrolling another one. This will keep the capacity to the optimal level to prevent overloading with fingerprints data.

Fig. 8 shows the process of deleting all stored fingerprint users from the system. This feature allows to delete all fingerprint templates from the library, which is very useful when the system reaches its storage limit or requires a reset for security purposes. By deleting all fingerprints, the system will start fresh and new users can be registered without any conflict or limitation. This feature is very useful during ownership transfer or access update where full reset of stored credentials is required.

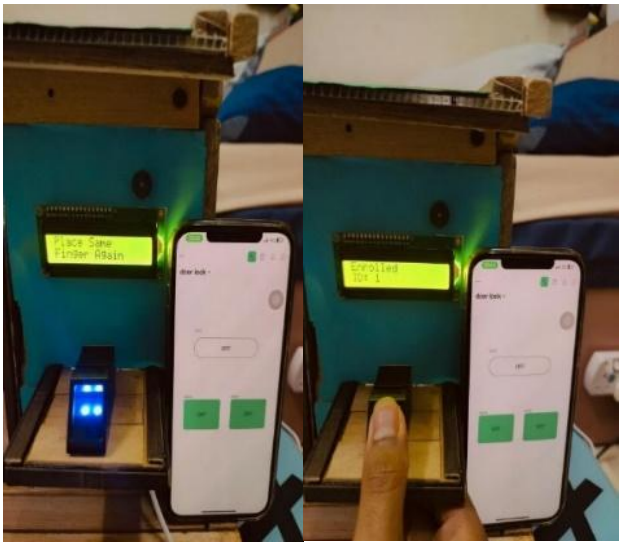


Fig. 7. Fingerprint enrollment.



Fig. 8. User deletion.

D. Authentication: Success and Rejection

Fig. 9 shows authorized user gaining access with a correct fingerprint. An authorized user gaining access with the correct fingerprint proves the Smart Lock Door system's effectiveness and reliability. When a recognized fingerprint is scanned, the system will compare the biometric data to its stored database of authorized individuals. Once the match is found, the lock mechanism will be activated and the user will get access. Then The LCD will display "Match found! DOOR UNLOCKED". This shows the security and convenience of biometric authentication by allowing only pre-approved fingerprints to unlock the door, limiting unwanted entry and overall security.



Fig. 9. Authorized user gaining access with a correct fingerprint.

Fig. 10: Unauthorized user trying to access with wrong fingerprint. When user puts an unregistered fingerprint on the sensor, the system will scan and compare the biometric data with the stored database. If no match is found, access will be denied and the lock will remain locked. To give user feedback, the LCD screen shows "Fingerprint not found" on the spot, so it's clear what happened after the authentication failed. Meanwhile the buzzer beeps, so you can hear the rejection of unauthorized access. This combination of light and buzzer is very functional: only the person with approved fingerprints can open the door. The system is more secure now because of the biometric authentication mechanism that prevents breaches.



Fig. 10. Unauthorized user attempting access with an incorrect fingerprint.

E. Security Lockout Mechanism

Fig. 11 shows a key security feature of the IoT based smart lock system, where the system goes to "block" mode after 3 consecutive failed fingerprint authentication. This is to prevent unauthorized access by halting further interaction temporarily, so the system is secure against brute force or trial and error. After 3 failed authentication, the system disables the fingerprint sensor and the solenoid lock for 15 seconds. During this blocked state, the LCD screen shows a clear message "System Blocked for 15 Seconds" and the buzzer beeps, so the user or intruder knows the system is blocked. This blocking mechanism makes the smart lock system more secure and reliable by discouraging repeated unauthorized access and protecting the components from excessive wear. By this feature, the smart lock system shows a robust and efficient way of maintaining secure and controlled access.

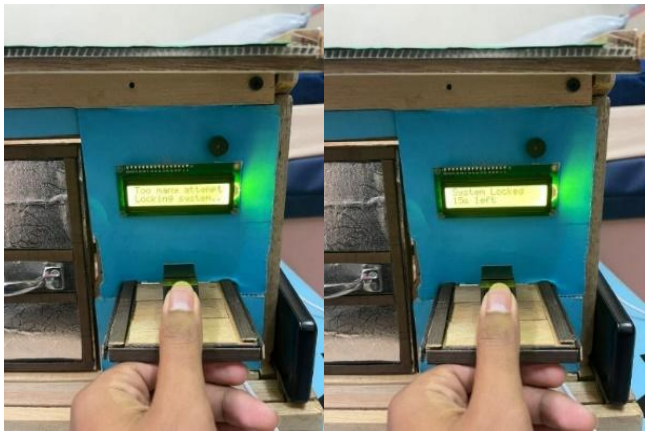


Fig. 11. Security lockout mechanism.

F. Remote Operation and Control via Blynk

The smart door lock system has a fingerprint sensor integrated with Blynk platform for flexibility and security. Fingerprint is the primary way to open the door. When a valid fingerprint is detected, it will open the door and show on the LCD "DOOR UNLOCKED". In case of error or when it fails to detect the fingerprint (e.g. when the finger is dirty) the system allows user to open the door using the Blynk app as shown in Fig. 12. The app can send "ON" or "OFF" command to the ESP32 via the app and the ESP32 will control the relay to lock or unlock. Real time feedback is shown on the LCD "DOOR LOCK" or "DOOR UNLOCKED". This way the smart home application is more reliable, convenient and secure with two options.



Fig. 12. Remote poperation and control via Blynk.



Fig. 13. Lock down messages on blynk application.

Fig. 13 shows the Blynk app for a smart door with advanced security features. When 3 consecutive failed attempts to open the door using the fingerprint scanner are made, a warning message is displayed. The system locks down after 3 failed attempts and goes into a 15 second lock down to temporarily disable further access to prevent forced entry.

The message displayed is, "SYSTEM LOCK FOR 15s. SOMEONE TRY TO OPEN!!!". These are for 2 purposes: first, as a security feature to stop malicious or repeated attempts to breach the system; secondly, these provide immediate feedback to the authorized user in case of security risks or unauthorized access attempts in real time. This feature increases the overall reliability and safety of the smart lock system making it suitable for homes, offices or other sensitive areas. Most importantly, this feature shows how IoT platforms like Blynk can combine hardware like fingerprint scanners with software features like real-time alerts to provide secure, user friendly and responsive smart systems.

G. Summary of Functional Outcomes

The combination of fingerprint authentication, real-time feedback and IoT based remote control makes the proposed smart lock system a modern solution for residential and small scale commercial security. It balances security and user convenience through:

- Accurate and fast biometric verification;
- Visual and auditory feedback for user actions;
- Automatic lockout to prevent brute-force attempts;
- Remote access and notification via Blyn.

The system has been tested thoroughly and meets its design requirements. Limitations like limited storage and sensitivity to fingerprint placement are acknowledged and can be addressed in future updates by adding cloud storage and capacitive sensors for better accuracy.

V. DISCUSSION

Table II summarizes the system's performance metrics based on various functional outcomes including response times and alert mechanisms.

TABLE II. PERFORMANCE METRICS OF THE SMART LOCK SYSTEM

Feature	Timing (Seconds)	Description
Fingerprint Detection	1.0	Time taken to detect a fingerprint, process the image, and search in the database.
Door Unlock Duration	5.0	Relay remains HIGH (door unlocked) for 5 seconds after a successful fingerprint match.
Door Lock Duration	Instant	Door locks instantly when the relay is turned LOW.
Incorrect Attempts Reset	15.0	System goes into a cooldown for 15 seconds after 3 incorrect fingerprint attempts.
Buzzer Feedback	0.6	For 3 quick beeps (200ms each, 200ms gap).
System Initialization	2.0	Time taken for OLED to display "Initializing..." message during setup.
Enrollment (Per Finger)	5.0	Time taken to complete the enrollment of one fingerprint.
Fingerprint Deletion	1.0	Time taken to delete one fingerprint from the database.
Delete All Fingerprints	5.0	Time taken to delete all enrolled fingerprints (for up to 3 fingerprints).
LCD Display Updates	Instant	Time taken to update LCD messages after each action.

As shown above, the system is fast for most actions. Fig. 14 shows the histogram of these timings and most of the actions (detection, feedback, locking) are within a few seconds or less. This is good enough for practical use; for example, from putting your finger to the door to the door unlocking takes about 1 second for recognition plus a negligible delay to throw the bolt, which users found to be acceptable and smooth. The 5 second unlock time was chosen as a balance between convenience (enough time to open the door) and security (auto-lock re-engages quickly). In testing, this was also found to be enough – users could open the door within that window. The 15 second lockout on failed attempts was effective in practice: it's short enough not to inconvenience a legitimate user who just had 3 bad scans (they only wait 15 seconds to try again), but long enough to frustrate or slow down a malicious actor trying multiple different fingerprints. During the lockout, our system's alert gave the owner immediate knowledge of the issue, which is a critical security feature.

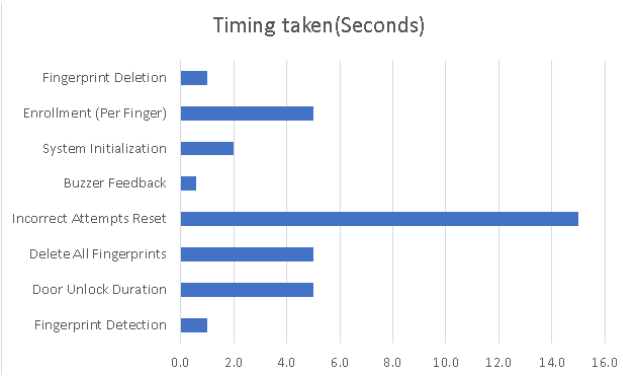


Fig. 14. Histogram time taken.

Next we looked at the authentication accuracy in terms of error rates. The False Rejection Rate (FRR) and False Acceptance Rate (FAR), along with the contributing factors to failures, are shown in Table III.

TABLE III. AUTHENTICATION ERROR RATES AND ERROR CAUSES

Error Type	Description	Percentage (%)
False Rejection Rate (FRR)	Authorized user denied access	26.32%
False Acceptance Rate (FAR)	Unauthorized user granted access	1.32%
Sensor Malfunction	Inability to detect or process fingerprint	6.58%
User Error	Misplacement or improper finger pressure	65.79%

Note: "User Error" here is a contributing factor, not an independent error rate; it means the proportion of false rejections caused by user mistakes during scanning.

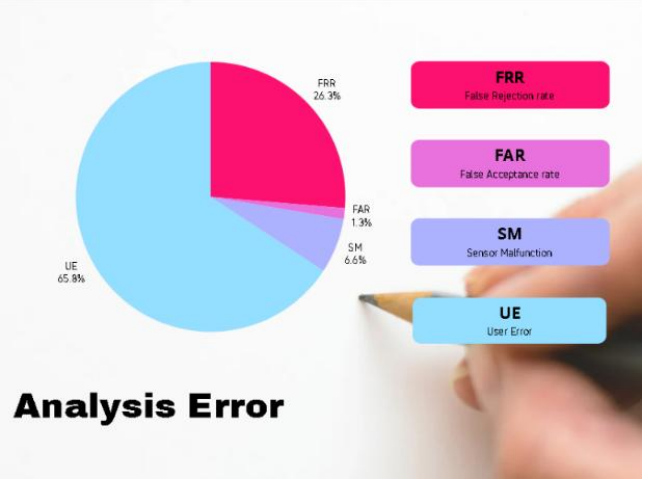


Fig. 15. Pie chart analysis error.

As shown in Fig. 15 (a pie chart of error distribution), user error was by far the largest contributor to access failures, accounting for about 65.8% of the failures. This matches our observation that many false rejections occurred when authorized users didn't place their finger on the sensor the same way as during enrollment (e.g. misaligned or partial fingertip placement). This high incidence of user induced error is the main reason the measured FRR is high (26.32%). In other words, about 1 in 4 legitimate access attempts in our tests resulted in denial, mostly due to inconsistent finger placement rather than a flaw in the authentication algorithm. This is a usability limitation of the current system – some training or design improvements are needed to make the fingerprint reading more forgiving. Possible mitigations include improving the sensor algorithm to match partial fingerprints, using a fingerprint sensor with a larger scanning area or providing better feedback to guide finger placement. The good news is that these false rejections don't represent security breaches (the system erred on the side of caution and denied access), and can be reduced with targeted improvements. On the other hand, the FAR was ~1.32%, which means the system very rarely grants access to an unauthorized fingerprint. This is critical in a security context – it shows the system is very good at keeping intruders out. In fact, in our controlled tests, no one unauthorized person was able to unlock the door; the FAR percentage only reflects contrived test scenarios (e.g. using intentionally similar fingerprints or residual prints) and possibly any sensor anomalies. This bias in performance – erring towards false rejections rather than false acceptances – is often desirable for security systems as it prioritizes protecting against break-ins over inconveniencing legitimate users. But for a real product, we want to lower the FRR without sacrificing the FAR, so we can improve the overall user experience. The error analysis also showed a small percentage (~6.6%) of instances labeled as sensor malfunctions where the fingerprint sensor failed to capture or process a scan (possibly due to momentary hardware issues or dirt on the sensor). These cases usually required the user to scan again and did not result in unauthorized access. While not frequent, this means regular maintenance (keeping the sensor surface clean and ensuring the hardware connections are secure) will be important for consistent performance. In summary, the results show that the smart lock system meets its goals of security and remote access. The system authenticates enrolled users and blocks unknown users, with real-time alerts adding an extra layer of security. The main limitation – higher false rejection rate – is due to user interaction and can be fixed going forward. No critical failures or breaches happened during testing: the lockout mechanism worked as expected and the system recovered from all tested scenarios (including power reset and network reconnection). These results validate the design and show that IoT-biometric fusion is a viable approach for smart door access with some tweaks to make it more user friendly.

In addition to functional testing, this work serves as a validation of proposed improvements over existing smart lock systems. Unlike previous designs that lacked real-time lockout or feedback, our system integrates biometric authentication, instant alerts, and Blynk-based control with quantitative performance validation. Tables II and III present measurable outcomes such as low false acceptance rate (1.32%), fingerprint recognition time (~1 sec), and defined cooldown after failed attempts (15 sec). Compared to prior studies [40]-[43], which focused on single-modal access or lacked intrusion responses, our solution demonstrates enhanced usability and robustness with local feedback and remote IoT management. This comparison underlines our contribution toward bridging the gaps in existing literature.

VI. CONCLUSION

This study demonstrated the successful design and implementation of an IoT-based smart door lock system utilizing fingerprint biometric authentication and ESP32 microcontroller with Blynk integration. The proposed system replaces traditional keys with a secure and user-friendly mechanism capable of local and remote access management. The system achieved a fingerprint recognition time of approximately 1 second, and a door unlock duration of 5 seconds, offering a good balance between speed and usability. The False Acceptance Rate (FAR) was measured at 1.32%, indicating strong protection against unauthorized access, while the False Rejection Rate (FRR) was 26.32%, primarily due to user misplacement errors rather than system flaws. Additional features, such as a 15-second security lockout after 3 failed fingerprint attempts, OLED display feedback, and buzzer alerts, contributed to a more secure and informative user experience. Real-time alerts and remote control via the Blynk mobile app further enhanced system utility and responsiveness. Despite its advantages, the system has limitations, such as storage for only

three fingerprints and sensitivity to user finger placement. However, these do not overshadow its practical viability. The system is low-cost, modular, and well-suited for residential and small commercial applications. Overall, this project provides a robust foundation for smart home access control. Its successful integration of biometric recognition, IoT communication, and security mechanisms demonstrates the potential of lightweight embedded systems to deliver real-time, user-centric, and secure access management solutions. Future improvements will focus on expanding biometric storage, integrating cloud capabilities, and enhancing sensor accuracy.

VII. Future Work

This study lays the groundwork for further enhancements in biometric access control systems. Future work will explore cloud-based biometric data storage to support additional user profiles, enable remote access recovery, and enhance data redundancy. The system can also be extended to incorporate multi-biometric inputs (e.g., fingerprint + facial recognition + PIN) for multi-factor authentication, improving both convenience and security. Another key direction is the expansion of the framework to support smart city access control infrastructures, such as IoT-enabled gates in residential compounds, office buildings, or campus environments. Integration with advanced IoT and cloud platforms will allow centralized monitoring and control, paving the way for scalable, secure, and efficient access management. Additionally, implementing features such as real-time access logs, time-based permissions, and low-power optimizations (e.g., sleep modes) will further improve usability, energy efficiency, and reliability. These enhancements would help in shaping the next generation of secure, intelligent, and user-centric smart lock systems for modern homes and cities.

Acknowledgment

The authors extend their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) and to the Ministry of Higher Education of Malaysia (MOHE) for their support in this research.

Authors' Contribution

The authors' contributions are as follows: "Conceptualization, J.A.J.A.; methodology, S.M.; software, J.A.J.A. and R.A.H.; validation, S.G.H.; formal analysis, M.F.Y.; investigation, J.A.J.A. and M.F.Y.; resources, S.M.; writing—original draft preparation, J.A.J.A. and S.G.H.; writing—review and editing, R.A.H.; funding acquisition, M.F.Y. and S.G.H.

Data Availability Statement

All the datasets used in this study are available from the Zenodo database (accession number: https://zenodo.org/records/15177144).

References

[1] S. B., A. K., and S. M., "Fingerprint door lock using Arduino UNO R3," International Journal of Multidisciplinary Research, vol. 5, no. 3, 2023. [Online]. Available: https://doi.org/10.36948/ijfmr.2023.v05i03.3819..

[2] S. Kaya, E. Aşkar Ayyıldız, and M. Ayyıldız, "Smart door lock design with Internet of Things," International Journal of 3D Printing Technologies and Digital Industry, vol. 6, no. 2, pp. 201–206, 2022. [Online]. Available: https://doi.org/10.46519/ij3dptdi.1074468.

[3] R. W. Tambunan, A. A. Ar-Rafif, and M. Galina, "Multi-security system based on RFID, fingerprint, and keypad to access the door," Elkha, vol. 14, no. 2, p. 125, 2022. [Online]. Available: https://doi.org/10.26418/elkha.v14i2.57735.

[4] C. Engineering and A. Sultana, "Embedded Linux based multiple features smart door unlock with fingerprint detection," International Journal of Engineering Research & Technology, vol. 8, no. 4, pp. 1032–1035, 2021.

[5] S. Jeong, "Design on novel door lock using minimizing physical exposure and fingerprint recognition technology," International Journal of Informatics and Visualization, vol. 6, no. 1, pp. 103–108, 2022. [Online]. Available: https://doi.org/10.30630/joiv.6.1.858.

[6] A. Tailor and V. Pandya, "IoT-based smart door lock system," International Journal of Research in Applied Science and Engineering Technology, vol. 11, no. 3, pp. 922–926, Mar. 2023. [Online]. Available: https://doi.org/10.22214/ijraset.2023.49470.

[7] S. Kadam, "IoT-based smart door lock system with ESPCAM-32," International Journal of Research in Applied Science and Engineering Technology, vol. 9, no. VI, pp. 4713–4716, Jun. 2021. [Online]. Available: https://doi.org/10.22214/ijraset.2021.35317.

[8] M. Almosawi, "IoT security applied on a smart door lock application," 2018, p. 56.

[9] B. Amoah, X. Wang, J. Zhang, S. Mao, S. C. Periaswamy and J. Patton, "Optimizing Adaptive Power Control for Enhancing Robustness in RFID Sensing," 2024 IEEE International Conference on RFID Technology and Applications (RFID-TA), Daytona Beach, FL, USA, 2024, pp. 157-160, doi: 10.1109/RFID-TA64374.2024.10965152.

[10] M. K. Hamzah, N. Sulaiman, M. Kassim, S. Saaidin and S. B. Kutty, "IoT Smart Door System with Motion Sensing and Facial Recognition," 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2024, pp. 178-183, doi: 10.1109/CSPA60979.2024.10525522.

[11] N. A. Al-Shareefi, S. A. Abbas, M. S. Alkhazraji, and A. A. R. Sakran, "Towards secure smart cities: Design and implementation of smart home digital communication system," Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 1, pp. 271–277, 2021. [Online]. Available: https://doi.org/10.11591/ijeecs.v21.i1.pp271-277.

[12] M. A. Al Rakib et al., "Fingerprint-based smart home automation and security system," European Journal of Engineering and Technology Research, vol. 7, no. 2, pp. 140–145, 2022. [Online]. Available: https://doi.org/10.24018/ejeng.2022.7.2.2745.

[13] C. B. Abdallah, M. K. Baazaoui, I. Ketata, S. Sahnoun, A. Fakhfakh and F. Derbel, "AI-Enhanced Wide Range Fingerprinting Localization for Low-Energy Wireless Sensor Networks," 2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia, 2025, pp. 738-742, doi: 10.1109/SSD64182.2025.10989883.

[14] D. Hercog, T. Lerher, M. Truntič, and O. Težak, "Design and implementation of ESP32-based IoT devices," Sensors, vol. 23, no. 15, 2023. [Online]. Available: https://doi.org/10.3390/s23156739.

[15] O. Dithologo, "The use ESP32 in home," Nov. 2023.

[16] Z. B. Zainon, "RFID door lock with realization of Internet of Things for smart home system," Universiti Sains Malaysia, 2017.

[17] E. Media, S. Rif'an, and M. Rif'an, "Internet of Things (IoT): BLYNK framework for smart home," KnE Social Sciences, vol. 3, no. 12, p. 579, 2019. [Online]. Available: https://doi.org/10.18502/kss.v3i12.4128.

[18] A. W. Y. Khang, S. J. Elias, N. Zulkifli, W. A. Indra, J. A. J. Alsayaydeh, Z. Manap, and J. A. M. Gani, "Qualitative Based QoS Performance Study Using Hybrid ACO and PSO Algorithm Routing in MANET," Journal of Physics: Conference Series, vol. 1502, 2020, doi: 10.1088/1742-6596/1502/1/012004.

[19] J. A. J. Alsayaydeh, M. F. Yusof, M. Z. Abdul Halim, M. N. S. Zainudin and S. G. Herawan, "Patient Health Monitoring System Development using ESP8266 and Arduino with IoT Platform" International Journal of Advanced Computer Science and Applications(IJACSA), 14(4), May 2023, pp. 617-624. http://dx.doi.org/10.14569/IJACSA.2023.0140467.

[20] Z. Hossain, A. K. M., N. B. Hassim, J. A. J. Alsayaydeh, M. K. Hasan, and M. R. Islam, "A tree-profile shape ultra wide band antenna for chipless RFID tags," International Journal of Advanced Computer

Science and Applications, vol. 12, no. 4, pp. 546-550, 2021, doi: 10.14569/IJACSA.2021.0120469.

[21] R. Sahtyawan, A. B. Saputra, and S. Arief, "NodeMCU microcontroller-based disinfectant fluid monitoring system using water level control sensor and ultrasonic sensor," Angkasa: Jurnal Ilmiah Bidang Teknologi, vol. 12, no. 2, pp. 141–150, 2020. [Online]. Available: https://doi.org/10.28989/angkasa.v12i2.770.

[22] N. A. Afifie, A. W. Y. Khang, A. S. B. Ja'afar, A. F. B. M. Amin, J. A. J. Alsayaydeh, W. A. Indra, S. G. Herawan, and A. B. Ramli, "Evaluation Method of Mesh Protocol over ESP32 and ESP8266," Baghdad Science Journal, vol. 18, no. 4, pp. 1398–1401, 2021. doi: 10.21123/bsj.2021.18.4(Suppl.).1397.

[23] V. Shkarupylo, I. Blinov, A. Chemeris, V. Dusheba, J. A. J. Alsayaydeh and A. Oliinyk, "Iterative Approach to TLC Model Checker Application," 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2021, pp. 283-287, doi: 10.1109/KhPIWeek53812.2021.9570055.

[24] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[25] V. V. Shkarupylo, I. V. Blinov, A. A. Chemeris, V. V. Dusheba, and J. A. J. Alsayaydeh, "On applicability of model checking technique in power systems and electric power industry," in Systems, Decision and Control in Energy III, Cham: Springer International Publishing, 2022, pp. 3–21, doi: 10.1007/978-3-030-87675-3_1.

[26] F. Rak and J. Wiora, "Comparison of ESP programming platforms," Computer Science and Information Technology, vol. 2, no. 2, pp. 77–86, 2021. [Online]. Available: https://doi.org/10.11591/csit.v2i2.p77-86.

[27] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[28] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[29] Y.-L. Cheng et al., "Liveness detection in biometrics," IntechOpen, vol. 11, no. Tourism, p. 13, 2016. [Online]. Available: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics.

[30] N. Neema, H. Kagiwala, J. Khavadia, R. Patel, and P. T. Patel, "Smart E-Meter," International Journal of Scientific Research in Engineering and Management, vol. 7, no. 2, pp. 1–4, 2023. [Online]. Available: https://doi.org/10.55041/ijsrem17824.

[31] J. Morai, "Performing open source WiFi mesh sensor network using ESP32," Null, vol. 8, no. 3, pp. 12–17, 2021.

[32] K. A. Hashim, H. H. Qasim, A. E. Hamzah, O. A. Hasan, and M. Al-Jadiri, "Door lock system based on Internet of Things and Bluetooth by using Raspberry Pi," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 5, pp. 2753–2762, 2023. [Online]. Available: https://doi.org/10.11591/eei.v12i5.5134.

[33] A. Poddar, S. Roy, S. Raha, K. Thakur, T. Dasgupta, and S. Maity, "Fingerprint door lock system with temperature sensor," Journal of Physics: Conference Series, vol. 1797, no. 1, 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/1797/1/012052.

[34] G. K. Verma and P. Tripathi, "A digital security system with door lock system using RFID technology," International Journal of Computer Applications, vol. 5, no. 11, pp. 6–8, 2010. [Online]. Available: https://doi.org/10.5120/957-1334.

[35] A. David, M. Chinaza, and J. O. Odinya, "Design and implementation of a door locking system using Android app," International Journal of Scientific and Technology Research, vol. 6, no. 8, p. 8, 2017. [Online]. Available: http://www.ijstr.org.

[36] M. A. Bonaventure, S. Priyadarshini, S. Nayak, and A. Ushadevi, "Smart key: Secure door lock system using NFC enabled smartphone," International Journal of Information Technology, vol. 6, no. 2, pp. 67–70, 2017. [Online]. Available: https://doi.org/10.5923/j.ijit.20170602.12.

[37] A. Ahmed et al., "Design and implementation of an IoT-based smart door lock system," in 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICMEAS58693.2023.10379324.

[38] A. Singh et al., "IoT based smart lock," International Research Journal of Modern Engineering and Technology and Science, no. 05, pp. 2582–5208, 2022.

[39] M. Nguyen et al., "Toward the Impact of Video Luminance on the Energy Consumption of OLED TVs and Viewer Perception," 2025 IEEE Green Technologies Conference (GreenTech), Wichita, KS, USA, 2025, pp. 118-122, doi: 10.1109/GreenTech62170.2025.10977708.

[40] M. Shen, "Smart Door Lock System Design Based on UWB Technology," 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS), Yanji, China, 2024, pp. 485-488, doi: 10.1109/EIECS63941.2024.10800017.

[41] N. R. S, R. Venkatasamy, J. A. Dhanraj, S. Aravinth, K. Balachandar and D. N, "Design and Development of IOT based Smart Door Lock System," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1525-1528, doi: 10.1109/ICICICT54557.2022.9917767.

[42] M. A. Khan et al., "Prototype Model of an IoT-based Digital and Smart Door Locking System with Enhanced Security," 2022 14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 2022, pp. 1-7, doi: 10.1109/MACS56771.2022.10023385.

[43] A. Saroha, A. Gupta, A. Bhargava, A. K. Mandpura and H. Singh, "Biometric Authentication Based Automated, Secure, and Smart IOT Door Lock System," 2022 IEEE India Council International Subsections Conference (INDISCON), Bhubaneswar, India, 2022, pp. 1-5, doi: 10.1109/INDISCON54605.2022.9862840.

Modelling Cloud Computing Adoption in the Malaysian Healthcare

A Resource-Based View and Technology-Organisation-Environment (TOE) Perspective

Normilah Mohd Noh[1], Nurhizam Safie Mohd Satar[2], Hasimi Sallehudin[3], Ibrahim Hassan Mallam[4], Surya Sumarni Hussein[5], Nur Azaliah Abu Bakar[6]
School of Digital Health, KPJ Healthcare University, Malaysia[1]
Faculty of Information and Science Technology, Universiti Kebangsaan Malaysia, Malaysia[1, 2, 3]
Library Department, Nasarawa State University, Nigeria[4]
Faculty of Computer Science and Mathematics, Universiti Technology MARA, Malaysia[5]
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia[6]

***Abstract*—Cloud computing is increasingly reshaping the global IT landscape, offering scalable and efficient solutions across industries, including the healthcare sector. This study investigates the determinants of cloud computing adoption in the Malaysian healthcare industry by integrating the Resource-Based View (RBV) and Technology-Organisation-Environment (TOE) frameworks. Emphasising internal organisational capabilities, the study excludes traditional Information Systems (IS) models to maintain theoretical coherence with RBV's strategic orientation toward firm-level resource advantages. Data were collected from 265 respondents across 127 healthcare organisations and analysed using Partial Least Squares Structural Equation Modelling (PLS-SEM). The study also proposes an extended taxonomy of cloud services contextualised for healthcare, strengthening the theoretical underpinnings and practical applicability of cloud adoption strategies in this domain. The findings reveal that among IT capabilities, managerial IT capability exerts the most substantial influence on adoption, followed by relational and technical capabilities. Within the TOE dimensions, regulatory support emerged as the most critical enabler, while business resources, change management, organisational culture, and vendor support also demonstrated significant positive effects. The results offer empirical validation for a comprehensive conceptual model grounded in RBV and TOE, providing both theoretical insights and practical guidance for healthcare organisations aiming to strengthen IT capabilities, optimise organisational readiness, and align with external institutional drivers for successful cloud migration.**

***Keywords*—Adoption; cloud computing; Malaysian healthcare; partial least squares-structural equation modelling; resource-based view (RBV)**

I. Introduction

In recent years, cloud computing has experienced significant global growth, evolving into a fundamental component of modern organisational IT infrastructures. As businesses face intensifying market pressures and increasing demands for agility, many have adopted cloud-based solutions not merely as a technological upgrade but as a strategic enabler of innovation and operational efficiency. Defined as the delivery of computing services, such as servers, storage, applications, and platforms over the internet, cloud computing builds upon the principles of distributed and parallel computing [1], offering a scalable and reliable environment for resource sharing and remote access [2] [3]. The National Institute of Standards and Technology (NIST) defines cloud computing as "a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., network, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [4]. The NIST definition reveals the main advantages of cloud computing as shareability, scalability, and cost reduction.

The transition to greater cloud computing adoption in some industries is propelled by firms utilising the benefits of cloud computing to maintain market competitiveness. Numerous pieces of evidence indicate that company agility and competitiveness, cost-effective disaster recovery, swift resilience, and rapid deployment are significant advantages of cloud usage [5]. In [6], the authors indicate that small organisations derive profits via enhanced collaboration with customers and staff, whereas medium enterprises achieve profitability through the establishment and expansion of new ventures, positioning themselves for global growth. In addition, the adoption of cloud computing aligns with the Fourth Industrial Revolution (IR4.0), and numerous entities, including those in the healthcare sector, have begun their journey towards complete digitisation. For healthcare organisations, this initiative is intended to enhance the quality of systems for managing patient or health treatment information [7]. In summary, a variety of research has shown how cloud computing's IT features enhance organisational value. Here are several specific ways in which adopting cloud computing can be advantageous for the healthcare sector:

1) It enables healthcare organisations to reduce capital expenditure by eliminating the need for costly on-premises data centres, which are typically expensive due to the mission-critical, lifesaving functions they must support [8]. Additionally, it minimises service downtime through high availability features essential for uninterrupted clinical operations and reduces operational inefficiencies through

scalable, on-demand resource provisioning tailored to dynamic healthcare demands.

2) Cloud computing also enhances information systems and business agility by enabling rapid execution, supporting critical healthcare operations through high availability, and facilitating swift recovery via built-in disaster recovery capabilities [9].

3) Strengthens the quality of patient health and well-being through innovative implementations such as Electronic Health Records (EHR), Personal Health Records (PHR), diagnostic and pharmaceutical cloud services, and support for first responders.

Although cloud computing offers benefits such as enhanced organisational agility, flexibility, and reduced IT infrastructure costs, many organisations, particularly in the healthcare sector, face challenges in developing effective strategies for cloud deployment, including internal readiness assessments and vendor selection (e.g., [9]). Moreover, the rapid pace of cloud adoption has led to shortened IT planning horizons, potentially undermining long-term implementation success [10]. These challenges are further compounded by the unique characteristics of the healthcare industry, which differ significantly from sectors such as manufacturing, transportation, and financial services, as reported in previous information systems studies. The operational complexities of the healthcare sector, along with the participation of constitutional bodies, result in a complex environment that significantly influences the adoption of cloud computing. The diverse organisational structures within healthcare, encompassing for-profit, non-profit, government-owned, and privately owned entities, each with distinct objectives and motivations, contribute to this complexity. This duality, characterised by medical professionals alongside administrative and support staff, further shapes the industry's unique administrative landscape. The reliance of private hospitals on insurance companies and external agencies for reimbursement introduces additional layers of complexity, as these entities often operate with their own agendas. Such complexities can impact the implementation of cloud computing solutions within the healthcare sector. The integration of cloud computing necessitates careful consideration of strategic planning to align technological advancements with organisational goals [11].

Despite the promise of cloud computing to improve organisational agility and operational flexibility, many healthcare organisations continue to face difficulties in devising coherent implementation and deployment strategies [12]. These challenges often result in delayed adoption, potentially placing them at a competitive disadvantage in an increasingly digital healthcare environment. To address this issue, the study proposes a conceptual model that integrates the TOE framework with IT capability, grounded in the RBV theory, to examine the factors influencing cloud computing adoption in the Malaysian healthcare sector.

The remainder of this study elaborates on the theoretical background underpinning the study, particularly the integration of the RBV and the TOE framework. It then defines the key factors influencing cloud computing adoption, along with the research hypotheses. This is followed by a description of the research methodology and a discussion of the empirical findings. The study concludes with a reflection on the study's limitations and suggests for future research possibilities.

II. Theoretical Background and Research Framework

A. Cloud Computing in Healthcare

Cloud computing has evolved into a foundational technology in the digital economy, enabling scalable, on-demand access to computing resources via the internet. It supports not only traditional IT functions but also emerging technologies such as big data analytics and the Internet of Things (IoT), which generate and process vast volumes of data in real-time. By abstracting physical infrastructure into a virtualised environment, cloud computing allows organisations to store, manage, and analyse data seamlessly, regardless of geographical location. According to [13], the taxonomy of cloud services is classified into six categories based on their characteristics from the perspective of IT vendors, cloud service users and software. This classification has been restructured by [14] into three basic groups based on the characteristics of the tree-structure taxonomy and is summarised in Table I.

Given the multifaceted nature and specialised demands of cloud computing in the healthcare industry, it differs significantly from its application in other sectors. Several taxonomies have been proposed to conceptualise its various dimensions. Among them, the taxonomy developed by [15] provides a foundational classification by organising cloud computing knowledge into two primary dimensions, which are principle knowledge and how-to knowledge. The principal knowledge includes service models, deployment types, and technical characteristics such as scalability, elasticity, ubiquity, cost efficiency, shareability, interoperability, and security, whereby the how-to knowledge relates to practical applications such as task support, user types, device integration, and the degree of patient data involvement. This framework offers a structured view of how cloud services are characterised and operationalised in healthcare environments. However, while it provides useful categorisation, it does not explicitly account for the organisational capabilities or contextual factors that influence adoption decisions. Therefore, this study seeks to advance through further theoretical integration.

B. IT Capabilities and Resource-Based View Theory (RBV)

The Resource-Based view or theory (RBV) is a well-established strategic management theory that emphasises the internal resources of an organisation as the foundation for achieving sustainable competitive advantage. In the context of information systems, RBV suggests that firms can enhance performances by effectively integrating external IT resources with internal IT capabilities [16]. This theoretical perspective has been widely applied in IT research to conceptualise how organisations derive value from technological investments and leverage IT assets to strengthen operational effectiveness and strategic positioning.

TABLE I COMPUTING TAXONOMY BY TREE-STRUCTURE FRAMEWORK AND STAKEHOLDER PERSPECTIVES

Taxonomy	Characteristics	Example
Cloud Architecture	is the design of software applications that leverage the internet to deliver on-demand services • Software-as-a-Service (SaaS) • Platform-as-a-Service (PaaS) • Developers implementing cloud applications • Infrastructure-as-a-service (IaaS) • [(Virtualization, Storage Network) as-a-Service] • Hardware-as-a-Service • Cloud Layered Architecture	
Virtualization Management	is a technology that abstracts the coupling between hardware and operating systems	Server, Storage, and Network virtualization
Service	Three different categories of cloud services: • Software-as-a-Service (SaaS) • Is a multi-tenant platform • Platform-as-a-Service (PaaS) • Supported programming languages and environment • Supported operating systems • Supported applications and frameworks • Infrastructure-as-a-service (IaaS) • Supported operating systems • Supported applications and frameworks • Provided infrastructure management tools • Virtualization technology	• Email services, Google Docs, and Dropbox usage – (SaaS) • Supports platforms for system/application development from a programming perspective – (PaaS) • Offer similar functions as data centre – (IaaS)
Fault Tolerance	Is the capability of a system to remain fully operational despite hardware or software failures, achieved through redundancy.	
Security	Is the focal concern in terms of data, infrastructure and virtualization etc.	
Other Issues	Relate to critical operational and architectural mechanisms, including: • Load balancing • Interoperability • Scalable data storage • Internal security • Cloud performance	

While resources such as physical assets, human capital, and organisational processes must be valuable, rare, and inimitable to confer competitive advantage [16], capabilities are distinct in that they refer to an organisation's ability to combine, mobilise, and deploy these resources effectively [17]. Capabilities enable organisations to synthesise knowledge, integrate technologies, and adapt to dynamic environments, making them harder for competitors to replicate [18]. For instance, [19] introduced the concept of IT capabilities as a form of organisational competency that supports core functions and strategy execution.

According to RBV, internal resources that contribute to organisational success must possess three essential characteristics, i.e., value, heterogeneity, and imperfect transferability [17]. Building on this premise, [20] examined how various organisational resources influence firm performance and adopted RBV to explore the role of modern IT resources in developing strategic core competencies. Their study further extended the IT competency model proposed by [19] by integrating organisational capacity dimensions and investigating the complementary effect of resources. The findings demonstrated that synergistic combinations of organisational capabilities and IT resources create complex, path-dependent assets that are deeply embedded in a firm's history, culture, and experience, making them extremely difficult to imitate. Furthermore, their research empirically validated the mediating role of organisational capability in the relationship between IT resources and firm performance. Organisational capability, in this context, reflects the firm's responsiveness to environmental changes and managerial challenges, reinforcing the idea that synergy between IT resources and organisational capabilities is crucial for superior performance.

In separate contemporary research, [21] state that different IT capabilities depend on different structural mechanisms, and this reinforces the concept that the value of various IT capabilities may derive from more complex relationships [19]. The variety of IT capabilities includes Management IT Capabilities, Technical IT Capabilities, and Relationship IT Capabilities.

C. TOE Framework

The TOE framework, created in 1990 by Tornatzky and Fleischer, contends that the firm's adoption and implementation of technological innovations are influenced by the technological context, the organisational context, and the environmental context. Technological context describes factors describing the adoption of an existing technology as well as new technologies relevant to the firm. The term organisational context describes how organisational factors like scope, organisational size, and the amount of available slack resources affect an organisation. Meanwhile, the environmental context refers to the influence of external factors, including those related to industry, competitors, and government, on the adoption of technology.

TOE has been widely recognised for its suitability in studying technology adoption due to its holistic perspective and flexibility in application across various industries, including healthcare. The healthcare sector, in particular, presents unique challenges in adopting cloud computing, such as compliance with data security regulations, complex stakeholder structures,

and critical operational sensitivity. Previous studies have validated the robustness of TOE in cloud adoption research, such as [6] [22] [23] and [24] which demonstrated its effectiveness in identifying influential factors in diverse organisational settings, including the Malaysian public sector. Furthermore, as cloud computing represents a significant paradigm shift in IT service delivery, the integration of TOE with complementary theories such as the RBV can offer deeper insights into how contextual conditions interact with internal capabilities to influence adoption outcomes. Numerous previous studies have utilised either the RBV or the TOE framework to examine cloud computing adoption, but with restricted integration or a lack of sector-specific emphasis. Table II indicates that although RBV-based research (e.g., [19], [18]) has highlighted the strategic significance of IT capabilities, the majority of studies lack application within the healthcare context. Conversely, TOE-based studies (e.g., [22], [12]) have investigated external and organisational factors influencing adoption, but neglecting the internal IT capability as a resource. This study, therefore, addresses these gaps by integrating IT capability (from RBV) into the TOE framework and contextualising the model specifically for the Malaysian healthcare sector.

D. Integrating RBV and TOE in Cloud Computing Taxonomy

To suit the theoretical orientation of this study, the original taxonomy of cloud computing services developed by [12] has been adapted to reflect the integration of IT capability within the RBV and contextualised using the TOE framework. While the original taxonomy classifies cloud computing knowledge into principle and how-to dimensions, the revised version introduces a strategic layer by aligning each element with relevant IT capability attributes and the three TOE contexts. This adaptation allows for a more structured analysis of cloud computing adoption, particularly in understanding how specific service characteristics interact with organisational capabilities and environmental influences within the healthcare sector. The extended taxonomy of cloud computing services in healthcare organisations, as shown in Table III, synthesises key constructs from the Resource-Based View and TOE frameworks, offering a structured lens through which cloud adoption determinants can be categorised. This taxonomy was developed to guide the identification of strategic and contextual factors relevant to cloud implementation in healthcare settings.

TABLE II COMPARATIVE REVIEW OF CLOUD ADOPTION STUDIES IN HEALTHCARE

Study	Domain	Framework / Theory	Focus Area	Healthcare Context	Key Contribution	Gap Addressed by Current Study
[20]	Telehealth	RBV	Telehealth systems and organisational performance	√	Telehealth system analysis	Does not extend TOE taxonomy
[18]	Cloud Computing	RBV	Cloud capability impact	X	Cloud capabilities and business impact	No TOE or healthcare link
[19]	General IS	RBV	IT capability and firm performance	X	Empirical study on IT as resource	No cloud or healthcare context
[17]	Strategic Management	RBV	Resource advantage	X	Foundational RBV theory	Not applied to IT or healthcare
[6]	SMEs	TOE	Cloud adoption & performance	X	Analytical model using TOE	Not applied to healthcare
[15]	Healthcare	TOE (contextual factors)	Adoption determinants	√	Systematic TOE-based review	Does not integrate RBV
[22]	Manufacturing firms	TOE	Usage and performance	X	Survey-based TOE adoption study	Lacks IT capability dimension
[11]	Public Sector	ROCCA and TOGAF	Strategic planning	X	Adoption planning framework	Not TOE or RBV-based
[25]	IT Outsourcing	Theoretical synthesis	Cloud-sourcing decisions	X	Literature reflection on IT sourcing	Lacks empirical healthcare context
[8], [26]	Healthcare	CHIS Utilization Determinants	HIS utilization	√	Empirical studies - behavioural & system-related determinants	No RBV/TOE integration
[9]	Healthcare	Systematic Review	Opportunities and issues	√	Comprehensive review of cloud healthcare	No model or framework
[21]	Cloud Computing	IT Capabilities	Success factors	X	Cloud delivery models and performance	No TOE or healthcare scope
[27]	Public Sector (General)	Extended UTAUT + Performance Impact Model	Cloud adoption drivers, IT personnel performance, organizational factors	X	Cloud adoption impact on IT staff in public sector	Does not examine cloud adoption in healthcare using an integrated RBV–TOE lens

TABLE III EXTENDED TAXONOMY OF CLOUD COMPUTING SERVICES ALIGNED WITH IT CAPABILITIES & TOE FRAMEWORK

Dimension 1: *Principal Knowledge* and its Characteristics		
Attribute	**IT Capability Alignment**	**TOE Alignment**
1a) Service Form (SaaS, PaaS, IaaS)	Technical IT Capability (ability to deploy/manage various cloud service models)	Technological (nature of technology adopted)
1b) Deployment Model (Public, private, hybrid, community)	Managerial IT Capability (decision-making, policy, and governance around deployment models)	Technological
1c) Targeted cloud advantage (Scalability, elasticity, ubiquity, cost efficiency, shareability, interoperability, security)	Technical & Managerial IT Capability (technical – achieving performance; managerial – aligning benefits to business goals)	Technological & Organisational
1d) Timeliness (Real time, not real time)	Technical IT Capability (infrastructure readiness)	Technological
Dimension 1: How-to Knowledge and its Characteristics		
Attribute	**IT Capability Alignment**	**TOE Alignment**
2a) Supported task (Clinical, administrative, strategy, research)	Managerial IT Capability (process integration, alignment with business/clinical goals)	Organisational
2b) User (Patient, medical staff, family member)	Relational IT Capability (coordination among stakeholders, user-oriented design)	Organisational & Environmental
2c) Service Delivery device (Independent, adapted, specialized)	Technical IT Capability (device-level implementation and configuration)	Technological
2d) Patient data involvement (Internal, external, no involvement)	Relational IT Capability (data sharing, collaboration with external entities)	Environmental

III. FACTORS DEFINITION AND RESEARCH ASSUMPTIONS

The development of this research model involves identifying critical factors that influence cloud computing adoption in the healthcare industry. The constructs and measures used in this study are informed by an extended taxonomy of cloud computing services adapted for the healthcare context (Table III), allowing for more context-sensitive operationalisation of cloud-related capabilities. Drawing upon validated constructs from prior literature, each factor is conceptually defined to reflect its relevance within the Malaysian healthcare context. These factors are categorised according to their theoretical origin, whether rooted in IT capability dimensions or TOE contextual elements, and operationally based on their expected influence on cloud adoption. The following section outlines these factors and their definitions, followed by the formulation of the research assumptions underpinning this study:

A. IT Capability

- Managerial IT capability

Managerial IT capability reflects an organisation's ability to leverage the knowledge and expertise of its IT leaders to align emerging technologies with strategic objectives. This capability encompasses both business acumen and technical proficiency, enabling IT managers to evaluate technological potential, orchestrate complex implementation efforts, and integrate new solutions into existing workflows efficiently and cost-effectively [19] [21]. Within the RBV, such capabilities are classified as strategic assets that are rare, valuable, and difficult to imitate, thereby contributing to sustained competitive advantage [20]. In the context of healthcare, where operational complexity and regulatory demands are high, managerial IT capability plays a critical role in guiding cloud adoption decisions. Past studies also suggest that strong managerial capability enhances technology assimilation by enabling better alignment of IT with organisational goals and mitigating risks such as obsolescence or resource misallocation [19] [21]. Accordingly, the following hypothesis is proposed:

H1: Managerial IT capability has a positive influence on the adoption of cloud computing.

- Technical IT capability

Technical IT capability refers to an organisation's ability to acquire, manage, and utilise both tangible and intangible IT assets to support operational and strategic needs. Tangible assets include infrastructure components such as servers, databases, and networking equipment, while intangible assets comprise specialised technical expertise, domain-specific knowledge, and collaborative IT practices across business units [19] [28]. Within the RBV framework, these technical resources serve as foundational enablers that, when effectively deployed, can enhance organisational agility, streamline business processes, and reduce operational cycle times [21] [28]. In the context of cloud computing adoption, especially within healthcare organisations where data sensitivity, interoperability, and infrastructure readiness are critical, strong technical capability enables seamless integration, scalability, and responsiveness. This study suggests that the extent to which healthcare organisations adopt cloud computing is significantly influenced by the maturity of their technical IT capabilities. We, therefore, made the hypothesis as follows:

H2: Technical IT capability has a positive influence on the adoption of cloud computing.

- Relational IT capability

Relational IT capability refers to an organisation's ability to cultivate enduring, trust-based relationships with stakeholders, including technology partners, cloud vendors, and inter-organisational collaborators [21]. Under the lens of the RBV, this capability is considered an intangible strategic resource that is valuable, rare, and difficult to replicate, thereby contributing to sustained competitive advantage [20]. In the context of cloud

computing adoption, especially in the healthcare sector where data privacy, regulatory compliance, and trust are critical, relational IT capability plays a vital role in reducing uncertainty, enabling knowledge-sharing, and enhancing mutual commitment between partnering organisations.

Trust is a core element of relational capability which facilitates seamless collaboration, minimises perceived risks, and lowers transaction costs, particularly when adopting externally hosted services. When healthcare organisations possess strong relational capabilities, they are better positioned to coordinate with cloud providers, resolve implementation challenges, and derive strategic value from the adoption process. Therefore, relational IT capability serves as an essential enabler of cloud computing adoption, consistent with RBV's proposition that organisational capabilities can be leveraged to secure competitive advantages.

H3: Relational IT capability has a positive influence on cloud computing adoption. That is, a relational IT capability characterised by trust contributes to the factor for cloud computing adoption.

B. TOE Framework

- Technological capability

Technological capability refers to the ability of a company to mobilise and apply technological resources in support of innovation and strategic objectives. Within the TOE framework, it is typically conceptualised through three critical dimensions, consisting of technology resources, human resources and business resources [22]. These dimensions encompass the technological infrastructure, the technical expertise of personnel, and the alignment of technology with core business functions.

In the context of the RBV, technological capability is considered a strategic organisational resource that can contribute to competitive advantage when it exhibits characteristics of value, rarity, inimitability, and non-substitutability. Particularly in the healthcare industry, where technology adoption involves complex integration and regulatory sensitivities, technological capabilities are key enablers of cloud computing adoption. By framing these capabilities through RBV, this study emphasises how the deployment of cloud solutions depends not only on the availability of technology but also on the organisation's internal capacity to exploit these resources effectively.

1) Technology resources: Technology resources refer to the foundational assets, including hardware, software, communication technologies, IT applications, and personnel, that enable the implementation of information systems. These resources underpin cloud computing environments by facilitating access, scalability, and data management. From the RBV perspective, the strategic value of such resources depends not merely on their availability but on how they are integrated and leveraged to create organisational advantage. Well-developed and effectively utilised technological infrastructures are expected to enhance operational agility and support innovation. Based on this rationale, the following hypothesis is proposed:

H4: Technology resources have a positive influence on cloud computing adoption.

2) Human resources: Human resources refer to the technical knowledge, skills, and experience of IT personnel and professionals within an organisation. Their ability to design, implement, and manage cloud-based systems is crucial in healthcare settings where data security, compliance, and service availability are critical [22]. Under RBV, these human competencies are intangible yet valuable resources that contribute to the effective use of technology and foster innovation. It is therefore highlighted here that human resource (as denoted by skills and capability to manage cloud services) has a significant effect on cloud computing adoption:

H5: Human resources have a positive influence on cloud computing adoption.

3) Business resources: Business resources are defined as the capabilities and assets, including relations with suppliers, training, planning and financial resources [22]. In other words, the drive for business resources relates to consumers' personal commitment towards issues faced by the suppliers and their individual-level activities that are intended to develop a conducive environment for effective usage of technology. They are important factors in the appropriate and continuing use of this technology. H6 thus describes the nature of the relationship between the use of cloud computing and business resources.

H6: Business resources have positive influence on cloud computing adoption.

- Organisational capability

Organisational capability in the TOE framework encompasses the internal characteristics of a firm that influence its readiness and ability to adopt new technologies [22]. In cloud computing adoption, particularly within the healthcare sector, organisational capability determines how effectively a firm can align its people, processes, and internal structures to facilitate change and integrate innovations. When viewed through the RBV, organisational capabilities such as adaptability, leadership support, and shared cultural values are considered intangible yet strategic resources that enable firms to transform technical potential into realised value. These capabilities affect the success of cloud adoption by shaping internal readiness, fostering collaboration, and promoting innovation acceptance.

4) Change management: Cloud-based services differ from other technologies as they require changes in business processes and the customisation of the cloud-based services to fit existing business processes. Cloud implementation involves a series of coordinated IT customisation and implementation projects [22]. This requires firms to acquire the capability to manage individual projects, as well as multiple projects operating in a coordinated and integrative manner. In the context of RBV, this ability represents a dynamic capability that allows the firm to reconfigure existing competencies and adapt swiftly to disruptive technologies. We contend that businesses that can modify and integrate cloud services as well as alter or

redesign business processes increase the use of cloud computing services. Therefore, on the basis of these discussions, H7 is developed for testing the relation between change management and cloud computing adoption:

H7: The cloud computing adoption will be negatively correlated with resistance to change.

5) Organisation culture: Organisational culture refers to the shared beliefs, values, and norms that influence behaviour within the firm. A culture that supports learning, innovation, and openness to technological advancement is believed to be critical to the success of technology implementation and usage [29]. From the RBV perspective, such a culture constitutes a firm-specific resource that is difficult to replicate and serves as a foundation for continuous improvement and competitive advantage in technology use. The purpose of H8 is to determine whether corporate culture has a significant impact on cloud computing adoption.

H8: Organisation cultures have a positive influence on cloud computing adoption.

- Environmental capability

The environmental context within the Technology-Organisation-Environment (TOE) framework captures external factors that shape an organisation's ability and readiness to adopt technological innovations [22]. In cloud computing, especially within regulated and mission-critical sectors like healthcare, these environmental enablers or barriers play a significant role in adoption decisions. From the perspective of the RBV, environmental capability is not traditionally classified as a firm-internal resource. However, when organisations are able to leverage external support, such as from technology vendors or regulatory bodies, effectively, they can transform these external enablers into strategic assets. The ability to cultivate strong partnerships and align with regulatory frameworks becomes a valuable dynamic capability that complements internal resources and contributes to sustained competitive advantage.

6) Vendor support: Vendor support refers to the assistance and collaboration provided by cloud service providers during implementation, integration, and maintenance. This includes technical support, training, and knowledge transfer, which are crucial for healthcare organisations with limited internal IT capabilities. As noted by [21], vendor competence, reliability, and credibility are instrumental in reducing uncertainty and facilitating adoption. When organisations can establish cooperative, trustworthy relationships with vendors, this inter-organisational trust and knowledge flow becomes a valuable relational resource aligned with the RBV, enhancing the organisation's ability to exploit technological innovation. H9 is an experiment to see if vendor support influences the cloud computing adoption:

H9: Vendor support has a positive influence on cloud computing adoption.

7) Regulatory support: Regulatory support pertains not only to external government initiatives but also to how organisations internally interpret, respond to, and operationalise such regulations. This includes maintaining compliance practices, ensuring vendor accreditation, and aligning operations with legal frameworks such as Malaysia's PDPA 2010. From the RBV perspective, an organisation's ability to systematically respond to regulatory requirements is by embedding compliance mechanisms, legal safeguards, and governance practices into internal routines, which can be considered a valuable and rare organisational capability. Such capability enables the organisation to mitigate legal risks, build legitimacy, and sustain operational continuity, thereby transforming external institutional pressure into strategic advantages. This view is reinforced by [22], who recognised regulatory support as a key enabler in cloud computing adoption. Thus, the following theories are put forth:

H10: Regulatory support has a positive influence on cloud computing adoption.

Grounded in the resource-based theory and aligned with the proposed hypothesis, the integration of IT Capability and the TOE framework informs the development of the proposed model, as depicted in Fig. 1.

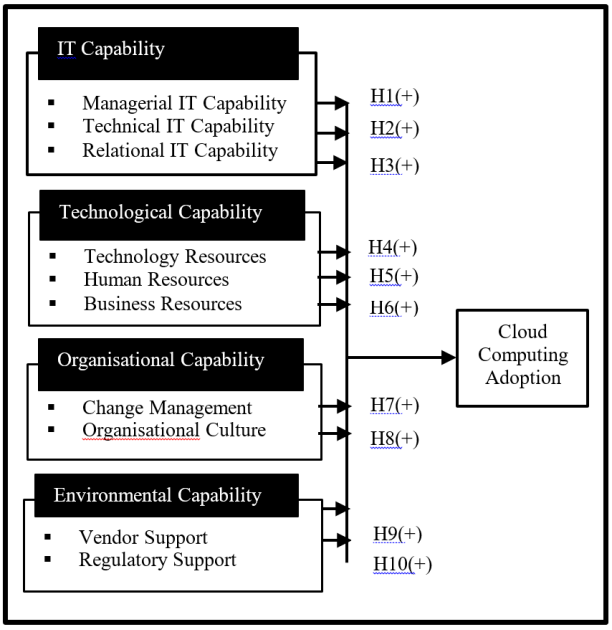


Fig. 1. Conceptual model.

IV. Methodology

A. Data Collection and Sample

This study employed a purposive sampling technique targeting key personnel involved in strategic IT decision-making within healthcare institutions. Respondents were selected based on their organisational roles that influence or oversee cloud computing adoption processes. These included Chief Executive Officers (CEOs)/ *Timbalan Pengarah Pengurusan*, Chief Financial Officers (CFOs)/ *Ketua Unit Kewangan*, Chief Information/ Technology Officers (CIOs/CTOs), IT/IS

Managers/ *Ketua Unit IT*, and IT Executives. Data were gathered through a combination of online surveys and the drop-off and pick-up method, allowing flexibility in participation while ensuring a high response rate.

The sample was drawn from three distinct groups of healthcare organisations, namely KPJ Specialist Hospitals, private healthcare facilities unaffiliated with KPJ, and tertiary public hospitals under the Ministry of Health Malaysia (MOH). According to the National Medical Research Register (NMRR), there were 238 registered healthcare organisations at the time of the study, of which 233 were active.

Out of 37 MOH tertiary hospitals, 28 (75.7%) participated, with 26 confirming active cloud usage. In the private sector, 130 out of 196 licensed private hospitals responded (66.3%), of which 56 reported cloud adoption. This includes a high response from KPJ Specialist Hospitals (26 of 29, or 89.7%), as well as contributions from other private networks such as Columbia Asia (9 of 20, or 45%) and IHH (3 of 18, or 16.7%). Among 129 standalone private hospitals, 92 (71.3%) responded to the survey.

In total, 127 healthcare organisations responded to the survey, yielding a 54.5% response rate. Of these, 82 organisations (64.6%) reported having adopted cloud computing to varying degrees. This moderate level of adoption reflects a gradual yet cautious progression toward digital transformation within Malaysia's healthcare sector. It further highlights the importance of examining the technological, organisational, and environmental conditions, as well as internal IT capabilities, that influence cloud computing adoption.

The research was ethically approved by the Medical Research and Ethics Committee (MREC) Malaysia under reference number NMRR-23-03142-NH7 (IIR). Respondents were asked to evaluate a conceptual model comprising both exogenous and endogenous variables influencing cloud computing adoption. The instrument included 29 measurement items adapted from established technology adoption frameworks and TOE literature [30], [22] – [31], alongside 12 items representing IT capability dimensions as conceptualised in the Resource-Based View (RBV). Responses were captured using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Describing the descriptive statistics, the positions of respondents in the survey reveal that the majority (59.9%) are IS senior executives or managers. This dominant group reflects a strong representation from operational and technology management roles within healthcare institutions. Additionally, 14.7% of the respondents are CEOs, while CFOs and CIOs/CTOs each constitute 12.7%. In terms of industry experience, a significant portion of the respondents (56%) have been in the industry for more than 10 years, indicating a mature and experienced respondent base. This is critical as it adds depth and credibility to the survey findings. Meanwhile, 23% have 6 to 10 years of experience, and only a small percentage (21.1%) have less than 6 years in the field. The operational longevity of hospitals in the survey is also notable. Over half (54.8%) have been in operation for more than 20 years, and another 42% fall within the 2 to 20-year category. Only 3.2% are newly established (less than 1 year). Regarding staff size, the distribution shows that a large number of hospitals (65%) employ fewer than 1,000 full-time staff, with 32.9% having less than 500 employees. Conversely, around 35% of hospitals are larger institutions with over 1,000 employees. The patient bed capacity further reinforces the predominance of smaller healthcare facilities in the sample. Approximately 61.1% of the hospitals have fewer than 250 beds, while 21.8% have between 250 and 400 beds. Only 17.1% of the hospitals are large facilities with more than 400 beds. In terms of financial strength, nearly half (49.2%) of the hospitals reported annual sales of less than RM100 million, and another 43.7% earned between RM100 and RM500 million. Only a small minority generated revenues above RM500 million.

B. Data Analysis

- The measurement model

SmartPLS 4.1.1.2 was used to analyse the research model in order to estimate the relationships within the structural model and evaluate the psychometric qualities of the measurement items. To get stable standard error estimates, a bootstrapping process using 5,000 resamples of the original dataset (n = 252) was used. The analysis began with an evaluation of reliability and validity. As shown in Table III, the composite reliability (CR) values ranged from 0.936 to 0.988, and Cronbach's alpha values ranged from 0.91 to 0.987, both exceeding the recommended threshold of 0.70 [32]. These results demonstrate that the constructs exhibit strong internal consistency. Furthermore, the Average Variance Extracted (AVE) values also surpassed the 0.50 benchmark, indicating that more than 50% of the variance in each indicator is explained by its corresponding latent construct, thereby supporting convergent validity [33].

Discriminant validity was also confirmed. As detailed in Table IV, the square root of the AVE for each latent construct (bolded along the diagonal) is greater than the corresponding inter-construct correlations, affirming that the constructs are statistically distinct. Additionally, all inter-item correlations were below the 0.90 threshold [33], further reinforcing construct distinctiveness.

Table V displays the item loadings and cross-loadings, confirming that each indicator aligned most strongly with its designated construct. Items that failed to meet the minimum recommended loading threshold of 0.70 [32] were excluded from subsequent analysis to ensure the integrity of the measurement model. The remaining indicators showed strong convergent validity, as evidenced by high t-statistics in the outer model, indicating significant contributions to their respective latent variables. To evaluate multicollinearity, the Variance Inflation Factor (VIF) values were examined for all ten latent constructs. As presented in Table VI, the VIF values ranged between 1.141 and 4.163, remaining well below the conservative cut off of 5.0 [34], thus confirming the absence of problematic multicollinearity. Furthermore, the f^2 effect size was computed to determine the relative influence of exogenous constructs on endogenous variables. Based on thresholds provided by [35], 0.02 (small), 0.15 (medium), and 0.35 (large), i.e., the f^2 values in this study ranged from 0 to 0.118, reflecting predominantly small effect sizes.

TABLE IV CONSTRUCT CORRELATIONS, CONSISTENCY AND RELIABILITY

Construct	CR	Alpha	AVE	1	2	3	4	5	6	7	8	9	10	11	12
Business Resources	0.963	0.942	0.897	**0.947**	**0.947**										
Change Management	0.983	0.979	0.907	0.727	0.727	**0.952**									
Cloud Computing Adoption	0.988	0.987	0.903	0.079	0.786	0.811	**0.951**								
Human Resources	0.973	0.959	0.924	0.786	0.804	0.694	0.731	**0.961**							
Managerial IT Capability	0.956	0.939	0.845	0.804	0.128	0.112	0.21	0.102	**0.919**						
Organisational Culture	0.936	0.91	0.786	0.128	0.469	0.483	0.558	0.395	0.044	**0.886**					
Regulatory Support	0.972	0.962	0.897	0.469	0.684	0.738	0.802	0.627	0.008	0.527	**0.947**				
Relational IT Capability	0.956	0.94	0.812	0.684	0.565	0.641	0.717	0.543	-0.033	0.452	0.758	**0.901**			
Technical IT Capability	0.972	0.956	0.919	0.565	0.439	0.553	0.597	0.425	-0.084	0.359	0.593	0.586	**0.959**		
Technology Resources	0.974	0.964	0.902	0.439	0.575	0.585	0.643	0.527	-0.034	0.423	0.678	0.683	0.614	**0.95**	
Vendor Support	0.986	0.982	0.934	0.575	0.692	0.575	0.695	0.639	0.177	0.467	0.611	0.53	0.341	0.477	**0.966**

*Bold numbers on the diagonal are the square root of the AVE

TABLE V LOADINGS AND CROSS-LOADINGS

	Business Resources	Change Management	Cloud Computing Adoption	Human Resources	Managerial IT Capability	Organisational Culture	Regulatory Support	Relational IT Capability	Technical IT Capability	Technology Resources	Vendor Support
MC1	0.145	0.116	0.19	0.112	**0.917**	0.021	-0.024	-0.076	-0.127	-0.029	0.181
MC2	0.093	0.061	0.156	0.091	**0.91**	0	-0.017	-0.067	-0.153	-0.067	0.16
MC3	0.128	0.113	0.194	0.096	**0.931**	0.044	0.029	0.006	-0.056	-0.037	0.179
MC4	0.101	0.113	0.221	0.078	**0.919**	0.084	0.034	0.002	-0.001	-0.004	0.133
TR1	0.594	0.61	0.677	0.545	0.027	0.397	0.676	0.652	0.598	**0.939**	0.497
TR2	0.538	0.535	0.578	0.484	-0.09	0.417	0.622	0.65	0.549	**0.948**	0.439
TR3	0.52	0.556	0.589	0.49	-0.031	0.398	0.642	0.663	0.588	**0.955**	0.435
TR4	0.522	0.512	0.586	0.475	-0.047	0.397	0.627	0.626	0.594	**0.956**	0.434
HR1	0.776	0.682	0.682	**0.96**	0.068	0.392	0.607	0.516	0.421	0.503	0.587
HR2	0.778	0.654	0.71	**0.957**	0.078	0.354	0.601	0.523	0.398	0.519	0.616
HR3	0.766	0.667	0.715	**0.966**	0.146	0.393	0.598	0.527	0.408	0.498	0.639
BR1	**0.935**	0.741	0.714	0.766	0.066	0.427	0.664	0.543	0.424	0.577	0.65
BR2	**0.95**	0.63	0.737	0.751	0.123	0.43	0.633	0.536	0.406	0.534	0.665
BR3	**0.956**	0.696	0.78	0.768	0.169	0.472	0.649	0.529	0.417	0.524	0.653
CM1	0.723	**0.934**	0.76	0.71	0.111	0.475	0.686	0.62	0.546	0.575	0.545
CM2	0.661	**0.948**	0.749	0.64	0.106	0.447	0.688	0.602	0.537	0.545	0.514
CM3	0.688	**0.955**	0.784	0.648	0.096	0.452	0.716	0.609	0.503	0.57	0.572
CM4	0.688	**0.964**	0.757	0.637	0.105	0.45	0.685	0.597	0.491	0.545	0.567
CM5	0.691	**0.954**	0.783	0.666	0.075	0.468	0.734	0.636	0.548	0.574	0.518
CM6	0.702	**0.959**	0.799	0.667	0.147	0.468	0.708	0.602	0.534	0.534	0.567
OC1	0.502	0.517	0.602	0.423	0.1	**0.89**	0.532	0.414	0.328	0.425	0.481
OC2	0.359	0.373	0.441	0.299	0.007	**0.881**	0.429	0.408	0.307	0.355	0.416
OC3	0.348	0.348	0.404	0.262	0.038	**0.867**	0.387	0.34	0.286	0.278	0.301
OC4	0.418	0.438	0.489	0.381	-0.006	**0.906**	0.489	0.429	0.344	0.415	0.424
VS1	0.711	0.614	0.746	0.68	0.18	0.5	0.66	0.575	0.388	0.518	**0.962**

VS2	0.629	0.534	0.654	0.59	0.186	0.44	0.557	0.471	0.303	0.432	**0.971**
VS3	0.666	0.538	0.644	0.593	0.177	0.434	0.564	0.501	0.312	0.443	**0.967**
VS4	0.663	0.534	0.652	0.593	0.122	0.445	0.592	0.523	0.342	0.483	**0.959**
VS5	0.67	0.548	0.649	0.622	0.187	0.429	0.57	0.48	0.297	0.419	**0.971**
TC1	0.415	0.528	0.582	0.404	-0.036	0.344	0.562	0.544	**0.961**	0.597	0.354
TC2	0.399	0.52	0.542	0.391	-0.108	0.322	0.543	0.535	**0.954**	0.568	0.281
TC3	0.447	0.542	0.59	0.428	-0.1	0.366	0.6	0.605	**0.961**	0.601	0.343
RS1	0.708	0.716	0.816	0.631	0.089	0.468	**0.936**	0.705	0.548	0.661	0.623
RS2	0.62	0.688	0.74	0.585	-0.031	0.489	**0.956**	0.733	0.58	0.647	0.539
RS3	0.645	0.704	0.743	0.593	0	0.527	**0.95**	0.722	0.558	0.626	0.569
RS4	0.612	0.687	0.731	0.56	-0.035	0.514	**0.947**	0.714	0.563	0.632	0.58
RC1	0.481	0.537	0.609	0.467	-0.033	0.375	0.657	**0.936**	0.505	0.603	0.429
RC2	0.497	0.521	0.617	0.495	-0.075	0.435	0.645	**0.932**	0.508	0.601	0.489
RC3	0.471	0.521	0.587	0.454	-0.052	0.358	0.631	**0.935**	0.518	0.584	0.438
RC4	0.51	0.543	0.631	0.479	-0.032	0.37	0.674	**0.94**	0.522	0.617	0.487
RC5	0.553	0.712	0.732	0.521	0.031	0.465	0.759	**0.748**	0.555	0.634	0.511
CA1	0.751	0.784	**0.964**	0.697	0.197	0.564	0.75	0.704	0.608	0.635	0.693
CA2	0.783	0.796	**0.952**	0.703	0.195	0.535	0.762	0.67	0.567	0.581	0.668
CA3	0.744	0.75	**0.94**	0.694	0.186	0.542	0.745	0.669	0.549	0.6	0.617
CA4	0.737	0.748	**0.941**	0.675	0.211	0.459	0.726	0.664	0.541	0.617	0.654
CA5	0.708	0.735	**0.939**	0.679	0.222	0.544	0.738	0.674	0.565	0.586	0.654
CA6	0.748	0.786	**0.956**	0.698	0.178	0.49	0.803	0.683	0.595	0.63	0.645
CA7	0.768	0.781	**0.956**	0.701	0.18	0.636	0.781	0.687	0.579	0.629	0.67
CA8	0.735	0.785	**0.951**	0.693	0.193	0.494	0.791	0.707	0.586	0.629	0.649
CA9	0.752	0.773	0.954	0.71	0.234	0.51	0.76	0.67	0.51	0.592	0.693

TABLE VI VARIANCE INFLATION FACTOR (VIF)

	VIF	F-Square
Business Resources → Cloud Computing Adoption	4.163	0.044
Change Management → Cloud Computing Adoption	3.276	0.105
Cloud Deployment → Cloud Computing Adoption	1.158	0.023
Human Resources → Cloud Computing Adoption	3.273	0.013
Managerial IT Capability → Cloud Computing Adoption	1.141	0.118
Organisational Culture → Cloud Computing Adoption	1.601	0.03
Regulatory Support → Cloud Computing Adoption	3.896	0.077
Relational IT Capability → Cloud Computing Adoption	3.007	0.032
Technical IT Capability → Cloud Computing Adoption	1.979	0.052
Technology Resources → Cloud Computing Adoption	2.578	0
Vendor Support → Cloud Computing Adoption	2.379	0.044

- The structural model

Table VII shows the results of the structural model, which shows that there are several statistically significant direct effects between the latent constructs and the use of cloud computing. Managerial IT capability exhibits a robust and highly significant influence on adoption outcomes ($\beta = 0.138$, $t = 5.028$, $p < 0.001$), underscoring the critical role of IT leadership in strategic alignment, planning, and decision-making processes. This finding backs up earlier research [21] [19] [20] which emphasises that effective managerial orchestration of IT resources substantially facilitates the assimilation of cloud technologies. As for the technical IT capability, although the empirical findings indicate that technical IT capability exerts a statistically significant yet modest influence on cloud computing adoption ($\beta = 0.121$, $t = 2.525$, $p = 0.012$), this does not diminish its theoretical importance. Within the RBV framework, technical IT capability remains a foundational enabler that supports system integration, scalability, and agility [28] especially in healthcare settings, where infrastructure robustness and data governance are critical. The relatively moderate effect size may suggest that while technical readiness is essential, it operates alongside other organisational and environmental factors in influencing adoption decisions. Hence, its strategic value may be more catalytic than direct, setting the groundwork

upon which other enablers exert stronger influence. Relational IT capability was found to have a statistically significant positive influence on cloud computing adoption ($\beta = 0.116$, $t = 2.672$, $p = 0.008$), underscoring the role of effective collaboration, mutual understanding, and trust between IT personnel and other organisational stakeholders. This finding reinforces the hypothesis that strong relational ties facilitate smoother coordination and reduce uncertainties associated with cloud adoption. However, the strength of this relationship appears to be moderate in the current study, differing from previous work by [21], which identified relational IT capability as the most dominant predictor in the context of Korean firms. This discrepancy may reflect contextual differences, particularly in Malaysia's healthcare sector, where relational dynamics are shaped by distinct regulatory, cultural, and operational factors. From an RBV standpoint, trust-based relationships are central to relational IT capability and represent a valuable and inimitable organisational asset that can foster integration with cloud service providers, support shared problem-solving and enhance adoption readiness. These results are also consistent with [12], who emphasised the role of trust and collaboration in promoting successful cloud initiatives.

These findings are consistent with earlier research [21] that affirmed the relevance of IT capabilities in shaping cloud computing adoption. Unlike prior studies that considered this relationship within the context of moderating variables such as deployment models, the current study offers a direct evaluation of the influence exerted by IT capabilities. This approach provides clearer insight into their standalone strategic importance. Viewed through the lens of the RBV, the results underscore that managerial, technical, and relational IT capabilities are not merely operational enablers but serve as critical organisational resources that drive cloud adoption in Malaysia's healthcare sector.

TABLE VII DIRECT EFFECT RESULTS

Hypothesis	Path	Coefficient	SD	T	Results
1	Managerial IT Capability → Cloud Computing Adoption	0.138	0.028	5.028	Supported
2	Technical IT Capability → Cloud Computing Adoption	0.121	0.048	2.525	Supported
3	Relational IT Capability → Cloud Computing Adoption	0.116	0.044	2.672	Supported
4	Technology Resources → Cloud Computing Adoption	-0.006	0.043	0.139	Not supported
5	Human Resources → Cloud Computing Adoption	0.076	0.041	1.877	Not supported
6	Business Resources → Cloud Computing Adoption	0.162	0.057	2.848	Supported
7	Change Management → Cloud Computing Adoption	0.222	0.091	2.446	Supported
8	Organisational Culture → Cloud Computing Adoption	0.082	0.03	2.721	Supported
9	Vendor Support → Cloud Computing Adoption	0.122	0.043	2.838	Supported
10	Regulatory Support → Cloud Computing Adoption	0.207	0.06	3.471	Supported

Among the TOE dimensions, technology resources exhibit a negative and non-significant relationship with cloud computing adoption ($\beta = -0.006$, $t = 0.139$, $p = 0.889$), suggesting that the availability of technological infrastructure alone does not directly influence adoption within the studied context. This contradicts the assumption that having sufficient IT assets automatically facilitates adoption. While prior research [25] reported a positive link between technological readiness and cloud uptake, the discrepancy here may stem from contextual factors such as organisational readiness or sector-specific constraints in Malaysian healthcare. From the RBV standpoint, this finding implies that technology resources, unless effectively configured, integrated, and leveraged, may lack the strategic value necessary to drive adoption. In other words, possessing IT assets is insufficient unless those resources are transformed into capabilities that are valuable and difficult to replicate, characteristics essential for competitive advantage and innovation. Although the analysis reveals that human resources have a positive but marginally insignificant effect on cloud computing adoption ($\beta = 0.076$, $t = 1.877$, $p = 0.061$), the direction of the relationship is theoretically consistent with prior expectations. This suggests that, while the presence of skilled and knowledgeable IT personnel contributes to adoption readiness, their influence alone may not be sufficient to significantly drive adoption decisions within the studied context. The result resonates with previous findings [22], [25], which emphasise the enabling but not singularly decisive role of human capital in technological innovation. From the RBV perspective, human resources remain a critical intangible asset; however, their strategic contribution to cloud adoption may depend on how effectively these competencies are integrated with other organisational and environmental enablers. In contrast to the insignificant effects of technology and human resources, business resources demonstrate a significant positive relationship with adoption ($\beta = 0.162$, $t = 2.848$, $p = 0.004$). This finding highlights the critical role of organisational capabilities and assets, such as supplier relationships, employee training, strategic planning, and financial resources, in driving adoption. The result is consistent with the findings of [22], [25], who emphasised business resources as key enablers for effective and sustained use of cloud technology.

The results also show that change management has a moderate yet statistically significant positive influence on cloud computing adoption ($\beta = 0.222$, $t = 2.446$, $p = 0.014$). This finding suggests that healthcare organisations with stronger capabilities to manage organisational change, such as adapting workflows, realigning processes, and coordinating implementation efforts, are more likely to adopt cloud services effectively. Given that cloud-based solutions often require

customisation and transformation of legacy processes, the ability to navigate change emerges as a key enabler. This supports the theoretical assertion that such adaptability constitutes a dynamic capability under the RBV lens, allowing organisations to reconfigure internal competencies and respond proactively to disruptive innovations. The evidence reinforces the hypothesis that overcoming resistance to change is instrumental in enhancing cloud adoption outcomes. Similarly, organisational culture shows a positive and statistically significant effect on cloud computing adoption ($\beta = 0.082$, $t = 2.721$, $p = 0.007$), underscoring the influence of a supportive internal environment in embracing technological change. A culture that encourages innovation, collaboration, and openness facilitates smoother adoption processes, particularly in complex sectors like healthcare. This finding supports the RBV perspective, where such culture is viewed as a valuable, non-replicable asset that enables firms to adapt and sustain technological advancements. The result affirms H8 and aligns with prior research [22] [25], highlighting culture as a critical determinant of adoption success.

With regard to environmental factors, vendor support ($\beta = 0.122$, $t = 2.838$, $p = 0.005$) and regulatory support ($\beta = 0.207$, $t = 3.471$, $p = 0.001$) both exhibit significant positive relationships with cloud computing adoption. This suggests that effective vendor collaboration through knowledge transfer and reliable partnerships, and supportive government policies, including regulations on privacy, security, and funding incentives, are key enablers of cloud implementation. These findings are consistent with [22] [25].

V. Results

A. Current Situation and Prospect of Cloud Adoption in the Healthcare Industry

The shift towards increasing the application of cloud computing in the healthcare industry is driven by the continuous efforts of the industry for the health and well-being of patients, that is to treat patients to be better and healthier, and also to prevent disease as much as possible, and ultimately to promote life peace. With cloud technology, many healthcare institutions benefit exponentially in terms of scalability, cost savings and cloud efficiency. Referring to [15], who studied cloud computing trends and the latest status in healthcare, the approach to using cloud healthcare increased exponentially. In Malaysia, Kumpulan Perubatan Johor, better known as KPJ, is the first private healthcare provider to use cloud computing. With 28 hospitals in Malaysia and two hospitals in Indonesia, they believe that cloud computing can lower their costs or enjoy savings of 30% to 40% of their IT expenses in the long run.

B. Main Influencing Factors of Cloud Computing Adoption

The analysis of structural relationships revealed that multiple factors significantly influence cloud computing adoption in the Malaysian healthcare industry. Among IT capability dimensions, managerial IT capability emerged as the strongest predictor, affirming that leadership proficiency in aligning IT with strategic goals facilitates technology adoption. This underscores the critical role of top management in orchestrating IT initiatives and integrating cloud technologies with organisational priorities.

Technical IT capability also showed a statistically significant, albeit modest, effect. While not the most dominant driver, its foundational role remains critical in enabling system integration, ensuring infrastructure readiness, and promoting cloud scalability, particularly in healthcare, where technological robustness and data sensitivity are paramount. Similarly, relational IT capability demonstrated a positive influence, highlighting the importance of inter-organisational collaboration and trust in navigating external dependencies such as vendor relationships and service reliability.

Within the technological context, human resources showed a near-significant effect, suggesting that while the availability of skilled personnel supports cloud readiness, its standalone impact may be limited unless complemented by broader organisational readiness. Conversely, technology resources showed no significant relationship, implying that infrastructure alone is insufficient to trigger adoption, potentially due to redundancy, underutilization, or a lack of strategic alignment.

For the organisational context, both change management and organisational culture were positively and significantly associated with adoption. These findings indicate that adaptability, readiness to restructure workflows, and an innovation-oriented culture are essential enablers of successful cloud integration. Such internal factors often mediate the effective deployment of technological investments, especially in highly regulated and process-intensive sectors, like healthcare.

Under the environmental context, vendor support and regulatory support emerged as significant external enablers. Effective vendor collaboration facilitates knowledge transfer and smoother system implementation, while robust regulatory frameworks enhance trust and reduce perceived risk, encouraging adoption. These findings affirm the role of institutional forces in shaping adoption decisions, especially in healthcare, where legal compliance, data governance, and policy alignment are critical.

VI. Discussion

The findings highlight the applicability of the integrated RBV-TOE framework in understanding cloud computing adoption in healthcare. The strong influence of managerial and relational IT capabilities supports the RBV view that firm-specific resources offer competitive advantages. These capabilities allow for the strategic alignment of IT initiatives and foster collaborative relationships that enhance cloud implementation outcomes.

In contrast, the limited role of technology resources indicates that hardware or infrastructure, in isolation, lacks strategic impact without corresponding organisational and managerial readiness. Similarly, human resource capability requires integration with broader organisational strategies to influence adoption meaningfully.

Within the TOE framework, organisational culture and change readiness emerged as essential internal enablers, reflecting healthcare institutions' need for adaptable and innovation-oriented environments. Meanwhile, vendor and regulatory support validate the significance of environmental conditions in shaping IT decisions, particularly in highly

regulated sectors where compliance, trust, and policy alignment are crucial.

Overall, a multi-dimensional readiness anchored in robust IT capabilities, conducive organisational environments, and supportive external ecosystems is essential for effective cloud adoption. These results offer actionable insights for healthcare leaders navigating digital transformation and highlight that success depends on coordinated efforts across internal and external domains.

VII. Recommendations

Based on the findings, the following recommendations are proposed:

1) Define Strategic IT Capability Profiles Tailored to the Healthcare Context.

Given the unique operational demands and regulatory sensitivities of the healthcare industry, it is essential for healthcare organisations to develop targeted IT capability profiles that support the adoption and integration of cloud computing technologies. Strategic emphasis should be placed on strengthening managerial capacity for aligning IT initiatives with clinical and administrative goals, investing in technical expertise to ensure secure and interoperable systems, and nurturing relational capabilities that facilitate collaboration across departments and with cloud service providers. This tailored capability development will allow organisations to capitalise on cloud technologies while managing sector-specific constraints such as data privacy, compliance, and institutional readiness.

2) A Common Standard Model for a comprehensive and specific Cloud Adoption Framework for the Healthcare industry.

To further advance cloud computing adoption and value realisation, future strategies should consider expanding internal IT capabilities through targeted IT Outsourcing (ITO) and IT Innovation (ITI) initiatives. While this study has established the significance of managerial, technical, and relational IT capabilities in influencing adoption, the inclusion of external IT partners via ITO can complement resource constraints and enhance service scalability, especially in settings where internal capacity is limited. From an RBV perspective, outsourcing allows healthcare organisations to access rare and specialised resources, effectively extending their resource base without diluting strategic focus. Similarly, promoting IT innovation, such as adopting new service delivery models or leveraging AI-driven cloud tools, can foster dynamic capabilities, enabling firms to continuously adapt and reconfigure existing IT assets to meet evolving patient care and regulatory demands. These recommendations are particularly relevant under the TOE framework's technological and environmental dimensions, where responsiveness to external trends and technological advancement is key to sustaining competitive advantage.

VIII. Limitations and Future Research

While this study offers meaningful insights into cloud computing adoption by integrating the RBV with the TOE framework, several limitations merit acknowledgement. First, the research is geographic and sector-specific, focusing solely on healthcare organisations within Malaysia. The country's regulatory environment, technological maturity, and public-private healthcare dichotomy may limit the generalisability of the findings to other sectors or regions with different institutional and infrastructural characteristics.

Second, the study employed a cross-sectional design, which restricts the ability to capture temporal shifts in cloud adoption behaviour, such as changes influenced by evolving regulations, technological advancements, or organisational digital strategies. Longitudinal research could better reveal how internal capabilities and external pressures interact over time.

Third, although the study incorporated both internal (IT capabilities) and contextual (TOE) enablers, it did not examine interaction or mediation effects between constructs, such as how IT capabilities may moderate or mediate the influence of environmental or organisational factors on adoption. Investigating these relationships could further enrich theoretical understanding and practical strategies for cloud integration.

Future research may explore multi-level models that consider clinical, administrative, and policy-level dynamics in cloud computing decisions. In addition, the integration of IT Outsourcing (ITO) and IT Innovation (ITI) constructs within the RBV–TOE framework may yield deeper insights into how firms balance internal capabilities with external service acquisition to enhance digital transformation. Expanding the scope to include deployment models, cloud maturity levels, and cross-country comparisons can also enhance the robustness and applicability of the proposed model.

Acknowledgment

This research was conducted as part of the Ph.D. program at Universiti Kebangsaan Malaysia and received full financial support for the research grant from the Ministry of Higher Education (MoHE) and Universiti Kebangsaan Malaysia (UKM) via the Fundamental Research Grant Scheme (FRGS/1/2024/TK01/UKM/02/2).

The author would like to thank the participating healthcare organisations for their cooperation and all respondents for their valuable time and insights. The study also received ethical approval from the Medical Research and Ethics Committee (MREC), Malaysia, under NMRR-23-03142-NH7 (IIR), and KPJ Clinical and Research Ethics Review Committee (CRERC), Ref. kpj_002/2023.

References

[1] A. Naseri and N. Jafari Navimipour, "A new agent-based method for QoS-aware cloud service composition using particle swarm optimization algorithm," J Ambient Intell Humaniz Comput, vol. 10, no. 5, pp. 1851–1864, May 2019, doi: 10.1007/s12652-018-0773-8.

[2] Y. Ebadi and N. Jafari Navimipour, "An energy-aware method for data replication in the cloud environments using a Tabu search and particle swarm optimization algorithm," Concurr Comput, vol. 31, no. 1, Jan. 2019, doi: 10.1002/cpe.4757.

[3] S. C. Park and S. Y. Ryoo, "An empirical investigation of end-users' switching toward cloud computing: A two factor theory perspective," Comput Human Behav, vol. 29, no. 1, pp. 160–170, 2013, doi: 10.1016/j.chb.2012.07.032.

[4] P. Mell and T. Grance, "The NIST-National Institute of Standars and Technology- Definition of Cloud Computing," NIST Special Publication 800-145, p. 7, 2011.

[5] L. Barthelus, "Adopting cloud computing within the healthcare industry: opportunity or risk?," Online Journal of Applied Knowledge Management, vol. 4, no. 1, pp. 1–16, 2016, doi: 10.36965/ojakm.2016.4(1)1-16.

[6] A. Khayer, M. S. Talukder, Y. Bao, and M. N. Hossain, "Cloud computing adoption and its impact on SMEs' performance for cloud supported operations: A dual-stage analytical approach," Technol Soc, vol. 60, no. September 2019, p. 101225, 2020, doi: 10.1016/j.techsoc.2019.101225.

[7] B. Nayak, S. K. Padhi, and P. K. Pattnaik, "Impact of cloud accountability on clinical architecture and acceptance of healthcare system," Advances in Intelligent Systems and Computing, vol. 701, pp. 149–157, 2018, doi: 10.1007/978-981-10-7563-6_16.

[8] A. Meri, M. K. Hasan, M. Danaee, M. Jaber, M. T. Jarrar, N. Safei, et al., "Modelling the utilization of cloud health information systems in the Iraqi public healthcare sector," Telematics and Informatics, vol. 36, pp. 132–146, 2019.

[9] O. Ali, A. Shrestha, J. Soar, and S. F. Wamba, "Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review," Int J Inf Manage, vol. 43, no. July, pp. 146–158, 2018, doi: 10.1016/j.ijinfomgt.2018.07.009.

[10] H. E. Newkirk, A. L. Lederer, and A. M. Johnson, "Rapid business and IT change: Drivers for strategic information systems planning?," European Journal of Information Systems, vol. 17, no. 3, pp. 198–218, 2008, doi: 10.1057/ejis.2008.16.

[11] N. Anggraini, Binariswanto, and N. Legowo, "Cloud computing adoption strategic planning using ROCCA and TOGAF 9.2: A study in government agency," in Procedia Computer Science, Elsevier B.V., 2019, pp. 1316–1324. doi: 10.1016/j.procs.2019.11.247.

[12] F. Gao and A. Sunyaev, "Context matters: A review of the determinant factors in the decision to adopt cloud computing in healthcare," Int J Inf Manage, vol. 48, no. July 2018, pp. 120–138, 2019, doi: 10.1016/j.ijinfomgt.2019.02.002.

[13] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," NCM 2009 - 5th International Joint Conference on INC, IMS, and IDC, pp. 44–51, 2009, doi: 10.1109/NCM.2009.218.

[14] C. N. Höfer and G. Karagiannis, "Cloud computing services: Taxonomy and comparison," Journal of Internet Services and Applications, vol. 2, no. 2, pp. 81–94, 2011, doi: 10.1007/s13174-011-0027-x.

[15] F. Gao, S. Thiebes, and A. Sunyaev, "Rethinking the meaning of cloud computing for health care: A taxonomic perspective and future research directions," J Med Internet Res, vol. 20, no. 7, pp. 1–16, 2018, doi: 10.2196/10041.

[16] S. Mithas, N. Ramasubbu, and V. Sambamurthy, "How information management capability influences firm performance," MIS Q, vol. 35, no. 1, pp. 237–256, 2011, doi: 10.2307/23043496.

[17] J. Barney, "Firm Resources and Sustained Competitive Advantage," J Manage, vol. 17, no. 1, pp. 99–120, 1991, doi: 10.1177/014920639101700108.

[18] X. (Robert) Luo, W. Zhang, H. Li, R. Bose, and Q. B. Chung, "Cloud computing capability: its technological root and business impact," Journal of Organizational Computing and Electronic Commerce, vol. 28, no. 3, 2018, doi: 10.1080/10919392.2018.1480926.

[19] A. S. Bharadwaj, "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation Bharadwaj/IT Capability and Firm Performance MIS Quarterly A RESOURCE-BASED PERSPECTIVE ON INFORMATION TECHNOLOGY CAPABILITY AND FIRM PERFORMANCE," Source: MIS Quarterly, vol. 24, no. 1, pp. 169–196, 2000.

[20] C.-H. Chen, Y.-L. Lan, W.-P. Yang, F.-M. Hsu, C.-L. Lin, and H.-C. Chen, "Exploring the Impact of a Telehealth Care System on Organizational Capabilities and Organizational Performance from a Resource-Based Perspective," International Journal of Environmental Research and Public Health Article, 2019, doi: 10.3390/ijerph16203988.

[21] G. Garrison, R. L. Wakefield, and S. Kim, "The effects of IT capabilities and delivery model on cloud computing success and firm performance for cloud supported processes and operations," Int J Inf Manage, vol. 35, no. 4, pp. 377–393, 2015, doi: 10.1016/j.ijinfomgt.2015.03.001.

[22] H. Gangwar, "Cloud computing usage and its effect on organizational performance," Human Systems Management, vol. 36, no. 1, pp. 13–26, 2017, doi: 10.3233/HSM-171625.

[23] H. Sallehudin et al., "Performance and key factors of cloud computing implementation in the public sector," International Journal of Business and Society, vol. 21, no. 1, pp. 134–152, 2020.

[24] H. Sallehudin, R. Che Razak, M. Ismail, A. F. Md Fadzil, and R. Baker, "Cloud Computing Implementation in The Public Sector: Factors and Impact," Asia-Pacific Journal of Information Technology and Multimedia, vol. 07, no. 02(02), pp. 27–42, 2018, doi: 10.17576/apjitm-2018-0702(02)-03.

[25] S. Schneider and A. Sunyaev, "Determinant factors of cloud-sourcing decisions: Reflecting on the IT outsourcing literature in the era of cloud computing," Journal of Information Technology, vol. 31, no. 1, pp. 1–31, 2016, doi: 10.1057/jit.2014.25.

[26] A. M. Kadhum and M. K. Hasan, "Assessing the determinants of cloud computing services for utilizing health information systems: A case study," Int J Adv Sci Eng Inf Technol, vol. 7, no. 2, pp. 503–510, 2017, doi: 10.18517/ijaseit.7.2.1814.

[27] N. Mahusin, H. Sallehudin, N. S. Mohd Satar, A. H. Mohd Aman, and F. Yahya, "Modeling cloud computing adoption and its impact on the performance of IT personnel in the public sector," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 2, 2025.

[28] T. Ravichandran and C. Lertwongsatien, "Effect of information systems resources and capabilities on firm performance: A resource-based perspective," Journal of Management Information Systems, vol. 21, no. 4, pp. 237–276, 2005, doi: 10.1080/07421222.2005.11045820.

[29] Z. Shao, Y. Feng, and L. Liu, "The mediating effect of organizational culture and knowledge sharing on transformational leadership and Enterprise Resource Planning systems success: An empirical study in China," Comput Human Behav, vol. 28, no. 6, pp. 2400–2413, 2012, doi: 10.1016/j.chb.2012.07.011.

[30] J. Ferreira and C. Fernandes, "Resources and capabilities' effects on firm performance: what are they?," Journal of Knowledge Management, vol. 21, no. 5, pp. 1202–1217, 2017, doi: 10.1108/JKM-03-2017-0099.

[31] A. Zaheer, B. McEvily, and V. Perrone, "Does Trust Matter? Exploring the Effects of Interorganizational and Interpersonal Trust on Performance," Organization Science, vol. 9, no. 2, pp. 141–159, 1998, doi: 10.1287/orsc.9.2.141.

[32] Matthew C. Howarda, Christian Nitzl Joe F. Hair Jr.a, "Assessing measurement model quality in PLS-SEM using confirmatory composite analysis," J Bus Res.

[33] C. Fornell and D. F. Larcker, "Fornell, C. and Larcker, D.F. (1981), "Evaluating structural equation models with unobservable variables and.pdf," Journal of Marketing Research, vol. XVIII, no. February, pp. 39–50, 1981.

[34] M. Sarstedt, C. M. Ringle, and J. F. Hair, Partial Least Squares Structural Equation Modeling. 2021. doi: 10.1007/978-3-319-05542-8_15-2.

[35] J. F. Hair, G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 3rd ed. Sage Publishing, 2022.

Potential Variables in Pharmaceutical Drug Prediction Research with Machine Learning Approach: A Literature Review

Gunadi Emmanuel[1], Yulyani Arifin[2], Ilvico Sonata[3], Muhammad Zarlis[4]
Computer Science Department-Binus Graduate Program-Doctor of Computer Science,
Bina Nusantara University, Jakarta, Indonesia[1, 2, 3]
Information System Management Department-Binus Graduate Program-Master of Information System Management,
Bina Nusantara University, Jakarta, Indonesia[4]

***Abstract*—As a downstream component of the drug supply chain, pharmaceutical installations often face uncertainty in drug demand. Predicting pharmaceutical drugs using a machine learning approach enables the development of new variables that can enhance the performance of medicine prediction. Amidst limited data and a choice of prediction algorithms, the accuracy of variable selection is significant for drug prediction performance. This study remaps the scope of variables from previous studies related to drug demand prediction and machine learning performance to develop further significant variables. Investigating research literature on significant variables in drug demand prediction with machine learning models published in 2020-2024. The systematic literature methodology uses the Kitchenham method. Mapping problems, discussion areas, and data availability result in ten categories of issue areas, each with its respective data needs and algorithm choices. A qualitative exploration of issue areas identifies potential variables for pharmaceutical drug prediction, including drug consumption, epidemiology, drug management, supply chain-patient domicile, and pharmacotherapy. Mapping potential variables facilitates the availability and integration of data relevant to local or regional characteristics, enabling further research on the characteristics of data and algorithm choices.**

***Keywords*—Drug demand; machine learning; pharmaceutical installations; prediction; potential variables**

I. INTRODUCTION

Drug availability is a crucial issue during the COVID-19 Pandemic. Hospital management and health services need to build an optimal drug inventory control system in the future. Amid the uncertainty of drug demand, the planning and procurement process becomes a determining stage in pharmaceutical logistics operational management [1]–[3]. The drug planning and procurement process in hospital pharmacy installations is related to the drug supply chain from pharmaceutical wholesalers (distributors) [4], [5].

Both manufacturers and distributors require high downstream demand prediction results to form the basis for drug supply planning [6]. Accurate downstream planning supports the analysis of supply chain visibility, lead time, inventory status, and supplier performance. Pharmaceutical logistics can calculate drug availability variables, lead time, and drug costs from suppliers [7]. These results will improve internal pharmaceutical supply chain performance and optimize inventory management and distribution activities [8]. The accuracy of demand predictions generated by algorithms anticipates elements of uncertainty or fluctuations in drug demand [7], [9]. In the supply chain, drug demand uncertainty is a category of external uncertainty and risk. These external risks are identified from events that deviate from the general pattern and may change the prediction performance [9], [10].

In the pharmaceutical industry, stakeholders must identify external variables of drugs in the demand prediction process. Prediction results do not only anticipate events that disrupt drug distribution and provision planning downstream in the supply chain. Furthermore, the results of medicine demand prediction maintain the balance of supply against drug demand and mitigate against events that deviate from the general pattern in determining the amount and pattern of consumption, epidemiology, and available budget [11], [12]. Stakeholders can leverage machine learning models for prediction, saving time and costs, and allowing for a more focused investment in clinical trials [13]. Machine learning models can capture variables of changing market conditions and consumer demand. However, machine learning requires the availability of good, reliable, and real-time features and data to achieve high performance [14].

Beyond drug variables, several variables in drug prediction may contribute to changes in drug demand, including strong seasonal patterns, uncertain demand, and a lack of historical data [15]. Each variable involved in the prediction may have different data characteristics. Prediction variables can exhibit non-linear correlations with several parameters of their dependent variables, such as historical sales data in conjunction with demographic and epidemiological data (including disease type, disease progression, or patient distribution level) [16].

The greater the data variation, the more the machine learning model obtains aspects of demand variability. The model can predict consumption frequency, consumption amounts over time, and consumption patterns based on geographic areas [17]. Increasing data dimensions, computational complexity, storage requirements, the ambiguous or noisy nature of data, and device capacity also demand higher model performance [18]. Among the many features used, machine learning will identify several significant

ones. Dimensionality reduction can impact the performance, complexity, time, and space requirements in machine learning. Similarly, some irrelevant or redundant features can significantly reduce classifier performance [19].

The selection of variables from epidemiological data of dengue fever patients considers variables that affect the development of dengue fever. In their research, Syahrir [20] built a model based on the development variables of each stage of the patient's disease. The model known as susceptible-expose-infectious-removal (SEIR) is used to predict drug needs in the face of dengue fever outbreaks in Indonesia. Based on data on the number of hospitalized patients with dengue fever from 2014 to 2016, the prediction model employed a Moving Average (MA), regression, and Exponential Smoothing (ES). In subsequent studies, they developed Single Exponential Smoothing (SES), Double Exponential Smoothing (DES), Linear Regression (LR), Quadratic regression, Bayesian regression, the winters method, Autoregressive Integrated Moving Average (ARIMA), and the SEIR model for prediction [21]. They predicted the number of patients and drug demand based on time variables while ignoring other epidemiological variables, such as climate and environmental health.

Meanwhile, Ho [22] employed machine learning algorithms, including the Decision Tree (DT), Deep Neural Network (DNN), and Logistic Regression, to identify cases of dengue fever. Using the algorithm, they identified four primary variables: age, body temperature, white blood cell count, and platelet count. Then, based on the doctor's recommendations and epidemiological characteristics, they added Hemoglobin (Hb) and gender parameters. The two additional parameters did not improve the prediction performance. With four main variables, the performance of the DT, DNN, and logistic regression models provided a sensitivity of 90%. Other research on dengue cases utilised climate variables [23]–[25], dengue case data, population demographics, entomology, and environmental factors [26], [27]. The use or addition of specific variables is not necessarily significant to the performance of the machine learning model until high prediction performance is validated.

In the development of the Earth Cognitive System for COVID-19 (ECO4CO) information system for handling the COVID-19 Pandemic, the display of notification services, alerts, and scheduled communications, as well as geographic localization for end-users, supports logistics management of the Pandemic's drug supply [28]. In general, the pharmaceutical industry can further optimize drug planning and procurement. Specifically, ECO4CO helps aggregate data in information systems to help pharmaceutical installations predict drug needs [29]. Integrated data will facilitate the acquisition of significant variables for prediction. Government policy encouragement, especially in developing countries, can accelerate the development of integrated health information systems, particularly when potential predictive variables are sourced from multiple government and private institutions.

Health disruption situations, such as the COVID-19 Pandemic, require efforts to identify variables that are not yet clearly significant (referred to as hidden variables) in building drug prediction models. Machine learning can help identify hidden variables and pinpoint significant factors in drug demand, even outside of historical data on drug demand or sales [9]. According to Koala [30], significant variables for predicting drug demand can vary from one case or place to another. If the prediction refers to changes in consumer attitudes, the time series approach will influence the determination of significant variables. Machine learning can perform a clustering analysis approach to identify customer demand behavior. In addition, regression analysis produces continuous-valued functions for products that cannot be stored for long in storage [31]. These approaches direct significant variables for predicting drug demand.

Researchers generally use correlation testing between predictive features to identify significant variables in machine learning models. In the case of the COVID-19 Pandemic, many studies have used new variables from the field. The use of new variables highlights the limited scope of previous similar research, particularly in identifying potentially significant variables in predicting drug demand in pharmaceuticals using a machine learning approach. Mapping previous research areas can contribute to the study of developing machine learning models for predicting drug demand in the pharmaceutical industry, utilizing potentially significant variables. Therefore, this study analyzes research related to pharmaceutical drug prediction using a Machine Learning approach. The study redefines the scope and area of variables in drug demand prediction research.

Mapping the use of variables in previous research suggests the direction for developing a machine learning approach to predict pharmaceutical drug demand. The implications of this study can encourage the government to develop policies related to the use of drug supply calculation methods and significant prediction variables. At the same time, accelerating the integration of national health data into the National Health Network increases the availability of data and the variability of drug prediction variables [32]. To sharpen the direction of research on potential variables in drug prediction with this machine learning approach, researchers formulated research objectives (RO): RO1. Identifying areas of drug demand, research issues, and data availability for prediction from previous studies. RO2. Examining the performance of using machine learning algorithms for drug demand prediction. RO3. Categorizing potential variables to be developed in predicting drug demand.

Next, we discuss literature studies related to drug demand design in pharmacy and machine learning methods commonly used to predict drug demand. Next, we will describe the systematic literature review method. Next, the results and discussions related to research issue areas and finding variables. The final section is the conclusion and the direction of further studies related to drug demand variables.

II. Related Studies

Generally, drug demand calculations in pharmacies utilize internal inventory data. Drug demand calculations are central to drug management and the pharmaceutical supply chain. Drug supply concerns often center on drug stock and demand. In a study of drug planning in several pharmaceutical institutions, Burinskiene [33] utilized historical drug sales data

from multiple pharmacies within a chain to compare prediction performance results at each pharmacy location. The prediction target was to improve accuracy at each pharmacy point. The computational process took into account seasonal data trends, outlier handling, and the handling of missing data. However, differences in pharmacy locations with their variable trends were not a consideration in the computation. In the field, medication supply planning considers many variables and data sources that can support data analysis and prediction. These data include variables such as historical demand, patient demographics, clinical information from Electronic Health Records (EHRs), epidemiology, supply chain, and external elements such as market dynamics and regulations [6].

Information on the extent of drug demand and the variables that influence changes in drug demand are external factors that have not received much attention. Drug demand prediction has a close causal correlation with drug consumption patterns [30]. Correlating external factors with increasing demand prediction accuracy requires careful analysis of market trends, customer behavior, suppliers, and technology [31]. According to Koala [30] factors that can influence drug demand prediction include socioeconomic factors such as gender, age, patient origin, annual income, and social class. Second, patient health factors include comorbidities, insurance, and lifestyle (smoking, alcohol, and obesity). The third factor includes institutional facilities and healthcare personnel. However, Koala does not provide a basis for external variables that remain attached to the dynamics of drug demand in a time series sequence.

Similarly, the behavior of drug supply chain actors focuses on predicting the public's need or demand for drugs to support health services, particularly timely access to drugs in the right amount, type, and quality [34], [35]. Supply chain management strives to ensure that provision and availability match demand [36]. Therefore, analyzing demand prediction variables is necessary to anticipate changes in drug demand [37].

Periods of increased risk of disease spread, such as dengue fever, can change drug demand during specific periods [34]. Likewise, in disaster areas, factors such as limited data on the number of disaster victims, distribution, and the increasing number of victims, and substitution of certain types of drugs (which cannot be replaced by others) can add variables to the prediction of drug needs [20]. In dealing with these issues, the machine learning approach can handle the problems of interdependent variables, complexity of causal factors, and non-linear relationship patterns [36]. Additionally, machine learning can analyze unstructured data sets and identify significant factors that impact supply chain performance [15]. In the case of predicting drug sales volume using sales promotion stimuli, Al-Gunaid [34] compared the performance of five algorithms (LR, Random Forest (RF), Neural Network, Support Vector Regression (SVR), and the Levenberg-Marquardt algorithm). During the analysis, previously unseen factors may emerge due to the increasing correlation between features. Marketing promotions have a high correlation coefficient because they are closely related to variables that influence consumption behavior. Consequently, sales predictions increase. However, the promotion variable as a planned stimulus for supply chain actors is only a small part of the other variables that influence the performance of drug demand prediction.

Dynamic market changes reveal hidden factors that necessitate methods to capture these factors, thereby enhancing data-driven predictions [38], [39]. Machine learning learns from all available data, including environmental parameters, to make informed decisions. Each time new historical data becomes available, the learning process continues to occur, enabling the making of new decisions [40]. Machine learning cannot solve all problems, but it can improve accuracy in certain cases and datasets [41]. Several machine learning algorithms are available for demand prediction based on historical data, including random forests, gradient boosting, and neural networks. Note that environmental changes can affect the performance of prediction algorithms over time [6]. A non-negligible variable is that drug evaluation often results in switching to different drugs to enhance pharmacological effects or mitigate side effects [34].

III. METHODOLOGY

Machine learning has become a widely used method for predicting demand for goods, including in the pharmaceutical sector, specifically for drug demand [30], [42]–[45]. This study examines the mapping of the scope and area of research variables for predicting drug demand in machine learning models. For this reason, this research uses a qualitative literature study method. A systematic review allows researchers to analyze the phenomenon of research on the use of drug demand prediction variables based on machine learning models. This literature study method adapts the steps of the Kitchenham literature review method, except that the data synthesis section is replaced by an analysis of the results and concludes with a report on the study results [46]. Fig. 1 shows the steps of this systematic review research.

1) Research identification: Uncertainty in drug demand necessitates the use of an accurate machine-learning approach with variables that are truly significant for predicting performance. Therefore, the investigation requires as many original articles from previous studies as possible. The original articles targeted are primary research related to machine learning models and drug prediction. The identification of potential variables in previous studies is based on the study's problem area, the context of the problem, data availability, and the performance of the machine learning algorithm used. Meanwhile, the scope of the research is multidisciplinary, covering drug supply chain information systems, pharmacy, and informatics. Therefore, the discussion of the research follows the questions:

- RQ1. What are the areas of drug demand research issues and data available for prediction from previous research?
- RQ2. How is the machine learning algorithm performed for drug demand prediction?
- RQ3. What are the potential variable findings from the selected articles identified in the categories that can build drug predictions with machine learning models?

Furthermore, this research requires quality primary (original) research articles. For this reason, article exploration is conducted in digital libraries, including Scopus, PubMed, and Google Scholar. Researchers use the Publish and Perish 8 Application to search for research literature sources. We also use Mendeley Desktop for documentation and article processing. The selection of search keywords refers to the research questions. Article searches on the application combine keywords in three searches by compiling search keywords for each search, namely: "drug demand AND prediction method AND machine learning"; "prediction AND drug AND pharmaceutical AND demand"; and "medicine OR pharmaceutical OR drug demand AND prediction method AND machine learning".

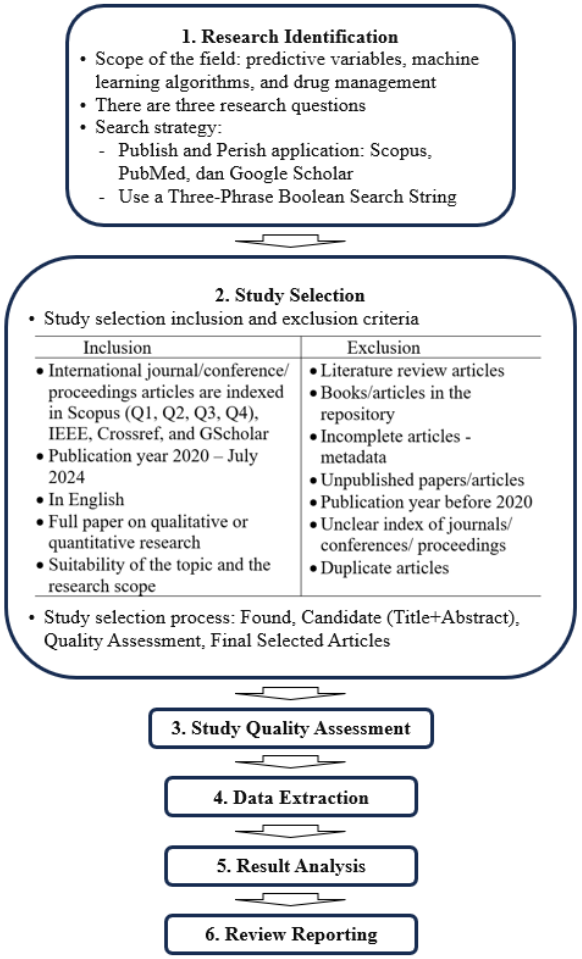


Fig. 1. Research stages - adapted from kitchenham [47], [46].

2) Study selection: After obtaining the articles, the following article selection process involves reading the title, keywords, and abstract to filter out articles that do not match the topic. The method of skimming the articles ensures that the discussion of the article aligns with the research criteria. Furthermore, the fulfillment of the inclusion and exclusion criteria helps in obtaining articles that align with the research objectives. The study requires that selected articles analyzed have a similar research scope, research design, and type of research. The similarity of selected articles suggests that the research articles are closely related to this study. The article has complete data regarding the information needed in this study. The article includes objective and valid measurement techniques for the research model. Differences in methods in presenting research results will follow the objectives of each study. In machine learning research, it is possible to use the same data sources, allowing this literature study to incorporate research literature sources that utilize the same data. Research questions help the stages of analysis of articles according to research topics.

3) Research quality assessment: The researcher evaluates whether the journal's quartile level can ensure a minimum level of study quality. However, the researcher does not group the quality of the articles. The researcher limits observations to variables used to reduce bias in the analysis.

4) Data extraction: The search results for selected articles can be presented in tables or graphs containing the researchers' names, the year of research, problems, data availability, research methods/designs, research variables, and research results. Selected articles can answer research questions that contain the name and year of research, the algorithm used, and the variables used in machine learning. Likewise, the order of appearance follows the discussion of the findings.

5) Result analysis: The analysis of the selected article review results is qualitative in nature. The findings are presented in tables and figures. The analysis of the results, including the research issue area, data availability, and machine learning model performance, leads to the mapping of potential variables to predict drug demand. This presentation is a form of descriptive research synthesis.

6) Review reporting: The final activity of this research is to communicate the results and contributions of the research effectively in international scientific journals. The reporting of this literature study is directed primarily at journals in the areas of computer science or information systems. This research topic is very close to the selection of variables used in machine learning studies.

IV. Results and Discussion

The first part of this session presents the results of the article selection process, quality article publishers, and the number of article topics that tend to increase. Next, we will present the results in the order of the research questions before discussing them. Mapping the problem areas of the selected articles and the variables used to predict drug demand helps organize and systematize the discussion into several subsections.

A. Selected Articles

The use of the Publish and Perish Application makes it easier to search for selected articles and confirms that these articles meet the qualifications as study material. The process of searching for articles with three keyword phrases obtained 886 articles. Selection according to the research stages follows the keyword phrase groups. At the quality assessment stage, the researcher eliminated duplicate articles. Ultimately, the article selection process yielded 23 selected articles (Table I).

TABLE I. STUDY ARTICLE SELECTION PROCESS

Boolean search string	Found	Candidate	Quality assessment	Final selected
		Title+Abstract		
1) "Drug Demand AND Prediction Method AND Machine Learning".	277	64	13	11
2) "Prediction AND Drug AND Pharmaceutical AND Demand"	279	45	15	4
3) "Medicine OR Pharmaceutical OR Drug Demand AND Prediction Method AND Machine Learning"	330	39	12	8
Total	**886**	**148**	**40**	**23**

Determining the scope of the study limits the number of articles that can be found before they are analyzed. Fig. 2 presents the mapping of twenty-one selected articles from the library database. The composition of the articles shows that the selected articles have met international standards and are worthy of analysis. The publisher is also reputable for publishing high-quality papers.

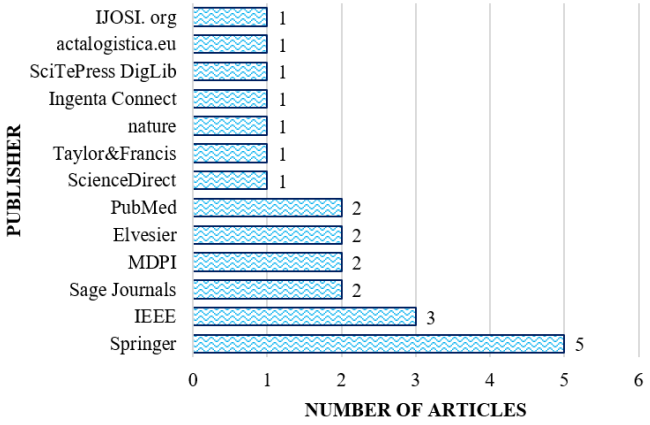


Fig. 2. Composition of selected article publications and the library database.

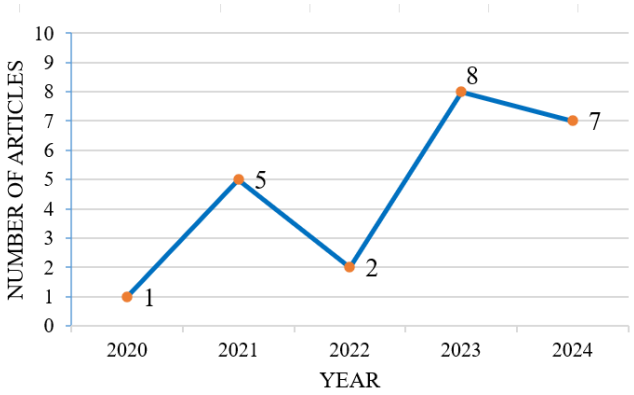


Fig. 3. Number of selected article publications 2020 - July 2024.

This study examines articles published between 2020 and July 2024 to demonstrate the novelty of the research and the article. During that period, the findings of the research topics on drug demand prediction and machine learning were quite encouraging, as the number of related research studies reached more than 50% compared to the previous year, as shown in Fig 3. This finding is interesting because there has been research interest in drug availability in pharmaceutical installations over the past two years.

B. Research Issue Areas and Data Availability (RQ1)

The answer to the first question (RQ1) of this study is completed by presenting the research issue areas and data availability from selected articles. Issues and data availability represent the perspectives of previous researchers on drug prediction to maintain drug availability in pharmacies. The issues raised by each researcher can show the perspective used to solve the problem. This study identified ten research issue areas, as discussed in Table II. Several research issue areas have relationships with others (combination of areas). Subudhi's research issue area [48] has a combination of issue areas, including Epidemiology, Patient Medical Record Data, and Pharmacy Drug Management. We learn to observe the correlation of the problem context and the availability of data that can be used to solve it. The combination of research issue areas raises discussions about data availability and the central point of the problem to be solved using machine learning.

Six selected articles discussing the research issue areas of drug supply chain and usage/sales history in Table II. Data availability in these areas includes discussions of time series, historical market consumption data, and inventory data. Real-time and integrated data are research challenges in the supply chain and usage/sales history area. Interestingly, the availability of data that affects the dynamics of drug demand has not been developed, particularly about disease factors (epidemiology) that determine the magnitude of drug demand. In addition to standing alone, epidemiological variables are also related to the research issue areas of Geography, Medical Records, Pharmaceutical Drug Management, and Disasters. In these areas, epidemiological variables are also considered to influence drug demand. In these areas, epidemiological variables are also considered to influence the prediction of drug demand.

The correlation between research issue areas requires an analysis of the significance of the predictive variables. The same variable that is not significant for drug prediction can change significantly, and vice versa. The combination of ten research issue areas illustrates the complexity of drug management. The more combinations of issue areas, the higher the data dimension, thus requiring more careful data handling for machine learning.

TABLE II. ISSUE AREAS, PROBLEMS, AND DATA AVAILABILITY

Research Issue Area	Problem Context	Data Availability	Ref.
Medical data record - pharmaceutical drug management	– The development of the number, diversity, and sources of data requires an appropriate prediction model to improve accuracy. – Inappropriate determination of significant variables in production, consumption, and storage stock predictions affects prediction accuracy.	– The electronic logistic management information system (eLMIS); for a five-year (2015–2019); available monthly or quarterly. Scope of research: Regional - Local Area. – Kerman Hospital data consisting of 283 features, covering more than 9351 different drugs and 121,690 patients (three years). Scope of research: Pharmacy.	[17], [49]
Epidemiology	– The level of emergency and spread of the disease affect the allocation of drugs. – The different and complex handling, distribution, storage, and administration methods for each drug make it a challenge to predict drug availability.	– Three major pharmaceutical companies producing insulin in the US. Scope of research: National companies. – Internal hospital data from Fast Healthcare Interoperability Resources (FHIR). Data collected from 2017 and 2021. Scope of research: Blood bank.	[55], [56]
Epidemiology - disaster	– The uncertainty of the impact of a disaster requires the prediction of drug supplies with a fluctuating cases number.	– Data on the number of COVID-19 infections in Shanghai. Period 20 January 2020 - 24 September 2022; Emergency Unit Scope of research: Hospital	[57]
Epidemiology - medical data record - pharmaceutical drug management	– Prediction models with the complexity of patient medical conditions, disease characteristics, and drug availability require more than one data source. The difficulty of designing a drug demand prediction model, especially for emergency units.	– Medical Data Record from the Mass General Brigham (MGB) Healthcare database. Emergency units from 14 hospitals. Scope of research: Local – Regional Hospital	[48]
Epidemiology - geographic	– For high-spread diseases such as COVID-19, epidemiological analysis becomes an important variable in predicting drug needs related to geographic distribution (from district level to national level).	– Data from: Lazio region health system (open datasets), Italian Medicine Agency (AIFA) (medicines and equipment), Italian National Civil Protection dataset (patient care/treatment - regional level), National Agency of Regional Health Services (AGENAS) (medical facilities). Scope of research: regional health services	[28]
History of usage/sales data	– Design of accurate drug demand prediction model for the needs of the drug supply pattern. – There is a demand for drugs/drug groups following seasonal trends, a prediction model is needed that is suitable for time series. – The need for a prediction approach method for complex pharmaceutical sales data sets.	– Monthly Sales Dataset from Kaggle. Transaction data of 57 drugs. – Weekly sales data for pharmaceutical products from the Kaggle. Transactions of 57 drugs with 600,000 transactions (2014–2019). Scope of research: algorithm comparison, Pharmacy, improving accuracy with different algorithmic approaches. – Historical sales data and market trends. A dataset of 600,000 transaction records from Kaggle (2014–2019). Scope of research: public data – eLMIS public health – Rwanda: to perform demand prediction of medicine. – Data from the analytical company AMS (Market Data in the Pharmaceutical Industry) in the Middle East and North Africa region, especially in Iraq. Scope of research: Time series dynamic	[35], [58]–[62]
History of usage/sales data - geographic	– A model approach to drug demand that does not only occur in one geographical point but also with the nearest drug demand point. – Non-communicable diseases have characteristics that require optimization of the management of essential drugs for non-communicable diseases (consumption or distribution).	– The National Health Service (NHS) in England Data. Prescription data in each practice August 2010 - December 2019. Scope of research: Adjacent pharmacy. – The electronic logistic management information system (eLMIS) data (2015–2019). Scope of research: Regional.	[63], [64]
Pharmaceutical drug management	– Enterprise planning (ERP) systems require an accurate demand forecasting model.	– Actual six different drug classes sales data (from 6 pharmacies in Pakistan). Data was collected from sources including injections, tablets, capsules, and suspensions of various potencies. Scope of research: Pharmaceuticals.	[65]
Drugs - patient behavior (self-medication)	– Patients have difficulty finding medicine for their illnesses independently (self-medication).	– The UCI Machine Learning Repository data - Drug Review with the subject Health and Medicine 6 features. Scope of research: Public data	[66]
Drug supply chain	– Ignoring that each region and demographic has its own drug needs – Manufacturers need accurate prediction of drug demand from downstream. On the other hand, pharmacies do not have visibility of drug supply – Data limitations: the real-time and patient prescription data for demand prediction – Data limitations, many hidden factors, changing market situations	– Sales data from PharmaGuide covering 22 pharmacies in Ontario, Canada (2014–2021) and historical drug shortage data from the Drug Shortages Canada website. Scope of research: Pharmacies – Simulation for application with sales data collected during 2014–2019. Scope of research: Pharmacies – Data from electronic data interchange (EDI). Weekly data on distributor-level sales volumes and inventory levels for 133 unique national drug codes (NDCs) from July 2007–August	[9], [50], [67]–[70]

Research Issue Area	Problem Context	Data Availability	Ref.
	– Variables determining demand prediction in the drug supply chain – Method of grouping and classifying products at the downstream level to improve the accuracy of drug demand prediction.	2017. Scope of research: Distributors – Database from an integrated services company, Coopservice Group, Italy. Scope of research: National companies – Data available for 519 samples containing historical information for 43 product types. Scope of research: Pharmacies	

Data availability relies on pharmaceutical logistics data sources, patient medical records, health services, independent health service units or hospitals, drug manufacturers and distributors, as well as pharmacies or hospital pharmacies. Complete and quality data helps find correlations between data variables. Complete and high-quality data help find correlations between data variables. The available dataset builds the initial assumption of the prediction variables. Nabizadeh [49] attempted to gather data sources from all departments but found the effectiveness of prediction to be limited when considering only one department. In contrast, Pall [50] used drug sales data from 22 pharmacies and historical drug shortage report data from Health Canada to obtain prediction results that help estimate drug demand. From this phenomenon, the correlation between the research issue area and the variables used determines the effectiveness of the available dataset.

The ten research area issues are divided into two categories. First, the main category encompasses areas such as pharmaceutical drug management, drug supply chain, and epidemiology. Second, additional categories occur due to research needs involving variables and data in geographic areas, medical records, and patient behaviour. Likewise, epidemiology can also develop in areas such as climatology and disaster management. Historical consumption and sales data can also involve areas such as supply chain management or pharmaceutical drug management. According to Mbonyinshuti [17], historical data on essential drug consumption extracted from eLMIS is an area of pharmaceutical drug management. In addition to historical data, patient origin data from medical records can help determine the geographic location of patients who require essential drugs. Finally, we understand that many studies predict the spread of COVID-19 to help distribute drugs more precisely [28], [51]–[54].

In the Epidemiology - Disaster, the hospital emergency unit is a frontline unit that requires drug resources according to the cases that occur. There is a correlation between disease prediction and the prediction of the type and amount of drugs required to treat the disease. Correlation can be observed in the analysis of patient data, diseases, and the medications they consume. Patient medical record data is a significant source for predicting patient drug needs [48]. The network of emergency units between many hospitals can produce data on the spread of infectious diseases as a significant variable. Collecting prescription data from several health service points provides information for analyzing drug demand in a particular distribution area [49], [64], [68]. With this analysis, distributors can formulate a drug distribution chain strategy. Drug demand information (from prescriptions) in the blockchain network can be used to identify the source of drug demand [71].

C. Machine Learning Algorithms for Drug Demand Prediction (RQ2)

The literature of this study, as shown in Fig. 4, indicates that ARIMA, LR, RF, eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbours, Support Vector Machine (SVM), ANN, multi-layer perceptrons (MLP), and LSTM algorithms are employed in more than two research studies. These algorithms are widely recognized for their performance. The selection of prediction algorithms is closely related to the performance and working method of the algorithm.

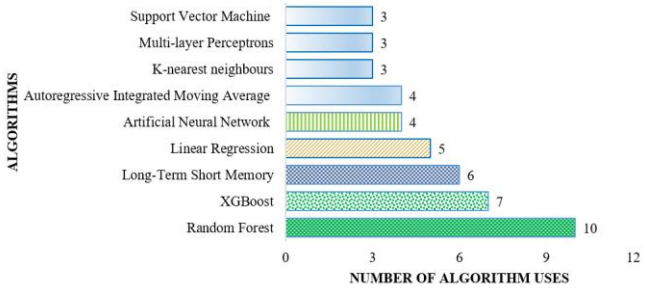


Fig. 4. The most widely used algorithm sequence in machine learning studies.

Additionally, the selection of machine learning algorithms is correlated with data availability factors and significant variables in prediction models. The choice of deep learning algorithms, such as ANN, MLP, and LSTM, has not dominated shallow machine learning algorithms. The complexity of the prediction research issue area is evident in its limited scope. To answer RQ2, Table III presents the performance of machine learning algorithm implementations for drug prediction based on selected articles.

Comparing multiple machine learning algorithms to improve prediction accuracy is a limited research area. The development of methods to handle cases of research area combinations presents an opportunity to enhance machine learning algorithms. The Machine Learning approach can simultaneously combine learning algorithms for big data and several causal variables with non-linear relationships. Machine Learning approaches can simultaneously combine learning algorithms for big data and several causal variables with non-linear relationships [36]. The study by Subudhi [48] took place in the Intensive Care Unit (ICU). The epidemiology, medical data record, and pharmaceutical drug management areas encouraged the use of ensemble models to handle cases of multi-source data. The average F1 score prediction results were higher than 0.83. Likewise, in the study by Zhang and Li [57], the disaster and epidemiology areas face challenges in processing disaster data and handling diseases that arise from disasters. The combination method produces R2= 0.9996 for the BILSTM-GASVR combination prediction model.

TABLE III. MEASUREMENT OF MACHINE LEARNING ALGORITHM PERFORMANCE IN SELECTED ARTICLES

Research issue area	Algorithm	Measurement	Ref.
Medical data record - pharmaceutical drug management	Linear regression, RF, ANN	RF (accuracy 76%)	[17]
	RF, neural network (NN)	RF model accuracy 83.3% (for a month) and 81.3% (for a season - two years data)	[49]
Epidemiology	RF, multiple regression, ANNs	ANN (the correlation coefficient value between output and response value is 93,1%)	[55]
	XGBoost, RF, dummy classifier, and Optuna	XGBoost (Accuracy 40,4%) (Specificity 98,75%)	[56]
Epidemiology - disaster	Bilateral long-short-term memory and genetic algorithm support vector regression (BILSTM-GASVR), bilateral long-short-term memory (BILSTM), genetic algorithm support vector regression (GASVR), informer	BILSTM-GASVR combined prediction model (R^2=99,96%)	[57]
Epidemiology - medical data record - pharmaceutical drug management	AdaBoost, bagging, gradient boosting, RF, XGBoost, ExtraTrees, logistic regression, decision tree, linear discriminant analysis, quadratic discriminant analysis, MLP Classifier, perceptron, passive AggressiveClassifier, linear SVC, Gaussian NB, K-neighbors, naive Bayes, SVM, NN	ICU admission: ensemble and tree-based models had significantly higher F1 scores than all other model types. Mortality: all ensemble-based models had average F1 scores higher than 0.83.	[48]
Epidemiology - geographic	Linear regression, K-nearest neighbours (KNN), recurrent neural networks (RNN)	The medicinal purchases prediction for the regions of Lazio and Piemonte has yielded a low RMSE 16%	[28]
History of usage/sales data	Linear regression, RF, SVM, XGBoost	XGBoost (accuracy: > 95% every drug sale)	[35]
	Probabilistic neural network, generalized regression neural network, radial basis function neural network (RBF-NN), long-term short memory neural network, stacked LSTM	Shallow neural networks. RMSE: Probabilistic Neural Network (6,21), Generalized Regression Neural Network (6,01), RBF-NN (6,59)	[58]
	ARIMA, LSTM	ARIMA (short term prediction), LSTM (long term and complex patterns)	[62]
	Model Bayesian, regresi proses gaussian (GPR): exponential squared, Revised Mat´ern, Rational Quadratic kernels	Kernel Ensemble MSE (9,09 × 10 −21), MAE (7,48 × 10 −11), RMSE (9,53 × 10 −11), Score R^2 (1,0)	[60]
	ARIMA, LSTM	The best LSTM accuracy RMSE 2.0 (training) and 2.043 (testing); R^2=0.952 (training) and 0.912 (testing).	[61]
	Multi-layer perceptrons (MLP), convolutional neural network (CNN), LSTM	The best sales LSTM performance with RMSE=1.28(k) and MEA=0.85(k); for USD Price RMSE=0.75, and MEA=0.44.	[59]
History of usage/sales data - geographic	ARIMA, XGBoost, gate recurrent unit (GRU), graph convolutional network (GCN), spatiotemporal graph convolutional network (STGCN), attention-based spatial-temporal graph convolutional network (ASTGCN)	ASTGCN (MAE 0,0957, RMSE 0,1241 – short period), STGCN (MAE 0,1139 RMSE 0,1465 – long period)	[64]
	RF	Accuracy: R-square 0.78 (training) and 0.71 (testing)	[63]
Pharmaceutical drug management	Autoregressive integrated moving average (ARIMA), Holts Winter, ARHOW	ARHOW (RMSE 18; MAPE 1,99)	[65]
Drugs - patient behavior (self-medication)	RF, naïve Bayes, long-short term memory	RF (Accuracy 83%)	[66]
Drug supply chain	Facebook prophet, LSTM neural networks, and XGBoost	XGBoost (has the lowest MAPE value of 5 out of 8 drug sales)	[69]
	XGBoost	able to address 59% of major drug shortages	[50]
	XGBoost	has the lowest MSE of 6 out of 8 types of drugs	[68]
	LR, moving average (MA), SVR, RF, basic exponential smoothing (ES), RNN	RNN (NMSE 2.69±0.37; NMAE 0.30±0.06)	[9]
	LR, RF, KNN, SVM, MLP	SVM (MSE 2,02 Medicine; 1,89 month)	[70]
	K-Means, Elbow method, Classification And Regression Trees (CART), MLP- neural network, RBF-NN, polynomial (or group method of data handling (GMDH) neural network	MLP-NN (cluster 1 is MLP2 and MLP4, cluster 2 is MLP4 and GMDH10, cluster 3 is MLP4, cluster 4 is MLP2, for all samples is MLP2.	[67]

Each algorithm has its own variation in prediction accuracy (see Table III). Using public data, Ankita et al. [62] applied different algorithms according to the prediction target, following both short-term and long-term seasonal trends. In other words, the complex pattern of historical drug consumption data affects accuracy. Using the same data source, Mbonyinshuti et al. [17] employed different significant variables for essential drug target prediction, as the predicted target variables are more specific to non-communicable diseases [63]. Another perspective on the same data source can provide new knowledge. Every change in data can change the pattern and prediction results. We need an analysis study to explore each algorithm in solving time series problems [58]. Analysis of low machine learning performance requires further research because generalization requires carefulness in approaching the data source [17], [50], [56].

The actual value factor also opens up the possibility of developing significant variables in predicting drug demand. In the case of Engelke [56], they evaluated data from multiple systems within a hospital to form the basis for predicting platelet concentrate (PC) transfusion. The low prediction recall value (0.3648) indicates a significant false negative (FN) value.

The patient is predicted not to need a PC, but in reality, the patient needs it. So, the availability of historical health data on the patient's condition from the EHR can affect the prediction performance. In addition, Mirshekari [60] utilized the power of Bayesian and Gaussian process regression (GPR) models for time series analysis, specifically the kernel ensemble, which can process low-dimensional variables and provide complete and high-quality data. This Bayesian technique optimization produced a prediction model with high accuracy. Again, real-time data on drug demand trends drives the algorithm to deliver better accuracy. Ultimately, the significant variables from the available variable data drive the performance of the combined prediction model algorithm [57], [68].

D. The Potential Variables in Pharmaceutical Drug Prediction (RQ3)

Manufacturers and distributors require accurate product predictions in production and distribution planning to ensure patients receive products with sufficient quality and quantity. Thus, variables related to drug and medical device products also include information on the area of need and the magnitude of the decision. In addition, epidemiological variables need parallel studies with predictions of disease spread. Table IV shows that the areas of research issues and prediction variables can continue to develop through machine learning.

TABLE IV. RESEARCH ISSUE AREAS AND RESEARCH VARIABLES

Research Issue Area	Variables		Ref.
Medical data record - pharmaceutical drug management	– Patient identity and demographics – Type of disease	– Doctor's prescription – Type and amount of medication	[17], [49]
Epidemiology	– Patient demographics – Disease diagnosis – Mortality – Obesity	– Cholesterol – Medication type – Patient condition	[55], [56]
Epidemiology - disaster	– Number of cases – Type and quantity of medication	– History of medication consumption	[57]
Epidemiology - medical data record - pharmaceutical drug management	– Patient's medical condition – Patient's medical history	– Treatment history – Laboratory test results – Mortality	[48]
Epidemiology - geographic	– Number of new cases – Geographic mortality	– Number of outpatients/inpatients	[28]
History of usage/sales data	– Drug sales history data – Drug types and quantity – Date/time of sale – Therapeutic group	– Company Type – Product Source – Number of Competitors – Price	[35], [58]–[62]
History of usage/sales data - geographic	– Drug sales history data – Drug types and quantity – Name and number of prescriptions	– Consumer region – Location of hospital/health service – Time of consumption	[63], [64]
Pharmaceutical drug management	– Name, type, and number of drugs		[65]
Drugs - patient behavior (self-medication)	– Name and condition of the drug, – Contents and properties of the drug	– Drug rating – Consumption time	[66]
Drug supply chain	– Drug sales history data – Seasonal trends – Number of prescriptions and patients – Number of sales per time – Name, type, and number of drugs – Number of orders per time – Timeliness of order arrival	– Supplier and number of orders – Number of drug stock – Storage qualifications – Frequency of consumption – Patient demographics – Product price – Type of disease	[9], [50], [67]–[70]

Variables spread across ten research issue areas assume that they have been tested in their research as significant variables in machine learning algorithms. The challenge is how the model performs when these variables are absent [50]. Likewise, expert judgment needs to provide comprehensive analysis of the variables' significance [65]. In general, the availability of data on demand variables for each type of drug can reveal the seasonal trend of the drug, allowing the model to produce more accurate pattern predictions, as consumption trend data is synchronous with seasonal patterns [62]. The addition of other variables becomes significant when they support the prediction pattern, including the average number of days of drug supply per patient, the duration of drug supply, previous drug shortages, drug hierarchies, and different therapeutic groups [50].

Determining research issue areas, especially those involving combinations, helps researchers identify potential variables for predicting drug demand. As a result, machine learning utilizes richer, multi-source data for learning. Multimodal data from various hospital information systems opens the discovery of potential variables from geospatial, epidemiological, and climatological data [56], [63]. Potential variables can be identified through the analysis of disease epidemiology spread. Real-time data and trends related to

geographic distribution (from local to regional and national levels) are essential data requirements for training machine learning prediction models [28]. Epidemiological variables are not limited to disease names but also include disease characteristics, disease sources, and drugs that control the disease. Variables predicting disease spread correlate with variables in drug prediction [49].

Furthermore, an analysis of variables is needed to determine which variables remain significant in the present and the future [55]. The development of complex modern pharmaceutical predictions requires careful analysis to explore significant variables [69]. Variable development requires analysis of the relationship between one variable and another. Finding the relationship between location variables, drug prices, and weather conditions can improve system performance [68]. As a development, classification methods can help drug reviews identify the relationship between disease names, drug names, and disease indications, enabling patients to find suitable drugs for their conditions [66], [67], [72].

Table IV also presents the findings of the same variables from several other studies. The need for the same variables arises. Fig. 5 shows the frequency of variables used in selected articles.

Fig. 5 shows that drug sales/consumption (quantity, type, time) (11) and drug data (name, condition, quantity, expired date, and price) (12) are the most widely used variables in drug prediction. Together with the doctor's prescription parameters (name and drug quantity) (2) and drug reviews (content and properties of the drug) (1), these variables are the most dominant among the other variables. The epidemiological variable group uses the most variables, namely the number of disease cases (3). The variables of health service location - distributor (5) and patient domicile (3) represent discussions related to the drug supply chain. Meanwhile, the prediction pattern using the susceptible-expose-infectious-removal (SEIR) method uses many patient identity and condition variables.

The various research variables found can be grouped into five categories: epidemiology, drug consumption, drug storage and ordering (drug management in pharmacies), supply chain and patient domicile (drug distribution - area - health service distribution), and patient identity and condition (pharmacotherapy). The third research question (RQ3) is answered. These five categories represent potential variables for predicting drug demand, as identified in selected articles. This study recommends the following five categories for selecting potential variables for drug prediction using machine learning. The five categories demonstrate that obtaining significant variables by measuring the proximity between variables using machine learning algorithms can overlook field conditions, such as the spread of disease and the distribution of drugs from a health service. However, the five categories of potential variables are identified as significant variables in machine learning models for drug demand prediction.

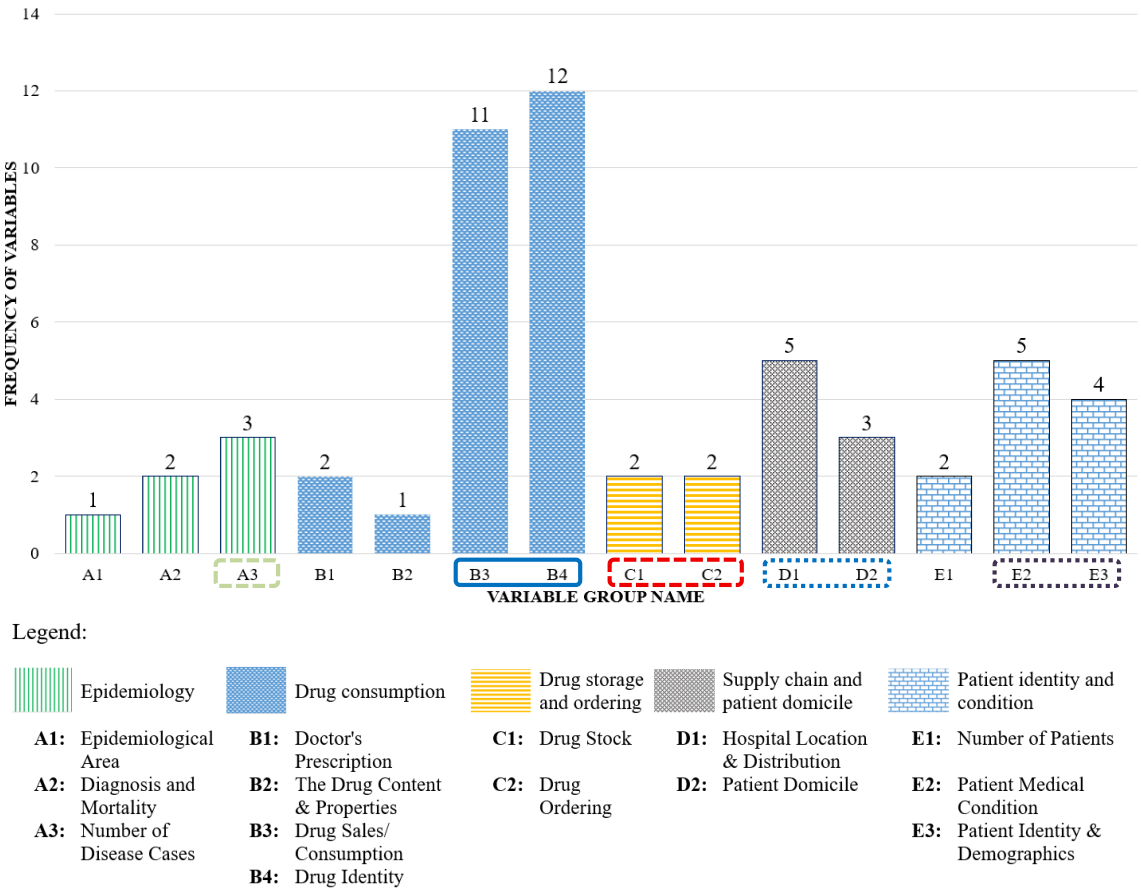


Fig. 5. The Frequency of variable use in research.

E. Integrated Datasets, Algorithms, and Potential Variables in Pharmaceutical Drug Prediction

The study begins by exploring potential variables by identifying issue areas. Classifying issue areas facilitates the study in sorting out the problem context and data availability in building machine learning models for drug demand prediction. Of the ten issue areas, the drug supply chain and drug consumption are closely related to the problem of predicting drug demand. This closeness arises because machine learning models aggregate historical data on drug consumption for prediction purposes. Increasing accuracy is the target of drug demand prediction research. Therefore, model development relies on algorithms that can enhance prediction accuracy. However, several studies have encountered obstacles due to limited data or restricted access to data (see Table II). The use of public data is one solution in building machine learning models for prediction [35], [58], [62], [69]. Indirectly, the development of drug demand prediction models ignores uncertain environmental conditions. Dynamic data changes, diverse data types, and varied data sources, which tend to increase data complexity, are not widely observed in this study. Therefore, further research is needed to understand how algorithms recognize data patterns and select significant features.

This study also examined the combination of medication consumption, disease, and treatment management. Dataset limitations often hinder the development of machine learning models. Potential variables can be a valuable option for developing machine learning algorithms, especially given the difficulty of obtaining certain types of data, which can reduce predictive performance. The results of this study, which paired the issue area with research variables (see Table IV), can be used as a consideration in developing machine learning models for regression, classification, and clustering. Selecting potential variables with close relevance can improve the performance of machine learning algorithms, particularly by reducing model complexity. Potential variables are rationally selected based on their relevance to the case and their proximity to the case, before being tested in machine learning algorithms using available data [73].

A limitation of this study is that many articles were included in the initial search because the search keywords were still too broad. The suitability of the search keywords captures more secondary research articles. Meanwhile, the discovery of primary research articles often involves fewer search keywords that match the contents of the primary research article. Selecting articles requires much careful effort to read the entire article. The number of selected articles is not yet representative of a particular continental region, thus requiring further study at the national level.

V. Conclusion

Mapping of research problem areas and significant variables in machine learning models for drug demand prediction produces ten research problem areas. The dynamics of data needs and algorithm choices for solving research problems are evident in the mapping. The challenge in the combination area is the increasingly complex and evolving data dimensions and algorithm choices. The selection of algorithms is still dominated by RF, XGBoost, and LSTM (for deep learning) machine learning algorithms. Efforts to improve accuracy dominate the research, providing a comparative perspective of various algorithms to improve predictions. Meanwhile, this study is in the area of significant variables that often depend on data limitations and algorithm performance.

This study identifies five potential variables in pharmaceutical drug prediction: drug consumption, drug management, epidemiology, supply chain-patient domicile, and pharmacotherapy. These five variables have not established themselves as a benchmark model or framework for designing drug demand predictions. However, mapping these variables can provide perspective for drug supply chain actors in developing machine learning models for downstream drug prediction.

Potential variables will not be significant in the model without the development of integrated data availability that is appropriate and relevant to the characteristics of the drug supply chain, endemic diseases, and public health services. Potential variables can be the basis for updating government policies related to health data integration and drug procurement guidelines in pharmacies. This research provides a basis for future research. The challenge of building a machine learning model for drug demand prediction involves working with multiple pharmacies across different geographic and demographic areas.

References

[1] Y. A. Mashuri et al., “The response to COVID-19 among drug retail outlets in Indonesia: A cross-sectional survey of knowledge, attitudes, and practices,” Lancet Reg. Heal. - West. Pacific, vol. 22, p. 100420, 2022, doi: 10.1016/j.lanwpc.2022.100420.

[2] W. Ying, Y. Qian, and Z. Kun, “Drugs supply and pharmaceutical care management practices at a designated hospital during the COVID-19 epidemic,” Res. Soc. Adm. Pharm., vol. 17, no. 1, pp. 1978–1983, Jan. 2021, doi: 10.1016/j.sapharm.2020.04.001.

[3] Mudriyan, A. Razak, Sukri, M. A. Arifin, A. Syam, and Y. Thamrin, “Analysis of Management of Consumable Medical Materials during the Covid-19 Pandemic in Pharmaceutical Installations Health Department of Baubau City,” NeuroQuantology, vol. 20, no. 8, pp. 926–938, 2022, doi: 10.14704/nq.2022.20.8.NQ44100.

[4] H. A. Sani, A. Asriwati, M. Nyorong, M. Sari, and I. Ginting, “Analysis of Drug Availability Management at the Regional General Hospital Pharmacy Installation,” J. La Medihealtico, vol. 4, no. 2, pp. 87–102, 2023, doi: 10.37899/journallamedihealtico.v4i2.837.

[5] M. H. Ambon, “Evaluation of Drug Planning and Procurement at the Pharmacy Installation of Regional General Hospital Dr. M. Haulussy Ambon,” Medico-Legal Updat., vol. 20, no. 4, 2020, doi: 10.37506/mlu.v20i4.1981.

[6] M. Niaz and U. Nwagwu, “Managing Healthcare Product Demand Effectively in The Post-Covid-19 Environment: Navigating Demand Variability and Forecasting Complexities,” Am. J. Econ. Manag. Bus., vol. 2, no. 8, pp. 316–330, 2023, doi: https://doi.org/10.58631/ajemb.v2i8.55.

[7] C. Franco and E. Alfonso-Lizarazo, “Optimization under uncertainty of the pharmaceutical supply chain in hospitals,” Comput. Chem. Eng., vol. 135, p. 106689, 2020, doi: 10.1016/j.compchemeng.2019.106689.

[8] K. Moons, G. Waeyenbergh, and L. Pintelon, “Measuring the logistics performance of internal hospital supply chains – A literature study,” Omega (United Kingdom), vol. 82, pp. 205–217, 2019, doi: 10.1016/j.omega.2018.01.007.

[9] X. Zhu, A. Ninh, H. Zhao, and Z. Liu, "Demand Forecasting with Supply-Chain Information and Machine Learning: Evidence in the Pharmaceutical Industry," Prod. Oper. Manag., vol. 30, no. 9, pp. 3231–3252, 2021, doi: https://doi.org/10.1111/poms.13426.

[10] M. Wang and F. Jie, "Managing supply chain uncertainty and risk in the pharmaceutical industry," Heal. Serv. Manag. Res., vol. 33, no. 3, pp. 156–164, Jun. 2019, doi: 10.1177/0951484819845305.

[11] L. P. L. Wulandari, A. A. S. Sawitri, and A. Hermansyah, "The potential roles of pharmacy medication sales data to augment the syndromic surveillance system in response to COVID - 19 and preparedness for other future infectious disease outbreaks in Indonesia," Int. J. Health Plann. Manage., vol. 37, no. 1, pp. 30–39, Jan. 2022, doi: 10.1002/hpm.3356.

[12] I. Vanany, M. H. Ali, K. H. Tan, A. Kumar, and N. Siswanto, "A Supply Chain Resilience Capability Framework and Process for Mitigating the COVID-19 Pandemic Disruption," IEEE Trans. Eng. Manag., 2021, doi: 10.1109/TEM.2021.3116068.

[13] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," Int. J. Intell. Networks, vol. 3, no. February, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.

[14] Tangi Ndakondja Angula and Abraham Dongo, "Assessing the impact of artificial intelligence and machine learning on forecasting medication demand and supply in public pharmaceutical systems: A systematic review," GSC Biol. Pharm. Sci., vol. 26, no. 2, pp. 140–150, Feb. 2024, doi: 10.30574/gscbps.2024.26.2.0071.

[15] D. Ni, Z. Xiao, and M. K. Lim, "A systematic review of the research trends of machine learning in supply chain management," Int. J. Mach. Learn. Cybern., vol. 11, no. 7, pp. 1463–1482, 2020, doi: 10.1007/s13042-019-01050-0.

[16] R. Alfred and J. H. Obit, "The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review," Heliyon, vol. 7, no. 6, p. e07371, 2021, doi: 10.1016/j.heliyon.2021.e07371.

[17] F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, "The prediction of essential medicines demand: A machine learning approach using consumption data in Rwanda," Processes, vol. 10, no. 1, pp. 1–13, 2022, doi: 10.3390/pr10010026.

[18] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," Appl. Intell., vol. 52, no. 4, pp. 4543–4581, Mar. 2022, doi: 10.1007/s10489-021-02550-9.

[19] 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. 1, pp. 56–70, 2020, doi: 10.38094/jastt1224.

[20] I. Syahrir, Suparno, and I. Vanany, "Drug Supplies Planning in Hospital for Epidemic Attack using SEIR Model," J. Phys. Conf. Ser., vol. 1179, no. 1, 2019, doi: 10.1088/1742-6596/1179/1/012150.

[21] I. Syahrir, Suparno, and I. Vanany, "A Proposed Model for Drug Demand Forecasting and Ordering Inventory System for Dengue Endemic," Oper. Supply Chain Manag., vol. 15, no. 1, pp. 69–78, 2022, doi: 10.31387/oscm0480331.

[22] T. S. Ho et al., "Comparing machine learning with case-control models to identify confirmed dengue cases," PLoS Negl. Trop. Dis., vol. 14, no. 11, pp. 1–21, 2020, doi: 10.1371/journal.pntd.0008843.

[23] N. A. M. Salim et al., "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," Sci. Rep., vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-020-79193-2.

[24] P. Siriyasatien, A. Phumee, P. Ongruk, K. Jampachaisri, and K. Kesorn, "Analysis of significant factors for dengue fever incidence prediction," BMC Bioinformatics, vol. 17, no. 1, pp. 1–9, 2016, doi: 10.1186/s12859-016-1034-5.

[25] F. Yavari Nejad and K. D. Varathan, "Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–12, 2021, doi: 10.1186/s12911-021-01493-y.

[26] S. F. McGough, L. Clemente, J. N. Kutz, and M. Santillana, "A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles," J. R. Soc. Interface, vol. 18, no. 179, 2021, doi: 10.1098/rsif.2020.1006.

[27] J. Ong et al., "Mapping dengue risk in Singapore using Random Forest," PLoS Negl. Trop. Dis., vol. 12, no. 6, pp. 1–12, 2018, doi: 10.1371/journal.pntd.0006587.

[28] S. Atek et al., "A predictive decision support system for coronavirus disease 2019 response management and medical logistic planning," Digit. Heal., vol. 9, 2023, doi: https://doi.org/10.1177/20552076231185475.

[29] M. Syaripuddin, R. Ayu, D. Sartika, M. Nadjib, and M. Syaripuddin, "Qualitative Study on the Drug Availability for National Health Insurance (JKN) in Hypertensive Patients : the Perspective of the Tender Winner Manufacturer," East African Sch. J. Med. Sci., vol. 4421, no. 11, pp. 674–678, 2019, doi: 10.36349/EASMS.2019.v02i11.018.

[30] D. Koala, Z. Yahouni, G. Alpan, and Y. Frein, "Factors influencing Drug Consumption and Prediction Methods," CIGI-Qualita Conférence Int. Génie Ind. Qual. - Grenoble, no. May 2021, pp. 1–8, Sep. 2021, doi: https://doi.org/10.48550/arXiv.2109.11779.

[31] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," J. Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00329-2.

[32] J. Y. S. Ng, R. V. Ramadani, D. Hendrawan, D. T. Duc, and P. H. T. Kiet, "National Health Insurance Databases in Indonesia, Vietnam and the Philippines," PharmacoEconomics - Open, vol. 3, no. 4, pp. 517–526, 2019, doi: https://doi.org/10.1007/s41669-019-0127-2.

[33] A. Burinskiene, "Forecasting Model: The Case of the Pharmaceutical Retail," Front. Med., vol. 9, no. August, pp. 1–16, 2022, doi: 10.3389/fmed.2022.582186.

[34] M. A. Al-Gunaid et al., "Analysis of drug sales data based on machine learning methods," Proc. 2018 Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2018, pp. 32–38, 2018, doi: 10.1109/SYSMART.2018.8746968.

[35] A. Manna, K. Kolpe, A. Mhalungekar, S. Pattewar, P. Kaloge, and R. Patil, "Comparative Study of Various Machine Learning Algorithms for Pharmaceutical Drug Sales Prediction," in 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), May 2023, pp. 1–6, doi: 10.1109/IConSCEPT57958.2023.10169964.

[36] J. Feizabadi, "Machine learning demand forecasting and supply chain performance," Int. J. Logist. Res. Appl., vol. 25, no. 2, pp. 119–142, 2022, doi: 10.1080/13675567.2020.1803246.

[37] B. K. Almentero, J. Li, and C. Besse, "Forecasting pharmacy purchases orders," in 2021 IEEE 24th International Conference on Information Fusion (FUSION), Nov. 2021, pp. 1–8, doi: 10.23919/FUSION49465.2021.9627017.

[38] X. Zhu, A. Ninh, H. Zhao, and Z. Liu, "Cross - Series Demand Forecasting using Machine Learning: Evidence in the Pharmaceutical Industry," Prod. Oper. Manag., no. March 2019, pp. 1‑54, 2021, doi: 10.1111/poms.13426.

[39] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2018th ed. Lean Publishing, 2018.

[40] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning, vol. 19. MIT Press book, 2017.

[41] Y. K. Ever, K. Dimililer, and B. Sekeroglu, "Comparison of Machine Learning Techniques for Prediction Problems," in Web, Artificial Intelligence and Network Applications - WAINA 2019. Advances in Intelligent Systems and Computing, 2019, vol. AISC, 927, pp. 713–723, doi: https://doi.org/10.1007/978-3-030-15035-8_69.

[42] B. A. Mousa and B. Al-Khateeb, "Predicting medicine demand using deep learning techniques: A review," J. Intell. Syst., vol. 32, no. 1, Aug. 2023, doi: 10.1515/jisys-2022-0297.

[43] X. Zhu, G. Zhang, and B. Sun, "A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence," Nat. Hazards, vol. 97, no. 1, pp. 65–82, 2019, doi: 10.1007/s11069-019-03626-z.

[44] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," Artif. Intell. Med., vol. 104, no. February, p. 101822, 2020, doi: 10.1016/j.artmed.2020.101822.

[45] A. M. Aamer, L. P. E. Yani, and I. M. A. Priyatna, "Data analytics in the supply chain management: Review of machine learning applications in demand forecasting," Oper. Supply Chain Manag., vol. 14, no. 1, pp. 1–13, 2021, doi: http://doi.org/10.31387/oscm0440281.

[46] S. Pizard, F. Acerenza, X. Otegui, S. Moreno, D. Vallespir, and B. Kitchenham, "Training students in evidence-based software engineering and systematic reviews: a systematic review and empirical study," Empir. Softw. Eng., vol. 26, no. 3, p. 50, May 2021, doi: 10.1007/s10664-021-09953-9.

[47] L. Koval et al., "Unlocking the Potential of Information Modeling for Root Cause Analysis in a Production Environment: A Comprehensive State-of-the-Art Review Using the Kitchenham Methodology," IEEE Access, vol. 12, no. May, pp. 80266–80282, 2024, doi: 10.1109/ACCESS.2024.3406020.

[48] S. Subudhi et al., "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19," NPJ Digit. Med., vol. 4, no. 1, p. 87, May 2021, doi: 10.1038/s41746-021-00456-x.

[49] A. H. Nabizadeh, M. M. Ghaemi, and D. Goncalves, "Predicting the pharmaceutical needs of hospitals using machine learning algorithms," Int. J. Data Sci. Anal., pp. 0–20, 2024, doi: https://doi.org/10.1007/s41060-024-00530-z.

[50] R. Pall, Y. Gauthier, S. Auer, and W. Mowaswes, "Predicting drug shortages using pharmacy data and machine learning," Health Care Manag. Sci., vol. 26, no. 3, pp. 395–411, 2023, doi: 10.1007/s10729-022-09627-y.

[51] N. Alballa and I. Al-turaiki, "Machine learning approaches in COVID-19 diagnosis , mortality , and severity risk prediction : A review," Informatics Med. Unlocked, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[52] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, and J. Chaki, "Forecasting Models for Coronavirus Disease (COVID - 19): A Survey of the State-of-the-Art," SN Comput. Sci., vol. 1, no. 4, pp. 1–15, 2020, doi: 10.1007/s42979-020-00209-9.

[53] E. Campillo-Funollet et al., "Predicting and forecasting the impact of local outbreaks of COVID-19: use of SEIR-D quantitative epidemiological modelling for healthcare demand and capacity," Int. J. Epidemiol., vol. 50, no. 4, pp. 1103–1113, Aug. 2021, doi: 10.1093/ije/dyab106.

[54] M. A. A. Al-qaness, A. A. Ewees, H. Fan, L. Abualigah, and M. Abd Elaziz, "Marine Predators Algorithm for Forecasting Confirmed Cases of COVID-19 in Italy , USA , Iran and Korea," Int. J. Environ. Res. Public Health, vol. 17(10), no. May, p. 3520, 2020, doi: 10.3390/ijerph17103520.

[55] Y. Haoudi, M. A. Yazdani, D. Roy, and S. Hennequin, "Demand Prediction Based on Machine Learning Algorithms for Optimal Distribution of Insulin," IFAC-PapersOnLine, vol. 56, no. 2, pp. 10174–10179, 2023, doi: 10.1016/j.ifacol.2023.10.893.

[56] M. Engelke et al., "Predicting Individual Patient Platelet Demand in a Large Tertiary Care Hospital Using Machine Learning.," Transfus. Med. Hemother., vol. 50, no. 4, pp. 277–285, Aug. 2023, doi: 10.1159/000528428.

[57] W. Zhang and X. Li, "A data-driven combined prediction method for the demand for intensive care unit healthcare resources in public health emergencies," BMC Health Serv. Res., vol. 24, no. 1, 2024, doi: 10.1186/s12913-024-10955-8.

[58] R. Rathipriya, A. A. Abdul Rahman, S. Dhamodharavadhani, A. Meero, and G. Yoganandan, "Demand forecasting model for time-series pharmaceutical data using shallow and deep neural network model," Neural Comput. Appl., vol. 35, no. 2, pp. 1945–1957, 2023, doi: 10.1007/s00521-022-07889-9.

[59] N. Qassrawi, M. Azzeh, and M. Hijjawi, "Drug sales forecasting in the pharmaceutical market using deep neural network algorithms," Int. J. Syst. Innov., vol. 8, no. 3, pp. 63–83, 2024, doi: https://doi.org/10.6977/IJoSI.202409_8(3).0006.

[60] S. Mirshekari, M. Moradi, M. Ensaf, H. Jafari, and M. Jafari, "Enhancing Predictive Accuracy in Pharmaceutical Sales Through an Ensemble Kernel Gaussian Process Regression Approach," SSRN Electron. J., vol. 18, no. 5, pp. 255–260, 2024, doi: 10.2139/ssrn.4860667.

[61] F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, "Health supply chain forecasting: a comparison of ARIMA and LSTM time series models for demand prediction of medicines," Acta Logist., vol. 11, no. 02, pp. 269–280, Jun. 2024, doi: 10.22306/al.v11i2.510.

[62] M. Ankita, A. Srinivas, A. Soni, G. Prajapati, and P. S. Manjunath, "Pharmaceutical Sales Data Prediction Using Time Series Forecasting," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 13s, pp. 681–696, 2024, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/4649.

[63] F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, "Application of random forest model to predict the demand of essential medicines for noncommunicable diseases management in public health facilities," Pan Afr. Med. J., vol. 42, p. 89, 2022, doi: 10.11604/pamj.2022.42.89.33833.

[64] X. Wang, L. Chen, J. Wang, and D. Zheng, "Machine Learning Approaches for Region-level Prescription Demand Forecasting," in 2023 IEEE Smart World Congress (SWC), 2023, pp. 1–6, doi: 10.1109/SWC57546.2023.10449058.

[65] R. Siddiqui, M. Azmat, S. Ahmed, and S. Kummer, "A hybrid demand forecasting model for greater forecasting accuracy: the case of the pharmaceutical industry," Supply Chain Forum, vol. 23, no. 2, pp. 124–134, 2022, doi: 10.1080/16258312.2021.1967081.

[66] S. Das, S. K. Mahata, A. Das, and K. Deb, "Disease Prediction from Drug Information using Machine Learning," Am. J. Electron. Commun., vol. 1, no. June, pp. 16–21, 2021, doi: https://doi.org/10.15864/ajec.1403.

[67] M. S. Amalnick, N. Habibifar, M. Hamid, and M. Bastan, "An intelligent algorithm for final product demand forecasting in pharmaceutical units," Int. J. Syst. Assur. Eng. Manag., vol. 11, no. 2, pp. 481–493, 2020, doi: 10.1007/s13198-019-00879-6.

[68] A. K. Bapatla, S. P. Mohanty, and E. Kougianos, "FortiRx 2 . 0 : Smart Privacy-Preserved Demand Forecasting of Prescription Drugs in," 2023, doi: 10.1109/OCIT59427.2023.10430944.

[69] K. P. Fourkiotis and A. Tsadiras, "Applying Machine Learning and Statistical Forecasting Methods for Enhancing Pharmaceutical Sales Predictions," forecasting, vol. 6, no. 1, pp. 170–186, 2024, doi: https://doi.org/10.3390/forecast6010010.

[70] M. B. de Oliveira, G. Zucchi, M. Lippi, D. F. Cordeiro, N. R. da Silva, and M. Iori, "Lead Time Forecasting with Machine Learning Techniques for a Pharmaceutical Supply Chain," in The 23rd International Conference on Enterprise Information Systems (ICEIS 2021), 2021, vol. 1, pp. 634–641, doi: https://dx.doi.org/10.5220/0010434406340641.

[71] A. Musamih et al., "A blockchain-based approach for drug traceability in healthcare supply chain," IEEE Access, vol. 9, pp. 9728–9743, 2021, doi: 10.1109/ACCESS.2021.3049920.

[72] M. M. H. Chowdhury and M. Quaddus, "Supply chain readiness, response and recovery for resilience," Supply Chain Manag., vol. 21, no. 6, pp. 709–731, 2016, doi: 10.1108/SCM-12-2015-0463.

[73] F. Orhan and M. N. Kurutkan, "Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors," BMC Health Serv. Res., vol. 25, no. 1, p. 366, Mar. 2025, doi: 10.1186/s12913-025-12502-5.

Scalable Graph Learning with Graph Convolutional Networks and Graph Attention Networks: Addressing Class Imbalance Through Augmentation and Optimized Hyperparameter Tuning

Chaima Ahle Touate[1], Rachid El Ayachi[2], Mohamed Biniz[3]
Information Processing and Decision Support Laboratory-Department of Computer Science-Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco[1]
Department of Computer Science-Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco[2]
Department of Mathematics and Computer Science, Sultan Moulay Slimane University, Beni Mellal, Morocco[3]

***Abstract*—In this study, we propose a graph-based node classification to address challenges such as data scarcity, class imbalance, limited access to original textual content in benchmark datasets, semantic preservation, and model generalization in node classification tasks. Beyond simple data replication, we enhanced the Cora dataset by extracting content from its original PostScript files using a three-dimensional framework that combines in one pipeline NLP-based techniques such as PEGASUS paraphrase, synthetic model generation and a controlled subject aware synonym replacement. We substantially expanded the dataset to 17,780 nodes—representing an approximation of 6.57x scaling while maintaining semantic fidelity (WMD scores: 0.27-0.34). Our Bayesian Hyperparameter tuning was conducted using Optuna, along with k-fold cross-validation for a rigorous optimized model validation protocol. Our Graph Convolutional Network (GCN) model achieves 95.42% accuracy while Graph Attention Network (GAT) reaches 93.46%, even when scaled to a significantly larger dataset than the base. Our empirical analysis demonstrates that semantic-preserving augmentation helped us achieve better performance while maintaining model stability across scaled datasets, offering a cost-effective alternative to architectural complexity, making graph learning accessible to resource-constrained environments.**

***Keywords*—Graph Convolutional Networks (GCN); Graph Attention Networks (GAT); hyperparameter tuning; data augmentation; PEGASUS; synonym replacement; optuna bayesian optimization; node classification; class imbalance**

I. Introduction

Text classification represents a foundational element within the domain of natural language processing (NLP), showcasing a broad scope of applications. Text classification systems must effectively map textual content to appropriate categories, a task complicated by the need to capture both local semantic patterns and global contextual relationships. In conventional methods, text is represented by hand-built features, typically lexical features (e.g., bags of words and n-grams) although they are efficient for particular tasks, semantic complex patterns are not well captured. The development towards deep learning models has been considerably applied to learn advanced textual representations, incorporating convolutional neural networks (CNNs) [1] and recurrent neural networks (RNNs) like long-term memory (LSTM) [2], such models perform well in terms of capturing semantic and syntactic insights within sequences of local consecutive words; however, they are still likely to disregard co-occurrence of the global words in a corpus that covers long-range, non-consecutive semantics dependencies that are important for overall text comprehension [3].

Lately, emerging graph-based approaches trends have gained significant attention to remedy these limitations, such as graph neural networks or graphical embeddings [4]. Unlike conventional sequence-based methods, GNNs allows for processing data in graphical interconnected network format. This paradigm shift enable the models to leverage both local characteristics of nodes and global graph topology, given that graphs are highly eloquent and boost calculation performance and extract semantic and syntactic information, innovative methods emerged for processing graph-structured data in machine learning, i.e., the graphical neural network, making its mark in numerous applications and proving itself as an effective and successful architecture.

Among the various GNN architectures, particularly Graph Convolutional Networks (GCNs) [5], shown remarkable success in node classification tasks. Based on the foundational work of spectral graph theory, the studies of (Bruna et al. 2014) [6], (Henaff et al. 2015) [7], and (Defferrard, Bresson, and Vandergheynst 2016) [8] set up the theoretical basis for carrying out convolution operations on graph-structured datasets, yielding competitive results in node classification seminal contribution introducing a localized, first-order method seminal contribution streamlined these spectral approaches by offering a localized, first-order approximation that maintains computational efficiency while achieving state-of-the-art performance across diverse graph-structured datasets, including the widely adopted Cora benchmark. Further applications include various domains from semantic role labeling in [9] to biomedical entity classification in [10].

Incorporating attention mechanisms into GNNs is yet a further significant advancement. The Graph Attention Network (GAT), introduced by Velickovic et al. (2018) [11], builds

upon the concept of attention, thereby enabling the model to prioritize the most significant nodes during the aggregation process. Graph Attention Networks (GATs) have been highly influential in enhancing the performance of GNNs by dynamically assigning adaptive attention weights to neighboring nodes based on their relevance allowing them to allocate computational resources to the richest connections, resulting in more nuanced and effective feature representations, notably in heterogeneous graph settings in which node and edge types diverge considerably.

Although there have been significant architectural advances, critical challenges persist that impede the practical application and scalability of graph-based textual classification models such as data scarcity, class imbalance, and semantic coherence preservation. Benchmark datasets are typically provided in a format that limits their potential for these type of models, Cora [12] an example of a standard dataset, while academically valuable it comes often in pre-vectorized formats, which detaches the learning process from the original semantic complexity, restricts certain feature engineering introducing a barrier in research from flexibility in exploration of different preprocessing pipelines. In addition to its small size and class imbalance which further hinders generalization rare classes are not well represented. Furthermore, the lack of access to the original text prevents any attempt at implementing effective augmentation strategies. These common limitations prevent GNNs from fully using their ability to learn contextual representations from textual content.

Existing approaches exhibit a notable research gap: while important efforts have focused on architectural innovations to advance GNN capabilities, less attention has been paid to the research-based investigation of dataset quality and preparation. This gap can be seen in three practical challenges that limit how well we can evaluate performance. First, benchmark datasets like Cora are distributed in pre-vectorized formats, which limit access to original textual content and opportunities for controlled preprocessing or semantic augmentation. Second, the often-overlooked persistent class imbalance such as Neural Networks composed of 30% of samples while only 6% for Rule Learning can make the evaluation less accurate. Third, semantic preservation metrics for augmented graph data remain inconsistent, this makes it hard to make sure that synthetic relationships are meaningful. These data-related limitations suggest that focusing only on architectural complexity may not solve foundational challenges in graph-based text classification, underscoring the need for complementary data enhancement approaches. Although several studies acknowledge these issues, more careful investigations into semantic-preserving augmentation and class distribution mitigation remain limited. Addressing these challenges is essential to support and not replace architectural progress.

To address this gap, this study explores the following research question: How can semantic-preserving dataset augmentation serve as a cost-effective complement to architectural innovations in graph-based text classification, while maintaining semantic consistency and boosting performance without increasing computational demands? This question is motivated by the practical need to make advanced graph learning accessible to resource-constrained environments while tackling fundamental data quality issues that architectural solutions alone cannot resolve. To answer this question, we present a scalable framework for dataset augmentation that preserves semantic consistency, aimed at improving graph-based text classification using graph neural networks (GNNs). Our work places particular emphasis on Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). The framework responds to a frequent issue in this field—the discrepancy between the sophistication of theoretical models and the constraints placed on by real-world datasets. While much of the existing literature centers on architectural innovation, our findings suggest that focusing on data quality—specifically through targeted augmentation—can lead to performance gains that are comparable to those achieved through complex model redesigns. Our approach addresses several challenges:

1) Cost-effective: alternative to developing new architecture, making advanced graph learning accessible to researchers with limited computational resources.

2) Data augmentation: We introduce a hybrid augmentation pipeline that combines in a refined pipeline, NLP-based techniques like PEGASUS for text paraphrase [13] , controlled synonym replacement techniques [14], and a synthetic domain-specific text generation, we augment the dataset to enhance both the diversity and quantity of the training data and solve class imbalances while preserving semantic coherence while enabling in addition flexible and transparent text preprocessing.

3) Semantic quality assurance: We use semantic similarity checks employing the metric Word Mover's Distance (WMD) [15] to confirm that augmented samples maintain meaning and contextual relevance rather than depending just on cosine similarity operating on fixed vector representations and may overlook semantic changes introduced by paraphrasing.

4) Model optimization and evaluation: we use Bayesian-optimized Optuna [16] hyperparameter tuning, and k-fold cross-validation [17] to guarantee strong evaluation of model performance, providing a more reliable assessment of generalization across different data subsets. This method enhances the reliability of our systematically fair comparison of GCNs and GATs, in order to guarantee that the models reach their maximum potential on the augmented dataset.

Our approach yields GCN achieving an accuracy of 95.42% and GAT achieving 93.46% on 6.57x larger Cora dataset while supporting semantic fidelity (WMD scores: 0.27-0.34) showing the effectiveness of models with our flexible pipeline.

This paper is organized as follows: Section II reviews related work on GNNs and recent advancements in node classification. Section III details our methodologies. Section IV presents experimental results and analysis. Finally, Section V concludes with limitations and future research directions.

II. Related Works

Recent advancements in Graph Neural Networks (GNNs) have propelled the field of graph-based node classification forward in addressing various limitations. Key developments include:

1) Spectral graph convolutions: The groundwork for spectral graph convolutions was established by Bruna et al. (2014) [6], which was further advanced by Defferrard et al. (2016) [8] through ChebNet, utilizing Chebyshev polynomials to approximate convolutions on graphs. Kipf and Welling (2017) [5] simplified these methods by introducing Graph Convolutional Networks (GCNs), achieving state-of-the-art performance while maintaining computational efficiency.

2) Attention mechanisms in GNNs: Veličković et al. (2018) [11] introduced Graph Attention Networks (GATs), employing self-attention mechanisms to dynamically weigh the significance of neighboring nodes, facilitating more flexible feature aggregation. This model has shown particular efficacy in heterogeneous graph settings, enhancing model adaptability.

3) Scalability and efficiency: Addressing scalability issues, Chen et al. (2018) [18] proposed FastGCN, which employs importance sampling to minimize computational complexity. Huang et al. (2018) [19] further refined this with Adaptive Sampling GCN (AS-GCN), improving performance for large-scale graphs.

4) Handling heterogeneity: The Heterogeneous Graph Attention Network (HAN), proposed by Wang et al. (2019) [20], tackles challenges in heterogeneous graph data, demonstrating notable improvements in tasks like author identification and paper classification.

5) Mitigating over-smoothing: The issue of over-smoothing in deeper GNN architectures was addressed in study [21] with DeepGCNs, which integrated residual connections and dilated convolutions, facilitating the training of much deeper networks.

6) Data augmentation for graphs: Zhao et al. (2021) [22] proposed GraphSMOTE, a modification of the SMOTE algorithm tailored for graph-structured data to tackle class imbalance. Ding et al. (2022) [23] introduced a comprehensive data augmentation framework specifically for GNNs, showing significant improvements in performance across multiple tasks. Additionally, Zhang et al. (2020) [13] highlighted the effectiveness of PEGASUS for paraphrasing and synonym replacement in augmenting text data for GNN applications.

7) Hyperparameter optimization: Recent work by (Akiba et al., 2019) [16] emphasized automated hyperparameter tuning for GNNs, achieving notable enhancements in model performance across various datasets. R. Kohavi (1995) [17] further explored the application of k-fold cross-validation, emphasizing robust model evaluation practices necessary for ensuring generalizability in GNN contexts.

To overcome previously cited limitations and build upon these architectural advances. Our work focuses on combining architectural strength with data-centric approaches.

III. Methodology

A. Graph Convolutional Networks

GCN is a fundamental component of our approach to text classification using graph-based data. This section outlines the GCN architecture, explains the mathematical foundations, and highlights its role in our methodology.

1) Architecture overview: The GCN [1] operates like a layer that receives a set of input vectors representing the nodes, in conjunction with the graph structure, producing a new set of node mappings. In the context of a directed graph, denoted:

$$G = (V, E) \tag{1}$$

where V represents the set of vertices (nodes), while E the set of edges (see Fig. 1).

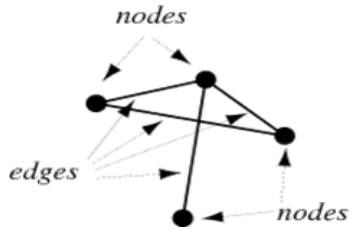


Fig. 1. Structure of a directed graph illustrating node-edge relationships in Cora dataset.

The purpose is to glean insights from the graph by the features function. This function uses as its input:

A description of the features xi for each node I, summed in a feature matrix N × D X (N: nodes, D: input features).

A graph structure description embodied in a matrix form; specifically, as an adjacency matrix A (or a function of A).

The output Z (an N × F matrix, F: output features at each node). Every layer unfurls as a unique non-linear function:

$$H^{(l+1)} = f(H^{(l)}, A) \tag{2}$$

H(0) initialized as X , as for I: the layers number. The patterns vary solely in choosing and parameterizing the function, so it can capture the transformation performed on the node representations in each layer, and A is a matrix attained through the normalization process applied to the adjacency matrix of the graph G.

2) Message passing: As its name suggests, it refers to the neighbor of the destination node. The general concept is to exchange message (insights) constantly with its neighbors to reach a steady balance. Fig. 2 exhibits an illustration of the message passing process that consists of two key steps:

a) Aggregation: each node transmits feature information (messages) to the target node from its neighbors.

b) Update: the functionality of each node according to the "Message" received to form an embedded representation (Embedded).

Message passing guarantees consistent dimensions for all representations, allowing easy further processing. A final integrated representation is achieved after many layers of message passing, capturing the complex relationships within the graph structure. Eq. (2) gives the rule for passing messages from a GCN to a layer for an undirected graph, G.

$$H = \sigma(AXW) \tag{3}$$

H represents the updated node representations, X the matrix that wraps the characteristics of the nodes, where W embodies weight parameters, and σ functions as the non-linear activation coordinator. Meanwhile, A appears as a matrix built using a process of normalization of the rows of the adjacency matrix of the complex graph to add richer feature representation of the model. The following Fig. 2 exhibits the message passing mechanism throw layers.



Fig. 2. Message passing mechanism.

In summary individual nodes aggregate information from their interconnected nodes, applying an aggregation function, which ensures that all representations result in equal size. Next, the vector undergoes a transformation through a dense neural network layer, together with a non-linear activation function to refine the vector's mapping. These steps are repeated according to how many layers are in the model. In particular, GCNs feature distinct representations at each layer:

At the zeroth layer, the representation is aligned with the node's specific characteristics.

At k layer, the node's mapping is calculated by going through its neighbors, aggregating their mappings from the preceding layer (k-1), averaging them and subjecting the result to a transformation via a parameter matrix. This process extends to include the node's proprietary messages from k-1 onwards. This value is then subjected to a non-linear function, such as ReLU in our case.

Eventually, when the node's mappings pass through the transformations in the hidden layers conclusive integration is achieved. In summary, GCNs work via iterative message transmission among nodes, allowing them to leverage representations capturing their structural context, both local and global, in the graph.

B. Graph Attention Networks

GATs offer an excellent opportunity for the development of graphical neural networks. With Graph Convolutional Networks (GCN), every neighbor has the same importance. Yet, some nodes are more essential than others. Graph Attention Networks solve this issue with the self-attention mechanism that regards significance of individual neighbors, an attention mechanism granting a weight coefficient per connection. The following section illustrates GAT architecture, provides an overview of its mathematical background, and outlines its contribution to our methodology.

1) Architecture Overview: GAT [11] architecture consists of various key elements, designed to contribute to its performance in graph-based neural networks:

Self-attention mechanism lies at the core of GAT, this mechanism assigns a weighting factor (attention score) to each connection, allowing the nodes to target the most relevant neighbors. GAT applies mathematical equations to aggregate node characteristics based on attention, the embedding of node 1 is calculated as the equation below shows, W: shared weight matrix:

$$h_1 = \alpha_{11} w\varkappa_1 + \alpha_{12} w\varkappa_2 + \alpha_{13} w\varkappa_3 + \alpha_{14} w\varkappa_4 \tag{4}$$

α_{ij} the attention scores across nodes i and j.

The calculation of these attention scores proceeds through three stages:

a) Linear transformation: To calculate the importance of each connection, pairs of hidden vectors are needed (see Fig. 3). A straightforward approach to forming these pairs involves concatenating the vectors of the respective nodes. Moving on from this step, a new linear transformation is applied using trainable attention vector a:

$$a_{ij} = a^t [W\, h_i || \, W h_j] \tag{5}$$

Where:

$W{\cdot}h_i$ and $W{\cdot}h_j$ are the transformed node features,

‖ Indicates vector concatenation,

a^t is a transposed attention vector (learned during training).

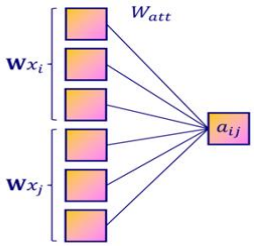


Fig. 3. Linear transformation process.

b) Activation function: Since the aim is to build a neural network, the activation is the second stage. In this context, the LeakyReLU function is added (see Fig. 4).

$$e'_{ij=} LeakyReLu(e_{ij}) \tag{6}$$

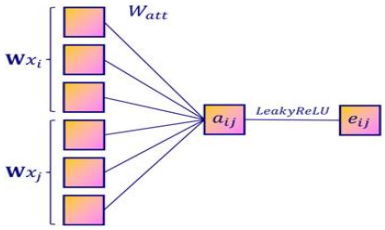


Fig. 4. Computing attention coefficient.

c) SoftMax normalization: The output of the neural network does not undergo normalization. To assess the relative relevance of node 2 to node 1 to node 3 ($\alpha_{12} > \alpha_{13}$), the same scale should be shared. A frequently employed method in neural networks involves the SoftMax function (see Fig. 5). The equation below shows how it is applied to every neighboring node:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ij})} \quad (7)$$

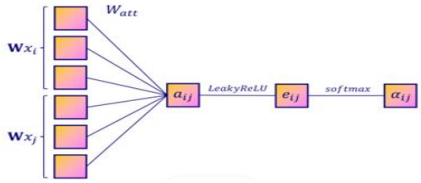


Fig. 5. SoftMax normalization.

We can calculate every α_{ij}. However, self-attention can exhibit instabilities. To enhance robustness, multi-head attention was introduced into the architecture of the transformer.

d) Multi-head attention: The multi-head attention (see Fig. 6) involves repeating three identical steps repeatedly, with the aim of calculating the average or recombine outputs. Here, rather than obtaining a unique hidden h_1, a separate hidden vector h_1^k is generated for each attention head. We can then apply one of two schemes:

Average: add together each of the h_i^k. Normalize via the total number n of heads of attention.

$$h_i = \frac{1}{n} \sum_{k=1}^{n} h_i^k \quad (8)$$

Concatenation: concatenate the different h_i^k.

$$h_i = ||_{k=1}^{n} h_i^k \quad (9)$$

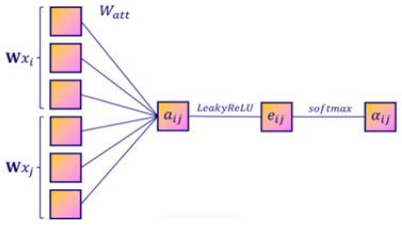


Fig. 6. Multi-head attention.

In summary, attention-graph networks define self-attention as a mechanism that allows nodes nearby to be accorded levels of importance. Mathematically, this self-attention is characterized by linear transformations, activation functions, and SoftMax normalization. Multi-headed attention adds to the model's performance by considering multiple perspectives during the aggregation process.

C. Data Preprocessing and Augmentation

1) Original Cora dataset description: The Cora dataset published by Andrew McCallum and his research group at the University of Massachusetts Amherst [12] (see Fig. 7) as a compressed zip file containing .ps files containing metadata, abstracts and citations for 2,708 academic publications distributed across seven classes. The accompanying citation grid comprises 5,429 links. Cora's established benchmark status in graph neural network literature enables meaningful performance contextualization, while its citation network structure provides the graph topology essential for GCN and GAT evaluation. We selected Cora for three reasons aligned with our objectives:

- Availability of original PostScript text to support sophisticated, meaning-preserving augmentation.
- Moderate size (2 708 nodes) that allows comprehensive k-fold validation without prohibitive compute demands.
- Pronounced class imbalance (Neural Networks: 30 %, Rule Learning: 6 %), making it an ideal benchmark for evaluating imbalance-aware graph learning strategies.

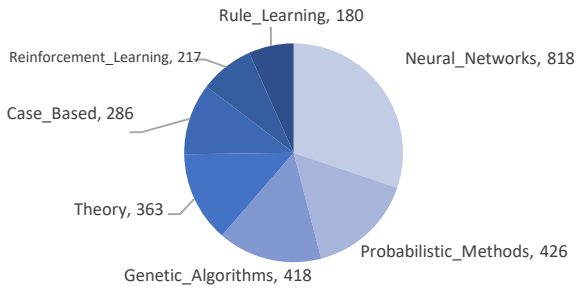


Fig. 7. Original Cora dataset classes distribution.

This dataset was subjected to a comprehensive preprocessing and augmentation protocol to enhance its utility and mitigate inherent limitations and was systematically expanded to a more robust dataset of 17,780 samples through a multifaceted augmentation strategy.

e) Semantic-preserving multi-modal augmentation framework: Theoretical motivation: data expansion traditional methods regularly introduce semantic leakage in which synthetic input samples deviate increasingly far from the domain's original pattern. This issue is also pronounced in datasets like Cora that are already heavily pre-processed to a simplified representation removing much of its contextual richness that GAT and GCN are meant to capture. To solve this problem, instead of developing a costly new augmentation algorithm we propose a combination of existing methods with a semantic control level mechanism following this pipeline:

- Preprocessing Cora original raw dataset: The initial phase was data preprocessing. It involved a meticulous examination and refinement of the raw Cora dataset, originally distributed in PostScript (ps) file format. This process was executed through a series of sophisticated steps to preserve data quality, relevance and semantic integrity.

Content Extraction and Metadata Removal: we isolate essential textual content directly from the PostScript file one by one using a custom rule-based string parser we implemented in

Python while systematically removing extraneous metadata, such as (e.g. author names and institutional affiliations, links, etc.). Abstracts were retained, as they serve as the primary textual representation of each paper.

Text Cleaning: A multi-tiered protocol to text cleaning has been used taking into consideration benchmark practices to not roughly eliminate semantic senses: custom regular expressions were employed to remove non-alphanumeric characters, normalize whitespace, and standardize formatting.

- Data augmentation Pipeline (Methodologies)

$$\text{Let } D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \quad (10)$$

D represent the original dataset, where x_i is the feature vector and y_i is the label that correspond to the feature in question. The generated D' is the result of the augmentation process such that |D'| > |D| while preserving the label condition distribution $P(x', y') \approx P(x,y)$.

PEGASUS-based Textual Transformation: We employed the PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) model, specifically fine-tuned for paraphrasing [13], as the primary method for generating diverse textual variations. This approach allows us to create semantically equivalent but linguistically diverse versions of the original text:

$$P = f_{\theta_p}(x) \quad (11)$$

where P is the paraphrased text, x is the paper textual content, and f_(θ_p)

is the PEGASUS paraphrasing function with parameters. This technique effectively doubles the dataset size while introducing linguistic variations that can improve model robustness.

Subject-Aware Lexical Substitution via Synonym Replacement: We used synonym replacement algorithm [13] while enforcing controlled lexical variations through filtering according to subject domain, thereby maintaining thematic coherence while expanding vocabulary:

$$x_{\text{aug}} = h(x_{\text{orig}}, \lambda) \quad (12)$$

Where x_aug is the augmented paper content, x_orig is the original or previously augmented content, h is the synonym replacement function, and λ is a parameter controlling the degree of replacement.

Conditional restrictions are applied so that only substitutions that met semantic and label constraint were retained (Samples exceeding the WMD threshold were rejected)

$$\text{WMD}(x_{\text{orig}}, x_{\text{aug}}) \leq \tau_{\text{wmd}} = 0.35 \quad (13)$$

Class-balanced custom Generation Pattern**:** we devised a sophisticated syntactic generation framework characterized by a meticulously curated lexicon of action-oriented verbs and descriptive adjectives. Enabling the formulation of new sentences that are both semantically rich and consistent with the underlying class distributions of the original text, this pattern was applied specifically to limited set of isolated nodes with no connection in the network and particularly those with the shortest text content that lack sufficient presentation.

$$N = g(X, V) \quad (14)$$

where N is the new synthetic sample, X is the original text, V encapsulates our curated vocabulary and advanced terminological construct, and g is our custom generation function that generates suitable terms from V while respect part-of-speech to maintain correctness. This combinatorial approach allows for controlled growth of even smaller contributor nodes in the dataset with semantic check through the threshold τ_{wmd}.

- Parameter Sensitivity: The choice of crucial parameters in our augmentation pipeline has an important impact on performance. While a lower Learning rate (Lr) produces nearly duplicate paraphrases with little advantage, a higher PEGASUS (Lr) introduces semantic drift that reduces classification accuracy. Raising the synonym rate beyond its optimal point disrupts domain-specific terminology, damaging coherence, while too low a rate limits new lexical variety. Temperature values above 0.8 produce semantically inconsistent text that misleads the classifier, and values below 0.6 result in repetitive, uninformative samples. While lowering the WMD threshold allows for excessive semantic divergence, raising it over-filters advantageous variations. These findings underscore how sensitive the augmentation process is to parameter choices and highlight the importance of systematic tuning, as even minor deviations from optimal settings can significantly degrade semantic quality.

- Methodological Validation: Semantic Preservation**:** For each newly augmented node inherits the original node's edges to preserve citation relationships so that the structural semantics of the graph remain intact after augmentation. To support the integrity of the augmented dataset, we implement: Automated intra-category semantic similarity checks between original and augmented document embeddings for each category. We computed the Word Mover's Distance (WMD) [15] and set threshold for strict filtering; this validation mechanism prevents the semantic drift commonly seen in synthetic data generation. In addition to analyzing repetitiveness in sentence structure and phrase.

Class Balance: The data were augmented to realize equal class distributions (approximately 14% each) drops the bias in the original Cora, where Neural Networks represented 30% of samples while Rule Learning only 6%, Neural Networks now present more equitable 16% reduced dominance. This balance guarantees fair model evaluation.

Computational Efficiency: compared to the improvement of ensemble methods or attention mechanism. This pipeline provides better graph learning preserved semantically, enabling models to achieve better performance without architectural modifications, reducing computational overhead and more accessible to any individual.

IV. Experimental Results

A. Dataset Description

In this study, we have created a customized augmented dataset from Cora [12] using various methods, each with critical parameter settings. The Pegasus model utilized a learning rate of 0.01 selected following several tests and monitoring continuously steady convergence without overshooting, a maximum input length of 512 tokens, a beam search size of 5, and was trained for 5 epochs to ensure effective learning. For synonym replacement, we set a replacement probability of 0.3 and a maximum of 3 synonyms per word, maintaining semantic integrity while promoting variability. In the text generation phase, we used a temperature of 0.7 for balanced creativity, top-k sampling with k=50 to restrict predictions to the most probable words, and to reduce repetitive phrases we used a repetition penalty of 1.2. These carefully chosen parameters, and many others significantly preserved the quality and diversity of the generated data.

The New customized Cora dataset (see Fig. 8) cumulative generated by the augmentation techniques resulted in the important expansion of our dataset from the original 2,708 samples to 17,780 samples, an increase in the factor of approximately 6,57x. This expansion not only increased the training data volume but also preserved the diversity of linguistic variations and synthetic samples, all while maintaining semantic coherence and domain.

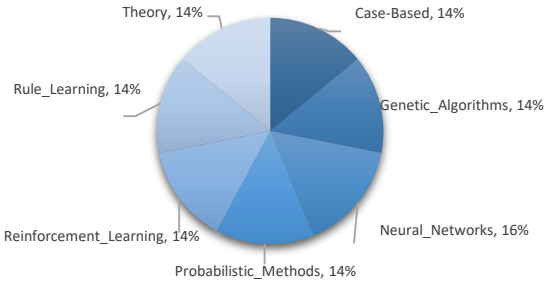


Fig. 8. Customized dataset classes distribution.

The quality and relevance of the augmented samples shows a similarity threshold of 0,8 was maintained. Additionally, WMD calculated scores intra classes (see Table I) values around 0.3 to 0.34 suggest moderate similarity between original and augmented texts within each category introducing meaningful variation.

TABLE I. WMD Scores

Category	WMD scores
Case-Based	0.3389
Genetic Algorithms	0.3049
Probabilistic Methods	0.3044
Reinforcement Learning	0.3309
Rule Learning	0.3283
Theory	0.2773

B. Hyperparameter Tuning Configuration

We adopted to conduct Naive Bayesian hyperparameters tuning with adopted Optuna [16], because of its empirical ability to efficiently navigate high-dimensional search spaces and typically generalize better, even in contexts like ours involving data imbalance and augmentation where stability is critical. The full set of optimized hyperparameter configurations for both GCN and GAT models is summarized in Table II, which lists ranges for learning rates, dropout, and architectural elements.

This probabilistic approach enabled us to model the relationship between hyperparameters and model performance. In addition to ensuring the robustness and generalizability of our hyperparameter configurations, we employed a k-fold cross-validation strategy. This approach partitioned our augmented dataset into k subsets, allowing for multiple training and validation cycles. The cross-validation procedure not only provided a more reliable estimate of model performance but also mitigated the risk of overfitting to specific data partition.

TABLE II. Optimized Hyperparameter Configurations

Elements	Values	Model
Optimization framework	100-400 trials	GCN-GAT
Hidden features	8 to 128	
Activation	["relu", "relu"], ["relu", "softmax"], ["elu", "softmax"]	
Dropout	0.1 to 0.8	
Learning rate	1e-5 to 1e-1 (Log-uniform)	
Weight decay	1e-6 to 1e-2 (Log-uniform)	
Cross-Validation	5-fold cross-validation	
Early Stopping	Patience of 20 epochs, monitored on validation set's F1 score	
Attn_heads	1 to 16	GAT
attn_dropout	0.2 to 0.8	

The hyperparameter tuning process aimed to maximize the mean cross-validated F1 score. The best-performing hyperparameters for each model were selected based on this metric. We then compared the optimized GAT and GCN models to determine which architecture was more suitable for our specific graph classification task.

C. Evaluation Metrics

Classification models are evaluated using well-established metrics to quantify their performance and effectiveness. These metrics are essential in determining how well a model can predict outcomes based on the given data. Below are the key metrics used in evaluating GCN and GAT classification [5] [11]:

True Positives (TP): Instances where the model correctly predicts positive outcomes (i.e., both the actual and predicted results are positive).

True Negatives (TN): Instances where the model correctly predicts negative outcomes (i.e., both the actual and predicted results are negative).

False Positives (FP): Occurrences where the model predicts a positive result, but the actual result is negative (also known as a Type I error).

False Negatives (FN): Occurrences where the model predicts a negative result, but the actual result is positive (also known as a Type II error).

Accuracy: Conversely, false negatives occur if the model predicted a negative result, while the real result proved positive**.**

$$\text{Accuracy} = \frac{\text{TP}_\text{i}+\text{TN}_\text{i}}{\text{TP}_\text{i}+\text{TN}_\text{i}+\text{FP}_\text{i}+\text{FN}_\text{i}} \quad (15)$$

Loss: Loss measures the discrepancy between predicted values and actual values. Cross-entropy is a common loss function in deep neural networks. It quantifies how well the predicted probabilities match the actual class labels.

$$\textbf{C}\text{rossEntropy} = -\sum_{i=1}^{n} log\left(P_{i,i}\right)y_{i,j} \quad (16)$$

where yi,j is the true value, i.e. 1 if sample I is in class j and 0 otherwise. Pi,j the likelihood forecast via the model that sample I is part of class j.

Precision: is the ratio between the True Positives and all the Positives.

$$\text{Precision} = \frac{\text{TP}_\text{i}}{\text{TP}_\text{i}+\text{FP}_\text{I}} \quad (17)$$

Recall: the extent to which our model correctly identifies true positives.

$$\text{Recall} = \frac{\text{TP}_\text{i}}{\text{TP}_\text{i}+\text{FN}_\text{I}} \quad (18)$$

F1 Score is the harmonic mean of Precision and Recall. It provides a balanced measure that considers both false positives and false negatives, making it particularly useful when dealing with imbalanced datasets. The F1 Score ranges from 0 to 1, with 1 indicating perfect precision and recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

D. Results and Discussion

1) Model settings: In this experiment we adapted across all models the same following configurations:

a) Data preparation: The textual preprocessed data was transformed into TF-IDF vectors, with a maximum of 1000 features for each record. Class labels were encoded using LabelEncoder, ensuring numerical representation for classification. We created an adjacency matrix from citation data, representing relationships between papers, and normalized it to ensure proper scaling in the graph model.

Stratified Data Splitting We employed stratified k-fold cross-validation with 5 splits to ensure balanced class distributions across training and validation sets. A random state with a fixed seed of 42 was used for reproducibility throughout the data splitting and model training. For each fold, data was split into training and validation sets, without using a k-fold throught cross-validation.

Following comprehensive hyperparameter tuning and testing, we identified a set of configurations for the foundational models as detailed in Table III. These configurations were meticulously selected to enhance the performance of the respective models.

The results yielded by the individual models, GCN and GAT, were rigorously analyzed. Fig. 9 depicts the outcomes of numerous testing and fine-tuning iterations, providing a summary of the average performance metrics from the 5 folds that are broken down per model, and summarized in Table IV including key metrics: validation accuracy, precision, recall, and F1 score.

The results of our analysis reveal that GCN is consistently better than GAT across all evaluated metrics. Specifically, GCN achieves a validation accuracy of 95.42%, which is a higher overall correct classification rate than GAT's 93.46%. In terms of precision, GCN achieves a score of 95.45%, compared to GAT's 93.50%, suggesting a lower false positive rate for GCN. In addition, GCN's recall of 95.42% above GAT's 93.46%. This shows that GCN's improved capability in identifying true positives. The F1 score shows GCN at 95.42%, while GAT scores 93.47% which reflects the balance between precision and recall.

TABLE III. GRAPH CONVOLUTIONAL NETWORK AND GRAPH ATTENTION NETWORK SETTINGS

Model	GCN Instance Model	GCN Average Model	GAT Instance Model	GAT Average Model
Model architecture	layer_sizes =108 activated by ReLU	layer_sizes = 116 activated by ReLU	layer_sizes = 106 activated by ELU	layer_sizes = 106 activated by ELU
Epochs	200	200	200	200
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	0.04401160472	0.06507184668	0.00066108710 step_size=50, gamma=0.5	0.000661087 Step_size=50, gamma=0.5
Early Stopping	Stagnant F1 score validation for 200 epochs	Stagnant F1 score validation for 200 epochs	Stagnant F1 score validation for 200 epochs	Stagnant F1 score validation for 200 epochs
Patience	20	20	20	20
Activation	ELU (hidden layers), Log Softmax (output)	ELU (hidden layers), Log Softmax (output)	ELU (hidden layers), Log Softmax (output)	ELU (hidden layers), Log Softmax (output)
Dropout Rate	0.44934028857 (45%)	0.53708280317 (50%)	0.29109068673 (30%)	0.2910906867 (30%)
Weight Decay	3.602702935945202e-05	3.602702935945202e-05	0.00010625787460334993	0.00010625787460334993
Attention Heads	None	None	16	16

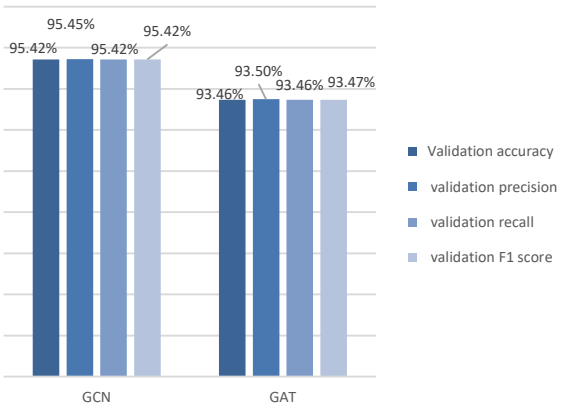


Fig. 9. Performance metrics comparison.

The GCN model performed better than the GAT model across all evaluation measures, with accuracy being approximately 1.96% higher (95.42% vs. 93.46%) and statistical significance with the paired t-tests showing ($p < 0.05$, $t = 2.89$, $df = 4$) also shows a 95% confidence interval: (94.77%, 96.07%) vs. (92.83%, 94.08%). The strong performance of both models, with low standard deviations, indicating stable learning over different data distributions.

TABLE IV. COMPARATIVE PERFORMANCE ANALYSIS OF GCN AND GAT MODELS ACROSS KEY EVALUATION METRICS

Model	Folds	Validation accuracy	Validation precision	Validation recall	Validation F1 score
GCN	F1	95,05%	95,06%	95,05%	95,04%
	F2	96,26%	96,27%	96,26%	96,26%
	F3	95,02%	95,09%	95,02%	95,03%
	F4	95,61%	95,61%	95,61%	95,60%
	F5	95,16%	95,23%	95,16%	95,18%
GAT	F1	92,69%	92,73%	92,69%	92,70%
	F2	93,64%	93,69%	93,64%	93,65%
	F3	93,25%	93,30%	93,25%	93,27%
	F4	93,98%	94,01%	93,98%	93,99%
	F5	93,73%	93,75%	93,73%	93,74%

Based on cross-validation findings, detailed performance showed that the GCN maintains an accuracy of 95.42% (±0.47%) with a precision of 95.45% (±0.45%), a recall of 95.42% (±0.47%) and an F1 score of 95.42% (±0.47%). While GAT also manifests the corresponding performance levels of 93.46% (±0.45%), 93.50% (±0.45%), 93.46% (±0.45%) and 93.47% (±0.45%). We kept track of validation loss across folds to determine model convergence and confidence level were balanced. GCN showed a consistently lower average validation loss of ~0.18 compared to GAT's ~0.21. Both models demonstrated small standard deviations, indicating a good level of stability as well as reproducibility across different data partitions.

However, there is a variation between the study datasets and data refinement methodologies, confronted with these limitations the establishment of internal reference measures was undertaken. To this end, we trained GCN and GAT on the original Cora dataset (2,708 samples) using identical architecture but without augmentation. The resulting accuracy rates were 83.39% and 82.24%, respectively, aligning with existing literature on the subject such as [5][11], where standard Cora benchmarks rarely go above 85% in the absence of architectural refinement. While this comparison does not represent a strictly controlled ablation, it is still useful to understand the potential impact of our semantic preservation, class balanced augmentation pipeline. The augmented and optimized models exhibited an accuracy rate of over 95%, these results show how important it is to preserve the semantics of the dataset during the engineering process. The findings of this study provide an indicative—but not strictly controlled—comparison, which is developed in the limitations section.

- Class-wise Performance Analysis: To further investigate per-class performance, we computed F1-scores for each class across all folds (see Table V).

TABLE V. AVERAGED F1-SCORES RESULTS ACROSS THE FIVE FOLDS FOR EACH CLASS WITH STANDARDS DEVIATIONS

Class	*GCN F1-Score*	*GAT F1-Score*	*(GCN-GAT)*	*Support*
Case-Based	0.9575 ± 0.0073	0.9378 ± 0.0029	+0.0197	500
Genetic_Algorithms	0.9726 ± 0.0037	0.9584 ± 0.0039	+0.0142	500
Neural_Networks	0.9194 ± 0.0097	0.8898 ± 0.0115	+0.0296	556
Probabilistic_Methods	0.9504 ± 0.0081	0.9284 ± 0.0054	+0.0220	500
Reinforcement_Learning	0.9733 ± 0.0057	0.9608 ± 0.0135	+0.0125	500
Rule_Learning	0.9737 ± 0.0068	0.9590 ± 0.0081	+0.0147	500
Theory	0.9364 ± 0.0096	0.9140 ± 0.0086	+0.0224	500
Weighted Avg	0.9542 ± 0.0047	0.9347 ± 0.0045	+0.0195	3556

As detailed in Table V the analyses by class highlight notable aspects. Overall, all seven classes are consistently well managed by the GCN model, with a performance variance between 1.25% (Reinforcement Learning) and 2.96% (Neural Networks). The Neural Networks category shows the most difficulties in classification with both models, particularly due to the conceptual overlap and semantic ambiguity inherent in the literature on neural networks affecting many areas of artificial intelligence. Probabilistic Methods and Theory exhibit relatively significant performance differences (2.20% and 2.24%), suggesting that GCN's aggregation process is effective in capturing the particularities of these domains. However, the relatively smaller difference between the two models in reinforcement learning (1.25%) clearly illustrates how well the two models deal with this specific domain.

The plots shown in Fig. 10 and Fig. 11 provide an overview of the learning history that gives a further indication of the convergence and stability of GCN and GAT performance over the course of the learning process with second fold as the representative fold. We can observe that the learning process for both models GAT and GCN presents some notable differences, particularly during the early training stages. GAT, for instance, is more unstable at the beginning, especially in the initial starting epochs. This instability likely comes from the way attention mechanism that GAT uses, as it requires more

time to adjust and optimize the importance of edges within the graph. Learning how to prioritize different node connections adds more complexity, which in turn causes performance to initially fluctuate.

How GAT manages isolated nodes is one of its main challenges. While GCN works best in treating all neighboring nodes equally, GAT's effectiveness is more dependent on the topology of the graph. The presence of isolated nodes, which lack neighboring information, makes it harder for GAT to fully benefit from its attention mechanism. This has likely contributed to the early performance issues and slower convergence compared to GCN.

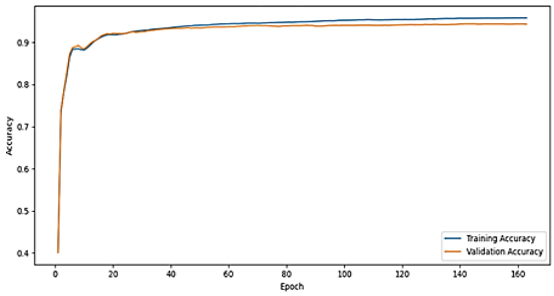


Fig. 10. GAT Fold 2 training and validation accuracy fit history.

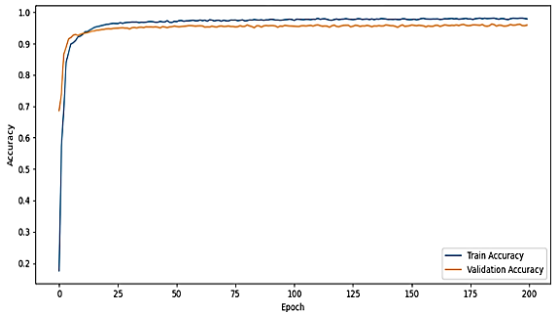


Fig. 11. GCN Fold 2 training and validation accuracy fit history.

Despite these challenges at the beginning, both models eventually ended up achieving high-performance levels. While it takes longer to converge for GAT, that frequently matches or even outperforms GCN in the end. This suggests that although GAT works in a more complex optimization environment, its ability to learn which edges are most important can offer important advantages as training continues.

Overall, the consistent results across the five cross-validation folds show that both models are robust, especially GCN, which was less variable across different data splits. The slight differences in GAT's performance could be due to its sensitivity to specific graph structures within each fold. Compared to baseline models, both GCN and GAT perform very well on the expanded dataset, suggesting their architecture's strength in handling complex graph data. This analysis underlines the unique strengths of each model and the importance of giving GAT sufficient training time to fully optimize its performance, particularly in scenarios with diverse graph connectivity.

V. Conclusion

This study represents a notable advancement in graph-based node classification, effectively addressing key challenges related to data quality in preserving semantic aspect, class imbalance, and model performance generalization constraints. Our approach enhanced the Cora dataset by extracting content from its original PostScript files and applying a systematic implementation of a tripartite augmentation framework combination of benchmark techniques: a combination of textual transformation PEGASUS-based, subject aware synonym replacement and customized class-balanced generation pattern function on isolated nodes to expand the dataset to 17,780 nodes, achieving a 6.57x scaling factor while maintaining semantic fidelity as evidenced by Word Mover's Distance scores ranging from 0.27 to 0.34. In addition the use of Optuna-driven hyperparameter optimization using Naïve Bayesian with k-fold cross-validation, to make sure achieving optimal model performance. Both the GCN and GAT models performed effectively, achieving accuracy of 95.42% and 93.46% respectively. This analysis highlights how is consistent is GCN alongside GAT more intricate learning pattern, where early fluctuation gave way to competitive results. These results show the potential of attention mechanisms to capture complex graph dynamics with sufficient training. The models' resilience, validated across five folds, further demonstrates motivation of our approach. Importantly, this work illustrates how advanced data augmentation and optimization techniques can boost the performance and scalability of graph-based models by preserving important semantic aspect that can be lost in rough engineering, also shows the ability of these models on larger datasets without compromising accuracy.

Despite overall good classification performance, there are some limitations that should be acknowledged, particularly in distinguishing between overlapping classes such as neural networks. The attention mechanism of GAT also introduced computational costs without corresponding gains. In conclusion, results demonstrate that performance obtains equivalent to architectural complexity alone can be obtained through strategic dataset augmentation, this principle provides a cost-effective alternative, making advanced graph neural networks accessible to researchers with limited computational resources.

A. Limitations and Future Work

While our approach has shown good results as discussed earlier in the previous section, we acknowledge different limitations that present opportunities for future improving:

1) Domain specificity: Our validation focuses on academic paper classification that the Cora dataset provides. Future work should explore the practical applicability across variety of text domains and languages.

2) Comparative evaluation: Direct comparison with existing methods is challenging due to the characteristics of the new dataset and the difference in engineering approaches. Future work should establish a more standardized evaluation protocol.

REFERENCES

[1] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[2] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, and R. Ward, "Semantic modelling with long-short-term memory for information retrieval," arXiv preprint arXiv:1412.6629, 2014.

[3] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, and Q. Yang, "Large-scale hierarchical text classification with recursively regularized deep graph-cnn," in Proc. 2018 World Wide Web Conf., 2018, pp. 1063-1072.

[4] H. Cai, V. W. Zheng, and K. C. C. Chang, "A comprehensive survey of graph embedding: problems, techniques, and applications," IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1616-1637, Sept. 2018.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with Graph Convolutional Networks," arXiv preprint arXiv:1609.02907, 2017.

[6] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," arXiv preprint arXiv:1312.6203, 2014.

[7] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," arXiv preprint arXiv:1506.05163, 2015.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in Advances in Neural Information Processing Systems, vol. 29, 2016.

[9] D. Marcheggiani and I. Titov, "Encoding sentences with graph convolutional networks for semantic role labeling," arXiv preprint arXiv:1703.04826, 2017.

[10] Y. Li, R. Jin, and Y. Luo, "Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs)," J. Amer. Med. Inform. Assoc., vol. 26, no. 3, pp. 262-268, 2019.

[11] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," arXiv preprint arXiv:1710.10903, 2018.

[12] McCallum, A., Rennie, J., & Nigam, K. (2000). Cora Research Paper Classification Dataset. University of Massachusetts Amherst. Available at: https://people.cs.umass.edu/~mccallum/data/cora-classify.tar.gz

[13] J. Zhang, Y. Zhao, M. Saleh, and J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020.

[14] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in Proc. EMNLP IJCNLP, Hong Kong, Nov. 2019, pp. 6382–6388.

[15] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From Word Embeddings to Document Distances," in Proc. 32nd Int. Conf. Machine Learning (ICML), 2015, pp. 957–966.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and Z. Kira, "Optuna: a next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2019.

[17] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI), vol. 2, pp. 1137–1143, 1995.

[18] J. Chen, T. Ma, and C. Xiao, "FastGCN: Fast learning with graph convolutional networks via importance sampling," arXiv preprint arXiv:1801.10247, 2018.

[19] M. Yoon Huang T. Zhang, Y. Rong, J. Huang "Performance-Adaptive sampling towards fast and accurate graph neural networks," in Proc. 27th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD), 2021, pp. 2046–2056.

[20] X. Wang, H. Ji, C. Shi, B. Wang, P. Cui, P. Yu, Y. Ye, "Heterogeneous graph attention network," in Proc. World Wide Web Conf., 2019.

[21] G. Li, M. Müller, G. Qian, C. Delgadillo, A. Abualshour, A. Thabet, B. Ghanem, "DeepGCNs: Making GCNs Go as Deep as CNNs" in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, Oct. 2019, pp. 9266–9275. DOI: 10.1109/ICCV.2019.00936

[22] T. Zhao, X. Zhang, S. Wang, "GraphSMOTE: imbalanced node classification on graphs with graph neural networks" in *Proc. ACM WSDM*, 2021, pp. 833–841. DOI: 10.1145/3437963.3441720 (arXiv:2103.08826)

[23] K. Ding, Z. Xu, H. Tong, and H. Liu, "Data Augmentation for Deep Graph Learning: A Survey," arXiv preprint arXiv:2202.08235, 2022.

Enhancing Banking Data Classification Through Hybrid L2 Regularisation and Early Stopping in Artificial Neural Networks

Khairul Nizam Abd Halim[1, 3], Abdul Syukor Mohamad Jaya[2]*, Fauziah Kasmin[2], Azlan Abdul Aziz[3]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[1]
Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[2]
Fakulti Sains Komputer dan Matematik, Universiti Teknologi MARA, Malaysia[3]

***Abstract*—The demand for robust data-driven classification (DDC) techniques remains critical in banking applications, where accurate and efficient decision-making is paramount. Artificial Neural Networks (ANNs), particularly Multi-Layer Perceptrons (MLPs), are widely used due to their strong learning capabilities. However, their performance often depends on effective hyperparameter tuning and regularisation strategies to avoid overfitting. This study aims to enhance the efficiency of the MLP training process by introducing a hybrid approach that integrates L2 regularisation with Early Stopping (ES) into the hyperparameter tuning procedure. The key contribution lies in embedding both techniques within a grid search framework, thereby streamlining the search for optimal hyperparameters. The proposed method was evaluated using three real-world banking datasets: two related to loan subscription (16 and 20 features) and one concerning credit card default payment (23 features). Experimental results demonstrate that the hybrid approach reduces hyperparameter tuning time by over 90% while achieving high classification performance. Notably, the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) scores of 93.89% and 91.21% were achieved on the loan datasets, and 73.28% on the credit card dataset, surpassing previous benchmarks. These findings highlight the potential of the L2ES hybrid method to improve both the accuracy and computational efficiency of DDC in financial applications.**

***Keywords*—Artificial neural networks; L2 regularisation; early stopping; banking; classification**

I. Introduction

The financial sector's digital transformation has intensified the demand for accurate and efficient data analysis techniques. Financial institutions continuously seek innovative strategies to enhance operational efficiency, particularly in loan processing, risk assessment, and fraud detection. In this context, data-driven classification (DDC) methods have emerged as essential tools for analysing large volumes of banking data to support informed decision-making. Among these, machine learning (ML)-based DDC techniques stand out because they can learn from complex, nonlinear data patterns and improve predictions over time. Their application spans multiple domains, including medicine, agriculture, and manufacturing, with particularly significant adoption in banking for tasks such as loan subscription analysis and credit card default prediction.

Artificial neural networks (ANNs), particularly the multi-layer perceptron (MLP), have gained recognition as robust classifiers within the DDC domain. Numerous studies have demonstrated the superior performance of MLPs in banking-related classification tasks. For example, MLPs have been effectively used to model customer decisions on loan subscriptions and to predict credit card default risks, as summarised in Tables I and II of this study. The robustness and adaptability of MLPs are further evidenced in domains such as bioinformatics. AbdElsalam et al. [1], for instance, significantly improved MLP performance in gene classification tasks by integrating the ADASYN technique to address class imbalance—an approach yielding high sensitivity and accuracy. These developments reaffirm that MLPs, when appropriately optimised, remain highly effective across various domains, including finance.

However, overfitting remains a critical challenge in MLPs, particularly when working with complex or imbalanced datasets. Regularisation techniques, such as L2 regularisation, have been widely adopted to mitigate this issue. Prior research by Aldelemy et al. [2] and Grosicki [3] demonstrates the efficacy of L2 regularisation in MLP-based classification for banking loan datasets. L2 regularisation has also been successfully applied to other ML models, including logistic regression and XGBoost, across both loan and credit card datasets [4]–[6]. Despite these applications, current implementations often overlook the considerable time and effort required for hyperparameter tuning in neural network models—an essential yet burdensome aspect of model optimisation.

This study identifies a gap in the literature: while L2 regularisation has been implemented to reduce overfitting in MLPs, limited attention has been given to streamlining the hyperparameter tuning process. Most existing approaches apply L2 within fixed or manually guided tuning routines, without exploring strategies to expedite this phase effectively. This oversight hinders the deployment of efficient ML systems in time-sensitive financial applications.

To address this, this paper proposes a hybrid approach that combines L2 regularisation with early stopping (ES) within a grid search framework. This technique aims to retain the regularisation benefits of L2 while simultaneously accelerating the hyperparameter tuning process via early termination of

*Corresponding Author.

underperforming configurations. By integrating ES into the grid search routine, the proposed method significantly lowers computational overhead and facilitates faster convergence to optimal settings, thus making the model development process more efficient.

Unlike conventional methods that treat L2 regularisation and hyperparameter optimisation as independent steps, our hybrid strategy embeds both into a unified workflow. This integration is particularly suitable for banking datasets, where rapid model development is critical. The proposed approach not only curbs overfitting but also enhances the efficiency of hyperparameter exploration—an aspect often neglected in previous studies.

TABLE I. LIST OF RESEARCH ON THE APPLICATION OF MLPS FOR PREDICTING LOAN SUBSCRIPTION THROUGH BANK TELEMARKETING CAMPAIGNS

Research	ROC-AUC	Other Metrics	Features
Moro *et al.*, [7]	0.79	-	22
Marinakos and Daskalaki [8]	0.87	Accuracy 79.11%, Precision 74.99%, Recall 81.44%, F1 77.86%	16
Farooqi and Iqbal [9]	0.89	Accuracy 89.83%, Precision 55.50%, Recall 49.00%, F1 52%	20
Ghatasheh *et al*., [10]	-	Accuracy 84.18%, TP 61.4%	16
Panigrahi and Patnaik [11]	-	Accuracy 90.02%	16
Mokrane [12]	-	Accuracy 98.93%, F1 0.95	20
Dutta and Bandyopadhyay [13]	-	Accuracy 88.32%	16
Masturoh *et al*., [14]	0.90	Accuracy 94.27%	20
Aldelemy and Raed A. Abd-Alhameed [2]	0.92	-	16

TABLE II. LIST OF RESEARCH ON THE APPLICATION OF MLPS FOR PREDICTING CREDIT CARD DEFAULT PAYMENTS

Research	ROC-AUC	Other Metrics	Features
Singh and Aggarwal [15]	-	Recall 0.849, Precision 0.866	23
Liu [16]	-	Accuracy 82.27%, F1 0.46	23
Vishwakarma *et al.*, [17]	0.5000	-	23
de Campos Souza and Torres [18]	0.6506	-	23
Almajid [19]	0.7184	Accuracy 76.50%	23
Jiang *et al*., [20]	-	Accuracy 77.35%	23
Shazly and Khodadadi [21]	-	Accuracy 89.45%, Recall 0.9967, Precision 0.6667, F1 0.9288	23
Idrees *et al*., [22]	-	Accuracy 81.96%, Recall 0.820, Precision 0.803	23
Yash *et al*., [23]	-	Accuracy 80.50%	23

Accordingly, this research evaluates the performance of the hybrid L2ES (L2 + Early Stopping) approach in banking data classification. The primary objectives are to (i) assess the effectiveness of integrating L2 regularisation within MLP-based classifiers, (ii) incorporate L2 into the grid search procedure for hyperparameter tuning, and (iii) demonstrate how the combination of L2 and ES within grid search can enhance training efficiency without compromising model accuracy.

The remainder of this paper is structured as follows: Section II introduces the proposed L2ES methodology, detailing the experimental setup and the grid search configuration. Section III presents the empirical results and offers a comparative analysis with conventional approaches. Section IV concludes the paper and discusses directions for future research.

II. METHODOLOGY

A. Materials

The datasets utilized in this study are derived from the banking sector, specifically focusing on loan and default credit card datasets. These datasets are publicly accessible through the UCI Machine Learning Repository. Table III presents a detailed overview of the loan subscription dataset, as documented by Moro et al. [24]. This dataset encompasses 16 features and one label, the latter serving as the target variable. The dataset comprises 45,211 records featuring a mix of data types, including integers, binary and categorical. The binary target variable is imbalanced: 'yes' labels account for 5,289 records, while 'no' labels constitute 39,922.

TABLE III. OVERVIEW OF THE LOAN SUBSCRIPTION DATASET DOCUMENTED BY MORO ET AL. [24]

Num.	Variable Name	Role	Type
1	age	Feature	Integer
2	job	Feature	Categorical
3	marital	Feature	Categorical
4	education	Feature	Categorical
5	default	Feature	Binary
6	balance	Feature	Integer
7	housing	Feature	Binary
8	loan	Feature	Binary
9	contact	Feature	Categorical
10	day_of_week	Feature	Date
11	month	Feature	Date
12	duration	Feature	Integer
13	campaign	Feature	Integer
14	pdays	Feature	Integer
15	previous	Feature	Integer
16	poutcome	Feature	Categorical
17	y	Target	Binary

Table IV outlines the characteristics of another loan subscription dataset, also studied by Moro et al. [7]. This contains 20 features and one label. It includes a mix of integer, float, and categorical data types for its features. Like the first, it exhibits a binary target with an imbalanced distribution among

its 41,188 records: 4,640 are labelled 'yes', and 36,548 are labelled 'no'.

Table V details the dataset concerning bank credit card default payments sourced from Yeh and Lien [25]. This dataset comprises 23 features and one label, encompassing 30,000 records. The features are of integer data type, while the label is binary. The dataset exhibits label imbalance, with 6,636 records labelled '1' (indicating default) and 23,364 records labelled '0' (indicating no default).

TABLE IV. CHARACTERISTICS OF THE SECOND LOAN SUBSCRIPTION DATASET STUDIED BY MORO ET AL. [7]

Num.	Variable Name	Role	Type
1	age	Feature	Integer
2	job	Feature	Categorical
3	marital	Feature	Categorical
4	education	Feature	Categorical
5	credit	Feature	Categorical
6	housing	Feature	Categorical
7	loan	Feature	Categorical
8	contact	Feature	Categorical
9	month	Feature	Categorical
10	day of week	Feature	Categorical
11	duration	Feature	Integer
12	campaign	Feature	Integer
13	pdays	Feature	Integer
14	previous	Feature	Integer
15	poutcome	Feature	Categorical
16	emp.var.rate	Feature	Float
17	cons.price.idx	Feature	Float
18	cons.conf.idx	Feature	Float
19	euribor3m	Feature	Float
20	nr.employed	Feature	Integer
21	y	Target	Binary

TABLE V. DATASET ON BANK CREDIT CARD DEFAULT PAYMENTS SOURCED FROM YEH AND LIEN [25]

Num.	Variable Name	Role	Type
1	X1: limit	Feature	Integer
2	X2: gender	Feature	Integer
3	X3: education	Feature	Integer
4	X4: marital	Feature	Integer
5	X5: age	Feature	Integer
6	X6: past payment	Feature	Integer
7	X7: past payment	Feature	Integer
8	X8: past payment	Feature	Integer
9	X9: past payment	Feature	Integer
10	X10: past payment	Feature	Integer
11	X11: past payment	Feature	Integer
12	X12: bill statement	Feature	Integer
13	X13: bill statement	Feature	Integer
14	X14: bill statement	Feature	Integer
15	X15: bill statement	Feature	Integer
16	X16: bill statement	Feature	Integer
17	X17: bill statement	Feature	Integer
18	X18: previous payment	Feature	Integer
19	X19: previous payment	Feature	Integer
20	X20: previous payment	Feature	Integer
21	X21: previous payment	Feature	Integer
22	X22: previous payment	Feature	Integer
23	X23: previous payment	Feature	Integer
24	y	Target	Binary

B. Preparation

The experimental setup adhered to past and present research's standard data-driven classification methodologies. Specifically, this study adopted the methodology illustrated in Fig. 1, adapted from A. Almajid [19]. All experiments were conducted within the Google Colab server environment, with relevant specifications for 2023 to 2025. The MLP model development commenced with a data cleaning phase, employing mode imputation to address missing data. Subsequently, the dataset was split using an 80:20 ratio into training and test sets, with the training set used to construct the MLP model. Before model creation, the training set transformed one-hot encoding for categorical features and standard scalar scaling for numerical features. The MLP classifier was selected as the machine learning algorithm to fulfil the primary objective of this paper, which is to augment the performance of the MLP classifier. The baseline model incorporated L2 regularisation to mitigate overfitting during the hyperparameter tuning process via Grid Search. Conversely, the innovative model combined L2 regularisation and ES to address overfitting concerns. The rationale behind this hybridisation was the anticipation of improved MLP classifier performance.

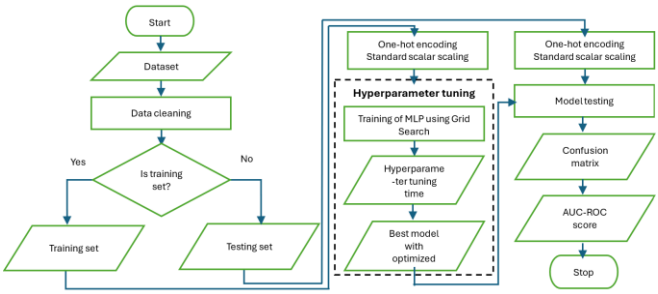


Fig. 1. Methodology for MLP model development and experimentation process.

Table VI outlines an algorithm for the hyperparameter tuning process that employs L2 regularisation in MLP models using Grid Search. This process is initiated by defining a hyperparameter grid, including the regularisation parameter, number of hidden layers, neuron range, activation function, and solver parameters.

A Grid Search Loop with 5-fold Cross-Validation follows, systematically exploring diverse hyperparameter combinations. Each iteration involves initialising and training an MLP model with L2 regularisation and monitoring training loss for convergence. The model's performance is assessed using the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) on held-out folds, with the average ROC-AUC

calculated across all folds. Each iteration records and accumulates training time from the previous iteration. Ultimately, the hyperparameter combination yielding the highest average performance metric is selected, resulting in the best MLP model with optimised hyperparameters.

Table VII introduces an innovative algorithm for the hyperparameter tuning process in MLP models utilising L2 regularisation and ES through Grid Search. Initially, a hyperparameter grid specifies parameters such as regularisation strength, hidden layer configuration, and activation function.

TABLE VI. ALGORITHM FOR HYPERPARAMETER TUNING WITH L2 REGULARISATION IN MULTI-LAYER PERCEPTRON (MLP) USING GRID SEARCH

Input	: Training set
Output	: Best MLP model with optimized hyperparameters
Step 1	Define Hyperparameter Grid for MLP: - Set regularisation parameter λ for L2 regularisation to 0.0001. - Set number of hidden layers to 1. - Define a range for the number of neurons per hidden layer from 5 to 100. - Set activation function to ReLU. - Set solver to Adam. - Set batch size (minibatch) to the minimum of 200. - Set initial learning rate to 0.001. - Set maximum iterations to 5000. - Set shuffle sample to True. - Set exponential decay rate for estimates of the first moment vector in Adam to 0.999. - Set exponential decay rate for estimates of the second moment vector in Adam to 1e-8.
Step 2	Grid Search Loop with 5-Fold Cross-Validation: - For each combination of hyperparameters in the defined grid: - Initialize an MLP model with the specified architecture and hyperparameters. - Divide the training data into 5 folds for cross-validation. - For each fold: o Train the MLP model with L2 regularisation. o Monitor the training loss during training and stop when the training loss does not improve by more than 0.0001 for 10 consecutive passes over the training set. o Evaluate the model's performance on the held-out fold using ROC-AUC. - Calculate the average performance metric (i.e., ROC-AUC) across 5 folds of the combination of hyperparameters. - Record and accumulate training time from the previous loop.
Step 3	Select Best Model: - Select the combination of hyperparameters that resulted in the highest average performance metric.

A Grid Search Loop with 5-fold Cross-Validation is implemented after that. An MLP model is instantiated and trained on five folds of the training data for each hyperparameter combination, with 10% reserved for validation. Training halts using ES when the validation score fails to improve by at least 0.0001 for ten consecutive epochs. Model performance is assessed using ROC-AUC on held-out folds, with the average metric calculated across all folds. Each loop records and accumulates training time from the previous loop. Finally, the hyperparameter combination yielding the highest average performance metric is selected, resulting in the best performing MLP model with optimised hyperparameters.

TABLE VII. ALGORITHM FOR HYPERPARAMETER TUNING WITH L2 REGULARISATION AND EARLY STOPPING IN MULTI-LAYER PERCEPTRON (MLP) USING GRID SEARCH

Input	: Training set
Output	: Best MLP model with optimized hyperparameters
Step 1	Define Hyperparameter Grid for MLP: - Set regularisation parameter λ for L2 regularisation to 0.0001. - Set number of hidden layers to 1. - Define a range for the number of neurons per hidden layer from 5 to 100. - Set activation function to ReLU. - Set solver to Adam. - Set batch size (minibatch) to the minimum of 200. - Set initial learning rate to 0.001. - Set maximum iterations to 5000. - Set shuffle sample to True. - Set exponential decay rate for estimates of the first moment vector in Adam to 0.999. - Set exponential decay rate for estimates of the second moment vector in Adam to 1e-8.
Step 2	Grid Search Loop with 5-Fold Cross-Validation: - For each combination of hyperparameters in the defined grid: - Initialize an MLP model with the specified architecture and hyperparameters. - Divide the training data into 5 folds for cross-validation. - For each fold: o Set aside 10% (using the stratified approach) of training data as validation. o Train the MLP model with L2 regularisation. o Terminate training using the Early Stopping approach, i.e. when the validation score is not improving by at least 0.0001 for 10 consecutive epochs. o Evaluate the model's performance on the held-out fold using ROC-AUC. - Calculate the average performance metric (i.e., ROC-AUC) across 5 folds of the combination of hyperparameters. - Record and accumulate training time from the previous loop.
Step 3	Select Best Model: - Select the combination of hyperparameters that resulted in the highest average performance metric.

The Cross-Entropy Loss function was employed for binary classification in the MLP model, as depicted in Eq. (1). This loss function quantifies the discrepancy between predicted and actual class labels, penalising misclassifications with logarithmic terms.

$$Loss(x_i, y_i) = -[y_i \, log(\hat{y}_i) + (1 - y_i) \, log(1 - \hat{y}_i)] \quad (1)$$

During the hyperparameter tuning phase using Grid Search, each cycle involves adjusting parameters such as the number of neurons (N) within a range from 5 to 100, alongside L2 regularisation. Eq. (2) represents the L2 regularisation term, penalising large weights to prevent overfitting.

$$n = \frac{\lambda}{2} \sum_{i=1}^{N} w_i^2 \quad (2)$$

Eq. (3) illustrates the cost function combining the Cross-Entropy Loss with the L2 regularisation term, where lambda (λ) is set to 0.0001. This cost function aims to balance minimising classification error and controlling model complexity, enhancing the model's predictive accuracy and

robustness. For the innovative model, the cost function includes L2 regularisation hybridised with ES, terminating training when the validation set score does not improve by at least 0.0001 for ten consecutive epochs.

$$Cost_{reg}(x_i, y_i) = Loss(x_i, y_i) + n = \frac{\lambda}{2} \sum_{i=1}^{N} w_i^2 \quad (3)$$

C. Testing

In this study, the development of MLP models was driven by two primary objectives: reducing the hyperparameter tuning time to identify the optimal set of hyperparameters for maximising classification accuracy and enhancing the model's ability to distinguish between classes. These objectives were evaluated using various metrics, including tuning time, confusion matrix analysis, and the ROC-AUC score. Tuning time refers to the duration from the commencement of the Grid Search process to its completion. It aims to enhance the model by pinpointing the most effective hyperparameters for achieving maximal accuracy in classification tasks. This step is critical for refining the model's configuration to improve its predictive performance. Furthermore, the evaluation process entailed an analysis of the confusion matrix to gauge the model's accuracy in classifying instances correctly across different categories. Additionally, the ROC-AUC score provided more profound insights into the model's discriminative capacity between classes.

Accuracy is a foundational metric, offering a high-level overview of the model's correctness. It is derived from the confusion matrix, as shown in Eq. (4), and calculates the proportion of instances correctly classified, irrespective of their class. This metric is handy when all classes are equally important, and the dataset is balanced. However, in cases of class imbalance, accuracy alone may not fully capture the model's effectiveness, as the prevalence of the dominant class may skew it. Hence, while accuracy is a valuable initial metric for evaluation, it is often essential to supplement it with additional metrics to obtain a more comprehensive picture of model performance, particularly in the presence of uneven class distribution.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (4)$$

with *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative.

Precision offers a more detailed examination of the model's performance, concentrating specifically on positive predictions. It is defined as the ratio of correctly predicted positive instances to the total number of predicted positive instances, as illustrated in Eq. (5). Precision is paramount in scenarios with significant consequences or costs associated with false positives.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (5)$$

with *TP* is True Positive and is False Positive.

In contrast, recall emphasises the model's capacity to identify all positive instances within the dataset. It calculates the ratio of correctly predicted positive instances to the total actual positive instances, highlighting its critical role in situations where overlooking positive instances (false negatives), could have serious repercussions. Eq. (6) presents the formula for calculating recall derived from the confusion matrix.

$$Recall/Sensitivity = \frac{(TP)}{(TP+FN)} \quad (6)$$

with *TP* is True Positive and is False Negative.

The F1 score, as presented in Eq. (7), amalgamates precision, and recall into a singular metric, offering a balanced evaluation of the model's performance. It computes the harmonic mean of precision and recall, providing a unified measure that accounts for false positives and false negatives, thereby preventing inflated scores when a significant discrepancy exists between precision and recall. The F1 score is particularly beneficial in scenarios involving class imbalance or when the costs associated with false positives and false negatives vary. By integrating precision and recall, the F1 score aids in guiding decision-making and model optimisation, ensuring a thorough classification performance assessment.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

The ROC-AUC score is widely recognised as a crucial metric for evaluating the performance of binary classification models. It depicts the likelihood that a randomly selected positive instance will be ranked higher than a randomly selected negative instance. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold settings. A higher ROC-AUC score signifies superior discrimination between positive and negative classes, with a score of 1 denoting a perfect classifier and 0.5 indicating a model no better than random guessing. The ROC-AUC score's robustness to class imbalance and applicability across different threshold configurations make it an invaluable tool for evaluating binary classifiers' overall performance in practical scenarios.

III. Result and Discussion

Table VIII illustrates the performance comparison of the MLP model with and without the employment of L2 regularisation. Implementing L2 regularisation with a 0.05 alpha value in the MLP model significantly enhanced its performance compared to the model devoid of L2 regularisation. This enhancement was particularly evident across three banking datasets, where the influence of L2 regularisation on the MLP model's performance was substantial. For instance, in a loan dataset comprising 16 features, the ROC-AUC score increased from 0.9133 to 0.9244. Similarly, for a loan dataset with 20 features, the ROC-AUC score improved from 0.9285 to 0.9312, and for a banking dataset focusing on credit card applications, the ROC-AUC score rose from 0.7464 to 0.7696. As observed in this study, the emphasis on the ROC-AUC metric in the analysis underscores its utility in providing a nuanced interpretation of model performance, especially in datasets with imbalanced labels. Additionally, results derived from the confusion matrix indicated an improvement in the MLP model's performance. These findings corroborate the significance of L2

regularisation in mitigating overfitting when developing MLP models, aligning with insights from prior research.

TABLE VIII. COMPARATIVE PERFORMANCE OF MLP MODELS WITH AND WITHOUT L2 REGULARISATION ACROSS BANKING DATASETS

Performance Matrices	Bank Loan 16 features		Bank Loan 20 features		Bank credit card default payments	
	No L2	L2	No L2	L2	No L2	L2
ROC-AUC	0.9133	0.9244	0.9285	0.9312	0.7464	0.7696
Accuracy	0.8979	0.9027	0.9052	0.9048	0.8021	0.8145
Precision (Positive)	0.5884	0.6373	0.5812	0.5839	0.5618	0.6072
Recall (Positive)	0.4384	0.4017	0.5375	0.5114	0.3855	0.4017
F1 Score	0.5024	0.4928	0.5585	0.5452	0.4572	0.4835

Table IX showcases the impact of integrating L2 regularisation into a MLP model, specifically employing a hyperparameter value setting as outlined in Table IV for the hyperparameter tuning process. This model exhibited remarkable performance, consistent with earlier observations that L2 regularisation could bolster MLP model performance. Specifically, Table IX outlines the results concerning ROC-AUC and the confusion matrix metrics. The bank loan dataset, consisting of 16 features, achieved ROC-AUC of 0.9221. A separate dataset, also related to bank loans but encompassing 20 features, registered ROC-AUC of 0.9455, while the bank credit card dataset demonstrated the ROC-AUC of 0.7814.

TABLE IX. IMPACT OF L2 REGULARISATION ON HYPERPARAMETER TUNING TIME ACROSS BANKING DATASETS

Dataset	Tuning Time (Sec.)	ROC-AUC	Accuracy	F1 Score
Bank Loan 16 features	26,787.75	0.9221	0.9005	0.5233
Bank Loan 20 features	23,714.07	0.9455	0.9189	0.5917
Bank credit card default payments	23,087.06	0.7814	0.8260	0.4847

The application of L2 regularisation has proven to enhance the MLP model's efficacy in classification tasks. Nonetheless, identifying the optimal hyperparameter settings with L2 regularisation application has been revealed to be a time-intensive endeavour. The hyperparameter tuning phase for all three datasets necessitated upwards of twenty-three thousand seconds, raising concerns regarding the practicality of deploying such models in real-time applications that demand frequent updates with new data entries.

Table X displays the performance of the MLP model employing a hybrid L2 regularisation and Early Stopping (L2ES) approach. Both techniques aim to mitigate overfitting, yet the experimental results indicate a significant reduction in hyperparameter tuning time with the hybrid L2ES approach. For the bank loan dataset with 16 features, tuning time was reduced by 91.44%, from 26,787.75 seconds to 2,424.88 seconds. Similarly, the dataset with 20 features saw a reduction of 90.95%, from 23,714.07 seconds to 2,030.71 seconds. The bank credit card dataset experienced a tuning time reduction of 93.57%, from 23,087.06 seconds to 1,485.15 seconds. The efficacy of the hybrid L2 and ES approach lies in its ability to shorten hyperparameter tuning time, primarily through the advantage offered by ES. This feature ceases the training process upon the absence of a minimum improvement of 0.0001 in the validation set score over ten successive epochs. This method is in stark contrast to the termination criteria of the baseline model, which discontinues training based on the absence of a notable improvement in the training loss over ten successive iterations, without consideration for the validation set score. Concurrently, L2 regularization is crucial in maintaining the classification model's accuracy by averting overfitting.

TABLE X. IMPACT OF HYBRID L2 REGULARISATION AND EARLY STOPPING ON HYPERPARAMETER TUNING TIME ACROSS BANKING DATASETS

Dataset	Tuning Time (Sec.)	ROC-AUC	Accuracy	F1 Score
Bank Loan 16 features	2,424.88	0.9121	0.8966	0.5351
Bank Loan 20 features	2,030.71	0.9389	0.9137	0.5854
Bank credit card default payments	1,485.15	0.7328	0.7877	0.4323

Fig. 2 presents a comparative analysis of hyperparameter tuning time between the MLP model using L2 regularisation and the proposed hybrid L2ES approach. The results are visually displayed using a bar chart, clearly illustrating a substantial reduction in tuning time across all datasets. Specifically, the hybrid L2ES method reduced tuning time by more than 90% compared to the L2-only approach, which was used as the baseline in this study. This represents a notable improvement, as such a reduction directly contributes to the practicality of implementing data-driven classification (DDC) models in real-world banking systems, particularly in scenarios requiring rapid model updates and continuous learning.

Fig. 3 compares the classification performance of the two approaches regarding ROC-AUC scores. The performance gap between the L2 and L2ES models is minimal, indicating that the significant gains in computational efficiency do not come at the cost of model accuracy. For the bank loan dataset with 16 features, the ROC-AUC difference is only 0.70%, while the 20-feature loan dataset shows a difference of 1.08%. The bank credit card dataset exhibits a slightly larger difference of 6.22%, which remains within an acceptable range of less than 10%. Although the L2ES approach yields marginally lower ROC-AUC scores in some cases, it consistently outperforms previous studies in overall performance benchmarks, particularly in hyperparameter tuning efficiency. These findings support the feasibility and value of the proposed hybrid approach for time-sensitive banking applications.

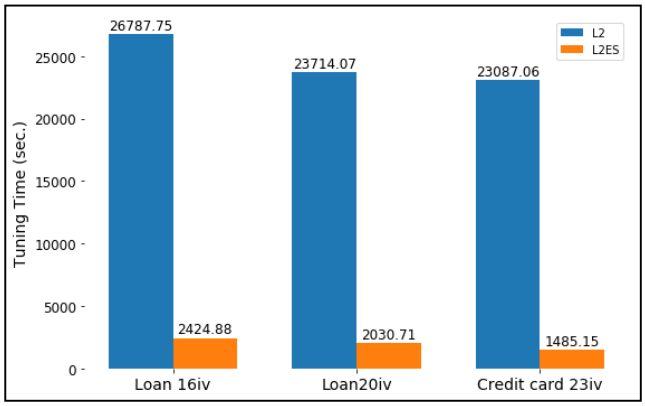


Fig. 2. Comparison of hyperparameter tuning time: L2 regularisation vs hybrid L2ES approach,

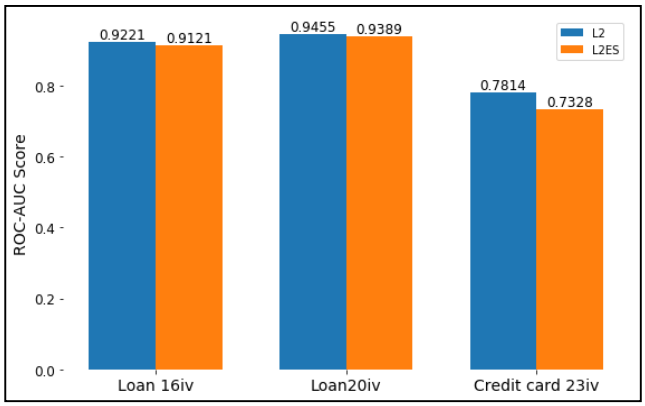


Fig. 3. ROC-AUC performance comparison: L2 regularisation vs hybrid L2ES approach across datasets.

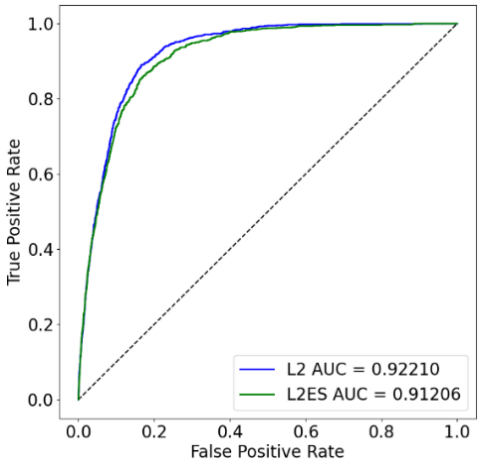


Fig. 4. ROC-AUC comparison for bank loan dataset with 16 features: L2 vs. hybrid L2ES approach.

Fig. 4 and Fig. 5 present the ROC-AUC curves for the bank loan datasets with 16 and 20 features, respectively, providing an alternative visual comparison between the L2 and hybrid L2ES approaches. Both graphs clearly show that the performance of the MLP model under the two conditions is very similar, with no significant differences observed in the shape or position of the curves.

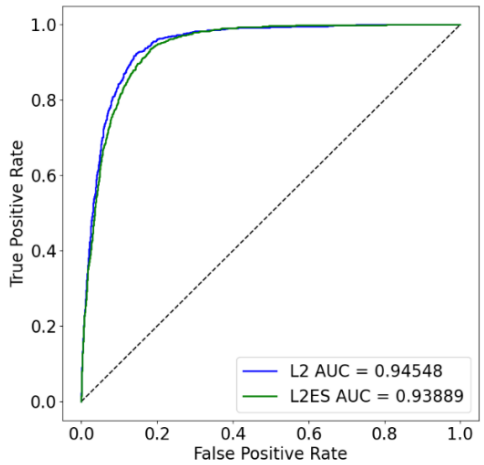


Fig. 5. ROC-AUC comparison for bank loan dataset with 20 features: L2 vs. hybrid L2ES approach.

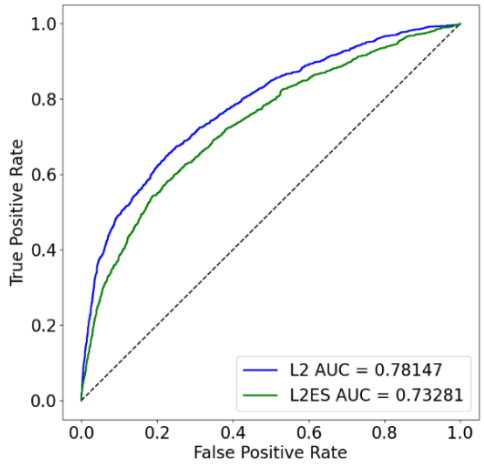


Fig. 6. ROC-AUC comparison for bank credit card dataset with 23 features: L2 vs. hybrid L2ES approach.

Fig. 6 displays the ROC-AUC curve for the bank credit card dataset. In this case, the hybrid L2ES approach shows a slightly lower curve than the model using L2 regularisation alone, indicating a marginal reduction in performance. However, this difference is relatively minor and does not significantly affect the overall effectiveness of the hybrid method. Notably, the L2ES approach outperforms previous studies, achieving a ROC-AUC score of 0.7328, compared to the values reported in Table II, where previous models did not exceed a ROC-AUC of 0.7200.

IV. Conclusion

This study evaluated the effectiveness of hybridising ES with L2 regularisation to reduce the time required for hyperparameter tuning—conducted via grid search—in developing MLP models, while preserving classification performance. The proposed L2ES approach significantly reduced tuning time without compromising model accuracy, as demonstrated by consistently high ROC-AUC scores comparable to those obtained using L2 regularisation alone. These results indicate that the L2ES technique offers a practical solution for streamlining hyperparameter tuning in

MLP models, thereby enhancing the efficiency of developing data-driven classification models.

Nevertheless, this study is not without limitations. The evaluation was limited to a set of structured banking datasets, and further validation is needed across diverse data domains and model architectures. Additionally, while grid search was used for tuning, other optimisation methods such as random search or Bayesian optimisation could offer further improvements.

Future research will aim to extend the proposed hybrid approach to other neural network models and investigate its compatibility with alternative tuning strategies. Enhancing generalisability, automating the hybridisation process, and applying it to unstructured or real-time financial data are promising directions for continued research.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from Universiti Teknikal Malaysia Melaka (UTeM) and the voluntary collaboration between research teams at UTeM and Universiti Teknologi MARA (UiTM), driven by a shared commitment to advancing data-driven classification methods.

REFERENCES

[1] A. AbdElsalam, M. Abdallah and H. Refaat, "Predicting Human Essential Genes Using Deep Learning: MLP with Adaptive Data Balancing" International Journal of Advanced Computer Science and Applications, vol. 16, no. 4, 2025.

[2] A. Aldelemy, and R. A. Abd-Alhameed, "Binary Classification of Customer's Online Purchasing Behavior Using Machine Learning." Journal of Techniques, vol. 5, no. 2, pp. 163-86, 2023.

[3] M. Grosicki. "Application of artificial neural networks to prediction of success of bank telemarketing campaign." 11-th European Conference of Young Researchers and Scientists, no. 3, pp. 6-8. 2015.

[4] A. Vitório and G. Marques. "Impact of imbalanced data on bank telemarketing calls outcome forecasting using machine learning." IEEE 2021 International Conference on Data Analytics for Business and Industry (ICDABI), pp. 380-384, 2021.

[5] T. C. Hsu, S. T. Liou, Y. P. Wang, and Y. S. Huang. "Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction." ICASSP 2019 IEEE International Conference on Acoustics, pp. 1572-1576, 2019.

[6] A. Arram, M. Ayob, M. A. A. Albadr, A. Sulaiman, and D. Albashish. "Credit card score prediction using machine learning models: A new dataset." arXiv preprint arXiv:2310.02956, 2023.

[7] S. Moro, P. Cortez, and P. Rita. "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems, vol. 62, 22-31, 2014.

[8] G. Marinakos, and S. Daskalaki. "Imbalanced customer classification for bank direct marketing." Journal of Marketing Analytics, no. 5, pp. 14-30, 2017.

[9] R. Farooqi and N. Iqbal. "Performance evaluation for competency of bank telemarketing prediction using data mining techniques." International Journal of Recent Technology and Engineering, vol. 8, no. 2, pp. 5666-5674, 2019.

[10] N. Ghatasheh, H. Faris, I. AlTaharwa, Y. Harb, and A. Harb. "Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks." Applied Sciences, vol. 10, no. 7, pp. 2581, 2020.

[11] A. Panigrahi, and M. C. Patnaik. "Customer deposit prediction using neural network techniques." International Journal of Applied Engineering Research, vol. 15, no. 2, pp. 253-258, 2020.

[12] S. Mokrane. "Predicting the success of bank telemarketing using Artificial Neural Network." International Journal of Economics and Management Engineering, vol. 14, no. 1, pp. 1-4, 2020.

[13] S. Dutta and S. K. Bandyopadhyay. "Recommender System for Term Deposit Likelihood Prediction Using Cross-Validated Neural Network." South Asian Journal of Social Studies and Economics, vol. 11, no. 3, pp. 21-28, 2021.

[14] S. Masturoh, F. S. Nugraha, S. Nurlela, M. R. Ramadhan Saelan, D. U. Eka Saputri, and R. Nurfalah. "Telemarketing Bank Success Prediction Using Multilayer Perceptron (MLP) Algorithm with Resampling." Jurnal Pilar Nusa Mandiri, vol. 17, no. 1, pp. 19-24, 2021.

[15] R. Singh and R. R. Aggarwal. "Comparative evaluation of predictive modeling techniques on credit card data." International Journal of Computer Theory and Engineering, vol. 3, no. 5, pp. 598, 2021.

[16] R. Liu. "Machine learning approaches to predict default of credit card clients." Modern Economy, vol. 9, no. 11, pp. 1828-1838, 2018.

[17] Vishwakarma, S. Kumar, A. Rasool, and G. Hajela. "Machine Learning Algorithms for Prediction of Credit Card Defaulters—A Comparative Study." Proceedings of International Conference on Sustainable Expert Systems Expert Systems: ICSES 2020, pp. 141–149, 2021.

[18] S. de Campos, P. Vitor, and L. C. Bambirra Torres. "Extreme wavelet fast learning machine for evaluation of the default profile on financial transactions." Computational Economics, vol. 57, no. 4, pp. 1263-1285, 2021.

[19] A. Almajid. "Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction." Journal of Advances in Information Systems and Technology, vol. 3, no. 2, pp. 67-74, 2022.

[20] J. Jiang, X. Meng, Y. Liu, and H. Wang. "An enhanced TSA-MLP model for identifying credit default problems." SAGE Open, vol. 12, no. 2, 2022.

[21] K. Shazly, and N. Khodadadi. "Credit card clients classification using hybrid guided wheel with particle swarm optimized for voting ensemble." Journal of Artificial Intelligence and Metaheuristics vol. 2, pp. 46-54, 2023.

[22] M. Q. Idrees, H. Naeem, M. Imran, A. Batool, & N. Tabassum. "Identifying Optimal Parameters and their Impact for Predicting Credit Card Defaulters Using Machine-Learning Algorithms." Lahore Garrison University Research Journal of Computer Science and Information Technology, vol. 6, no. 1, pp. 1-21, 2022.

[23] H. Yash, Affan, K. Saurav, and S. S. Dhanda. "Credit Card Default Prediction Using Machine Learning Models." Proc. - 2023 3rd Int. Conf. Innov. Sustain. Comput. Technol. CISCT 2023, pp. 1–5, 2023.

[24] S. Moro, R. Laureano, and P. Cortez. "Using data mining for bank direct marketing: An application of the crisp-dm methodology." The 2011 European Simulation and Modelling Conference, pp. 117–121, 2011.

[25] I-C. Yeh, and C-H. Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert systems with applications, vol. 36, no. 2, pp. 2473-2480, 2009.

Empirical Validation and Enhancement of ADiBA: A Framework for Big Data Analytics Implementation

Norhayati Daut[1], Naomie Salim[2], Sharin Hazlin Huspi[3], Anazida Zainal[4], Chan Weng Howe[5], Muhammad Aliif Ahmad[6], Siti Zaiton Mohd Hashim[7], Masitah Ghazali[8], Mohd Adham Isa[9], Rashidah Kadir[10], Nuremira Ibrahim[11], Norazlina Khamis[12]

Faculty of Computing and Informatics, Universiti Malaysia Sabah, Sabah, Malaysia[1, 12]
UTM Big Data Centre, Universiti Teknologi Malaysia, Johor, Malaysia[1, 2, 9]
Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia[3, 4, 5, 6, 7, 8, 10]
Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, Johor, Malaysia[11]

***Abstract*—The implementation of Big Data Analytics (BDA) in organisations requires a structured approach to ensure alignment with strategic goals and infrastructure readiness. This study presents an enhanced version of the previously published ADiBA (Accelerating Digital Transformation Through Big Data Adoption) framework that aimed at guiding organizations through critical components necessary for successful BDA implementation. The initial framework was developed based on systematic literature review. To validate and refine the framework, a mixed-methods survey was conducted among domain experts using a five-point Likert scale and open-ended questions to assess the relevance of each framework component. Quantitative responses were analysed using the Content Validity Index (CVI), with a threshold of 0.78 adopted as the minimum acceptable I-CVI score for each item. Complementing the quantitative analysis, qualitative feedback from the open-ended survey responses, Focus Group Discussions (FGDs), and in-depth interviews were examined through thematic analysis, revealing key themes related to framework's clarity and operational aspects. Insights from both analyses informed the refinement of several components. The resulting framework is a validated and empirically-informed guide designed to support effective BDA implementation in organizational contexts.**

Keywords*—*Adoption process; big data; big data analytics; framework; framework validation; expert survey; content validity index; thematic analysis; organizational implementation; digital transformation

I. Introduction

The implementation of BDA within organizations has emerged as an important factor in driving digital transformation and informed decision-making. Despite its significance, many organizations encounter challenges in effectively integrating BDA into their operations, often due to the absence of structured implementation frameworks. These challenges include data quality and integration, infrastructure and scalability issues, and shortage of skilled professionals [1].

While several BDA implementation frameworks has been proposed, a significant number lack empirical validation, limiting their practical applicability. This gap underscores the need for frameworks that are not only theoretically sound but also validated through expert insights to ensure relevance and effectiveness in real-world settings. In response to this, this paper aims to validate and refine a previously developed framework named ADiBA [2], designed to adopt and implement BDA in organizations. Although the initial construction of the framework was based on literature and theoretical model, validation is highly advisable to ascertain its relevance to practical settings.

To facilitate this, expert feedback obtained through structured surveys and FGDs. The quantitative data from the survey were analysed using CVI to assess the relevance of each framework component, while qualitative insights from open-ended responses and FGDs were examined through thematic analysis. The analyses were then used to determine which aspects require changes to achieve the intended objectives.

The primary objective of this study is to enhance the proposed ADiBA framework's validity and practical utility by integrating expert insights, thereby contributing a more robust tool for organizations navigating BDA implementation. By addressing the existing gap in empirically validated frameworks, this research offers a refined, expert-informed model that can serve as a valuable resource for practitioners and research alike.

To achieve these objectives and present our findings comprehensively, the remainder of this paper is structured as follows: Section II provides the background of the previously developed ADiBA framework, including the theoretical models that guided its initial construction. Section III details the mixed method methodology employed for framework validation, covering participant selection and data analysis techniques. Section IV presents the quantitative and qualitative results derived from the expert feedback. Section V outlines the enhancements made to the ADiBA framework based on these findings. Finally, Section VI concludes the paper by summarizing key contributions and discussing avenues for future research.

II. Background

The implementation of BDA within organizations has emerged as an important factor in driving digital transformation and informed decision-making. Despite its significance, many organizations encounter challenges in effectively integrating BDA into their operations, often due to the absence of structured implementation frameworks.

Numerous studies highlight the critical need for proper frameworks, models, or methodologies to successfully implement BDA. For instance, optimizing BDA process is challenging, as it involves combining diverse data assets to produce actionable insights [3]. and improved decision making [4], [5]. Similarly, a business-driven reference architecture is a viable solution to address the high failure rate of BDA implementation projects in public sector organizations, emphasizing the need to systematically consider business use cases and processes before assembling technical aspects [6]. As highlighted in [7], many BDA projects fail or do not deliver promised business value, often due to technical issues (data consolidation, quality, visualization), managerial challenges (inability to gain insights and link them to business problems), and organizational obstacles (changes in decision-making, culture, and clear direction of initiatives). The lack of a structured approach not only can jeopardize decision-making, but also prevent organizations from capitalizing on market opportunities, ultimately compromising their capacity to create value [8]. Furthermore, a significant gap between existing frameworks and the integration of big data into various business functions, explicitly stating the need for developing new models and implementation frameworks for better insights and patterns [9].

While the emerging interest in BDA has led to the proposal of various BDA implementation frameworks, a significant gap persists in the exiting literature. Many of these proposed frameworks primarily focus on the technical aspects of BDA implementation, often overlooking crucial dimensions such as people, organizational structure, culture, and strategic alignment. This narrow focus limits their holistic applicability and can lead to incomplete or unsustainable BDA initiatives within complex organizational environments. Furthermore, a substantial number of these studies propose frameworks without subjecting them to rigorous validation or verification processes, raising questions about their practical utility and real-world effectiveness.

In response to this gap, the ADiBA framework was developed to provide a structured approach to BDA implementation. This framework was initially formulated through a systematic literature review (SLR), which identified key components, steps, and activities essential for successful BDA implementation. The components of ADiBA framework (Fig. 1) are:

1) Prepare immersion of analytic culture: Ensure that organizations can fully leverage data-driven insights for decision-making and strategic growth.

2) Business understanding: Define clear objectives, identifying challenges, and ensuring that analytical efforts contribute to business value

3) Data management and governance: Ensure that data is accurate, secure, compliant, and valuable for decision-making

4) Project planning: Define project objectives, allocate resources effectively, and assess the potential risks and benefits.

5) Data understanding: Ensure that the data used is relevant, accurate, and suitable for analysis and modelling.

6) Data preparation: Transform the data into usable format for analysis and modelling.

7) Tool, infrastructure and technology procurement and presentation: Acquire and deploy necessary technological resources.

8) Business analytics modelling: Create analytical models that interpret data to provide actionable insights.

9) Data analytics product development: Transform insights derived from analytics into dashboards, and reports.

10) Evaluation of model and products: Assessing the performance of analytics modelling and visualization.

11) Data analysis product deployment: Transition the product from development phase to live production environment.

12) Monitoring, maintenance and upgrades: Monitor the performance, perform regular maintenance, and implement upgrades.

13) Inculturation of big data analytics into business: Foster data-driven decision-making, improves operational efficiency, drives innovation, supports strategic planning, and encourages continuous improvement.

The theoretical grounding of ADiBA framework draws upon established model such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) model [10], Kotter's 8-step model [11], DAMA-DMBOK Data Management [12], Data lifecycle [13], and Project Management lifecycle [14]. These models were specifically selected and integrated based on their demonstrated utility in various studies that proposed frameworks for BDA implementation, technology adoption, and organizational change, ensuring a robust foundation for ADiBA.

The CRISP-DM model is a widely adopted, structured methodology for managing data mining projects. It encompasses six iterative phases to provide a comprehensive framework that ensures a systematic approach to data analytics. The Kotter's 8-step model is designed to assist leaders in effectively managing organizational transformation by identifying potential sources of resistance and developing strategies to address them, thereby facilitating a smoother change process. The DAMA-DMBOK framework outlines best practices across key data management areas such as data governance and quality, ensuring robust data management practices. The Data lifecycle models describe the progression of data from creation to disposal, ensuring effective data management throughout its lifespan. Finally, the project management lifecycle offers a structured framework that guides projects from initiation to completion, enhancing the likelihood of success.

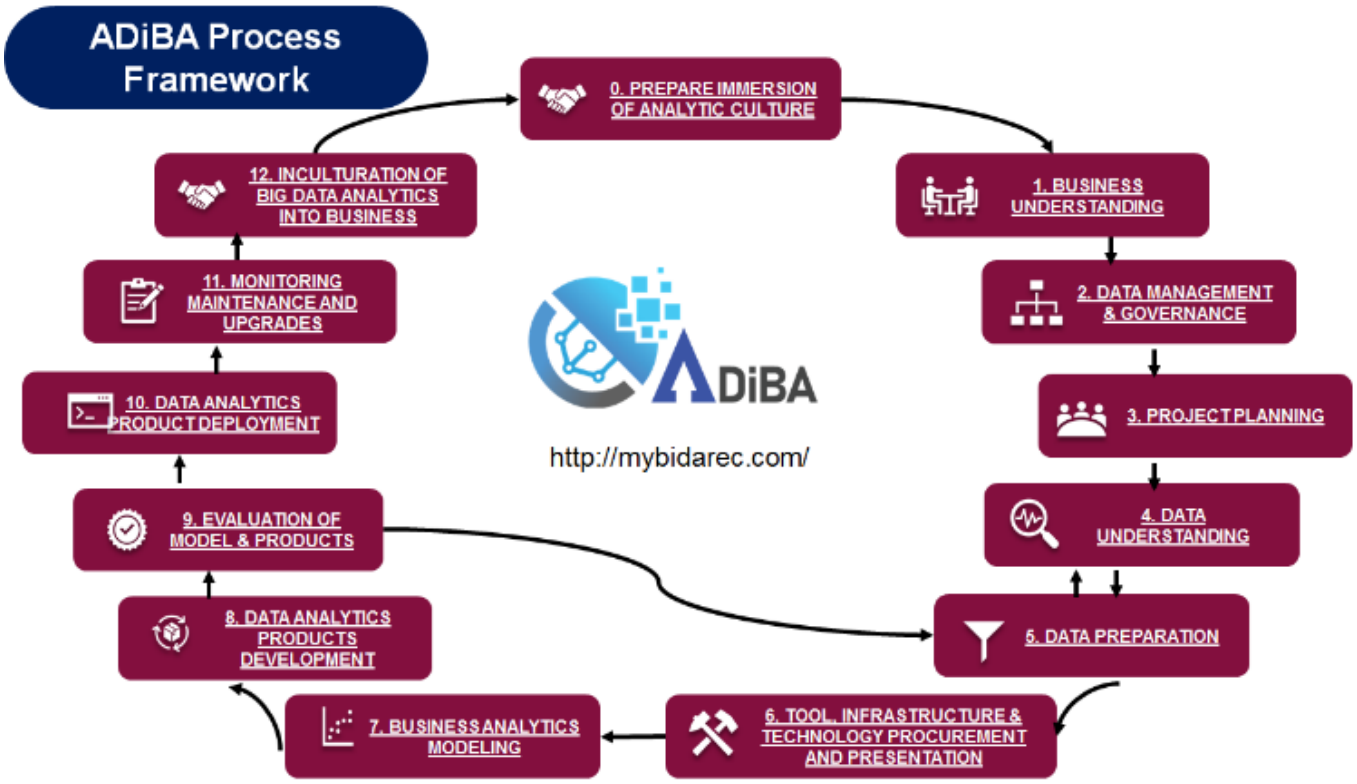


Fig. 1. ADiBA framework.

By synthesizing these models, the ADiBA framework offers a holistic approach to BDA implementation, deliberately addressing both technical execution and the crucial non-technical dimensions of strategic alignment, data management, and organizational change. The ADiBA framework has a step-by-step approach with tools and guidelines that can be used by organizations to implement BDA. These steps can be tailored and adopted selectively based on size, types, resources, and analytics goals of organizations.

Building upon the theoretical foundations and the structured components of the ADiBA framework, this study seeks to validate and refine the framework to ensure its practical applicability in organizational settings. To achieve this, a comprehensive methodological approach was employed, combining quantitative assessments with qualitative insights. The following section details the research design, data collection methods, and procedures to evaluate the relevance of the ADiBA framework in facilitating BDA implementation

III. Methodology

This study adopted a mixed-methods research design to validate and refine the ADiBA framework. The section details the development of the evaluation instrument, the selection criteria for domain experts, the procedures for data collection, and the quantitative (CVI) and qualitative (thematic analysis) approaches used for data analysis. The following subsections elaborate on each methodological phase.

A. Development of Evaluation Instrument

The survey was crafted to assess the relevance of each component within the ADiBA framework. It comprises both quantitative and qualitative elements to capture a holistic view of expert opinion.

For the quantitative assessment, a series of 4-point Likert scale questions were developed for each task (item) under the framework's component that ranged from 1 to 4, where 1 represents "not relevant", 2 represents "somewhat relevant", 3 represents "relevant", and 4 represents "highly relevant". Experts can rate the degree to which each task is applicable and essential within their organizational context. The objectives of assessing the tasks of each component are summarized in Table I.

Complementing the quantitative data, the survey incorporated open-ended questions aimed at eliciting qualitative insights. For each component, experts were invited to suggest any revisions or provide comments on the respective tasks and components. Additionally, a concluding open-ended question sought their overarching thoughts and recommendations concerning the entire ADiBA framework. The questions addressed topics such as the necessity of a structured methodology to implement BDA in organizations and the potential challenges organizations might encounter. These qualitative responses were subjected to thematic analysis, enabling the identification of common themes and nuanced perspectives that could inform further refinement of the framework.

TABLE I. OBJECTIVES FOR TASKS OF ADIBA FRAMEWORK COMPONENTS

Framework Component	Task	Objective
PREPARE IMMERSION OF ANALYTICS CULTURE	0.1	To assess the relevance of creating urgency among members of organization
	0.2	To assess the relevance of building a guiding coalition among members of organization
	0.3	To assess the relevance of creating a vision for change among members of organization
	0.4	To assess the relevance of communication the vision among members of organization
	0.5	To assess the relevance of removing barriers that can hinder the change among members of organization
	0.6	To assess the relevance of creating short-term wins for the organization
	0.7	To assess the relevance of building on the change among members of organization
	0.8	To assess the relevance of anchoring the change in the organizational culture
BUSINESS UNDERSTANDING	1.1	To assess the relevance of identifying the organization's business goal
	1.2	To assess the relevance of assessing the organization's current situation
	1.3	To assess the relevance of defining the organization's data analytics goals or insights
DATA MANAGEMENT & GOVERNANCE	2.1	To assess the relevance of defining the data governance engagement framework
	2.2	To assess the relevance of defining the data governance organization
	2.3	To assess the relevance of developing data security, privacy, sharing, ethics and compliance governance framework
PROJECT PLANNING	3.1	To assess the relevance of identifying organization's business use cases
	3.2	To assess the relevance of estimating the resources required for data analytics project
	3.3	To assess the relevance of performing cost-benefit analysis before starting any data analytics project
	3.4	To assess the relevance of developing a project plan for data analytics project
DATA UNDERSTANDING	4.1	To assess the relevance of defining sources of the data required
	4.2	To assess the relevance of designing and developing data sandbox
	4.3	To assess the relevance of describing the data required
	4.4	To assess the relevance of developing engines for data curation
	4.5	To assess the relevance of verifying the quality of data
DATA PREPARATION	5.1	To assess the relevance of extracting the data required
	5.2	To assess the relevance of transforming the extracted data
	5.3	To assess the relevance of exploring and visualizing the data
	5.4	To assess the relevance of modifying the data
TOOL, TECHNOLOGY, AND INFRASTRUCTURE PROCUREMENT & PRESENTATION	6.1	To assess the relevance of identifying the required tool, technology and infrastructure
	6.2	To assess the relevance of evaluating the required tool, technology and infrastructure
	6.3	To assess the relevance of procuring the required tool, technology and infrastructure
BUSINESS ANALYTICS MODELING	7.1	To assess the relevance of identifying key variables for data analytics modelling
	7.2	To assess the relevance of selecting the suitable modelling techniques
	7.3	To assess the relevance of designing the testing and training dataset
	7.4	To assess the relevance of building the data analytics model
	7.5	To assess the relevance of assessing the data analytics model
	7.6	To assess the relevance of managing the data analytics model
	7.7	To assess the relevance of deploying the data analytics model
DATA ANALYTICS PRODUCT DEVELOPMENT	8.1	To assess the relevance of pre-designing the data analytics product
	8.2	To assess the relevance of designing and developing dashboards
	8.3	To assess the relevance of designing and developing business reports
	8.4	To assess the relevance of developing alerts
EVALUATION OF MODEL AND PRODUCT	9.1	To assess the relevance of performing the data analytics product testing
	9.2	To assess the relevance of performing pilot test

Framework Component	Task	Objective
EVALUATION OF MODEL AND PRODUCT	9.3	To assess the relevance of preparing test report
	9.4	To assess the relevance of determining the next course of actions
DATA ANALYTICS PRODUCT DEPLOYMENT	10.1	To assess the relevance of planning the data analytics product deployment
	10.2	To assess the relevance of planning the monitoring and maintaining the data analytics product
	10.3	To assess the relevance of reporting the final results
	10.4	To assess the relevance of reviewing the data analytics project
MONITORING, MAINTENANCE AND UPGRADES	11.1	To assess the relevance of monitoring the performance of data analytics product
	11.2	To assess the relevance of correcting any errors occurred
	11.3	To assess the relevance of enhancing dashboards, reports, and alerts if required
	11.4	To assess the relevance of replacing or discarding the data analytics product if obsolete
INCULTURATION OF BIG DATA ANALYTICS INTO BUSINESS	12.1	To assess the relevance of generating short-term wins
	12.2	To assess the relevance of sustaining the change in the organization culture
	12.3	To assess the relevance of anchoring the change in the organization culture
	12.4	To assess the relevance of assessing the impact of big data analytics on organization

B. Selections of Experts

To ensure a comprehensive evaluation of the ADiBA framework, domain experts with substantial experience in BDA implementation were selected. These experts were identified through professional networks, industry conferences, and academic collaborations. Invitations outlining the study's objectives and the importance of their participation were sent via email. Table II shows the number of selected experts and their organization.

Even though many experts do not have a formalized framework or methodology for implementing BDA in their organizations, their hands-on experience enables them to determine which components are appropriate and effective when developing and executing BDA projects. This practical knowledge ensures that their evaluations are grounded in real-world applicability, contributing to their organization's progression toward becoming data-driven.

TABLE II. LIST OF EXPERTS

Organization	No of Experts	Government/non-government	Domain
Jabatan Digital Negara	3	Government	ICT
Malaysia Digital Economy Corporation (MDEC)	2	Government	ICT
Malaysia International Shipping Corporation Berhad (MISC)	1	Government	Shipping
Ebdesk Malaysia Sdn. Bhd	1	Non-Government	ICT
Malaysia Marine and Heavy Engineering Holdings Berhad (MMHE)	2	Government	Oil and Gas, Construction
Johor Land Berhad (JLAND)	1	Government	Real Estate
REFREX Sdn. Bhd.	1	Non-Government	ICT
Rebaie Analytics Group	1	Non-Government	ICT
Big Data Institute	1	Non-Government	ICT

C. Data Collection

Data were collected through a structured approach during FGDs sessions, integrating both quantitative and qualitative methods. Two FGDs were conducted as part of the study. The first session took place in Selangor, Malaysia and include representatives from Jabatan Digital Negara, MDEC, MISC, and Ebdesk Malaysia. The second FGD was held in Johor, Malaysia with participation from MHHE, JLAND, and REFREX. Each FGD session commenced with the administration of the framework evaluation survey. This was followed by a facilitated discussion encouraging participants to share their perspectives, allowing them to provide qualitative feedback, suggest revisions, and discuss the applicability of the framework's component in real-world organizational setting.

For the surveys, responses were recorded directly into a digital format, ensuring ease of analysis. The data were then utilized to calculate the CVI. For the qualitative components, a systematic note-taking and reconciliation process was used.

During the Focus Group Discussions (FGDs), two assistant moderators were assigned to systematically document participant inputs using structured note-taking techniques to capture key points, detailed arguments, and verbatim quotations. Immediately following each session, the assistants collaboratively reviewed and reconciled their notes. This process involved cross-verifying observations, resolving discrepancies, and consolidating insights to produce a single, comprehensive and finalized set of notes for each FGD.

Similarly, for the two in-depth interviews conducted in Lisbon, Portugal, one researcher led the conversation while the second focused on detailed note-taking. This collection of finalized, detailed notes from both the FGDs and in-depth interviews constituted the complete qualitative dataset. This dataset was then organized and served as the primary data for thematic analysis.

D. Data Analysis

1) CVI Analysis: The CVI analysis was employed to quantitatively assess the relevance of each task within the framework. Experts rated each task using a 4-point Likert scale, where 1 indicated "not relevant" and 4 indicated "highly relevant". Each task rated as 3 or 4 was converted to '1' to show that the rater finds the task relevant, and tasks rated as 1 or 2 were converted to '0' to show that the rater considers the task irrelevant.

For each task, the I-CVI was calculated using Eq. (1) by dividing the number of experts who rated the task as 1 by the total number of experts. The S-CVI/Ave as shown in Eq. (2) can be determined by averaging the I-CVI scores across all tasks [15]. The following are the formulas for the content validity index calculation:

$$I-CVI = \frac{Number\ of\ experts\ rating\ the\ item\ as\ "Relevant"}{Total\ number\ of\ experts} \quad (1)$$

$$S-CVI/Ave = \frac{\sum I-CVI}{Number\ of\ items\ in\ the\ component} \quad (2)$$

I-CVI with value of 0.78 or higher was considered acceptable for at least 9 raters, following the recommendations by Lynn [16]. For S-CVI/Ave value of 0.90 or above indicated excellent content validity, values between 0.80 and 0.89 suggested the need for minor revision, and values below 0.80 signalled the need for major revision or potential removal [15].

2) Thematic analysis: To explore participant's perspectives and experiences in depth, thematic analysis was conducted following the six-phase approach by Braun and Clarke [17]. Initially, the data were familiarized by reading and re-reading the transcribed FGDs and in-depth interviews. Next, initial codes were generated to identify significant features of the data relevant to the aim of the study and the enhancement of the proposed framework.

These codes were then organized into potential themes, which were reviewed and refined to ensure coherence and distinctiveness. Each theme was clearly defined and named to capture its essence accurately. Finally, a comprehensive report was produced, illustrating the themes with representative data extracts and providing a nuanced understanding of the participant's insights. This systematic approach facilitated a rich and detailed interpretation of the qualitative data.

IV. Result

This section presents the findings derived from the expert evaluation of the proposed ADiBA framework. The analysis incorporates both quantitative and qualitative data collected through the framework evaluation survey, FGDs, and in-depth interviews. The quantitative aspect was analyzed using the CVI to assess the relevance of each component and task within the framework. Meanwhile, qualitative insights obtained from open-ended survey responses and discussion sessions were subjected to thematic analysis to identify key themes, suggestions, and patterns in expert feedback. Together these findings informed the refinement and validation of the framework.

A. CVI Analysis

As per the established criteria, the content validity of the framework was assessed using the I-CVI and S-CVI/Ave. Table III presents the calculated I-CVI for each task, along the aggregated S-CVI/Ave for each component, and their corresponding interpretations based on the predefined threshold.

All the tasks achieved I-CVI values above 0.85, indicating strong agreement among experts regarding their relevance – except for two. The tasks "Remove Barriers" from Prepare Immersion of Analytical Culture component and "Develop Data Security, Privacy, Sharing, Ethics and Compliance Governance Framework" from Data Management and Governance component each scored an I-CVI of 0.77. However, given the minimal difference of 0.01 from the threshold, these tasks are still considered important and relevant. Therefore, rather than being removed, they are retained with minor revisions to improve clarity and better align them with their respective components.

TABLE III. Summary of I-CVI and S-CVI/Ave Values with Interpretation for Framework Components

Component	Task	Task Description (Task)	I-CVI	Interpretation
PREPARE IMMERSION OF ANALYTICS CULTURE	0.1	Create Urgency	0.85	Acceptable
	0.2	Build a Guiding Coalition	0.92	Acceptable
	0.3	Create a Vision for Change	0.92	Acceptable
	0.4	Communicate the vision	0.92	Acceptable
	0.5	Remove Barriers	0.77	Needs revision
	0.6	Create Short-Term Wins	0.85	Acceptable
	0.7	Build on the Change	0.85	Acceptable
	0.8	Anchor the change in the Organization's Culture	0.85	Acceptable
	S-CVI/Ave		0.87	Consider Minor Revision
BUSINESS UNDERSTANDING	1.1	Identify Business Goal	1.00	Acceptable
	1.2	Assess Situation	1.00	Acceptable
	1.3	Define Data Analytics goals or insights	1.00	Acceptable
	S-CVI/Ave		1.00	Component is Acceptable

DATA MANAGEMENT & GOVERNANCE	2.1	Data Governance Engagement Framework	0.85	Acceptable
	2.2	Define Data Governance Organization	0.85	Acceptable
	2.3	Develop Data Security, Privacy, Sharing, Ethics and Compliance Governance Framework	0.77	Needs revision
	S-CVI/Ave		0.82	Consider Minor Revision
PROJECT PLANNING	3.1	Identify Business Use Case	1.00	Acceptable
	3.2	Estimate Resources Required	0.92	Acceptable
	3.3	Perform Cost-Benefit Analysis	0.85	Acceptable
	3.4	Develop Project Plan	0.92	Acceptable
	S-CVI/Ave		0.92	Component is Acceptable
DATA UNDERSTANDING	4.1	Define Data Sources	1.00	Acceptable
	4.2	Design and Develop Data Sandbox	0.85	Acceptable
	4.3	Describe Data	1.00	Acceptable
	4.4	Develop Data Curation Engine	0.85	Acceptable
	4.5	Verify Data Quality	0.92	Acceptable
	S-CVI/Ave		0.92	Component is Acceptable
DATA PREPARATION	5.1	Extract Data	1.00	Acceptable
	5.2	Transform Data	1.00	Acceptable
	5.3	Explore and Visualize Data	0.92	Acceptable
	5.4	Modify Data	0.92	Acceptable
	S-CVI/Ave		0.96	Component is Acceptable
TOOL, TECHNOLOGY, AND INFRASTRUCTURE PROCUREMENT & PRESENTATION	6.1	Identify Required Tools, Infrastructure and Technology	0.92	Acceptable
	6.2	Evaluate Tools, and Technology	0.92	Acceptable
	6.3	Procure Tools	1.00	Acceptable
	S-CVI/Ave		0.95	Component is Acceptable
BUSINESS ANALYTICS MODELING	7.1	Identify Key Variables	0.92	Acceptable
	7.2	Select Modelling Techniques	0.92	Acceptable
	7.3	Design Test	0.92	Acceptable
	7.4	Build Model	0.92	Acceptable
	7.5	Assess Model	0.92	Acceptable
	7.6	Manage Model	0.92	Acceptable
	7.7	Deploy Model	0.92	Acceptable
	S-CVI/Ave		0.92	Component is Acceptable
DATA ANALYTICS PRODUCT DEVELOPMENT	8.1	Pre-Design Stage	0.92	Acceptable
	8.2	Design and Develop Dashboards	1.00	Acceptable
	8.3	Design and Develop Business Reports	1.00	Acceptable
	8.4	Develop Alerts	0.92	Acceptable
	S-CVI/Ave		0.96	Component is Acceptable
EVALUATION OF MODEL AND PRODUCT	9.1	Perform Data Product Testing	1.00	Acceptable
	9.2	Perform Pilot Test	1.00	Acceptable
	9.3	Prepare Test Report	1.00	Acceptable
Component	**Task**	**Task Description (Task)**	**I-CVI**	**Interpretation**
EVALUATION OF MODEL AND PRODUCT	9.4	Determine Next Course of Actions	1.00	Acceptable
	S-CVI/Ave		1.00	Component is Acceptable
DATA ANALYTICS PRODUCT DEPLOYMENT	10.1	Plan Deployment	1.00	Acceptable
	10.2	Plan Monitoring and Maintenance	1.00	Acceptable
	10.3	Report Final Results	1.00	Acceptable
	10.4	Review Project	1.00	Acceptable
	S-CVI/Ave		1.00	Component is Acceptable

MONITORING, MAINTENANCE, AND UPGRADES	11.1	Monitor Performance	1.00	Acceptable
	11.2	Correct Error	1.00	Acceptable
	11.3	Enhance Dashboard, Reports and Alerts	1.00	Acceptable
	11.4	Replace or Discard System if Obsolete	0.92	Acceptable
	S-CVI/Ave		0.98	Component is Acceptable
INCULTURATION OF BIG DATA ANALYTICS INTO BUSINESS	12.1	Generate Short-Term Wins	1.00	Acceptable
	12.2	Sustain: Build on the Change	1.00	Acceptable
	12.3	Anchor the Changes in the Organizational Culture	1.00	Acceptable
	12.4	Assess Impact of Big Data Analytics	1.00	Acceptable
	S-CVI/Ave		1.00	Component is Acceptable

All components recorded S-CVI values of 0.92 and above indicating excellent content validity, with the exception of two components. These two components – Prepare Immersion of Analytical Culture and Data Management and Governance had S-CVI/Ave values of 0.87 and 0.82, respectively. Based on the established interpretation range, these values fall within the 0.80-0.89 band, suggesting the need for minor revisions to enhance clarity or alignment with expert expectations.

From the result, it can be seen that experts provided almost 100% agreements for tasks toward the end of the framework component. It is uncertain whether these agreements reflect genuine perceptions of task relevance, or if they may have resulted from experts becoming fatigue and thus responding less critically or attentively. This introduces a potential bias in the data, as some responses may not fully reflect the experts' considered judgment.

While potential response fatigue cannot be entirely ruled out, the overall results of the CVI analysis indicate that the framework is considered relevant and appropriate by the expert panel. The high I-CVI and S-CVI/Ave values across the majority of tasks and components reflect strong agreement among experts on the framework's content validity.

B. Thematic Analysis

Thematic analysis was conducted to explore expert insights and experiences relevant to each component of the proposed framework. Qualitative data were drawn from open-ended survey responses, FGDs, and in-depth interviews. The thematic analysis revealed specific implementation challenges, enablers, and suggestions, which were mapped against the framework's components to guide further refinement. To ensure clarity and relevance, the thematic analysis findings are organized by the components of the proposed framework, enabling detailed exploration of expert perspectives within each specific area.

1) Thematic analysis by components: For Prepare Immersion for Analytics Culture component, experts highlighted a range of critical improvements and areas requiring clarification. Several experts found that the tasks in this component closely overlap with those in the next component (Business Understanding), leading to confusion, and suggested either merging the two or switching the order to better reflect their purpose. A key recommendation was to establish the data governance team at the outset, before any framework activity begins, to provide strategic direction and oversight. Creating a sense of urgency was another frequently mentioned point, with recommendations to include lessons learned or prior failures as motivation. There was concern over when to conduct maturity and adoption assessments, and a suggestion that these be used at multiple points – before and during framework application as a way to tailor next steps. In summary, Prepare Immersion for Analytics Culture requires clearer structure, stronger leadership components, an emphasis on culture and urgency, and better integration with the broader framework cycle.

Expert feedback on Business Understanding component revealed critical insights on refining the strategic foundation and ensuring the component sets a strong direction for the entire framework. Many experts suggested switching Business Understanding component with Prepare Immersion for Analytics Culture component, as Business Understanding is seen as an essential first step. There was a strong push to include a comprehensive maturity assessment covering both people and infrastructure technology to better inform the organization's readiness for BDA. Experts also recommended defining BDA roadmap under this component and starting with clear enterprise-level goals before diving into specific roles or committee structures. Several comments stressed the importance of clearly identifying the pain point, using tools such as SWOT analysis and impact analysis to understand business challenges and opportunities. Benchmarking with competitors in similar industries was encouraged to help organizations understand where they stand and what best practices to adopt. Feedback also suggested shortening unnecessary early-stage activities such as analysis of contractors and consultants, and instead focusing on the critical success factors and a forward-moving plan. Sustainability and cultural buy-in were also emphasized, with a need to motivate ongoing engagement. Overall, Business Understanding needs clearer guidance, stronger business alignment, and a practical approach to managing constraints while still establishing a solid business case for data-driven transformation.

For Data Governance and Management component, expert feedback emphasized the importance of establishing a clear and adaptive governance structure. Experts suggested that the data organization model whether centralized, decentralized, or hybrid must be well-defined, including distinctions between ICT-driven and business driven models, as well as the roles of data engineers versus data scientists. Additionally, there is a need to provide clear references and guidance on how to structure and assign roles within data governance councils or teams, such as owners, stewards, and risk officers, in alignment

with organizational data maturity. There was also concern about ownership and accountability, particularly regarding the role of data owners and the need to clarify whether this overlap with roles like a data sponsor. Ownership of the data governance framework whether it belongs to the strategy team or data owners, was seen as critical, especially in government contexts where BDA initiatives typically span multiple ministries and agencies. In multi-agency environments, initial agreements often break down when data sharing becomes contentious. This highlighted the need for strategic teams to oversee framework implementation and anticipate political and administrative challenges.

The Project Planning component received expert suggestion focusing on the need for refining task structures and sequencing. There were suggestions to reorganize subtasks – some could be moved to other components, while others could be combined or removed if already addressed elsewhere. Additionally, experts emphasized the importance of well-defined project implementation plan, including risk assessment analysis, which would improve preparedness and help mitigate uncertainties during the execution of BDA projects. A key point raised was the importance of prioritizing platform versus data readiness. Experts questioned which should come first in BDA projects, urging the need for clear guidance based on project context.

In relation to the Data Understanding component, experts pointed out the importance of the foundational role of infrastructure readiness, with multiple reviewers asserting that data understanding activities can only proceed effectively once platforms and data architecture are clearly defined. The need to define data source characteristics and behaviour emerged strongly, including guidance on metadata management, crawler/scraper tools, and the handling of data based on SOPs and business process. Several experts stressed the importance of data profiling and exploratory analysis before attempting data quality checks, citing practices such as measuring missing values and data health status. Data curation engine was recommended to be reframed as “digitalization”, and data ingestion was suggested as a better term, than merely defining data sources. In addition, experts emphasized developing small-scale lake and ensuring that machine-readable formats are prioritized for open data and sharing purposes. Lastly, confusion about the term “data sandbox” was raised, suggesting a need to revise or clearly define it in the framework.

Feedback from experts on the Data Preparation component emphasized the importance of adopting a more agile and iterative approach, with some suggesting that integrating DataOps principles could enhance the adaptability of this phase. A major point raised was the logical sequencing of tasks – experts noted that data processing should be addressed before initiating data exploration and visualization. There were also concerns about task allocation, with the observation that feature engineering might be more appropriate under the Business Analytics Modelling component rather than the current stage. Experts encouraged refining this component to avoid redundancy, especially concerning data validation and missing values, which they noted are already examined in the data understanding phase. These insights point toward the need to streamline the Data Preparation process and ensure that each sub-task is placed in the most relevant component to avoid duplication and confusion.

For the Tool, Technology, and Infrastructure Procurement and Presentation component, experts highlighted a key gap in the component: while the current focus is on tools, there is an evident lack of attention to infrastructure and technology elements. There was a string consensus on the need to expand this component to include implementation activities such as setting up the big data platform after tool procurement, and proof of concept execution using real business cases to validate feasibility. In light of current digital threats, experts recommended integrating cybersecurity requirements both during tool evaluation and procurement, ensuring security compliance and closing potential loopholes. Additionally, while some suggested merging this component with Project Planning, others were open to keeping it separate, provided that the scope is clearly defined and comprehensive.

Experts input on Business Analytics Modelling component suggested the importance of continuous improvement and clarity of expected outcome. One key theme that emerge was the need to ensure that the analytics models remain accurate and relevant when applied newly incoming datasets. Additionally, experts highlighted the need to include expected results or anticipated outcomes of each modelling task.

For the Data Analytic Product Development, experts pointed three key areas of enhancement. First, they noted that time-series visualizations are integrated and aligned with industry standards. Second, experts recommend exploring non-linear dashboard layouts, which break away from traditional grid-based or sequential flows. Lastly, there was a suggestion to incorporate data modelling directly within the dashboard development process.

For the Evaluation of Model and Product, experts revealed several areas of confusion and suggestions for improvement. A recurring theme was the structural clarity of the evaluation component, particularly the placement of model evaluation and product evaluation. Experts questioned whether model evaluation should be integrated as a subphase under Business Analytics Modelling component, as they are closely connected. Similarly, product evaluation, if it refers to dashboard, was suggested to be part of the Data Analytics Product Development component. Furthermore, the importance of practical evaluation strategies was emphasized, recommending the inclusion of a proper test plan, diverse testing approaches, and migration strategies to ensure smooth deployment.

For the Data Analytic Product Deployment component, experts suggested that the inclusion of a clear roll-out plan to ensure the smooth implementation of big data initiatives across the organization.

Feedback on the Monitoring, Maintenance, and Upgrades component emphasized the importance of providing a scalable infrastructure that can accommodate future growth in data volume and processing needs.

Finally, for the Inculturation of Big Data Analytics into Business component, experts highlighted the need for continuous efforts to embed analytics practices within

organizational culture. Experts suggested conducting a post-maturity assessment to monitor progress, implementing ongoing campaigns or awareness programs focusing on data-related challenges, and organizing regular checkpoint meetings by the steering or working committee to track developments. Moreover, performing impact analysis simulations and identifying "quick wins" alongside medium-term plans were highlighted to sustain momentum and demonstrate early value in the adoption of BDA.

The summarized thematic analysis for each component according to themes can be seen in the following Table IV.

TABLE IV. SUMMARY OF THEMATIC ANALYSIS BY THEME AND ITS COMPONENTS

Theme	Relevant Component	Key Issues	Expert Suggestions
Sequencing and Structural Clarity	Prepare Immersion for Analytics Culture, Business Understanding, Project Planning, Evaluation of Model and Product	Overlap between components; unclear starting point; misplaced tasks or evaluations	Switch or merge components 1 & 2; clearly define sequence of tasks; move evaluation activities under appropriate phases (e.g., modelling or dashboard development)
Governance and Role Clarity	Prepare Immersion for Analytics Culture, Business Understanding, Data Governance and Management,	Undefined roles; lack of ownership and accountability; unclear governance models	Define centralized/decentralized/hybrid models; assign clear roles (owners, stewards); establish governance team early
Readiness and Maturity Assessment	Prepare Immersion for Analytics Culture, Business Understanding, Inculturation of Big Data Analytics into Business	Readiness not assessed early or revisited	Conduct people + tech maturity assessments at start and mid-cycle; use post-maturity assessments after implementation
Cultural Embedding and Change Management	Prepare Immersion for Analytics Culture, Business Understanding, Inculturation of Big Data Analytics into Business	Weak emphasis on change culture and ongoing motivation	Create urgency using past lessons; embed campaigns and checkpoint meetings; track quick wins and momentum
Infrastructure and Platform Readiness	Data Understanding, Tool, Technology, and Infrastructure Procurement and Presentation, Monitoring, Maintenance, and Upgrades	Underdeveloped infrastructure focus; unclear timing for platform readiness	Address infrastructure before data work; add infra setup post-procurement; ensure scalability for growth
Agile and Iterative Practice	Data Preparation, Business Analytics Modelling, Data Analytic Product Development	Phases too linear or redundant; lack of agility	Apply agile/DataOps principles; reposition feature engineering; make components more iterative and streamlined
Security and Risk Management	Project Planning, Tool, Technology, and Infrastructure Procurement and Presentation,	Cybersecurity missing; risk planning weak	Add cybersecurity criteria to procurement; include risk analysis in planning
Output Clarity and Evaluation	Business Analytics Modelling, Data Analytic Product Development, Evaluation of Model and Product	Lack of clarity on expected outcomes; unclear evaluation scope	Define expected model results; create test/migration plans; relocate product evaluation to correct phase
Visualization and Usability	Data Analytic Product Development	Rigid layout; missed integration opportunities	Adopt time-series and non-linear dashboard options; embed modelling directly into visual design
Deployment Planning and Execution	Project Planning, Data Analytic Product Deployment	No clear roll-out plan; poor execution readiness	Define deployment steps; align platform/data readiness before execution
Data Quality and Profiling	Data Understanding, Data Preparation	Redundancy and poor sequencing of data checks	Emphasize data profiling before validation; avoid duplicate quality tasks across components

1) Thematic analysis of overall framework: Overall, experts revealed several critical themes that reflect both structural and practical challenges. One of the most pressing concerns was clarity and connectivity within the framework. Experts noted that the framework appeared disjointed, with unclear linkages between components. This lack of cohesion made it difficult for users to understand how the process flows from one stage to the next or how long each stage might take. The absence of a defined timeline or lifecycle for the framework components further compounded this issue, leading to suggestions for a clearer depiction of the overall process, including estimated durations and interdependencies.

The usability and agility of the framework were also scrutinized. Experts emphasized the importance of incorporating flexible decision-making mechanisms, particularly for context where data availability is limited or governance issues arise. This aligns with concerns that BDA adoption often fails due to an overemphasis on technology, while underestimating organizational and cultural aspects. To counter this, there were calls for more agile and adaptable framework that could accommodate varying entry points depending on an organization's maturity level or existing infrastructure.

Experts highlighted that the tools and instruments included in the framework appeared standard and more appropriate for beginners. However, they recommended that these tools be better integrated across the components to ensures continuity and support scalability. A particular emphasis was placed on improving the clarity and usability of some instruments, which were seen as difficult to interpret or complete without guidance.

Another recurring theme was the importance of infrastructure readiness. Multiple experts emphasized that infrastructure should be fully in place prior to initiating BDA projects, particularly in environments where phased implementation may not be viable. This was tied to the broader issue of platform and use case development, both of which were deemed essential before launching analytics activities.

Organizational roles and structure emerged as a key theme, with experts stressing the need for two distinct but collaborative working teams – data scientists and data engineers. Moreover, a dedicated BDA team should be

assigned exclusively to analytics work, without being burdened by unrelated tasks.

A strong leadership structure was considered vital, with top-down instructions and the appointment of BDA champions cited as enablers of successful implementation. In the absence of leadership or designated champions, enculturation and sustained adoption were identified as areas of concern.

Lastly, experts noted that the "early components" of the framework particularly problem identification were the most challenging. This was attributed to organizations lacking clarity in articulating business problems or use cases. Some experts proposed incorporating an "as-is/to-be" study at the beginning to define current capabilities and desired outcomes more effectively.

The summarized thematic analysis for overall framework according to themes can be seen in the following Table V.

TABLE V. SUMMARY OF THEMATIC ANALYSIS OF THE OVERALL FRAMEWORK

Theme	Expert Insights and Suggestions
Clarity and Connectivity	Framework components appeared disjointed; unclear linkages and lack of flow between stages. Suggested inclusion of process flow, interdependencies, and estimated timelines.
Usability and Agility	Need for more flexibility in the framework to accommodate varied contexts, such as low data maturity or limited infrastructure. Recommends support for different entry points.
Tools and Instruments	Tools included were seen as basic; need to enhance integration across components and improve usability. Some instruments required more guidance or simplification.
Infrastructure Readiness	Strong emphasis on ensuring infrastructure and platforms is in place before initiating BDA work. Phased implementation is not always suitable.
Organizational Structure	Clear roles needed: separate data scientists and data engineers, with a dedicated BDA team. Leadership and BDA champions essential for sustained adoption and success.
Leadership and Culture	Top-down support and cultural readiness critical. Lack of leadership seen as a barrier to adoption and long-term success.
Early-Stage Complexity	Early framework stages, especially problem identification, were most difficult. Suggested including an "as-is/to-be" study to better define current state and future goals.

The expert feedback underscores the importance of refining the framework to enhance its practical applicability and ensure it can support diverse organizational contexts and maturity levels. These insights directly inform the subsequent framework enhancement process.

V. ADIBA FRAMEWORK ENHANCEMENT

Based on the CVI and thematic analyses, the ADiBA framework underwent significant enhancements. While the framework still retained its thirteen core components, these revisions strategically refined its overall structure and significantly modified the underlying components and tasks, thereby improving its suitability for broader application. Specifically, feedback informed meticulous modifications, with components and tasks being added, revised, or removed to enhance clarity, eliminate redundancy, and integrate actionable assessment points. The details of these integrated enhancements are provided below, with the finalized visual representation of the ADiBA framework and its update components presented in Table VI and Fig. 2 at the end of this section.

For the first component, the name Prepare Immersion of Analytics Culture was changed to Enterprise Big Data Analytics Preparation for better and clearer purposes. Two new tasks were added to the component: Set up Strategic Planning Team [18], and Perform Internal Situation Analysis [10], [19], [20], [21], [22], [23], [24]. Setting up teams responsible for steering, planning and monitoring Big Data Analytics adoption and implementation in organizations is essential to ensure data initiatives align with overall business objectives. The team will consider the people, processes, and data infrastructure involved and guide the organization toward data-driven goals. Although the experts suggested that setting up the data governance team must be established early before any framework activities can be conducted, it is more important to set the strategic planning team to oversee the whole BDA implementation. The task "Performing an Internal Situation Analysis" is introduced to address the expert's recommendation to define specific points in the framework for conducting maturity and adoption assessment, especially before its application. This task includes sub-tasks such as administering surveys to assess the organization's structure and to identify key personnel who can provide accurate insights into the organization's data and analytics maturity. Additionally, it enables the organization to pinpoint existing gaps in skills and capabilities. This is crucial because possessing technical skills alone does not guarantee effective execution; some individuals may have the required skills but the lack of contextual understanding or practical ability to perform tasks effectively. This task was also introduced to address redundancy with a similar task "Assess Situation" in the Business Understanding component. The overlapping task in Business Understanding component will be removed to avoid duplication and ensure a clearer, more streamlined activities within the framework. To strengthen the leadership aspects within this component, a subtask related to securing leadership commitment was added under task "Build a Guiding Coalition" to ensure active commitment and visible support from executive leaders to champion the BDA initiatives. Based on the CVI analysis, the "Remove Barriers" task received a rating below the acceptable threshold, indicating that some experts questioned its clarity and relevance. Rather that removing the task entirely, it was retained and refined to address these concerns. The task now includes a more specific focus on the types of barriers that commonly hinder BDA readiness, and outlines actionable strategies for overcoming these barriers. To improve more on the clarity and focus of this component, two tasks "Build on the Change" and "Anchor the Change in the Organization's Culture" were removed, as they were redundant with tasks included in the final component of the framework – Data Analytics Product Enculturation.

TABLE VI. ENHANCED ADIBA FRAMEWORK COMPONENTS AND TASKS

No.	Enhanced Framework		
	Component	***Task***	
1	ENTERPRISE BIG DATA ANALYTICS PREPARATION	BDAP.1	Set up Strategic Planning Team
		BDAP.2	Perform Internal Situation Analysis
		BDAP.3	Create a Sense of Urgency
		BDAP.4	Build a Guiding Coalition
		BDAP.5	Form Strategic Vision and Initiatives
		BDAP.6	Enlist a Volunteer Army
		BDAP.7	Remove Barriers
		BDAP.8	Generate Short-Term Wins
2	BUSINESS UNDERSTANDING	BU.1	Identify Business Goal
		BU.2	Analyse Strategic Context and Organizational Structure
		BU.3	Define Data Analytics Goals or Insights
3	PROJECT PLANNING	PP.1	Identify Business Use Case
		PP.2	Perform Cost-Benefit Analysis
		PP.3	Develop Project Plan
4	DATA GOVERNANCE	DG.1	Define Data Governance Organization
		DG.2	Develop Data Security, Privacy, Sharing, Ethics and Compliance Governance Framework
		DG.3	Define Data Quality Process
5	DATA AND ANALYTICS REPOSITORY MANAGEMENT	DARM.1	Data Ingestion
		DARM.2	Design and Develop Enterprise Single Source of Truth Master Data Management
		DARM.3	Develop Enterprise Metadata Hub
		DARM.4	Verify and Implement Data Quality
		DARM.5	Incorporate New Data Request
		DARM.6	Maintain Data Analytics Inventories
6	TOOL, TECHNOLOGY, AND INFRASTRUCTURE PROCUREMENT & PREPARATION	TTIP.1	Identify Required Tools, Technology, and Infrastructure
		TTIP.2	Evaluate Tools, Technology, and Infrastructure
		TTIP.3	Procure and Prepare Tools, Technology, and Infrastructure
7	DATA ANALYTICS VISUALIZATION DEVELOPMENT	DAVD.1	Pre-Design Stage
		DAVD.2	Design and Develop Dashboards
8	DATA ANALYTICS PRODUCT MODELING	DAPM.1	Identify Key Variables
		DAPM.2	Select Modelling Techniques
		DAPM.3	Perform Data Sampling
		DAPM.4	Feature Engineering
		DAPM.5	Build Model
		DAPM.6	Assess Model
		DAPM.7	Manage and Deploy Model
9	DATA PREPARATION	DP.1	Identify Data Requirements and Extract Data
		DP.2	Explore and Transform Data
		DP.3	Publish Data
10	DATA ANALYTICS PRODUCT EVALUATION	DAPE.1	Perform Dashboard Testing
		DAPE.2	Prepare Evaluation Report and Next Course of Action
11	DATA ANALYTICS PRODUCT DEPLOYMENT	DAPD.1	Deploy Dashboard
		DAPD.2	Perform Stress Test

		DAPD.3	Document the Data Analytics Product
12	DATA ANALYTICS PRODUCT MONITORING, MAINTENANCE AND UPGRADES	MMU.1	Plan the Performance Monitoring
		MMU.2	Correct Any Errors
		MMU.3	Enhance Dashboards
		MMU.4	Replace or Discard Dashboards if Obsolete
No.	**Enhanced Framework**		
	Component	***Task***	
13	DATA ANALYTICS PRODUCT ENCULTURATION	DAE.1	Sustain the Changes
		DAE.2	Anchor the Changes in the Organizational Culture
		DAE.3	Assess Impact of Big Data Analytics

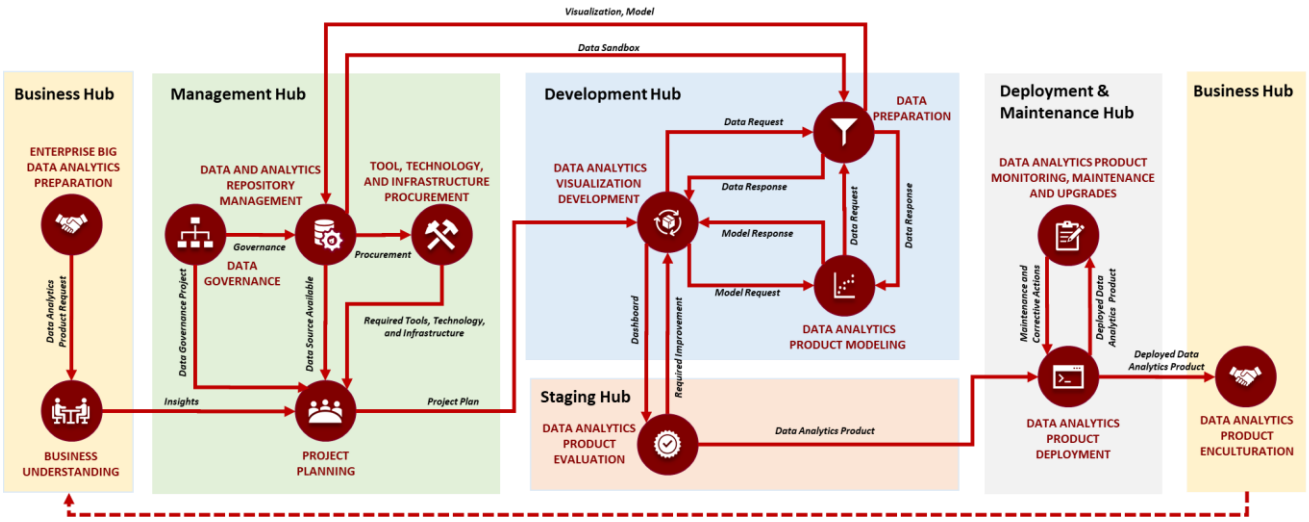


Fig. 2. Enhanced visual representation of the ADiBA framework.

For the Business Understanding component, the task previously titled “Assess Situation” was renamed to “Analyse Strategic Context and Organizational Structure” to avoid confusion with another existing task, “Perform Internal Situation Analysis” in Prepare Immersion of Analytics Culture component. Although both tasks involve forms of assessment, their focus and content are distinct. “Perform Internal Situation Analysis” task emphasizes understanding the organization’s current structure, readiness, and maturity in data and analytics, often through surveys or interviews. In contrast, the renamed task focuses on analysing strategic elements such as the organization’s goals and objectives, SWOT analysis, roles and responsibilities, customer charter, and key performance indicators. Following expert feedback, the task “Identify Business Goals” was enhanced to include the definition of critical success factors. The more extensive development of solid business cases as recommended by experts, has been comprehensively integrated into the Project Planning component.

The Data Management and Governance component was changed to Data Governance as the tasks were more aligned with governance-related activities. Additionally, the “Data Governance Engagement Framework” task was removed due to overlap with “Define the Data Governance Organization”, particularly in role and structure definition. Based on expert feedback, several enhancements were made to strengthen the “Define the Data Governance Organization” task. This task now incorporates outlining the roles of data engineers versus data scientists within the governance structure, and assigning responsibilities within data governance councils or teams. Furthermore, it addresses ownership and accountability, especially in distinguishing the roles of data owners and sponsors, and defines who should own the governance framework – whether it is the strategy team or data owners in both single and multi-agency contexts. Political and administrative challenges in multi-agency environments are acknowledged, and the role of strategic teams in overseeing governance framework implementation is emphasized. The task "Develop Data Security, Privacy, Sharing, Ethics and Compliance Governance Framework" received a CVI rating below the threshold, indicating a need for refinement. Although no explicit justification was provided by the experts, the low rating may be dues to the task’s broad scope and lack of practical guidance. To address this, several subtasks were included such as identifying applicable data protection laws and regulations; define policies for data access, classification, and retention; and establish ethical guidelines for analytics. A new task, “Define the Data Quality Process” [10] was added to strengthen the component’s foundation. The task focuses on data quality requirements and improvement programs and preparing data quality tests that cover all data quality dimensions: completeness, conformity, consistency, accuracy, uniqueness, integrity, validity, and timeliness. This is to ensure

data quality is a critical element of data governance, as it supports reliable data for analytics and decision-making.

The Project Planning component has been strengthened to provide a more comprehensive approach to developing project plans. This was achieved by combining the formerly separate "Estimate Resources Required" and "Perform Cost-Benefit Analysis" tasks into a unified process. This is to ensure that resource allocation is directly informed by a thoroughly financial and practical assessment. Furthermore, the updated "Develop Project Plan" task now explicitly incorporates a well-defined project implementation plan and integrates comprehensive risk assessment analysis, ensuring all potential challenges are identified and addressed proactively within the planning phase.

The Data Understanding component has undergone a comprehensive transformation, now formally established as Data and Analytics Repository Management. This strategic evolution underscores a deeper focus on not just comprehending data, but actively managing it within a robust repository environment, which emphasizes on infrastructure readiness for all subsequent data activities. This component's tasks have been meticulously refined and expanded. The initial task, formerly known as "Define Data Sources" has been reconceptualized as "Data Ingestion", which now provides guidance on defining intricate data source characteristics and behaviour, and strategically incorporates the use of crawler/scraper tools for efficient and structure data acquisition. The "Design and Develop Data Sandbox" task was renamed to "Design and Develop Enterprise Single Source of Truth Master Data Management" [25]. This is to detail the design and development of enterprise-wide data consistency frameworks, including data warehouses, data marts, data lakes, and integrated enterprise master data management with robust integrity checks. The "Describe Data" task was changed to "Develop Enterprise Metadata Hub" [26] to define data's metadata and incorporate advanced metadata management strategies for enhanced discoverability and governance. The experts suggested to change the "Develop Data Curation Engine" task to "Digitalization", but it seems to be too broad and abstract for a specific task. The subtasks described under this task align closely with what is now covered under the "Data Ingestion" task. Therefore, to ensure clarity and avoid redundancy within the component, the most logical approach would be to fully incorporate the development of data extraction engines / data input portals / data capturing systems directly into the "Data Ingestion" task. Significant enhancements have been made to the task "Verify and Implement Data Quality" as detailed by [10], [20], [22], [23], [27], [28], [29], [30], [31], and [32]. Organizations can identify the data quality requirements and their problems and implement the data quality measures and problem rectification. Two new tasks were added to this component: Incorporate New Data Requests and Maintain Data Analytics Inventories. Both designed to ensure continuous and managed growth of the data and analytics repository. Throughout all tasks, there is a clear prioritization of machine-readable formats to facilitate open data and efficient sharing, further enhancing the utility and accessibility of the managed data assets.

The tasks remained the same for the Tool, Technology, and Infrastructure Procurement & Presentation component. The only changes that were made were to the component and task names to make the purposes much more straightforward. The scope was broadened to include infrastructure and technology elements, moving beyond just tools to encompass he entire foundational environment.

The Data Analytics Visualization Development component, formerly "Data Analytics Product Development", has been refined to focus on creating impactful visual outputs. The tasks "Design and Develop Business Reports" and "Develop Alerts" were merged into "Design and Develop Dashboards" task.to further enhance this component, the "Design and Develop Dashboards" tasks now explicitly ensures that time-series visualizations are integrated and aligned with industry standards, providing clear insights into trends over time. This task also encourages exploring and suggesting non-linear dashboard layouts to create more intuitive and engaging user experiences.

The Data Analytics Product Modelling component, previously known as Business Analytics Modelling, has undergone a focused enhancement to better reflect its purpose and incorporate crucial best practices. While retaining its core responsibilities, the component now explicitly includes a vital new task: Feature Engineering [33]. This process encompasses feature extraction, selection, or reduction, and data augmentation as needed, significantly enriching the data preparation for model building. The "Design Test" task has also been refined and renamed to "Perform Data sampling", where data from the prepared sandbox is strategically selected and split into training and testing sets, with option to perform bootstrapping if the sample size is insufficient. To ensure models remain accurate and relevant over time, the need for continuous improvement within the "Manage and Deploy Model" task is emphasized. This ensures ongoing monitoring and recalibration as new data emerges. Additionally, the "Select Modelling Techniques" and "Build Model" tasks now require the definition of expected results or anticipated outcomes.

For the Data Preparation component, the initial "Extract Data" task has been expanded and renamed to "Identify Data Requirements and Extract Data", a task highlighted by [10], [34], and [35], emphasizing the crucial step of defining data needs before extraction. The workflow is streamlined by merging the "Transform Data" and "Explore and Visualize Data" into a single, comprehensive task: "Explore and Transform Data". This task is now central to preparing data for machine readability, which includes steps such as exploring data, cleaning data, transforming data, identifying and treating missing values, detecting and treating outliers, normalizing data, and tagging and annotating data. While "Feature engineering" has been re-evaluated and primarily moved to the Data Analytics Product Modelling component to avoid redundancy, data validation and the handling of missing values remain firmly within the "Explore and Transform Data" to prevent overlap with the Data and Analytics Repository Management component. The redundant "Modify Data" task was removed, and a new "Publish Data" task was introduced to

ensure prepared data is readily available to requesters in a data sandbox.

The Data Analytics Product Evaluation component, previously known as "Evaluation of the Model and Product". Has been refined to streamline its processes and enhance clarity. We've consolidated the original tasks into two comprehensive steps: "Perform Dashboard Testing" and "Prepare Evaluation Report and Next Course of Action". The "Perform Dashboard Testing" task now integrates component testing, analytic product testing, and crucial user acceptance and satisfaction testing, ensuring a thorough assessment. These diverse testing approaches include a broader range of validation techniques. While model evaluation is now considered for integration into the Data Analytics Product Modelling component, and dashboard-specific product evaluation might align with Data Analytics Product Development to avoid redundancy, this component retains its focus on overarching product validation. For the "Prepare Evaluation Report and Next Course of Action" task, defects are recorded, potential actions identified, and a final decision is made regarding product deployment. The task also includes essential guidance on migration strategies for smooth deployment, ensuring a seamless transition for the product.

For the Data Analytics Product Deployment component, the "Plan Monitoring and Maintenance" task was strategically moved to a subsequent component focused on ongoing monitoring activities, allowing this stage to concentrate purely on deployment. The "Deploy Dashboard" task includes the development of a clear roll-out plan for organization-wide implementation. A new critical task, "Perform Stress Test" was added to evaluate the dashboard's performance and stability under extreme or heavy loads. Additionally, the "Report Final Results" and "Review Project" tasks were merged into a single, comprehensive task: "Document the Data Analytics Product", ensuring all relevant information is captured.

The Monitoring, Maintenance, and Upgrades component has been refined and renamed to Data Analytics Product Monitoring, Maintenance, and Upgrades. This change clarifies that its focus solely on the data analytic product, not the entire ADiBA framework. While the core responsibilities remain consistent, the task names were updated for better clarity. The further enhance this component, the "Plan the Performance Monitoring" task now emphasizes the importance of a scalable infrastructure to accommodate future growth. The other tasks, "Correct any Errors", "Enhance Dashboards", and "Replace or Discard Dashboards if Obsolete", continue to ensure the long-term reliability and effectiveness of the data analytics product.

The Inculturation of Big Data Analytics into Business component has been refined and renamed to Data Analytics Product Enculturation, more accurately reflecting its purposes of embedding data analytics within the organization. While the core tasks remain, "Generate Short-Term Wins" was removed to avoid repetition with the earlier component. The finalized tasks are now: "Sustain the Changes", "Anchor the Changes in the Organizational Culture", and "Assess Impact of Big Data Analytics". The "Sustain the Changes" task now emphasizes continuous efforts to embed analytical practices and includes organizing regular checkpoint meetings by the steering and working committee to maintain momentum. For "Anchor the Changes in the Organizational Culture" task, the need for ongoing campaigns or awareness programs focusing on data-related challenges was integrated, fostering a data-driven mindset throughout the organization. Finally, the "Assess Impact of Big Data Analytics" task is strengthened by the inclusion of conducting a post-maturity assessment and performing impact analysis simulations, allowing a comprehensive evaluation of the analytics product's effectiveness and organizational transformation.

The ADiBA framework has undergone substantial refinement, building on its initial reorganization. While the framework still retains its thirteen core components, significant changes were made to their individual tasks and broader component definitions to enhance its effectiveness. These components are now strategically grouped into five distinct hubs – Business, Management, Development, Staging, and Deployment & Maintenance – to differentiate the phases and gather related functions, thereby improving process flow and logical alignment. The visual representation of the ADiBA framework can be seen in Fig. 2.

The revised framework begins with the Enterprise Big Data Analytics Preparation component, which outlines steps to prepare the organization for a data-driven culture. This preparatory stage is now implicitly strengthened by the expert emphasis on strong leadership, top-down instructions, and the formation of a dedicated BDA team (comprising distinct data scientists and data engineers), all crucial for effective enculturation. The deliverable of this component is the data analytics product requested by the organization. Next, the Business Understanding component helps organizations articulate their goals and problems, ensuring BDA delivers desired values and empowers decision-making. Critically, to address challenges in early problem identification, this component is now enhanced by incorporating an "as-is/to-be" study to more effectively define current capabilities and desired outcomes.

The Data Governance, Data and Analytics Repository Management, and Tool, Technology, and Infrastructure Procurement & Preparation components can be conducted concurrently, established and managed independently of specific BDA projects. These activities generate the overarching data governance project, detailed insights on available data assets, and the necessary tools, technologies and infrastructure. They thus serve as enabling pillars that support any subsequent BDA initiatives integrated within this framework. Furthermore, based on expert feedback, the framework now explicitly states that infrastructure should be fully in place prior to initiating BDA projects, particularly in environments where phased implementation isn't viable.

Organizations then use this strengthened information to plan the project through the Project Planning component. The project plan moves into the Development Hub, which includes Data Analytics Visualization Development, Data Analytics Product Modelling, and Data Preparation. These processes are iterative, ensuring the efficient development of the dashboard along with its underlying data and analytics model. This iterative approach now explicitly embodies the framework's

enhanced usability and agility, integrating flexible decision-making mechanisms particularly relevant for context with limited data availability or governance issues, and allowing for varying entry points based on organizational maturity.

Following development, dashboards undergo various testing in the Data Analytics Product Evaluation component before being deployed via the Data Analytics Product Deployment component. The Data Analytics Product Monitoring. Maintenance, and Upgrades component handles ongoing corrections or enhancements. Finally, the Data Analytics Product enculturation component outlines steps to embed the data analytics into business practices. This final stage is directly supported by the framework's emphasis on fostering strong leadership and the role of the BDA champions, ensuring the sustained adoption and cultural anchoring of BDA within organization.

Ultimately, these enhancements directly address the challenges highlighted by experts, forging a BDA implementation framework that is clearer, more agile, and more comprehensively integrated. The result is a more practical and effective guide for organizations navigating their data analytics journey.

VI. Conclusion

Recognizing the necessity for a structured pathway in BDA implementation in organizations, this study undertook the enhancement of the existing ADiBA framework. Initially developed through systematic literature review, the framework's refinement process involved a comprehensive mixed-methods survey. Domain experts provided quantitative feedback via a five-point Likert scale, which was then assessed using the CVI against an I-CVI threshold of 0.78. parallel to this, qualitative data from open-ended questions, FGDs, and in-depth interviews were subjected to thematic analysis, revealing key insights into the framework's clarity and practical application. The integration of these empirical findings directly guided the strategic adjustments made to the framework's various components.

This study makes a significant contribution by transforming a theoretically derived framework into an empirically validated and refined guide for BDA implementation. The enhances ADiBA framework offers organizations a robust and expert-informed tool to navigate the complexities of becoming data-driven, potentially accelerating digital transformation efforts and improving decision-making process. Theoretically, this research contributes to the body of knowledge on framework development and validation methodologies in the context of information systems.

While this study refined the ADiBA framework based on expert insights, it is important to acknowledge that the validation was primarily based on expert perception. Practical application and studies over time are essential for a full assessment of its efficacy.

Looking ahead, the next crucial step involves a more comprehensive verification of the enhanced ADiBA framework through real-life case studies within actual organizations. The plan is to identify several organizations currently undertaking BDA projects and guide them through the enhanced ADiBA framework's components. By integrating the framework into their existing processes, its practical utility and impact on their journey to becoming data-driven entities can be observed. From these case studies, in-depth feedback will be collected from organizational stakeholders regarding the framework's applicability, ease of use, and effectiveness in addressing their BDA challenges. This practical application will provide valuable insights into its efficacy and adaptability in diverse operational contexts, further strengthening its utility for BDA implementation and leading to subsequent rounds of refinement.

Beyond these immediate plans, potential avenues for future exploration include extending the ADiBA framework to incorporate emerging technologies like Artificial Intelligence (AI) and Machine Learning (ML) integration strategies. Further research could also investigate the framework's applicability across different industry sectors or organizational sizes, as well as developing a digital tool or assessment instrument based on the framework to aid organizations in self-assessing their BDA readiness.

References

[1] P. Ataei, S. Thrope, S. Regula, and D. Staegemann, "Why Big Data Projects Fail? A Systematic Literature Review," Assoc. Inf. Syst. AIS Electron. Libr. , no. January, pp. 1–13, 2024.

[2] N. Daut et al., "ADiBA Big Data Adoption Framework: Accelerating Big Data Revolution 5.0," no. Data, pp. 549–556, 2022, doi: 10.5220/0011351700003269.

[3] D. U. Thakar and D. D. Pandya, "A Comprehensive Framework for Big Data Analytics : Core Elements , Implementation , and Strategic Insights," pp. 291–296, 2025.

[4] B. Mago and N. Khan, "A Proposed Framework for Big Data Analytics in Higher Education," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, pp. 685–691, 2021, doi: 10.14569/IJACSA.2021.0120778.

[5] J. C. Morales-Arevalo and C. Rodríguez, "A Systematic Review of the Benefits and Challenges of Data Analytics in Organizational Decision Making," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 2, pp. 200–209, 2025, doi: 10.14569/IJACSA.2025.0160221.

[6] M. Didas, F. Chali, N. Elisa, and J. Matogoro, "Requirements for Building a Business-Driven Reference Architecture for Implementation of Big Data Analytics by Public Sector Organizations: A Case Study of Uganda," 1st Int. Conf. Adv. Artif. Intell. African Context. AAIAC 2023, vol. 14, no. 69, 2023, doi: 10.1109/AAIAC60008.2023.10465496.

[7] J. Maritz, S. Eybers, and M. Hattingh, Implementation Considerations for Big Data Analytics (BDA): A Benefit Dependency Network Approach, vol. 12066 LNCS. Springer International Publishing, 2020. doi: 10.1007/978-3-030-44999-5_40.

[8] G. Alghafri, T. Albalushi, and M. Rahim, "Understanding Big Data Adoption and its Impact on Decision-Making Quality: A Perspective from Public Sector IT Managers in Oman," J. Syst. Manag. Sci., vol. 14, no. 11, pp. 247–274, 2024, doi: 10.33168/jsms.2024.1113.

[9] N. U. Shahid and N. J. Sheikh, "Impact of Big Data on Innovation, Competitive Advantage, Productivity, and Decision Making: Literature Review," Open J. Bus. Manag., vol. 09, no. 02, pp. 586–617, 2021, doi: 10.4236/ojbm.2021.92032.

[10] P. Chapman et al., CRISP-DM 1.0. Step-by-step data mining guide. SPSS Inc, 2000.

[11] J. P. Kotter, "8 Steps to Accelerate Change in Your Organization," Accelerate: Building Strategic Agility for a faster-Moving World, 2014. https://www.kotterinc.com/8-steps-process-for-leading-change/ (accessed Apr. 24, 2022).

[12] DAMA International, DAMA - DMBOK: Data Management Body of Knowledge, 2nd Editio. Technics Publications, 2017.

[13] S. I. H. Shah, V. Peristeras, and I. Magnisalis, DaLiF: a data lifecycle framework for data-driven governments, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00481-3.

[14] Project Management Institute, A Guide to the Project Management Body of Knowledge, 5th Editio., vol. 34, no. 03. Project Management Institute, Inc., 2013. doi: 10.5860/choice.34-1636.

[15] D. F. Polit and C. T. Beck, "The content validity index: are you sure you know what's being reported? Critique and recommendations.," Res. Nurs. Health, vol. 29, no. 5, pp. 489–497, Oct. 2006, doi: 10.1002/nur.20147.

[16] M. R. Lynn, "Determination and quantification of content validity," vol. 35, no. 6, pp. 382–385, 1986, [Online]. Available: https://journals.lww.com/nursingresearchonline/citation/1986/11000/determination_and_quantification_of_content.17.aspx

[17] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qual. Res. Psychol., vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.

[18] J. Anderson, Data Teams: A Unified Management Model for Successful Data-Focused Teams. Reno: Apress, 2020. doi: 10.1007/978-1-4842-6228-3.

[19] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model," Procedia CIRP, vol. 79, pp. 403–408, 2019, doi: https://doi.org/10.1016/j.procir.2019.02.106.

[20] M. Z. Kastouni and A. Ait Lahcen, "Big data analytics in telecommunications: Governance, architecture and use cases," J. King Saud Univ. - Comput. Inf. Sci., 2020, doi: https://doi.org/10.1016/j.jksuci.2020.11.024.

[21] C. Ponsard, M. Touzani, and A. Majchrowski, "Combining Process Guidance and Industrial Feedback for Successfully Deploying Big Data Projects," Open J. Big Data, vol. 3, no. 1, pp. 26–41, 2017, [Online]. Available: http://www.ronpub.com/ojbd

[22] V. Sharma, A. Stranieri, J. Ugon, P. Vamplew, and L. Martin, "An agile group aware process beyond CRISP-DM: A hospital data mining case study," ACM Int. Conf. Proceeding Ser., vol. Part F1302, pp. 109–113, 2017, doi: 10.1145/3093241.3093273.

[23] J. Baijens, R. Helms, and D. Iren, "Applying Scrum in Data Science Projects," Proc. - 2020 IEEE 22nd Conf. Bus. Informatics, CBI 2020, vol. 1, pp. 30–38, 2020, doi: 10.1109/CBI49978.2020.00011.

[24] W. Qadadeh and S. Abdallah, "An improved agile framework for implementing data science initiatives in the government," Proc. - 3rd Int. Conf. Inf. Comput. Technol. ICICT 2020, pp. 24–30, 2020, doi: 10.1109/ICICT50521.2020.00012.

[25] B. van Gils, "Reference and Master Data Management BT - Data in Context: Models as Enablers for Managing and Using Data," B. van Gils, Ed. Cham: Springer Nature Switzerland, 2023, pp. 135–140. doi: 10.1007/978-3-031-35539-4_13.

[26] M. Massmann, M. Meyer, M. Frank, S. von Enzberg, A. Kühn, and R. Dumitrescu, "Framework for data analytics in data-driven product planning," Procedia Manuf., vol. 52, pp. 350–355, 2020, doi: 10.1016/j.promfg.2020.11.058.

[27] S. Angée, S. I. Lozano-Argel, E. N. Montoya-Munera, J. D. Ospina-Arango, and M. S. Tabares-Betancur, "Towards an improved ASUM-DM process methodology for cross-disciplinary multi-organization big data & analytics projects," Commun. Comput. Inf. Sci., vol. 877, no. July, pp. 613–624, 2018, doi: 10.1007/978-3-319-95204-8_51.

[28] Y. Li, M. A. Thomas, and K. M. Osei-Bryson, "A snail shell process model for knowledge discovery via data analytics," Decis. Support Syst., vol. 91, pp. 1–12, 2016, doi: 10.1016/j.dss.2016.07.003.

[29] A. Al-Badi, A. Tarhini, and A. I. Khan, "Exploring big data governance frameworks," Procedia Comput. Sci., vol. 141, pp. 271–277, 2018, doi: 10.1016/j.procs.2018.10.181.

[30] S. Ahmadi, M. M. Tavana, S. Shokouhyar, and M. Dortaj, "A new fuzzy approach for managing data governance implementation relevant activities," TQM J., vol. ahead-of-p, no. ahead-of-print, Jan. 2021, doi: 10.1108/TQM-01-2021-0015.

[31] D. Blazquez and J. Domenech, "Big Data sources and methods for social and economic analyses," Technol. Forecast. Soc. Change, vol. 130, no. March 2017, pp. 99–113, 2018, doi: 10.1016/j.techfore.2017.07.027.

[32] S. Orenga-Roglá and R. Chalmeta, "Framework for Implementing a Big Data Ecosystem in Organizations," Commun. ACM, vol. 62, no. 1, pp. 58–65, Dec. 2018, doi: 10.1145/3210752.

[33] M. Kuhn and K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press, Taylor \& Francis Group, 2020. [Online]. Available: https://books.google.com.my/books?id=wOhExQEACAAJ

[34] L. Sfaxi and M. M. Ben Aissa, "DECIDE: An Agile event-and-data driven design methodology for decisional Big Data projects," Data Knowl. Eng., vol. 130, no. September, p. 101862, 2020, doi: 10.1016/j.datak.2020.101862.

[35] S. Mathrani and X. Lai, "Big data analytic framework for organizational leverage," Appl. Sci., vol. 11, no. 5, pp. 1–19, 2021, doi: 10.3390/app11052340.

An In-Depth Analysis of Security Flaws in Advanced Authentication Protocols for the Internet of Medical Things

Haewon Byeon
Department of Future Technology, Korea University of Technology and Education (KOREA TECH), Cheonan 31253, South Korea

***Abstract*—This study evaluates a four-factor authentication protocol designed for IoT healthcare systems, identifying several key vulnerabilities that could compromise its security. The analysis highlights risks associated with node cloning, insider threats, biometric data security, session management, and scalability. To address these vulnerabilities, the study proposes a series of enhancements, including the implementation of Physical Unclonable Functions (PUFs) to prevent node cloning and the use of advanced encryption techniques, such as homomorphic encryption, to protect biometric data. Additionally, the adoption of role-based access control (RBAC) and attribute-based access control (ABAC) systems can mitigate insider threats by limiting user permissions. Optimizing session management through strict expiration and key rotation policies can maintain session integrity, while lightweight cryptographic algorithms and adaptive power management techniques enhance scalability and resource utilization. Future research directions include exploring quantum-resistant cryptographic algorithms and developing adaptive security policies leveraging artificial intelligence. These efforts are essential for maintaining the protocol's resilience against evolving threats and ensuring the secure operation of IoT-based healthcare systems.**

***Keywords*—Four-factor authentication; IoT Healthcare security; physical unclonable functions; quantum-resistant cryptography; biometric data protection**

I. Introduction

The Internet of Medical Things (IoMT) represents a significant advancement in the convergence of healthcare and technology. By interconnecting medical devices and healthcare systems through the internet, IoMT enables real-time monitoring, diagnosis, and treatment, ultimately improving patient outcomes and operational efficiency (Fig. 1). These technologies facilitate continuous patient monitoring and provide healthcare professionals with timely data, reducing the need for frequent hospital visits and enabling early intervention in critical situations [1].

As IoMT becomes more integrated into healthcare infrastructures, the security and privacy of sensitive medical data have emerged as paramount concerns. The data generated and transmitted by IoMT devices is highly sensitive, often containing personal health information that, if compromised, could lead to severe consequences for both patients and healthcare providers [2]. This necessitates robust security measures to protect against unauthorized access, data breaches, and cyber threats.

Authentication protocols play a critical role in securing IoMT systems by ensuring that only authorized users and devices can access sensitive data. Traditional authentication methods, typically based on single or dual-factor systems, have proven inadequate in the face of sophisticated cyber-attacks [3]. As a result, there is a growing need for more advanced authentication protocols that can provide comprehensive security in IoMT environments.

The development of four-factor authentication protocols presents a promising solution to these challenges [4]. By incorporating multiple authentication factors—such as knowledge-based (passwords), possession-based (smart cards), inherence-based (biometric data), and contextual (location or device-specific attributes)—these protocols offer enhanced security by requiring multiple forms of verification [5]. This layered approach significantly reduces the risk of unauthorized access, as an attacker would need to compromise all four factors to gain entry.

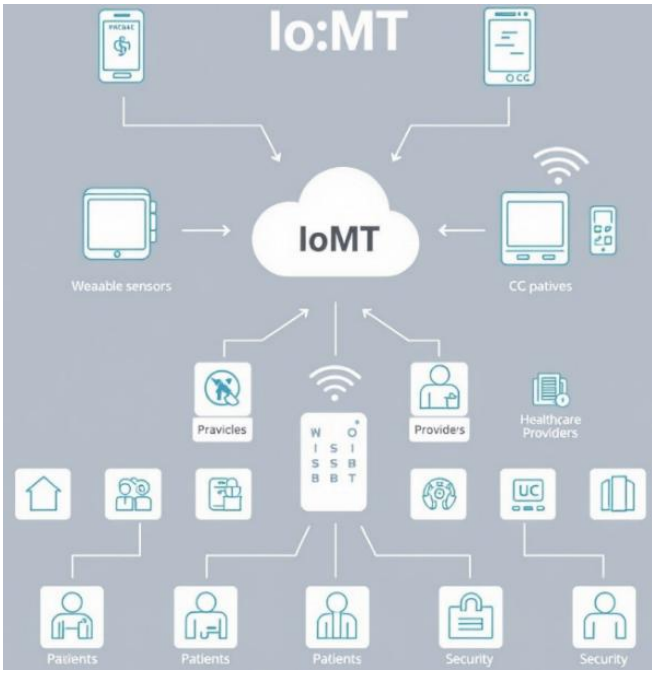


Fig. 1. The concept of Internet of Medical Things (IoMT).

Despite their potential, implementing four-factor authentication protocols in IoMT systems presents several challenges. The resource-constrained nature of many IoMT

devices limits their ability to execute complex cryptographic operations, necessitating the development of efficient, lightweight protocols. Additionally, the dynamic nature of healthcare environments requires authentication systems to be adaptable, maintaining security even as devices frequently join and leave the network. Furthermore, ensuring compliance with stringent healthcare regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), is critical for maintaining patient trust and protecting healthcare providers from legal liabilities.

This research aims to conduct a comprehensive analysis of a proposed four-factor authentication protocol designed for IoMT [5]. The primary objective is to identify and scrutinize five key security vulnerabilities within the protocol that could compromise its effectiveness. By examining the protocol's architecture and operational phases, this study seeks to uncover potential weaknesses and propose strategies for enhancement.

The paper is structured as follows: Section I presents a literature review of existing authentication protocols and their limitations. Section II outlines the methodology employed for vulnerability detection, including the analytical methods and tools used for cryptanalysis and security testing. Section III discusses the identified vulnerabilities in detail, while Section IV proposes improvements to enhance the protocol's security and efficiency. Finally, Section V and VI concludes with a summary of the findings and their implications for IoMT authentication systems. Through this investigation, the study aims to contribute to the development of more secure and efficient authentication protocols that effectively protect sensitive medical data while supporting the continued integration of IoMT in healthcare.

A. Literature Review

1) Existing authentication protocols: The proliferation of the Internet of Things (IoT) in healthcare has necessitated the development of robust authentication protocols to secure sensitive medical data. Traditional single-factor and two-factor authentication methods have proven inadequate against sophisticated cyber threats, prompting the exploration of more comprehensive solutions [6]. In response, four-factor authentication protocols have emerged, integrating multiple layers of security to mitigate unauthorized access and data breaches [7].

Four-factor authentication protocols typically combine knowledge-based factors (e.g., passwords), possession-based factors (e.g., smart cards), inherence-based factors (e.g., biometric data), and contextual or environmental factors (e.g., device location or usage patterns) [8]. This multi-layered approach significantly enhances security by requiring multiple forms of verification, thereby reducing the likelihood of successful attacks. For instance, even if an attacker obtains a user's password, they would still need to overcome other authentication barriers to gain access [9].

Recent advancements in authentication technologies have focused on optimizing these protocols for resource-constrained IoT environments. Lightweight cryptographic algorithms, such as elliptic curve cryptography (ECC) and lightweight encryption standards, have been employed to balance security and efficiency [10]. These algorithms provide strong encryption with reduced computational overhead, making them suitable for IoT devices with limited processing power and energy resources.

Despite their strengths, existing four-factor authentication protocols face several challenges. The integration of multiple authentication factors can lead to increased complexity and user friction, potentially impacting usability and compliance [11]. Additionally, the dynamic nature of healthcare environments, where devices frequently join and leave the network, requires authentication systems to be adaptive and resilient [12]. Ensuring user privacy and compliance with stringent healthcare regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), remains a critical consideration.

2) Related works: Recent studies [13, 14] have highlighted various security vulnerabilities within IoT-based healthcare authentication systems. One major concern is the risk of sensor node capture attacks, where adversaries physically access devices to extract sensitive information [13]. To address this, researchers have proposed the use of Physical Unclonable Functions (PUFs) as an additional layer of security [14]. PUFs leverage the intrinsic physical variations of semiconductor devices to generate unique identifiers, making them resistant to cloning and physical attacks.

Another significant vulnerability is the potential for replay attacks, where attackers intercept and retransmit valid authentication messages to deceive the system [15]. To mitigate this risk, protocols have incorporated time-stamping and nonce-based mechanisms to ensure message freshness and authenticity [16]. These measures help prevent attackers from reusing captured messages to gain unauthorized access.

Previous studies [17, 18] also underscores the importance of protecting biometric data, which, once compromised, cannot be easily replaced or revoked [17]. Advanced encryption techniques and secure transmission protocols have been proposed to safeguard biometric information against interception and misuse [18].

Interoperability and scalability are additional challenges that must be addressed to ensure seamless integration and operation of IoT healthcare systems [19]. The diverse range of IoT devices and platforms necessitates standardized protocols that can facilitate communication and data exchange across different systems.

In summary, the literature emphasizes the need for continued research and innovation in the development of secure, efficient, and user-friendly authentication protocols for IoT healthcare systems. By addressing the identified vulnerabilities and challenges, the healthcare industry can enhance the security and privacy of patient data, ultimately improving trust and outcomes in digital healthcare environments.

II. METHODOLOGY

A. Framework for Analysis

The methodology for analyzing the proposed four-factor authentication protocol for the IoMT [5] involves a comprehensive framework designed to identify and evaluate potential security vulnerabilities. This framework integrates theoretical analysis, mathematical modeling, and practical testing to ensure a thorough security assessment.

B. Analytical Methods for Vulnerability Detection

The initial phase of analysis involves a detailed examination of the protocol's architecture, focusing on its critical components: initialization, mutual authentication, session key establishment, and data protection. Formal security models and logical proofs are employed to assess the protocol's resilience against various attack vectors. One such method is the application of Burrows-Abadi-Needham (BAN) logic, which helps verify the authenticity and freshness of messages exchanged within the protocol:

$$P \mid\equiv X \quad (P \text{ believes } X)$$

$$P \mid\Rightarrow X \quad (P \text{ has jurisdiction over } X)$$

$$(X) \quad (X \text{ is fresh})$$

These logical expressions formalize the assumptions made during the protocol's execution, ensuring that it meets its intended security objectives.

C. Cryptanalysis and Security Testing Tools

For cryptanalysis, the study utilizes advanced tools such as the Automated Validation of Internet Security Protocols and Applications (AVISPA). AVISPA provides a platform for modeling the protocol's interactions and simulating potential attacks, including replay, man-in-the-middle, and impersonation attacks. The tool employs a high-level protocol specification language (HLPSL) to evaluate the protocol's security properties.

In addition, the Random Oracle Model (ROM) is used to ensure the integrity of cryptographic functions, particularly for verifying hash-based operations and ensuring collision resistance. This involves modeling cryptographic hash functions as random oracles to simulate their ideal behavior:

$$H(m) = RO(m)$$

where $H(m)$ represents the hash of message (m), and (RO) denotes the random oracle.

D. Mathematical Modeling of Protocol Operations

The analysis incorporates mathematical modeling to evaluate key cryptographic operations, such as key generation and exchange. The protocol employs elliptic curve cryptography (ECC) for secure key exchanges:

$$K_{session} = g^{ab}p$$

where (g) is a generator point, and (a) and (b) are private keys of the communicating entities. This session key ensures secure communication between devices.

E. Analysis of the Proposed Protocol

The proposed four-factor authentication protocol for IoT healthcare systems [5] is designed to provide robust security through a multi-layered authentication process, incorporating initialization, mutual authentication, session key establishment, and data protection phases.

III. OVERVIEW OF THE PROTOCOL

A. Initialization Phase

The protocol begins with the initialization phase, where cryptographic parameters are established. Each device generates a unique identifier and a corresponding key pair using elliptic curve cryptography (ECC). The public key is calculated as:

$$K_{public} = g^{k_{private}}p$$

where (g) is the generator point on the elliptic curve, $(k_{private})$ is the private key, and (p) is a prime modulus. This setup ensures that each device has a secure cryptographic foundation for subsequent operations.

B. Mutual Authentication Phase

During this phase, devices authenticate each other using a combination of ECC and hash functions. Each device computes an authentication token by hashing its identity and a session-specific random nonce:

$$AuthToken = h(ID \parallel Nonce)$$

where *h( )* is a secure hash function. This token verifies the device's authenticity, ensuring that only legitimate devices can participate in the network.

C. Session Key Establishment

Once mutual authentication is successful, a secure session key is established between the device and the server. The session key is derived from the ECC-based key exchange process:

$$K_{session} = g^{ab}p$$

where (a) and (b) are private keys of the communicating entities. This ensures that each session is secured with a unique key, providing confidentiality and integrity for data exchanged between the device and the server.

D. Data Protection Phase

The final phase involves securing data transmission using the established session key. Messages are encrypted as follows:

$$C = E_{K_{session}}(M)$$

where (C) is the ciphertext and (M) is the plaintext message. This encryption ensures that sensitive medical data remains confidential and protected from unauthorized access.

E. Identified Vulnerabilities

Vulnerability to Node Impersonation using Session Key Leakage

1) Problem: While the paper emphasizes mutual authentication, it inadvertently provides a vector for node impersonation. An adversary can intercept public messages and

compute the session key.

a) The paper states that the gateway computes session key sks = (sk⊕ N3)⊕ R3, which implies a dependency of the sensor's session key on the gateway's variables and exposes a way to derive it. Note that the doctor also derives the key as sk * = (sk* ⊕ N2) ⊕ PWD.

b) The adversary can intercept M and derive R to compute the session key using sk which has been used by gateway to produce sks.

2) Impact: The adversary can impersonate legitimate nodes, leading to unauthorized access, data manipulation, or disruption of medical services.

3) Mathematical formulation: The paper states that the doctor will calculate session key sk as sk * = (sk* ⊕ N^2) ⊕ PWD. If we take another perspective of generating session key, the gateway calculates the sensor node session key sks = (sk⊕ N)⊕ R, where sk is the session key, N is a random number and R is another derived random number. But since the attacker has eavesdropped and learned G' which has been generated as G = RG (RSNRG) + N, which will help the attacker to compute R= and the session key sks by guessing the password.

4) Improvement: Session keys must be computed with a hash operation using all participating entities secrets, a key agreement approach must be followed to make it secure for each party. Also, public key cryptography can be considered to enhance security for all involved entities.

F. Password Vulnerability (Re-used Random Numbers)

1) Problem: The paper states that HPWD = h(PWDR). If an attacker gets the value HPWD for a user then, it is possible to guess the password using a brute force method. The paper doesn't specify a different process of updating the password, which means there is a chance that the random values used with the password remain constant, leading to the possibility that HPWD value of the user remains static. As a result, it becomes vulnerable to attacks such as dictionary or brute force password guessing.

2) Mathematical formulation: Since random number R, remains the same during each authentication process, if the attacker compromises HPWD and knows R, he can use brute force attack to guess PWD. For example, when an attacker tries all password values from his/her dictionary list and computes the resultant HPWD value from HPWD = h(PWD || R), when they have achieved the matching value, the password of the user is compromised.

3) Impact: A compromised password leads to full control over the user account and the ability to compromise the system.

4) Improvement: Each user has to generate a new random number with each new session. A password update process must be added and made mandatory to change the password periodically. This ensures that the password is not leaked and remains fresh. Multi-factor authentication can also help to prevent password leakage attacks.

G. Susceptibility to Impersonation and Data Modification Attacks

1) Problem: The protocol, as written, is susceptible to impersonation and data modification attacks since, all information such as temporary keys and identities are exchanged over a public channel, and those values can be used by an attacker to impersonate. An attacker can intercept and use public messages to act as legitimate users or devices.

2) Mathematical formulation: The protocol does not use Message Authentication Codes (MAC) or digital signatures. Suppose the sensor node receives a message {G, DTID,GW, Gw, and SKs }, and the user receives {μ,η,sku,GW}. The attacker can intercept any of these messages, and they can either replay, drop, or modify them and, since message integrity is not protected, the recipient has no way to identify it.

3) Impact: An adversary can intercept and replay, drop, or modify messages. This can lead to several types of attacks.

a) An attacker can act as a gateway and can obtain the data, and can then, modify or drop the information.

b) An attacker can manipulate the information received by the user device and the sensor node device.

4) Improvement: Adding MACs or digital signatures to each message transmitted or received can ensure integrity. Moreover, an additional time stamp or nonce can also be used to verify the freshness of the messages.

H. Lack of Proper Revocation Mechanism

1) Problem: While the paper mentions a revocation phase, it doesn't specify how devices are removed or deactivated when compromised, nor how the data or secret keys associated with these nodes are handled. The removal process lacks proper security and can be exploited by adversaries if they have compromised a valid user. Also, the addition of new nodes can be attacked if an intruder starts acting as a legitimate user or a sensor node. This leaves the system vulnerable to a situation where compromised nodes can persist.

2) Impact: The lack of proper revocation can compromise the integrity and reliability of the whole system. Once a node is compromised, it should not be able to perform future operations or compromise the network. If proper revocation methods are not implemented, an attacker can easily compromise the whole system.

3) Improvement: A well-defined, secure, and well-implemented revocation mechanism is essential for managing compromised nodes. The security mechanism must delete/remove all previous keys and credentials from the network and the node. The new node added to the system should have different identities, credentials and keys, different from the compromised one.

This paper presented an approach towards secure authentication in the medical IoT field. The proposed protocol has potential to provide a secure environment and is more lightweight compared to existing solutions. However, the vulnerabilities mentioned above can be exploited by an adversary and can lead to severe security breaches. Addressing these flaws by implementing stronger cryptographic practices, adding proper message integrity checks, and better key

management practices is necessary for ensuring real-world security for medical devices.

4) Proposed improvements: To enhance the security and performance of the proposed four-factor authentication protocol for IoT healthcare systems, several improvements are recommended. These enhancements aim to address the identified vulnerabilities and ensure robust protection against unauthorized access and data breaches.

IV. Security Enhancements

A. Node Cloning/Replication Prevention

To mitigate the risk of node cloning attacks, the protocol should incorporate mechanisms for device attestation and verification. This could involve using Physical Unclonable Functions (PUFs) to generate unique, hardware-based identifiers for each device. PUFs leverage the inherent physical variations of semiconductor devices to create unclonable identifiers, providing a strong defense against node replication attacks. Additionally, implementing regular device health checks and network monitoring can help detect and respond to unauthorized node activity.

B. Strengthening Insider Attack Defenses

To protect against insider threats, the protocol should enhance its access control measures by implementing role-based access control (RBAC) and attribute-based access control (ABAC) systems. These systems limit user permissions based on their roles and attributes, ensuring that users can only access the resources necessary for their duties. Furthermore, incorporating user behavior analytics and anomaly detection can help identify suspicious activities and potential insider threats, enabling timely intervention.

C. Biometric Data Security

Enhancing the security of biometric data involves implementing advanced encryption techniques and secure transmission protocols. Homomorphic encryption allows computations to be performed on encrypted data without revealing the data itself, providing strong protection for biometric information. Additionally, secure multi-party computation (SMPC) can enable collaborative data processing while preserving privacy. These techniques ensure that biometric data remains confidential and protected against interception during transmission and storage.

D. Improving Session Management

To prevent unauthorized access and maintain session integrity, the protocol should implement strict session expiration and key rotation policies. Session keys should be refreshed regularly, and sessions should automatically expire after a predetermined period of inactivity. This approach reduces the risk of session hijacking and ensures that only active, authenticated sessions are maintained. Furthermore, incorporating ephemeral key exchanges can provide additional security by ensuring that session keys are unique and short-lived.

E. Optimizing Scalability and Resource Utilization

Addressing scalability and resource constraints requires optimizing cryptographic operations to reduce computational overhead. Implementing lightweight cryptographic algorithms, such as the Advanced Encryption Standard (AES) in its lightweight form, can significantly decrease energy consumption and processing time. Additionally, employing adaptive power management techniques, such as duty cycling and dynamic voltage scaling, can extend battery life and maintain device performance. These optimizations ensure that IoT devices can efficiently perform necessary tasks without draining resources.

V. Future Research Directions

A. Exploring Quantum-Resistant Cryptography

As quantum computing capabilities advance, exploring quantum-resistant cryptographic algorithms will be essential to future-proof the protocol against potential quantum threats. Algorithms such as lattice-based cryptography and hash-based signatures could offer security that withstands quantum attacks, ensuring long-term protection of sensitive data.

B. Developing Adaptive Security Policies

Leveraging artificial intelligence and machine learning to develop adaptive security policies could provide real-time threat detection and response. By analyzing network traffic patterns and user behavior, AI-driven systems can dynamically adjust security measures, offering a proactive approach to maintaining security.

C. Enhancing Privacy-Preserving Technologies

Further exploration of privacy-preserving techniques, such as differential privacy and secure enclave technologies, will enhance user privacy and data protection. These techniques can ensure that sensitive information remains confidential, even in the face of sophisticated attacks.

D. Addressing Interoperability Challenges

Research into standardized protocols and cross-platform communication frameworks could facilitate seamless integration and operation of diverse IoT healthcare systems. Ensuring interoperability across different devices and platforms is essential for creating a cohesive and efficient IoT ecosystem.

By implementing these proposed improvements and pursuing ongoing research directions, the protocol can provide robust protection against evolving threats and maintain the integrity and confidentiality of IoT-based healthcare systems. These efforts offer a comprehensive approach to addressing current vulnerabilities and preparing for future security challenges in the digital healthcare landscape.

VI. Conclusion

In conclusion, the analysis of the proposed four-factor authentication protocol for IoT healthcare systems has identified several critical vulnerabilities, including node cloning risks, insider threats, biometric data security concerns, session management weaknesses, and scalability challenges. By implementing the recommended security enhancements, such as the use of Physical Unclonable Functions, advanced encryption techniques, and optimized session management, the protocol can significantly improve its security posture. These

measures ensure robust protection of sensitive medical data, fostering trust in IoT-enabled healthcare environments. As the IoT landscape continues to evolve, ongoing research into quantum-resistant cryptography and adaptive security measures will be crucial in maintaining the protocol's resilience against emerging threats.

ACKNOWLEDGMENT

This research through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- RS-2023 00237287).

REFERENCES

[1] Monteiro, A. C. B., França, R. P., Arthur, R., and Iano, Y. An overview of the internet of medical things (IoMT): Applications, benefits, and challenges. Security and Privacy Issues in Internet of Medical Things, pp. 83-98, 2023.

[2] Hireche, R., Mansouri, H., and Pathan, A. S. K. Security and privacy management in Internet of Medical Things (IoMT): A synthesis. Journal of Cybersecurity and Privacy, vol. 2, no. 3, pp. 640-661, 2022.

[3] Khatiwada, P., and Yang, B. An overview on security and privacy of data in IoMT devices: performance metrics, merits, demerits, and challenges. pHealth 2022, pp. 126-136, 2022.

[4] Singh, A. K., and Garg, A. Authentication protocols for securing IoMT: current state and technological advancements. Securing Next-Generation Connected Healthcare Systems, pp. 1-29, 2024.

[5] Khajehzadeh, L., Barati, H., and Barati, A. A lightweight authentication and authorization method in IoT-based medical care. Multimedia Tools and Applications, pp. 1-40, 2024.

[6] Suleski, T., and Ahmed, M. A data taxonomy for adaptive multifactor authentication in the internet of health care things. Journal of Medical Internet Research, vol. 25, p. e44114, 2023.

[7] Rangwani, D., and Om, H. Four-factor mutual authentication scheme for health-care based on wireless body area network. The Journal of Supercomputing, vol. 78, no. 4, pp. 5744-5778, 2022.

[8] Aburbeian, A. M., and Fernández-Veiga, M. Secure Internet Financial Transactions: A framework integrating multi-factor authentication and Machine Learning. AI, vol. 5, no. 1, pp. 177-194, 2024.

[9] Awadh, W. A., Alasady, A. S., and Hashim, M. S. A multilayer model to enhance data security in cloud computing. Indonesian Journal of Electrical Engineering and Computer Science, vol. 32, no. 2, pp. 1105-1114, 2023.

[10] Thakor, V. A., Razzaque, M. A., and Khandaker, M. R. Lightweight cryptography algorithms for resource-constrained IoT devices: A review, comparison and research opportunities. IEEE Access, vol. 9, pp. 28177-28193, 2021.

[11] Rangwani, D., and Om, H. A robust four-factor authentication protocol for resource mining. Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 1947-1971, 2023.

[12] Sodhro, A. H., Awad, A. I., van de Beek, J., and Nikolakopoulos, G. Intelligent authentication of 5G healthcare devices: A survey. Internet of Things, vol. 20, p. 100610, 2022.

[13] Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., and Hossain, M. S. An intelligent healthcare system using IoT in wireless sensor network. Sensors, vol. 23, no. 11, p. 5055, 2023.

[14] Al-Meer, A., and Al-Kuwari, S. Physical unclonable functions (PUF) for IoT devices. ACM Computing Surveys, vol. 55, no. 14s, pp. 1-31, 2023.

[15] Nadeem, M., Arshad, A., Riaz, S., Zahra, S. W., Dutta, A. K., and Almotairi, S. A secure architecture to protect the network from replay attacks during client-to-client data transmission. Applied Sciences, vol. 12, no. 16, p. 8143, 2022.

[16] Babu, E. S., Dadi, A. K., Singh, K. K., Nayak, S. K., Bhoi, A. K., and Singh, A. A distributed identity-based authentication scheme for internet of things devices using permissioned blockchain system. Expert Systems, vol. 39, no. 10, p. e12941, 2022.

[17] Yang, W., Wang, S., Sahri, N. M., Karie, N. M., Ahmed, M., and Valli, C. Biometrics for internet-of-things security: A review. Sensors, vol. 21, no. 18, p. 6163, 2021.

[18] Mihailescu, M. I., and Nita, S. L. A searchable encryption scheme with biometric authentication and authorization for cloud environments. Cryptography, vol. 6, no. 1, p. 8, 2022.

[19] Pathak, N., Misra, S., Mukherjee, A., and Kumar, N. HeDI: Healthcare device interoperability for IoT-based e-health platforms. IEEE Internet of Things Journal, vol. 8, no. 23, pp. 16845-16852, 2021.

Enhancing Patient Health Through Smart IoT Technologies in Healthcare

Monica Bhutani[1]*, Osman Elwasila[2], Rajermani Thinakaran[3], Yonis Gulzar[4]*
Bharati Vidyapeeth's College of Engineering, New Delhi, India[1]
Department of Management Information Systems-College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia[2, 4]
Faculty of Data Science and Information Technology, INTI International University, Nilai, 71800, Negeri Sembilan, Malaysia[3]

***Abstract*—Health care has been revolutionized by this rapid change in the field of the Internet of things that enable smart connected devices that provide better patient monitoring, diagnosis and treatment. IoT technologies facilitate collection of real time health data and remote patient monitoring, as well as prediction, thus improving overall healthcare outcomes. The ideas and concepts related to Chronic diseases, Diseases, Emergency, Detection and its management are greatly transformed by the presence of Wearable sensors, Smart Hospital infrastructures and AI powered Analytics. Meanwhile, healthcare systems driven by IoT are more efficient, reduce number of hospital readmissions, and provide telemedicine services. But, IoT in healthcare comes with a lot of challenges such as IoT security risks, patient privacy issues and multitude of interoperability problems. In this paper a comprehensive review of smart IoT technologies in healthcare, their applications and the type of benefits that weigh them for the patient care is provided. Along with exploring the role of data analytics in the IoT based decision making, data handling ethical implications and security threats in IoT healthcare systems introduced here. Moreover, beyond this, they discuss future directions including integration of AI, 5G enabled telemedicine, and blockchain for secure patient data management. This makes IoT an ideal candidate for healthcare transformation — one that addresses existing challenges and capitalize on emerging innovations to go from more efficient, more accessible, and definitely more patient-centric healthcare.**

Keywords*—IoT in Healthcare; smart technologies; remote patient monitoring; data security; AI in healthcare; telemedicine; healthcare analytics

I. Introduction

IoT integration in healthcare has revolutionized patient care by making patient care smart and connected systems that offer the real time monitoring, the diagnostic accuracy and treatment efficiency. The IoT offers the global healthcare systems a way out of the increasing problems like increased aged populations, high cost of living and prevalence of chronic conditions. The technology in an IoT driven world enables healthcare professionals to remotely monitor patients, real time analysis of health data and user interface to make decisions that result in prompt interventions. Such innovations in wearable sensors and smart hospital infrastructures, to AI powered analytics, brings a fresh perspective to the healthcare, offering personalized, data driven treatments and improved overall medical efficiency [1]. These technologies are also benefiting those who receive them, since they permit patients to conduct continuous health monitoring, reduce hospital visits, and push proactive health management.

IoT in health exist in the universe of connection devices, cloud computing and artificial intelligence coexisting to construct an intelligent health system. By wearing smartwatches and biosensors, ones the most critical health parameters like heart rate, blood pressure, oxygen levels, blood glucose levels are being tracked and these critical parameters can give real time insights to concern one medical condition. RPM has also evolved into a go to for patients with chronic illnesses, and on top of that, remote patient monitoring (RPM) can do just that.. IoT helps healthcare professionals keep track of patients at a distance and prevent hospital admissions, so that the required medical attention is given before they are in a worse condition. Another very major application of IoT is that of telemedicine where the IoT is used to bridge the gap between patients and the healthcare providers, making it easier and faster for medical consultations [2].

Due to the adoption of IoT in hospitals, medical environments with a smart approach have emerged that are based on automation for improving medical workflow process. Smart beds in the hospitals connected via IoT, automated medication dispensers and RFID based inventory management systems are offered with improved efficiency and lesser human error and thereby higher patient safety. The impact of IoT and artificial intelligence comes to play even bigger as artificial intelligence further takes advantage of the high volume of patient data to which it has access to predict health trends, personalize treatment plan and identify at risk patients earlier than critical situation happens [3]. Through implementing IoT and AI and machine learning into the healthcare systems, healthcare systems can be tipped towards predictive and preventive medicine and ultimately reduce healthcare costs and improve the welfare of patients.

Despite its huge potential, implementing IoT in healthcare sector poses problems, specifically concerning security, privacy and interoperability [4]. Cyberattacks and data breaches on connected medical devices have become an increasing threat as reliance on connected medical devices grows. The patient health data that has to be processed here is very sensitive and any compromise in the security would have disastrous consequences. To secure patients' trust and data on the patients' personal health information, the providers of healthcare should also adhere to strict data protection regulations such as HIPAA

Corresponding authors: monica.bhutani@bharatividyapeeth.edu(M.B.); ygulzar@kfu.edu.sa (Y.G.)

and GDPR. The second issue with IoT devices is the tremendous variety of standards and lack of standardization; this poses problems for compatibility between various manufacturers as well as with system. In the absence of universal standards on integrating IoT devices seamlessly into current healthcare framework, it remains a big hurdle. Furthermore, the cost of taking IoT to the healthcare industry is an impediment for most of the healthcare institutions, including developing regions. This makes it even more difficult for widespread adoption as you need advanced infrastructure, high-speed internet and bear the heavy burden of secure cloud data.

And as IoT is still evolving, future innovations will solve these issues and open several new doors for healthcare. One would expect 5G technology would greatly improve the speed and reliability of the IoT enabled healthcare applications with the ability to transmit data in real time as well as to conduct remote surgeries. In addition, blockchain technology is also attracting a solution to enhance data security and to keep the health record transcripts accurate, transparent and HALON free [5]. Additionally, AI driven IoT systems will improve upon predictive analytics to a degree that will enhance the accuracy and porosity of personalized medicine. As edge computing grows, data processing will become more and faster, with less latency in important cases of healthcare.

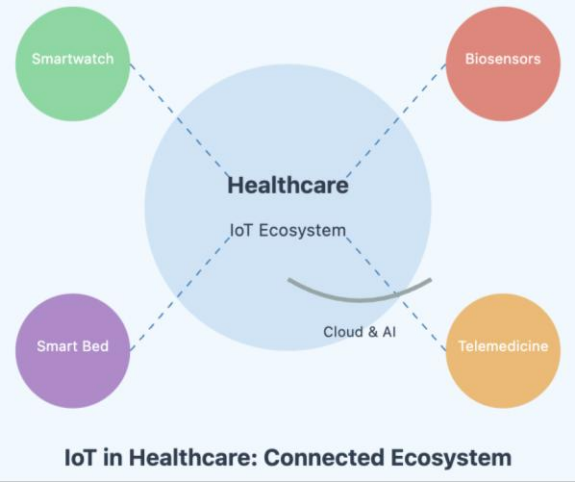


Fig. 1. IoT Healthcare ecosystem - Interconnected devices transforming patient care.

It is in Fig. 1 that the Internet of Things (IoT) has the potential to transform healthcare, as a dynamic ecosystem of connected technologies, represented in Fig. 1. Four key healthcare IoT devices, namely smartwatches, biosensors, smart hospital beds and telemedicine centers, respectively interlock around a main healthcare hub to indicate the omnidirectional and integrated concept of modern medical service. IoT based healthcare system is represented by the dotted line which shows the flow in the data and communication flow between the devices without interruption. Essentially, Cloud & AI technologies are the intelligent brains that take real time health data, process and analyze it to make it possible for predictive diagnostics, personalized treatment and proactive health management. Essentially, this paper's main argument is encapsulated in this visualization: IoT is revolutionizing healthcare by making it a much more connected, streamlined and patient centered way of monitoring and intervening in medical care [6].

In this paper, we discuss how smart IoT technologies can be applied to benefit the healthcare and the challenges that are impacted due to their implementation. IoT in Realtime Health Monitoring, Predictive Analytics and Remote Patient Care explores the role of IoT in real time health monitoring, predictive analytics and remote patient care by discussing how these are changing the modern healthcare. Another part of the study looks into security risks and ethical considerations that are associated with IoT driven medical systems and recommends building strong regulatory framework. Finally, it concludes with the future direction that IoT will take on in the healthcare domain, especially with the coming of the technologies that are AI, blockchain, and 5G that will add to the efficiency and security of the IoT powered healthcare systems. Stakeholders will be able to leverage smart IoT technologies to create better, more connected, efficient and patient-centric health care ecosystem by knowing opportunities and challenges.

The novel contributions of this paper are:

- In this paper we showcase a detailed description of the impact that IoT technologies bring to healthcare by real time patient monitoring, smart hospital infrastructure, and with the help of AI diagnostics. Unlike previous reviews that are focused on individual aspects of IoT in healthcare, this work presents the synergy between wearable devices, telemedicine, predictive analytics and hospital automation.
- To date, as adoption of IoT in healthcare surges, security and ethical matters continue to severely impede IoT adoption. This paper assesses emerging cybersecurity threats, data privacy risks and ethical assessment of IoT healthcare applications in a novel way. In this study, gaps exist in existing studies, and it proposes a framework that is based on a security through blockchain integration combined with AI based anomaly detection and regulatory compliance.
- The idea of this research is to explore how artificial intelligence helps in IoT based healthcare by making it possible to predict a disease, do an automated decision making and suggest personalised treatment. Unlike the normal studies that treat AI and IoT as two separated things, this paper focuses on the merged potential of the AI and IoT in disease forecasting, patient risk evaluation, and personalized healthcare interventions.
- Next, this study provides a forward view on the future of IoT in healthcare and how the next generation technologies like 5G, edge computing, and blockchain overcome the current one. This paper, by presenting a structured roadmap, is a guide for the practitioners of healthcare, researchers and policymakers to optimize IoT implementations to improved patient outcomes and system efficiency.

This paper is organized as follows. Section II describes IoT in healthcare through a general discussion of its architecture, a list of its key components, and the associated technological

enablers. In Section III, we cover on the various application of smart IoT technologies in the healthcare area: wearable health monitoring, telemedicine, and hospital automation. In Section IV, IoT's role in healthcare data analytics and decision making is discussed including AI driven predictive modelling and real time health insights. IoT driven Healthcare System: Section V studies the IoT driven health care systems security, privacy, and ethical challenges showing risks and their counter strategies. In the last Section VI, the conclusions of this study are drawn, a summary of the main findings and their implications are presented. In the last section, Section VII portrays the future directions and what these emerging technologies, like 5G and edge computing and blockchain, bring to the future of IoT based healthcare solutions.

II. Overview of IoT in Healthcare

The Internet of Things (IoT) has revolutionized healthcare, as it links medical devices with doctors and patients in the most efficient way or you can say it enhances the connectivity and makes the link seamless. Integration of smart sensors, real time data analytics and cloud based platforms has made a remarkable difference to patient care, operational efficiency and medical decision making through IoT. The arrival of IoT driven healthcare solutions has led to new possibilities of real-time monitoring, predictive diagnostics and remote treatment, which has changed the way healthcare was delivered. Wireless communication, artificial intelligence, as well as big data analytics advancements drive the growing adoption of IoT in healthcare that make medical practice automated as it relies on data [7]. However, this further increased the demand for IoT oriented system that provides proactive and efficient healthcare services as a result of the increasing demand for personalized healthcare and the growing prevalence of the chronic diseases.

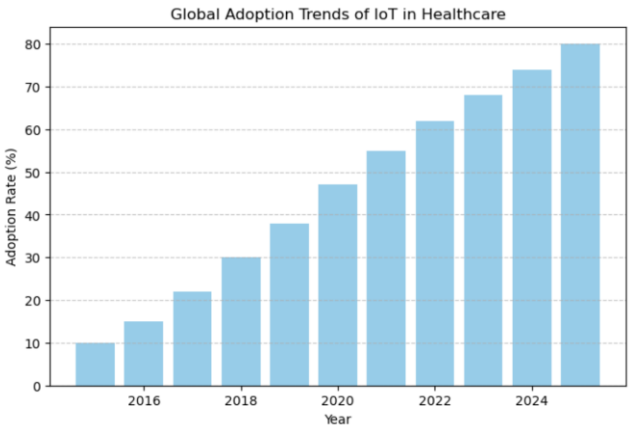


Fig. 2. Global adoption trends of IoT in healthcare (2016-2024).

Fig. 2 shows a very steep upward trajectory of Internet of Things (IoT) in healthcare from 2016 to 2024. The graph starts at around 10-15 % adoption rate in 2016 and progressively and greatly increases to almost 75-80 % in 2024. This visualization powerfully showcases how only one decade of healthcare developments has seen excessive expansion of the IoT technologies in this area. The steady year on year growth is on account of the rising perception of IoT's revolutionary capability in medical monitoring, patient treatment, and health system adaptation, as similarly observed in agriculture sector adoption [8]. But each successive bar bears not only technological advancement but also waning of institutional initial lack of trust in IoT's potential to improve diagnostic precision, patient outcomes, and operational efficiency in healthcare institutions. While a compelling visual testament to what is occurring in the global space where healthcare systems are all going through a digital transformation around innovations in connected medical technologies, the graph also sets the tone, period.

A. Architecture and Components of IoT in Healthcare

IoT healthcare architecture is responsible for multiple connected layers working hand in hand to deliver the information seamlessly among the patient, medical devices, healthcare providers, and the cloud computing platform. The basic premise of this ecosystem lies in smart sensors and medical devices that are producing patient health data like heart rate, blood pressure, oxygen level and glucose levels continuously. All of these devices are embedded with wireless communication technologies – Bluetooth, Wi-Fi or 5G – to be able to transmit data in real time. The collected data is then processed through the edge computing or cloud computing systems where the artificial intelligence algorithms are used to analyze patterns, identify anomalies and generate actionable insights. This data can be accessed by healthcare professionals through hospital information systems, mobile applications, and can be used for medical interventions at the time occasioned and remote patient monitoring.

Interoperability is an important part of IoT in health care as it supports between different medical devices and systems. The standardized protocols and secure APIs permit no language and no platform barrier of exchanging the information between different platforms and hence removing the risk of information silos and making the medical records in better accuracy. Further, cloud computing contributes to storing and managing huge amounts of health data as needed, thereby, making it real time available to the medical professionals. In addition to improving data security, data security provides a tamper proof ledger for medical records in IoT healthcare system blockchain integration adds further security. Because healthcare providers want to be able to scale IoT in their system, the overall architecture of IoT in healthcare is designed to be scalable, keeping new technology from interfering with existing workflows.

B. Role of IoT in Healthcare Transformation

IoT is one of the key drivers of transforming the classic healthcare eco-system to be automated, to be involved in remote observation, and to create usage of predictive analytics. The remote patient monitoring (RPM) is one of the most important pieces of the IoT in healthcare because it lets patients treat their medical issues in the convenience of their own houses [9]. More specifically, this technology is valuable to persons with chronic illnesses like diabetes, hypertension, and respiratory disorders since it can track vital signs on a continuous basis without needing to visit the hospital frequently. Patient health data is collected by wearable devises and biosensors and transmitted to healthcare providers for the purpose of making informed decisions and adjusting treatment plans. Not only does this improve patient outcomes but it also lessens the burden on the hospitals by minimizing hospital readmission.

IoT also has an impact on emergency medical response, which is another kind of transformation. Smart ambulances have real time communication systems that enable the transfer of data on patient to hospitals before the patient arrives, so that medical teams are able to be ready for immediate treatment. In case of cardiac arrests or strokes, real time monitoring and auto alerts could help a lot in survivals by allowing rapid medical interventions. Apart from that, it has improved hospital management by optimizing resource allocation, tracking medical equipment, and proper flow of patients. IoT for Smart hospitals rely on using IoT enabled RFID tags to supervise the availability of beds in a hospital and also the medical inventory and streamline administrative processes to enhance overall efficiency in healthcare services.

The good news is that besides patient monitoring and hospital automation, IoT has also started to change drug management and personalized medicine [10]. IoT-enabled smart pill dispensers can provide alerts to patients when to take their medication at the accurate time and decrease chances of missed doses and medication errors. Along with this, predictive analytics powered by IoT help the pharmaceutical companies monitor the efficacy of the prescribed treatments in real time and enable the development of personalized medicine targeted at individual patient needs. Using IoT insights, the healthcare providers can now take a more proactive approach towards disease prevention and move away from reactive treatment to the preventive care strategies.

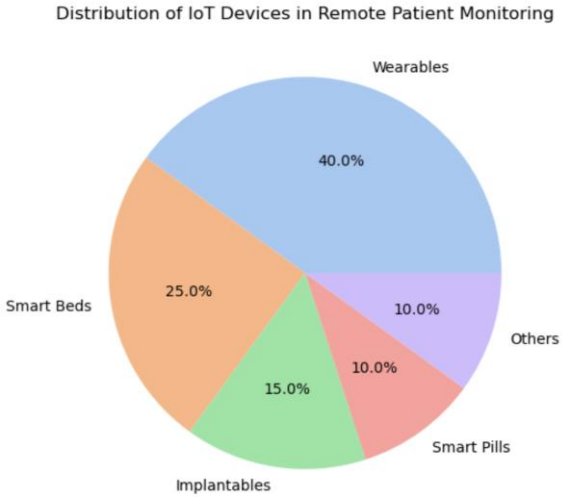


Fig. 3. Distribution of IoT devices in remote patient monitoring.

Fig. 3 shows a complete classification of IoT device types used in remote patient monitoring, revealing the general research environment for the modern healthcare. 40 percent of IoT devices are wearables due to its popularity along with ease in continuous health monitoring. About 25 percent of the ecosystem are smart beds, and these serve a critical purpose of monitoring patient conditions in hospital and home care settings. The 15 percent of the distribution of implantable devices indicates that medical technology is becoming increasingly sophisticated at measuring health from within the body [11]. The shares of smart pills and all other miscellaneous IoT devices is 10% each, representing emerging and innovative measurement of the patients. In the chart, you can see the many ways that IoT in healthcare works — for example, multiple technologies being combined to provide complete, real time patient care for many different medical contexts.

C. Challenges and Barriers to IoT Adoption in Healthcare

Although IoT has so many benefits, there are hurdles that block its widespread adoption in healthcare settings. Data security and privacy remain one of the main concerns, since the presence of IoT devices brings along with it an enormous amount of sensitive patient information that can be an easy target to cyber threats. It is a big risk for patients and healthcare providers to have unauthorized access, data breach, and ransomware attacks. Robust cybersecurity measures can be the answer for the above threats, such as end to end encryption, multi factor authentication, and AI driven threat detection system. Data protection regulations including the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) need to be complied with to protect patients' privacy and ensure proper way of handling their medical data.

A third barrier to healthcare IoT adoption is interoperability as there is often proprietary communication or protocols among medical devices and platforms and cannot easily exchange data says. Unfortunately, the absence of standardized frameworks means that it is a challenge for healthcare providers to realize integration of IoT solutions for existing health care systems. To solve this issue, it's going to take cooperation from the industry level in order to define universal standards to allow these different healthcare technologies to be interoperable. Further, the development of adaptive algorithms that interpret multiple sources of information increases the possibility for improvements in data harmonization based on artificial intelligence and machine learning.

Challenge for the widespread IoT deployment in healthcare, especially in developing region comes from cost and infrastructure limitations. The development and implementation of IoT solution requires high costs to acquire hardware, cloud infrastructures, and network connectivity, which makes it difficult for resource limited healthcare center to adopt this technologies. Moreover, rural or remote areas with poor connectivity or interrupting high speed network limits one of the reliable data transmission to real time data via continuous internet connectivity. Using faster data processing and reduced reliance on single cloud centers as being offered by 5G technology and edge computing, there are potential solutions that may emerge [12].

III. Applications of Smart IoT Technologies in Healthcare

Smart IoT technologies are now applied in healthcare, have revolutionized the patient care, hospital management and medical decision making. In the context of IoT, these medical devices, real-time health monitoring and AI-powered analytics have made healthcare more accessible, efficient and personalized. These technologies permit continuous monitoring of the patient, immunological accuracy, optimization in the hospital work processes, and provide remote healthcare services [13]. As IoT adoption in healthcare technologies grows, medical practitioners now have better sources of informational feedback

to make and patients acquire proactive healthcare follow. Weighting the benefits of data integration into healthcare systems from a practical viewpoint, the integration of IoT in healthcare systems has improved patient outcome as well as reduced operational costs and minimized the hospital admissions.

A. Remote Patient Monitoring and Wearable Healthcare Devices

Remote patient monitoring (RPM) is one of the most impactful applications of IoT in healthcare which allows healthcare providers to monitor the health status of a patient in real time without the need for him to visit a medical facility. In particular, RPM has proved to be extremely useful in managing some chronic conditions such as diabetes, hypertension and cardiovascular disorders by enabling early detection of anomalies, and consequent timely medical interventions. They are wearable healthcare device such as a smart watch, biosensor and a fitness tracker that measure the vital signs, continuously like heart rate, blood pressure, blood glucose level, oxygen saturation and other body temperature [14]. Data transmitted by these devices reaches cloud-based platforms where AI algorithms apply to the information and find possible health risks. Thus, healthcare professionals can use this data remotely to suggest personalized treatment and early intervention before the disease has advanced to a critical state [14].

Additionally, wearable devices powered by IoT have been of great importance for postoperative care and rehabilitation. At home, patients can be monitored after surgeries so that they do not have to stay in the hospital for prolonged periods. IoT sensors are used by smart rehabilitation systems to track a patient's progress and ensure they adhere to their prescribed exercises and recovery protocols. Furthermore, medication adherence with IoT is enhanced by such IoT driven medication adherence systems that remind patients to take the medications at the right time thereby reducing chances of missing the dose and hence good treatment adherence. In other words, smart home IoT solutions (such as the fall detection sensor and emergency alert systems) give real time assistance to elderly and disabled individual; hence improving the safety and quality of life.

B. Smart Hospitals and IoT-Enabled Healthcare Infrastructure

Apart from caring for an individual patient, the IoT has revolutionized hospital management through optimization of workflows, improved resource utilization and reduced operational inefficiencies. IoT has made its presence in smart hospitals by deploying IoT enabled medical devices, RFID based tracking systems and automated workflows to make healthcare process an easy ride. For instance, IoT enabled smart beds automatically change in order to provide the patient with maximum comfort while monitoring vital parameters like heart rate and respiratory rate. Later, should the beds detect irregularities, the staff get alerted to act in time. IoT enabled RFID and GPS tracking systems have also brought great enhancement in medical inventory and asset management [15]. With hospitals currently monitoring the availability and location of medical equipment from ventilators to defibrillators and infusion pumps to ensure their proper use and reduce delays in such critical situations, this time is proving far too soon. IoT integrates with smart medication dispensers that can track drug usage and avoid medication shortage, error. Inventory management in hospitals is automated for minimizing wastage and efficiency, as well as overall cost.

In hospitals one of the most crucial IoT application is infection control. Environmental sensors that are smart monitor air quality, humidity, and temperature to guarantee the execution of hygiene standards and decrease the spread of episodes [16]. IoT is being used to keep track of sanitation schedules and to run automated disinfection systems to make certain hospital rooms and surgical instruments are cleaned punctually. Furthermore, IoT based facial recognition and biometrics' access control system can also improve hospital security to restrict the access of unauthorized individuals to critical areas like ICUs and operating rooms.

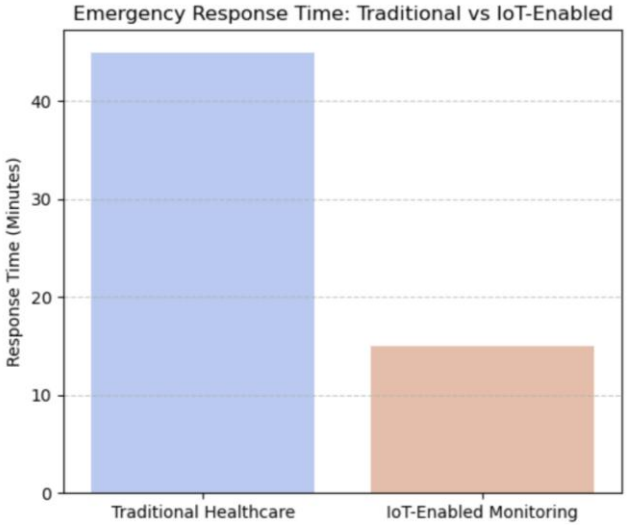


Fig. 4. Emergency Response Time Comparison - Traditional vs. IoT-Enabled Healthcare

The Fig. 4 shows a compelling graphical contrast in emergent response time of traditional medical care system and the one that is IoT enabled. The respective traditional healthcare area is represented by a blue bar with an almost 45 minute response time while IoT enabled monitoring area has an almost 15 minute response time. IoT technology in healthcare finds its efficiency in this visualization powerfully. Realtime monitoring, rapid speed of information transmission, and initially proactive applications of the Internet of Things devices allow for what would be an almost 70% reduction of emergency response time. IoT helps in faster detection of medical issues, bringing immediate alerts, and faster intervention thus making emergency medical response no more reactive but rather predictive and preventive. The graph itself proves on why adopting IoT in healthcare to save time, reduce complications and ultimately improves patient outcomes.

IV. IoT Data Analytics and Decision-Making in Healthcare

Though, integration of IoT in healthcare has not only enhanced connectivity in between healthcare devices and healthcare providers but also made a way for data driven

decision making. Real time patient data is created by IoT devices in huge volumes and when properly analysed, can enrich the diagnostic data, predict disease progression and will help in better optimizing treatment plans and improving healthcare outcome. These data enable healthcare providers to use advanced data analytics techniques, for example: machine learning, artificial intelligence and cloud computing to extract meaningful insights about the particular data [17][18]. IoT driven analytics lets hospitals and medicals professionals transition from reactive to proactive and predictive healthcare which provides better patient care, reduced operational cost and time.

A. Real-Time Data Collection and Processing in IoT Healthcare

For example, healthcare environments constantly 'spout' vast amounts of patient related data: whether it is vital signs and activity levels, environmental factors and medication adherence etc., from IoT devices deployed. Continuous real time health data streams are being produced from wearable devices, biosensors, smart implants and remote monitoring systems that are transmitted to centralized cloud platforms or edge computing systems for immediate processing. This real time data collection allows health care providers to collect the patient conditions in real time, detect anomalies and take proper actions in time to avoid medical complication.

Edge computing is critical to making IoT generated data go through the processing in event, and with as small latency as possible. Proceeding with critical health alerts which can be instantaneous at, say, the edge of the algorithm, on a wearable device or even a local hospital server, rather than pushing data up to the cloud, allows for crucial data to be processed closer to the source. Imagine, for example, a patient with a heart problem that causes irregular heart rhythms, and the use of IoT-enabled pacemaker to suddenly call for the help of the healthcare professionals and trigger an emergency response. Similarly, real time glucose monitoring systems can give alerts to diabetic patients about the time of the insulin and it also saves from medical emergencies.

Interoperability in data processing from the IoT is another critical issue. In the healthcare domain, the data from the sources such as electronic health records (EHRs), medical imaging systems, and wearable sensors should be integrated by the healthcare IoT systems. It should be standardized for interoperability in these systems as the protocols should be standardized and a robust data governance frameworks should be in place. In the cloud your healthcare platform relies on application programming interfaces (APIs) to have EHRs communicated seamlessly from one provider to the other. By enhancing clinical decision making, this interoperability gives doctors the holisitic possiblity to see a patients medical history and the present health status and potential risks [19].

B. Predictive Analytics and AI-Driven Decision Support Systems

However, the true power of IoT in healthcare truly lies in the ability it can give towards enabling predictive analytics, and AI and decision making. Machine learning and deep learning algorithms, widely applied in both healthcare and agricultural disease detection [20–22], can be applied to vast quantities of IoT generated health data in order to generate predictive models for the prediction of disease progression [23], for identifying high risk patients, and as recommendations for personalized treatment strategies [24]. Health care providers with good predictive analytics will change from a reactive mode of treating illnesses after symptoms are present to a proactive mode in which diseases are discovered and handled before they become very severe [25].

This, for example, are AI powered IoT systems that process data from heart rate monitors, blood pressure sensors etc. and predict the probability of cardiovascular diseases. Patient and doctor's with these systems can be warned about the possibility of risk as to intervene early and minimize the possibility of heart attacks or strokes. Like in oncology, IoT based predictive analytics is used extensively in the use of artificial intelligence models to run on genetic data, medical imaging scans, patient histories to predict the cancer progression and formulate a personalized treatment regimen [26].

The other role played by IoT analytics in hospitals is to improve hospital management by optimizing resource use and patient floor. These systems analysis patient's admission rate, bed occupation rate and emergency department's workflow, and predict the hospital resource demands. Forecasting the patient influx allows hospitals to put in staff, medical supplies, and ICU beds in order to minimise wait time and give the best patient care. During public health crises, pandemics for instance, predictive capability is very important to managing healthcare resources in a strategically limited way to avoid overcrowding and failure of the system.

On the other hand, AI is being used in IoT system which supports the personalized medicine by tailoring the treatment plan based on study of individual patient data. But, IoT driven analytics gives healthcare an approach that is more customised compared to what is possible with traditional healthcare approaches. AI takes genetic data, lifestyle habits, as well as real time health metrics and can suggest personalized drug dosage's, lifestyle changes, and rehabilitation plans. Such level of precision medicine improves treatment efficiency by minimizing side effects and adverse reactions.

V. Security, Privacy, and Ethical Challenges

It has definitely pushed healthcare into the Era of IoT integration where patient monitoring, diagnosis and treatment have been improved greatly [27]. Yet, it has also raised a number of security, privacy and ethical issues that need to be resolved in order for these technologies to be used safely and effectively. IoT devices have huge amounts of sensitive patient data collected and transmitted to them; and as a result, they are an obvious cyber threat, unauthorized and data breach targets. Furthermore, on top of that, one can make ethical arguments regarding the patient consent, data ownership and how the potential biased AI driven healthcare decisions might make the use of IoT in medical environment even more difficult. To address these challenges, we must employ advanced and secure cybersecurity measures along with strict regulatory compliance on ethical measures that strike the perfect balance between technological significant advancement and patient's right protection.

A. Security Threats and Cyber Risks in IoT Healthcare

Connected medical devices and data transmission networks in the IoT in healthcare are one of the serious issues concerning the system's vulnerability to cyberattacks. Cloud platforms, wireless communication, and interconnected devices characterize IoT healthcare systems, which along with a multitude of other entry points, make IoT healthcare systems very vulnerable to cybercriminals. Security loopholes irresponsibly open up for hackers to gain unauthorized access to sensitive patients' data as well as manipulate medical equipment and launch ransomware attacks against healthcare institutions. Such cyber threats have unfortunate implications, including various types of undermined patient safety, financial loses, and loss of trust in health care technologies.

One of the main security vulnerabilities in IoT healthcare is the interception of medical data in transit, right in the process of transloading the data between devices. Patients expose themselves to identity theft or insurance fraud if health records including diagnostic reports, treatment histories and genetic data are exposed. Also, these unsecure IoT devices can be forced to read false, modify medication dosages, or can lead to interference with lifesaving medical interventions. One example of a compromised insulin pump, for example, might mistake the variable for a dose and administer incorrect amounts to a patient, putting a patient's life in jeopardy. What further amplifies these risks? The lack of regular security update for IoT medical devices which leaves most of these devices vulnerable for exploits even after they are being deployed as many devices have security exploits that go undetected for long periods of time.

To counter these threats, there has to be robust cybersecurity frameworks that are adopted in IoT healthcare networks. Measures that can help safeguard the patient data and also device functionality include the end to end encryption, secure authentication protocols, and intrusion detection system. The promise of Blockchain technology in protecting IoT healthcare transactions using data integrity as well as decentralized control has also been realized in this [28][29]. On top of that, anomaly detection using AI can be used to identify unusual activities within IoT networks in advance in order to issue alerts to security teams once breaches may escalate to severe incidents.

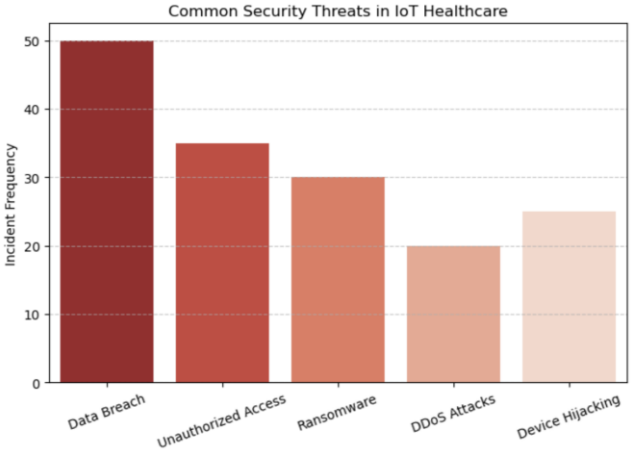


Fig. 5. Common security threats in IoT healthcare systems.

Fig. 5 provides an all-around picture of cybersecurity threats to IoT healthcare technologies with data breaches as the most frequent issue [30]. Fifteen security challenges are depicted on the visualization, where bar heights indicate the frequency of their occurrence with smallest being most frequent. Nearly 50 incidences belong to data breaches, 35 unauthorized access incidences and about 30 ransomware incidences. DDoS attacks and the device hijacking both represent significant, though lower occurring threats with frequencies of 20 and 25, respectively. The graphic points out the importance of having robust cybersecurity in place in an IoT healthcare system, noting that there can be many ways of having digital vulnerability. Dark to light red gradation of colors highlights the disparity in level of threat which are a visual cue to the plaguing risks with regard to the protection of sensitive medical data and of the healthcare infrastructure in a more and more connected ecosystem.

B. Privacy Concerns and Data Ownership in IoT Healthcare

Due to the concerns of privacy within the use of IoT, in healthcare, it is critically important to design a system that collects, stores and shares the patient data safely. Real time health information like biometric data, vital signs and medical history are collected by the IoT devices continuously. This data is critical to the improvement of patient care, but especially when stored in cloud platforms and healthcare databases, allows the risk of unauthorized and misuse of such data to rise. Who patients are and how they can control how their data is shared are ethically awkward issues of consent, data ownership and privacy.

Unlawful transfer of patient information to third parties including pharmaceutical companies, insurers, or research organizations is the truly greatest one. Healthcare institutions can share anonymized data without explicit patient consent for commercial purposes that can ultimately cause discrimination in the relative premiums on insurance or better employment possibilities. Moreover, when the health care systems are breached and patient records are exposed to malicious actors, there are further privacy violations.

These privacy concerns need to be addressed by healthcare organizations by adopting very strict data governance policies and compliance with the global regulatory frameworks such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). It is the mandate of such patients, to make them explicit consent for collecting data and organizations have to follow strict controls on the access to that data and the encryption mechanism. Both differential privacy and homomorphic encryption will also aid in the analysis of patients data while maintaining confidentiality. Blockchain and all its capabilities allow patients to have more control on their data through blockchain based personal health records (PHRs) and enhance transparency and trust in IoT healthcare solutions.

Fig. 6 shows an evolutionary outlook for the technology of IoT, on a healthcare base, from the present state to some main markers until 2034. A progressive and exponential way of moving into the IoT healthcare technologies is shown in the upward trajectory. Beginning from the current IoT implementation, the graph demonstrates improvements, along with their technological breakthroughs such as AI integration,

5G telemedicine, blockchain security and the smart hospital implementation. Significant milestones along the function of becoming a digital healthcare leader represent strategic steps in digital healthcare transformation with a steepening curve that indicates greater speed in the field of technology. The figure incorporated in the paper visualizes its forward looking view and shows how the development of the emerging technologies will keep shifting healthcare delivery. To emphasize that the every technological milestone embarks on a smooth, ascending line and each is built on the foundation of the previous innovations to the creation of more sophisticated, efficient and patient centric healthcare ecosystem.

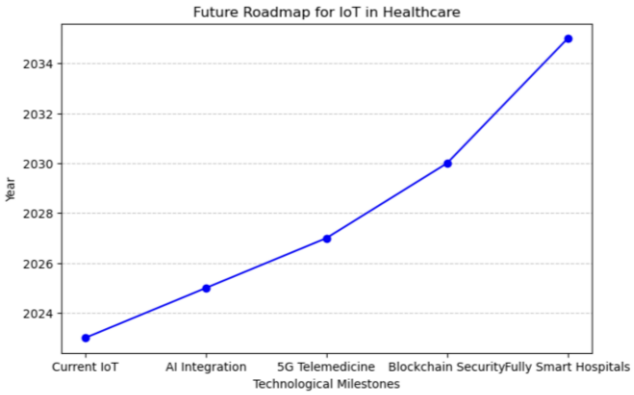


Fig. 6. Future roadmap for IoT in healthcare.

C. Ethical Considerations in AI-Driven IoT Healthcare

But aside from security and privacy, IoT healthcare systems need to be carefully probed on ethical ground. The concerns around bias, accountability and transparency arises as AI and machine learning algorithms are being built and used all the time with IoT healthcare data [31]. If the AI model was trained with a biased dataset, there is a potential that it will generate a discriminatory outcome, the amounts may vary between one patient group and another group based on race, gender or socio-economic status. For instance, an AI-based diagnosis system that is trained primarily on Caucasian patient population will have a lower sensitivity in detecting diseases in non-Caucasian people, thus creating differences in the process of healthcare delivery.

Automating the medical decision making is another ethical concern. However, while AI and IoT help with diagnostic accuracy and treatment planning, it would be a crime if those algorithms would be blamed for questionable diagnosis or poor treatment recommendation. When critical error occurs during an AI-powered IoT system and harms a patient, it becomes a legal and ethical problem as to who is legally and ethically responsible — if it is the healthcare provider, software developer or device manufacturer.

Transparency and explainability should be prioritized to bring up ethical AI deployment in IoT healthcare. The way AI algorithms work should be designed so as to be interpretable so that medical professionals can understand how decisions are made instead and don't blindly rely on recommendations generated by machines. Moreover, discrimination in AI based health care systems should also be detected and audited fairly regularly to eradicate it. It must be made clear to regulatory bodies that clear ethical guidelines are required to govern the use of AI in making medical decisions so that human oversight is always important.

The second important ethical question revolves around the digital divide in relation to enabling IoT-based healthcare solutions. Although IoT can hold great promise in bettering healthcare access especially in areas of low connectivity and reach, inequalities in technological infrastructure and digital learning may limit the physical and intellectual aspects of healthcare accessibility to those desiring to embrace the new technology. This implies that governments and healthcare organizations are lacking when it comes to investing in affordable IoT healthcare solutions and digital health education programs.

VI. Conclusion

Integration of IoT in healthcare has completely transformed patient care by allowing for real-time monitoring, predictive analytics and personalized treatment strategies. Ensuring that smart IoT technologies can be available and necessary for everyone to ensure availability of healthcare, operational efficiency and medical decision making. Faster, better and cheaper—this has been the case for healthcare providers who have been able to collect and analyse vast amounts of real time health data on patients, allowing detection of a disease early and enabling to use it for optimization of a treatment plan and improving the outcomes for the patients. Although these advancements are helpful, issues concerning security, privacy, interoperability, and ethical issues are still the most severe barriers to the wide adoption of them. Threats from cyber, data breaches and unauthorized access of sensitive patient information comes with huge risks, therefore, encryption, access control methods and regulatory compliance needs to be well implemented in order to maintain data security. Additionally, ethical issues such as biases in AI's decisions regarding healthcare should be addressed to prevent the mistreatment of any patient group. With the advancement of IoT, it is expected that the collaboration among healthcare professionals, technologists and policymakers would be indispensable to overcome these challenges and keep the IoT driven healthcare solutions secure, ethical, and viable. With the help of implementing stringent security, increasing AI transparency, and the adoption of standardized interoperability framework promises the future of IoT in health holds the potential of developing the IoT based medical services for healthcare on a more efficient, personalized, and patient centric base.

VII. Future Directions

In particular, nanotechnology and a combination of artificial intelligence, 5G connectivity, edge computing, along with blockchain technology are expected to shape the future of IoT in healthcare. With AI power, further development in disease forecasting, personalized treatment and early intervention strategy will move from reactive to proactive healthcare. 5G networks deployability will facilitate real time data transmission, reduce latency in remote patient monitoring and/or medical devices connectivity in the cloud. Furthermore, edge computing will help eliminate the use of the centralized cloud infrastructures and will allow the real time data processing at the

device level and will result in faster response times and the reduced Bandwidth. With regard to the data security, blockchain will ensure transparency and will allow decentralized, private patient health records within the regulatory compliance. In addition, federated learning will reduce the adoption of AI models that can train on decentralized healthcare data while upholding the confidentiality of the patient. It will be crucial that the digital divide is addressed and that IoT healthcare solutions are accessible to everyone. The future research should research towards the development of standardized IoT frameworks, mitigation of cybersecurity threats and improving AI interpretability to generate safe and ethical IoT driven healthcare systems that are useful for the patients across the globe.

ACKNOWLEDGEMENT

This work was supported by the Deanship of Scientific Research, the Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia under the project KFU252566.

REFERENCES

[1] Ala, A.; Simic, V.; Pamucar, D.; Bacanin, N. Enhancing Patient Information Performance in Internet of Things-Based Smart Healthcare System: Hybrid Artificial Intelligence and Optimization Approaches. Eng Appl Artif Intell 2024, 131, doi:10.1016/j.engappai.2024.107889.

[2] Li, C.; Wang, J.; Wang, S.; Zhang, Y. A Review of IoT Applications in Healthcare. Neurocomputing 2024, 565, doi:10.1016/j.neucom.2023.127017.

[3] Dhiman, P.; Bonkra, A.; Kaur, A.; Gulzar, Y.; Hamid, Y.; Mir, M.S.; Soomro, A.B.; Elwasila, O. Healthcare Trust Evolution with Explainable Artificial Intelligence: Bibliometric Analysis. Information 2023, Vol. 14, Page 541 2023, 14, 541, doi:10.3390/INFO14100541.

[4] Poongodi, M.; Sharma, A.; Hamdi, M.; Maode, M.; Chilamkurti, N. Smart Healthcare in Smart Cities: Wireless Patient Monitoring System Using IoT. Journal of Supercomputing 2021, 77, doi:10.1007/s11227-021-03765-w.

[5] Reegu, F.A.; Abas, H.; Gulzar, Y.; Xin, Q.; Alwan, A.A.; Jabbari, A.; Sonkamble, R.G.; Dziyauddin, R.A. Blockchain-Based Framework for Interoperable Electronic Health Records for an Improved Healthcare System. Sustainability 2023, Vol. 15, Page 6337 2023, 15, 6337, doi:10.3390/SU15086337.

[6] Onasanya, A.; Elshakankiri, M. Smart Integrated IoT Healthcare System for Cancer Care. Wireless Networks 2021, 27, doi:10.1007/s11276-018-01932-1.

[7] Javaid, M.; Khan, I.H. Internet of Things (IoT) Enabled Healthcare Helps to Take the Challenges of COVID-19 Pandemic. J Oral Biol Craniofac Res 2021, 11, doi:10.1016/j.jobcr.2021.01.015.

[8] Jabbari, A.; Humayed, A.; Reegu, F.A.; Uddin, M.; Gulzar, Y.; Majid, M. Smart Farming Revolution: Farmer's Perception and Adoption of Smart IoT Technologies for Crop Health Monitoring and Yield Prediction in Jizan, Saudi Arabia. Sustainability 2023, Vol. 15, Page 14541 2023, 15, 14541, doi:10.3390/SU151914541.

[9] Settia, N.; Bhutani, M.; Saini, V. Exploring the Impact of Artificial Intelligence on Healthcare: A Comprehensive Review. In Digitalization and the Transformation of the Healthcare Sector; Nilmini Wickramasinghe, Ed.; IGI Global Scientific Publishing, 1AD; pp. 115–138 ISBN 9798369396438.

[10] Bhutani, M.; Gupta, M.; Jain, T.; Sharma, T.P.; Solanki, A.; Malhotra, A. Data-Driven Robotics: A Performance Analysis of Wearable Sensor-Based Wound Assessment System for Sensor Data Collection. In Bio-Inspired Data-driven Distributed Energy in Robotics and Enabling Technologies; Abhishek Kumar, H.K.S.A.K.D.V.G.D., Ed.; CRC Press, 2024; pp. 249–266 ISBN 9781040147009.

[11] Wang, W.H.; Hsu, W.S. Integrating Artificial Intelligence and Wearable IoT System in Long-Term Care Environments. Sensors 2023, 23, doi:10.3390/s23135913.

[12] Zeghib, N.E.I.; Alwan, A.A.; Abualkishik, A.Z.; Gulzar, Y. Multi-Route Plan for Reliable Services in Fog-Based Healthcare Monitoring Systems. International Journal of Grid and High Performance Computing 2022, 14, doi:10.4018/IJGHPC.304908.

[13] Patil, S.; Shankar, H. Transforming Healthcare: Harnessing the Power of AI in the Modern Era. International Journal of Multidisciplinary Sciences and Arts 2023, 2, doi:10.47709/ijmdsa.v2i1.2513.

[14] Ullah, M.; Hamayun, S.; Wahab, A.; Khan, S.U.; Qayum, M.; Ullah, A.; Rehman, M.U.; Mehreen, A.; Awan, U.A.; Naeem, M. Smart Technologies Used as Smart Tools in the Management of Cardiovascular Disease and Their Future Perspective. Curr Probl Cardiol 2023, 48.

[15] Singla, D.; Kumar, S.; Gulzar, Y.; Mir, M.S.; Gupta, D.; Jaziri, W.; Sassi, N.; Arora, S. Designing and Implementing a Resilient Immutability Mechanism for Enhanced Supply Chain Management in E-Healthcare Systems. Frontiers in Sustainable Cities 2024, 6, 1403809, doi:10.3389/FRSC.2024.1403809/BIBTEX.

[16] Ajegbile, M.; Olaboye, J.; CC Maha; Igwama, T.; Abdul, S. Integrating Business Analytics in Healthcare: Enhancing Patient Outcomes through Data-Driven Decision Making. World Journal of Biology Pharmacy and Health Sciences 2024, 19.

[17] Shiwlani, A.; Khan, M.; … A.S.-I.J. of; 2023, undefined Synergies of AI and Smart Technology: Revolutionizing Cancer Medicine, Vaccine Development, and Patient Care. International Journal of Social, Humanities and Life Sciences 2024, 1.

[18] Majid, M.; Gulzar, Y.; Ayoub, S.; Khan, F.; Reegu, F.A.; Mir, M.S.; Jaziri, W.; Soomro, A.B. Using Ensemble Learning and Advanced Data Mining Techniques to Improve the Diagnosis of Chronic Kidney Disease. International Journal of Advanced Computer Science and Applications 2023, 14, doi:10.14569/IJACSA.2023.0141050.

[19] Khang, A.; Ragimova, N.A.; Hajimahmud, V.A.; Alyar, A.V. Advanced Technologies and Data Management in the Smart Healthcare System. In AI-Centric Smart City Ecosystems: Technologies, Design and Implementation; Alex Khang, S.R.A.K.S., Ed.; CRC Press, 2022; pp. 261–270 ISBN 9781003252542.

[20] Seelwal, P.; Dhiman, P.; Gulzar, Y.; Kaur, A.; Wadhwa, S.; Onn, C.W.; Goyal, N.; Shanmugasundaram, H. A Systematic Review of Deep Learning Applications for Rice Disease Diagnosis: Current Trends and Future Directions. Front Comput Sci 2024, 6, 1452961, doi:10.3389/FCOMP.2024.1452961.

[21] Gulzar, Y. PapNet: An AI-Driven Approach for Early Detection and Classification of Papaya Leaf Diseases. Applied Fruit Science 2025 67:4 2025, 67, 1–11, doi:10.1007/S10341-025-01466-9.

[22] Gulzar, Y. Enhancing Soybean Classification with Modified Inception Model: A Transfer Learning Approach. Emirates Journal of Food and Agriculture 36: 1-9 2024, 36, 1–9, doi:10.3897/EJFA.2024.122928.

[23] Khan, F.; Gulzar, Y.; Ayoub, S.; Majid, M.; Mir, M.S.; Soomro, A.B. Least Square-Support Vector Machine Based Brain Tumor Classification System with Multi Model Texture Features. Front Appl Math Stat 2023, 9, 1324054, doi:10.3389/FAMS.2023.1324054.

[24] Soni, T.; Gupta, D.; Uppal, M.; Juneja, S.; Gulzar, Y.; Ghafoor, K.Z. Deep Neural Network Framework for Predicting Cardiovascular Diseases from ECG Signals. Recent Advances in Computer Science and Communications 2024, 18, doi:10.2174/0126662558346126241222214404.

[25] Tripathi, S.; Sukumaran, R.; Cook, T.S. Efficient Healthcare with Large Language Models: Optimizing Clinical Workflow and Enhancing Patient Care. Journal of the American Medical Informatics Association 2024, 31, 1436–1440, doi:10.1093/JAMIA/OCAD258.

[26] Muhammad, A.; Jin, Q.; Elwasila, O.; Gulzar, Y. Hybrid Deep Learning Architecture with Adaptive Feature Fusion for Multi-Stage Alzheimer's Disease Classification. Brain Sciences 2025, Vol. 15, Page 612 2025, 15, 612, doi:10.3390/BRAINSCI15060612.

[27] Rayan, R.A.; Tsagkaris, C.; Iryna, R.B. The Internet of Things for Healthcare: Applications, Selected Cases and Challenges. Studies in Computational Intelligence 2021, 933, 1–15, doi:10.1007/978-981-15-9897-5_1/FIGURES/4.

[28] Gulzar, Y.; Reegu, F.A.; Jabbari, A.; Sonkamble, R.G.; Mir, M.S.; Soomro, A.B. BlockMed: AI Driven HL7-FHIR Translation with Blockchain-Based Security. International Journal of Advanced Computer

Science and Applications 2025, 16, 307–316, doi:10.14569/IJACSA.2025.0160233.

[29] Reegu, F.A.; Ayoub, S.; Dar, A.A.; Hussain, G.; Gulzar, Y.; Fatima, U. Building Trust: IoT Security and Blockchain Integration. Proceedings of the 18th INDIAcom; 2024 11th International Conference on Computing for Sustainable Global Development, INDIACom 2024 2024, 1429–1434, doi:10.23919/INDIACOM61295.2024.10499070.

[30] Kaur, K.; Kaur, A.; Gulzar, Y.; Gandhi, V. Unveiling the Core of IoT: Comprehensive Review on Data Security Challenges and Mitigation Strategies. Front Comput Sci 2024, 6, 1420680, doi:10.3389/FCOMP.2024.1420680/XML.

[31] Pradhan, B.; Bhattacharyya, S.; Pal, K. IoT-Based Applications in Healthcare Devices. J Healthc Eng 2021, 2021, 6632599, doi:10.1155/2021/6632599.

Modeling an Adaptive and Collaborative E-Learning System with Artificial Intelligence Tools

Kawtar Zargane*, Hassane Kemouss, Mohamed Khaldi
A Research Team in Computer Science and University Pedagogical Engineering-(S2IPU) Normal School of Tétouan, Abdel Malek Essaadi University, Morocco

***Abstract*—In an educational environment that is undergoing digital transformation, the need to create smarter, learner-centred learning environments is becoming increasingly urgent. This article presents a conceptual modeling of an e-learning system that integrates the adaptive and collaborative dimensions, relying on the tools of artificial intelligence (AI), which occupies a central place, both as a dynamic adaptation engine and as a collaboration facilitator and automaton of certain pedagogical activities. This methodical and structured approach makes it possible to develop a hybrid environment capable of adjusting to individual needs while promoting the co-construction of knowledge between peers. Based on instructional design principles and the 2TUP (Two Tracks Unified Process) process, this approach aims to develop a systematic architecture, illustrated by UML (Unified Modeling Language) diagrams, of classes, use cases, activities and sequences, integrating AI (Artificial Intelligence) through adaptive learning, conversational agents and intelligent tutoring systems that make it possible to personalize learning, Provide targeted feedback, optimize learner performance, and guide learners more accurately. This combination of standardized modeling and AI improves the synergy between stakeholders and increases the efficiency of online learning environments. Finally, this model paves the way for a new era of more flexible, inclusive and responsive techno-pedagogical systems, capable of facing the contemporary challenges of online training.**

Keywords*—*Conceptual modeling of an online learning system; Adaptive system; Online collaborative learning; Artificial intelligence (AI); Educational software architecture; UML modeling

I. INTRODUCTION

In the digital age, in a context where online learning is constantly developing, and in the face of the diversity of learner profiles, the complexity of journeys and the growing importance of collaboration, it is becoming imperative to design smarter, more flexible and learner-centred learning systems, it is becoming essential to design educational experiences that are both effective, customized and technologically robust. This study highlights the importance of a rigorous approach in software engineering to design pedagogical devices capable of adapting to the needs of learners [1]. By combining instructional design principles with software development best practices, this allows us to develop a systematic and consistent architecture, modeled using UML diagrams, to effectively structure these e-learning projects. The integration of artificial intelligence (AI) into this architecture represents a powerful lever for the personalization of pathways, taking into account the pace, profile and specific difficulties of each learner [2,3].

In this perspective, the integration of collaborative learning highlights the social dynamics of knowledge construction, by promoting interaction between peers, the co-creation of resources and the development of transversal skills. At the same time, AI-supported adaptive learning makes it possible to personalize the educational experience by adjusting the content, pace and pedagogical strategies and to automate certain tasks, analyze learners' performance and offer precise pedagogical recommendations based on the specific characteristics of each learner [4, 5]. This combination can profoundly transform education [6]. Through the use of intelligent chatbots and adaptive tutoring systems, this innovative learning system is able to generate targeted feedback, detect the specific needs of each user and recommend personalized content [7]. This approach, which combines personalization [8], collaboration and artificial intelligence, is based on an intelligent tutor that adjusts content and activities in real time according to the learner's profile. The collaborative dimension is fully integrated through exchange spaces, online discussion tools, and group work functionalities [9]. Continuous monitoring of progress also allows for the adjustment of pathways as needed, ensuring a responsive, stimulating learning environment that is conducive to autonomy [10].

This article thus proposes a conceptual modeling of an online learning system that is both adaptive and collaborative, based on AI tools. The objective is to lay the foundations of an intelligent environment capable of reconciling individual personalization and collective intelligence, in a learner-centered, scalable and equitable learning perspective. To meet pedagogical and technical requirements, the adoption of a structured design approach based on the 2TUP methodology is necessary. This allowed the full potential of the UML modeling language to be fully exploited through two complementary perspectives: a static view, with class and use case diagrams describing the architecture of the system, and a dynamic view, illustrated by activity and sequence diagrams translating the interactions and behaviors of the system. These visual representations reinforce the overall understanding of the system and concretely demonstrate the articulation between modeling, AI and collaborative learning.

This model paves the way for a new generation of technopedagogical systems that are more flexible, inclusive, and responsive, capable of meeting the current challenges of online training.

II. Conceptual and Theoretical Framework

This conceptual and theoretical framework aims to inform the design of an e-learning system based on three complementary domains that intersect to structure the pedagogical, technological and methodological approach of the proposed system. The pedagogical foundation is based on collaboration between peers, inspired by the work of Vygotsky this approach considers that knowledge is built through social interaction, cooperation between peers and dialogue [11]. Then, the second theoretical pillar concerns artificial intelligence in education, and more specifically adaptive learning systems and intelligent agents. These technologies make it possible to personalise the courses according to the profile, pace and needs of each learner. Finally, the third axis is based on educational software engineering, which provides a rigorous methodology for designing complex systems. By articulating these three dimensions, this work is part of a systemic approach, laying the foundations for a fine modeling of an online learning environment that is both personalized, collaborative and intelligent, capable of meeting the current challenges of distance learning.

This theoretical framework therefore guides the technical design and architecture of the adaptive and collaborative learning system maintained by AI is proposed, detailed according to:

A. The Pedagogical Approach

1) Collaborative learning: Collaborative learning, based on the principles of social constructivism, is based on the idea that knowledge construction is optimized when it is done in interaction with others. According to Johnson and Johnson, the pillars of this approach include positive interdependence, individual responsibility, and learning-promoting interactions [12]. By placing cooperation at the heart of the pedagogical system, the learning environment must encourage active exchanges between peers, thus allowing the confrontation of ideas, the collective resolution of problems and the development of social skills,[13]. To support this dynamic, the system must integrate appropriate collaborative tools (discussion forums, wikis, or shared workspaces), while ensuring balanced group participation [14]. Beyond simple interaction, one of the major objectives of collaborative learning is the co-construction of knowledge, where learners produce resources or solutions with high added value together.

Scardamalia and Bereiter emphasize the importance of supporting the complex cognitive processes involved in this co-elaboration: collective regulation, integration of divergent points of view, and shared reflection. Learning systems must therefore offer explicit pedagogical scenarios, as well as mechanisms for mediation or assistance in structuring collaborative work. Features such as virtual whiteboards, collaborative editors, or lightweight project management tools can promote this co-construction [15].

Finally, the management of roles and contributions through traceability systems, participation indicators or recognition mechanisms (badges, comments, peer evaluations), is a crucial lever to guarantee equity, commitment and efficiency of group work. As Strijbos and De Laat, [16] point out, it can strengthen learners' motivation and encourage active participation [17]. By integrating these dimensions, a well-designed collaborative learning system becomes a real trigger for collective intelligence, in the online learning experience.

B. The Technological Approach

1) Adaptive learning: Adaptive learning is now one of the major levers for personalizing the online educational experience. It is a pedagogical approach that dynamically adjusts the content, pace, and activities based on each learner's needs, skill level, and learning style.[18]. This personalization is based on a fine modeling of the user, integrating variables such as prior knowledge, cognitive preferences or learning styles, as proposed by Felder and Silverman and discussed more recently by Pashler et al ,[19],[20]. The objective is to make learning more relevant, motivating and effective by adapting the course to each profile.

Intelligent Tutoring Systems (ITS), platforms such as DreamBox or Knewton, as well as adaptive digital environments, illustrate these principles. These systems are based on a continuous evaluation of the learner's progress, made possible through the real-time analysis of their interactions with the platform. Algorithms monitor performance trends and identify gaps, making it possible to recommend specific resources, modify the sequence of content, or propose targeted remediation activities [21], [22]. The formative approach, as advocated by Bloom [23], is implemented here through integrated assessments, oriented towards pedagogical adjustment.

A third essential pillar of adaptive learning is personalized feedback. Beyond simple correction, this feedback aims to support metacognitive regulation, helping the learner to become aware of mistakes, to understand poorly mastered concepts and to plan future efforts [24], [25]. It can take the form of contextual comments, detailed explanations, recommendations for additional activities, or tailored motivational messages. This feedback, based on a detailed analysis of the learning data, helps to strengthen engagement and autonomy.

Thus, an effective adaptive learning system is not limited to distributing content; it acts as an intelligent tutor, capable of observing, analyzing and intervening in a targeted manner, to maximize the learning potential of each individual [26]. Despite their potential, these systems still have limitations, particularly in terms of general learner modelling, algorithm and data transparency, all of which are challenges for a more reliable and inclusive integration of AI in education.

2) Artificial intelligence in education: Artificial intelligence (AI) is now central to modern online learning systems, not only as a technological tool, but as a vector for educational intelligence. Its integration is profoundly transforming the ways in which digital learning environments are adapted, collaborated and automated [27].

The first major contribution of AI lies in its ability to personalize the learning experience. Unlike traditional fixed-rules-based approaches, AI-enabled systems use advanced data

processing and machine learning techniques to analyze, in real time, learner interactions (behaviors, response time, recurring errors, engagement levels). These analyses make it possible to dynamically adapt the content, the level of difficulty, the pace or the materials offered, according to the individual profile and the evolution of each learner [28]. This fine adaptability, based on learning traces, reinforces the relevance of the educational paths and promotes commitment, self-regulation and autonomous progression.

Beyond individual personalization, AI also facilitates collaboration between learners. Thanks to natural language processing and social learning network analysis technologies, it is able to observe interactions in forums, contributions in shared spaces or roles in collective activities [29]. It can thus detect imbalances, propose strategic groupings or intervene as an intelligent mediator, by structuring exchanges, relaunching discussions or formulating open questions promoting the co-construction of knowledge. AI thus contributes to a more efficient, equitable and reflective management of collaborative work, by strengthening inclusion and group dynamics.

A third fundamental axis is the automation of repetitive pedagogical tasks, both for teachers and learners. Intelligent systems now allow the automatic generation of personalized quizzes, the correction of complex exercises, the synthesis of discussions, or the early warning in the event of a risk of disengagement or failure [30, 31]. These features reduce the cognitive and temporal load on teachers, while maintaining a quality of individualized follow-up. They also allow for instantaneous, contextualized feedback, which is essential for self-regulated learning.

Finally, although the contributions of AI are promising, its deployment in education systems raises important technical challenges. Among the major challenges are the transparency of algorithms, the explain ability of automated decisions and the fairness in the proposed recommendations [32, 33, 34]. It therefore becomes crucial to design educational AI systems that are not only efficient, but also accountable, learner-centred and inclusive.

Finally, artificial intelligence is now an essential lever for pedagogical innovation. It acts as a catalyst that, far from replacing the learner, reinforces the adaptive, collaborative and formative dimensions of learning, while laying the foundations for a more intelligent, equitable and sustainable educational ecosystem [2],[4].

C. The Software Approach

UML (Unified Modeling Language) modeling allows the structure and behavior of the system to be represented, while the 2TUP (Two Track Unified Process) method offers an iterative and scalable framework for software development [1]. Combined with a reflection on instructional design, these approaches ensure consistency between instructional objectives and technological choices.

1) The development process: The 2TUP process an evolution of the unified process, offers a rigorous and flexible framework for the development of information systems [35]. By separating the functional aspects, related to business needs, from the technical aspects, related to the system architecture [36]. The 2TUP makes it possible to better adapt to the constant evolution of technological environments and user requirements. This two-pronged approach fosters close collaboration between development teams and business lines, improving the quality of the final product. In addition, by incorporating regular iterations and project reviews, the 2TUP reduces the risks associated with late changes and enables continuous delivery of value [37].

Unlike more rigid methods, 2TUP offers greater agility while ensuring a high level of quality. For example, in the development of mobile applications, 2TUP has demonstrated its ability to adapt quickly to changes in platforms and user feedback, while maintaining functional consistency [38].

2) The UML modeling language: UML (Unified Modeling Language), a universal language for object modeling, offers a range of graphical tools to accurately and concisely represent the structure and behavior of a system [39]. Whether designing a software application, a database, or a business process, UML provides a common language for developers, analysts, and users [40]. Thanks to its various diagrams (use cases, classes, sequence, activity, etc.), UML allows you to visualize the different facets of a system, from user interaction to implementation details. This graphical representation greatly facilitates understanding, communication and collaboration within a development team [41]. In addition, UML is an evolving language that adapts to new technologies and changing project needs. In short, UML is an essential asset for any professional wishing to design and develop quality software systems [42].

Ultimately, UML remains a central modeling tool, offering rigor, clarity and shared vision throughout the life cycle of a project. Its ability to represent both the functional and technical aspects of a system makes it an essential lever for the design of robust and sustainable software solutions.

D. Critical Analysis of Existing Systems

In order to design a more effective e-learning system, it is essential to analyze the limitations of the solutions currently available, [43,44]. This section critically examines existing adaptive and collaborative systems, highlighting their contributions, but also their weaknesses in personalization, social interaction, and the integration of artificial intelligence. These findings will help guide the modelling of the proposed system.

1) Comparison of existing adaptive and collaborative e-learning systems: In this section, a critical analysis of adaptive and collaborative e-learning systems is conducted in order to identify their strengths and limitations. Adaptive systems personalise learning pathways according to learners' individual needs [43] while collaborative systems promote social interaction to strengthen motivation and co-construction of knowledge [45], [46]. However, these approaches are often implemented in isolation, without any real integration between personalization and collaboration [42]. In addition, their effectiveness is sometimes limited by a lack of granularity in

adaptation or weak orchestration of collective activities. This comparison aims to shed light on the design orientations of a more efficient technopedagogical system, harmoniously integrating artificial intelligence, individual adaptation and social collaboration [47].

TABLE I. COMPARISON OF EXISTING ADAPTIVE AND COLLABORATIVE E-LEARNING SYSTEMS

Aspect	Adaptive e-learning systems	Collaborative e-learning systems
Main Objective	Personalize the learning experience based on the individual characteristics of the learner.	Harness the collective intelligence and social benefits of group learning.
Key Mechanisms	Modeling of the learner, the domain, and pedagogical adaptation.- Adaptation of the content or course according to the prerequisites or performances.- Detection of gaps and proposal of remedies.	Interaction, sharing of ideas, co-construction of knowledge.- Communication and collaborative work tools.
Examples	ITS (Intelligent Tutoring Systems), Knewton (Alma), DreamBox Learning.	Forums, Wikis, EVA, LMS (Moodle, Canvas, Blackboard).
Bounds	Adaptation often superficial, based on rigid rules.- Difficulty in integrating complex aspects such as motivation or emotion.- Dependence on the quality and quantity of data collected.- Little or no integration of social dimensions.	Collaborative tools without intelligent facilitation (little intervention on group dynamics).- Difficulty in valuing and measuring individual contributions.- Little synergy with adaptive systems (operating in silos).

According to Table I, Adaptive learning systems, such as Knewton (now integrated into the Alma platform), DreamBox Learning or Intelligent Tutorial Systems (ITS), aim to personalize the learner's experience by dynamically adapting the content, difficulty and paths according to their performance and profiles [21]. However, these devices often have limited adaptation, based on static rules or simple statistical models, struggling to integrate more qualitative variables such as motivation or emotional state. At the same time, collaborative platforms such as Moodle, Canvas or forum environments offer tools for interaction (discussions, sharing, group work), but collaboration remains poorly guided: group dynamics, role coordination or the valuation of contributions are rarely handled in an intelligent way [13]. This double observation reveals a persistent compartmentalization between individualized adaptation and collaborative learning. Data from learner profiles are rarely used to form relevant groups or adapt social interactions, and there is no automated orchestration to synchronize adaptive and collaborative processes within the same system [48]. This separation hinders the emergence of truly intelligent learning environments, fully integrating the social and cognitive dimensions.

III. RESULTS AND DISCUSSION

A. UML Modeling of an Integrative Learning System

UML modeling is a powerful tool for clearly and consistently representing the essential components of a complex system. In this section, we present a functional modeling of an integrative e-learning system, combining adaptivity, collaboration, and artificial intelligence. The main objective is to describe, by means of UML diagrams, a modular architecture as highlighted by Ouelhadj et al (2020), as well as Chen et al. (2023), which is capable of personalizing learning paths, managing social interactions and automating certain pedagogical tasks thanks to AI [40], [28].

This modeling thus aims to lay the foundations of an intelligent system, centered on both the individual and the collective, while facilitating the analysis, design and implementation of an evolving and adaptable learning environment [33], [41].

Based on the 2TUP development model and using the UML language, our study relies on the StarUML tool to accurately model the complex interactions between the learner, the content and the artificial intelligence. This work makes it possible to design tailor-made learning paths, adapted to the specific needs of each user.

1) Static part: UML makes it possible to visualize the structure, behavior and interactions of the different components of a system, thus facilitating communication between stakeholders, both technical and non-technical. UML is widely used in software engineering to analyze functional and non-functional requirements, design robust software architectures, and document existing or developing systems [49],[50]. Through this article, we propose four diagrams, including two structural diagrams (class diagram and use case diagram) and two behavioral diagrams (sequence diagram and activity diagram).

a) Class diagram: For Rosenberger, M. (2022) and Chehida, S. et al. (2021), a class diagram is used to illustrate classes along with their attributes, methods, relationships (such as association, aggregation, composition, inheritance, etc.) and a constraint [39],[51].

b) Diagram interpretation: Fig. 1 presents the static structure of the system, it allows to understand how the objects interact, to visualize the organization of the data and to clearly define the responsibilities of each class in the overall functioning of the system, as follows:

- TutorIntelligent Class: The TutorIntelligent class manages personalized learner coaching via the evaluateLearnerLevel(), adaptContent(), generate Exercises(), provideFeedbackPersonalized (), and followProgress() operations. This central class coordinates the adaptation of learning
  - evaluateLearnerLevel (): Allows you to evaluate the learner's current level. For example, it can analyze the learner's answers to programming exercises to determine their level in Python.
  - adaptContent(): Allows you to adjust the learning content according to the learner's level and profile. For example, it can offer more complex exercises if the learner masters the basics well.

- generateExercises(): Allows you to create custom exercises. For example, it can generate math exercises adapted to the learner's current level.
- provideFeedbackPersonalized (): Allows you to give detailed feedback on the learner's responses. For example, it can explain why an answer is incorrect and suggest ways to improve it.
- followProgress(): Allows you to track the progress of learning. For example, it can identify concepts that require more practice.

- Learner Group Class: The Learner Group class manages groups of learners through the setObjectives(), planActivity(), assignTutor(), evaluateProgress(), and WorkOnTheTask() operations. It organises and coordinates collective learning.
  - setObjectives(): Allows you to set learning objectives for the group. For example, they can set common goals for a collaborative project.
  - planActivity(): Allows you to plan learning activities. For example, they can organize group work sessions or practical workshops.
  - assignTutor(): Allows you to assign a tutor to the group. For example, he can assign a specialized tutor depending on the needs of the group.
  - evaluateProgress(): Allows you to evaluate collective progress. For example, it can measure the group's progress against the set goals.
  - WorkOnTheTask(): Allows you to manage work on assigned tasks. For example, it can track the distribution and progress of tasks among members.

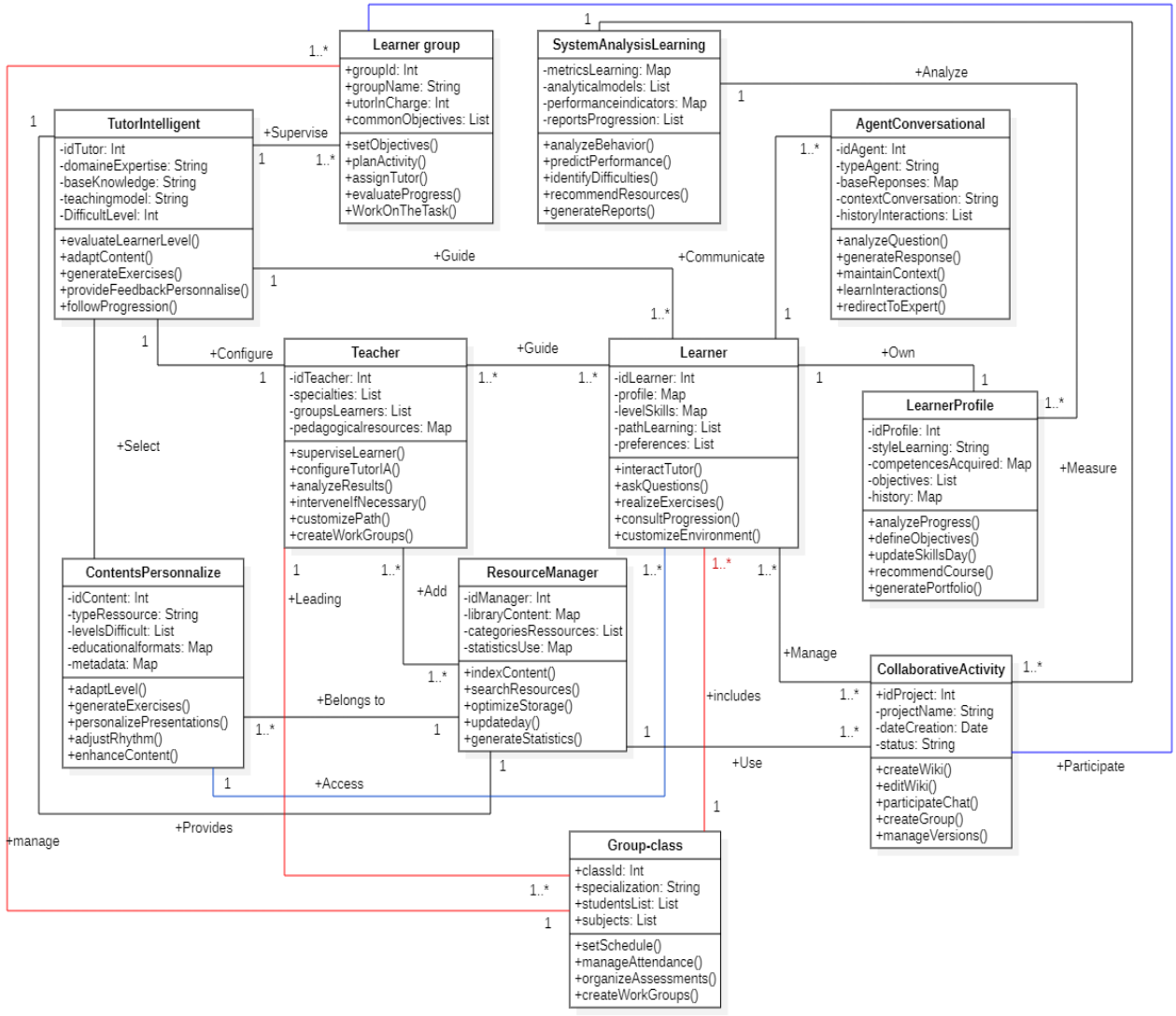


Fig. 1. Class diagram and static structure of online learning system.

- Teacher Class: The Teacher class oversees the learning process through the superviseLearning(), configureAITutor(), analyzeResults(), interveneIf Necessary(), and customizePath() operations. It represents the human expert who guides and adjusts the system.
  - superviseLearning(): Allows you to track the overall progress. For example, the teacher can observe the learning patterns of a group of learners.
  - configureAITutor(): Allows you to configure the tutorial system. For example, the teacher can adjust the evaluation criteria or difficulty thresholds.
  - analyzeResults(): Allows you to study learners' performance. For example, the teacher can identify concepts that cause problems for multiple learners.
  - interveneIfNecessary(): Allows you to provide human assistance. For example, the teacher can organize a remedial session for particularly difficult concepts.
  - customizePath(): Allows you to tailor learning paths. For example, the teacher can create specific learning paths for different learner profiles.
- Learner Class: The Learner class interacts with the system through the interactTutor(), askQuestions(), effectExercises(), consultProgress(), and customize Environment() operations. It represents the main user of the learning system.
  - interactTutor(): Allows the learner to interact with the smart tutor. For example, the learner may ask for additional explanations of a difficult concept.
  - askQuestions(): Allows the learner to ask questions to the system. For example, the learner may ask for clarification on a programming exercise.
  - realizeExercises(): Allows the learner to complete the proposed exercises. For example, the learner can solve math problems adapted to his or her level.
  - viewProgress(): Allows the learner to see their progress. For example, the learner can see their performance in different skills.
  - customizeEnvironment(): Allows the learner to adjust their learning preferences. For example, the learner can change the pace of learning or the type of exercises he or she prefers.
- Systems Analysis Learning class: The AnalysisLearningSystem class analyzes training data through the analyzeBehavior(), predictPerformance(), identifyDifficult(), recommendResources(), and generate Reports() operations. It provides information to optimize training.
  - analyzeBehavior(): Allows you to analyze learning patterns. For example, it can identify that the learner performs better in the morning or that they spend more time on geometry exercises.
  - predictPerformance(): Allows you to anticipate the learner's future performance. For example, it can predict whether the learner is ready to tackle a complex new concept.
  - identifyDifficulty(): Used to identify blocking points. For example, it can detect that the learner has systematic difficulties with fractions in mathematics.
  - recommendResources(): Allows you to suggest relevant resources. For example, it can suggest additional explainer videos about concepts that are not well understood.
  - generateReports(): Allows you to create detailed analysis reports. For example, it can produce a monthly summary of the learner's progress.
- AgentConvensional class: The Conversational Agent class handles conversational interactions through the analyzeQuestion(), generateAnswer(), maintainContext (), learnInteractions(), and redigerVersExpert() operations. It ensures natural communication with the learner.
  - analyzeQuestion(): Helps to understand the questions being asked. For example, it can identify whether the learner is asking a question about a mathematical concept or asking for clarification.
  - generateAnswer(): Allows you to formulate appropriate answers. For example, it can explain a complex concept with simple analogies.
  - maintainContext(): Helps keep the conversation consistent. For example, it can remember previous questions to provide contextualized answers.
  - learnInteractions(): Allows you to improve your future responses. For example, you can memorize the explanations that worked best.
  - writeToExpert(): Allows you to forward complex questions to the smart tutor. For example, it can redirect a technical question to an expert in the field.
- Learner Profile Class: The LearnerProfile class handles student information through the analyzeProgress(), setObjectives(), updateSkills(), recommendPath(), and generatePortfolio() operations. It maintains a complete view of the learning path.
  - analyzeProgress(): Allows you to evaluate the progress of learning. For example, it can track the improvement of programming skills over a period of time.
  - setGoals(): Allows you to set learning objectives. For example, it can set milestones to reach an advanced level in mathematics.
  - updateSkills(): Allows you to update the mastery level. For example, it can validate the acquisition of a new skill after passing exercises.

- recommendPath(): Allows you to suggest learning paths. For example, it can suggest an optimal sequence of modules to follow.
- generatePortfolio(): Allows you to create a portfolio of skills. For example, it can compile achievements and certificates obtained.

- ContentPersonnalize Class: The Content Personalize class handles adapting content through the adaptLevel(), generateExercises(), personalizePresentations(), adjustRhythm(), and enrichContent() operations. It helps ensure that the content matches the needs of each learner.
  - adaptLevel(): Allows you to adjust the difficulty of the content. For example, it can simplify programming exercises if the learner encounters too many difficulties.
  - generateExercises(): Allows you to create custom exercises. For example, it can generate math problems based on the learner's interests.
  - customizePresentations(): Allows you to adapt the format of the content. For example, it can favor visual aids for a learner with a visual learning style.
  - adjustPace(): Allows you to change the pace of learning. For example, it can slow down the progress of a module if the learner needs more time.
  - enrichContent(): Allows you to add additional resources. For example, it can incorporate concrete examples related to the learner's area of interest.
- Resource Manager class: The Resource Manager class organizes educational resources using the indexContent(), searchResources(), optimizeStorage(), update(), and generateStatistics() operations. It manages the system's resource library.
  - indexContent(): Allows you to catalog the available resources. For example, it can categorize educational videos by level and theme.
  - searchResources(): Allows you to quickly find relevant content. For example, it can locate all available exercises on loops in programming.
  - optimizeStorage(): Allows you to manage storage space efficiently. For example, it can archive infrequently used resources.
  - update(): Allows you to update existing resources. For example, it can update exercises with new use cases.
  - generateStatistics(): Used to produce usage analytics. For example, it can identify the most consulted resources.

The Resource Manager class:

- Interacts with the Learner class for participant management

- Contact SmartTutor for group work evaluation

- Uses Resource Manager to store collaborative content

- Integrates with the Learning Analysis System to measure participation

- Collaborative Activity Class: The Collaborative Activity class facilitates collaboration between learners through the createWiki(), editWiki(), participateChat(), createGroup(), shareResources(), and manageVersions() operations. It allows learners to work together and communicate effectively.
  - createWiki(): Allows you to create wiki pages for knowledge sharing. For example, learners can create a wiki page on the key concepts of object-oriented programming.
  - editWiki(): Allows collaborative editing of wiki content. For example, a learner can add practical examples to an existing wiki page on the unified modeling language.
  - participateChat(): Allows you to exchange in real time with other learners. For example, learners can discuss a complex problem or organize their group work.
  - createGroup(): Allows you to form work teams. For example, learners can create a group for a common software development project.
  - shareResources(): Allows you to exchange documents and resources. For example, a learner can share their notes or useful links with their group.
  - manageVersions(): Allows you to track changes made to the wiki. For example, learners can see the history of changes and revert to an earlier version if necessary.
- Group Class: The Group-class handles the academic aspects of a class through the setSchedule(), manageAttendance(), organizeAssessments(), and createWorkGroups() operations. It structures the pedagogical organization of a class group.
  - setSchedule(): Allows you to set the class schedule. For example, they can schedule class sessions and assessments.
  - manageAttendance(): Allows you to manage student attendance. For example, it can track attendance and generate attendance reports.
  - organizeAssessments(): Allows you to organize assessments. For example, they may schedule and coordinate ongoing reviews and assessments.
  - createWorkGroups(): Allows you to create workgroups. For example, he can form balanced teams for collaborative projects.

c) Use case diagram: The use case illustrates the exchanges between a system and its stakeholders (users, other systems, etc.) in terms of usage scenarios and functionalities offered by the system [52]. The identification of user needs and the main functionalities to be developed is often done upstream of the design phase using the use case diagram [53]. In the use

case diagram Fig. 2, the actors are symbolized by external blocks while the use cases are indicated by ellipses. The arrows symbolize the links between the actors and the usage scenarios [54].

- Diagram interpretation:

Learner: The learner is at the center of the system and interacts with several functionalities. He can:

- Interact with the tutor: This suggests a built-in tutoring system, likely AI-driven, that offers personalized assistance.
- Check your progress: The learner can track their performance and progress in the course.
- Customize the environment: The system offers customization options to tailor the learning experience to the learner's preferences.
- Create/Edit a wiki: A collaborative feature for learners to contribute to a common knowledge base.
- Participate in a chat: Facilitates communication and interaction between learners.
- Create a group: Allows learners to form workgroups for collaborative projects or studies.

Teacher: The teacher supervises the learning process and has access to the following functions:

- Set up the AI tutor: The teacher can adjust the settings of the AI tutor to optimize its effectiveness.
- Track learning: The teacher can track learners' progress and identify difficulties.
- Analyze the results: The teacher can examine the performance of the learners and evaluate the effectiveness of the teaching.
- Intervene if necessary: The teacher can intervene to provide additional support or personalized advice.

System Analysis Learning: This component represents the analytical backend of the system. It does the following:

- Analyze behavior: The system tracks and analyzes learners' actions in the learning environment.
- Predict performance: Based on the data collected, the system can predict the future performance of learners.
- Identify difficulties: The system can detect areas where learners are struggling.
- Recommend resources: Based on the needs identified, the system may recommend additional learning resources.

Chatbot: This actor represents a chatbot or virtual assistant. Its actions include:

- Analyze the question: Understand the questions asked by the learners.
- Generate Response: Provide relevant answers to questions.
- Maintain context: Maintain the context of the conversation for more natural interactions.

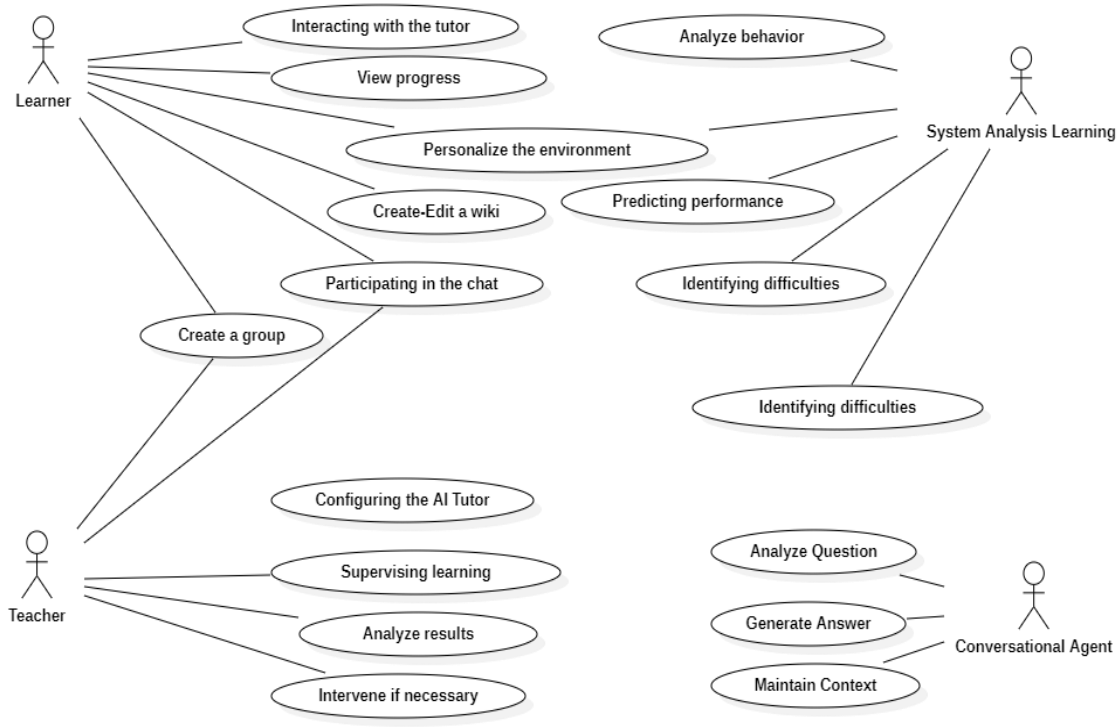


Fig. 2. Use case diagram and interactions between the different actors.

2) Dynamic part: UML dynamic modeling offers a complementary vision of the behavior of a system [55]. While sequence diagrams emphasize the chronological order of interactions between different objects, activity diagrams present a more holistic view of processes, breaking down actions into sequences, parallels, or conditions. They thus provide a better understanding of how the system works, identify potential problems and facilitate communication between the different actors of a project [56]. There are two types of diagrams: the sequence diagram and the activity diagram.

a) The Activity diagram: A UML activity diagram is a graphical representation that visualizes the control flow of a process, procedure, or algorithm [57]. It makes it possible to model in detail the different stages of a process [58]. The decisions to be made at each stage, the loops and the synchronizations.

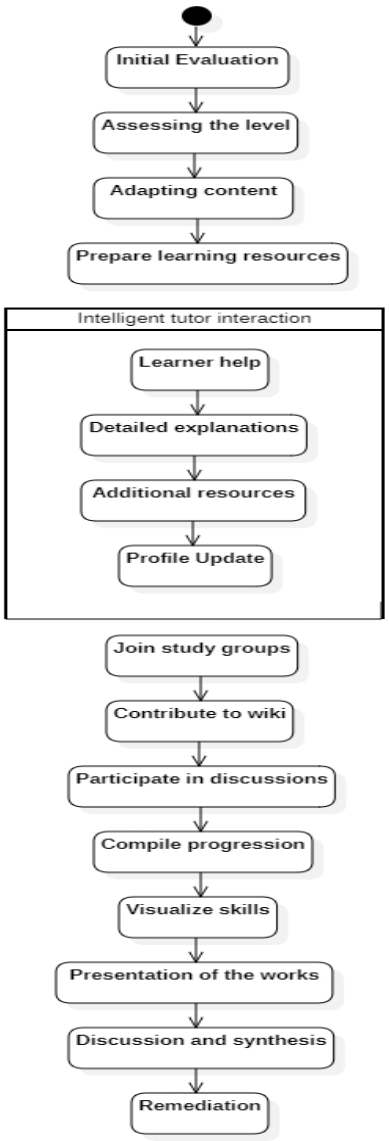


Fig. 3. Activity diagram describing the workflow within the learning system.

- Interpreting the intelligent learning system presented in this activity diagram (Fig. 3), illustrates a comprehensive and sophisticated learning process. It all starts with a crucial initial assessment phase that precisely identifies the learner's level, followed by an in-depth assessment to tailor the learning content to their specific needs. Once this foundation is established, the system prepares customized learning resources. Interaction with the smart tutor is at the heart of the system, offering targeted assistance, detailed explanations and additional resources, while keeping the learner's profile up to date. The system strongly encourages the collaborative dimension of learning by inviting learners to join study groups, contribute to a collective wiki, and actively participate in discussions. Progress is monitored rigorously, with a regular compilation of progress, a clear visualization of the skills acquired and presentations of the work carried out. The process concludes with discussion and synthesis sessions, allowing to consolidate learnings, with the possibility of implementing corrective actions if necessary. This methodical approach combines personalized learning, intelligent interaction, and social collaboration, creating a comprehensive and effective learning environment that continually adapts to the learner's needs.

b) A sequence diagram: According to Damy, the UML sequence diagram is a visual representation that illustrates the interactions between different objects in a system over time [59]. It makes it possible to model in detail the exchange of messages between these objects, following a precise chronology [60]. Each object is represented by a vertical lifeline, and the messages exchanged are represented by horizontal arrows. These diagrams are particularly useful for understanding the dynamics of a system, identifying synchronization points, and analyzing usage scenarios.

- Interpreting: The sequence diagram (Fig. 4), illustrates the dynamic and chronological interaction between the different actors in the learning system: the teacher, the learner, the group of learners, the class group and the system itself. The sequence highlights the message exchanges and communication flows between these entities. The learner initiates the process by registering and defining their profile, followed by interactions with the learning content through learning activities and assessments. The system plays a pivotal role in managing resources, processing responses, and providing automated feedback. The collaborative dimension is represented by the interactions between the learner and his or her group, as well as with the whole class, especially during discussions and shared activities. The teacher intervenes at key moments for validation and final evaluation. This sequential representation places particular emphasis on the temporal aspect and the sequencing of interactions, showing how each action triggers a specific reaction or response from other actors in the system, thus creating a coherent and structured learning flow.

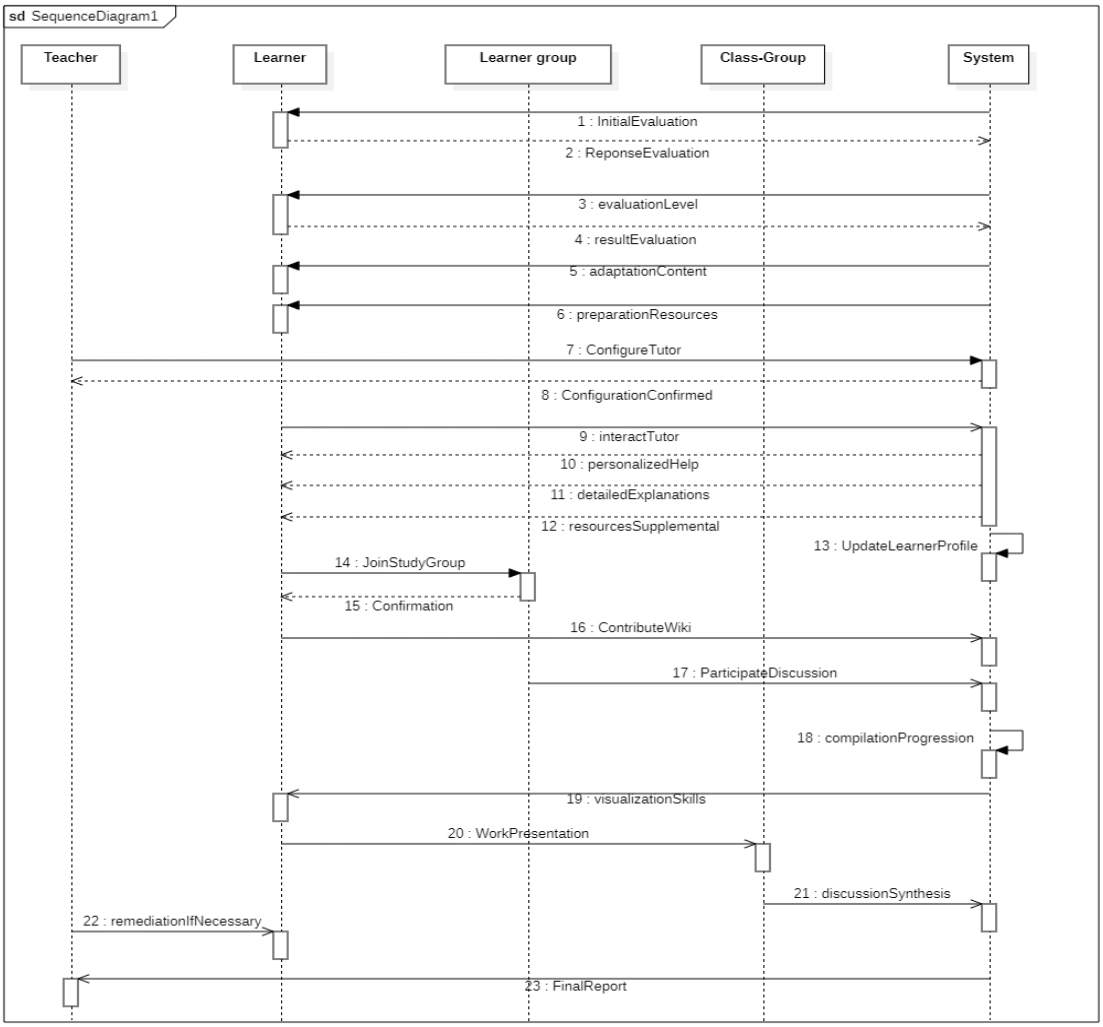


Fig. 4. Sequence diagram and interactions between system objects.

From the foregoing, the UML modelling of an integrated learning system combining adaptability, collaboration and artificial intelligence is a significant advance in the field of digital educational environments. This model demonstrates how it is possible to harmonize the personalization of learning paths with the activation of social interactions, two aspects frequently examined separately in current systems. The clear and modular structure offered by UML not only facilitates design, but also improves communication between the different actors, ranging from creators to teachers, as well as technical developers. This strategy better meets the challenging requirements of personalized, collaborative learning at scale.

The incorporation of artificial intelligence as a central element of the system represents a real lever for optimizing educational pathways and perfecting support. By examining learners' behavioral and interactive data in real time, AI can refine content, organize relevant collaborations, and provide tailored feedback. This capacity for dynamic adjustment far exceeds that of traditional adaptive systems anchored in rigid rules, offering a more responsive and resilient framework. In addition, the automation of repetitive tasks by artificial intelligence relieves teachers of certain administrative obligations, allowing them to focus more on supporting and designing innovative pedagogies. However, while promising, this modeling also presents challenges: the complexity of managing various adaptive and collaborative modules simultaneously requires a robust and scalable technical architecture that can adjust to technological and pedagogical advances. The modular approach adopted offers essential flexibility, opening up prospects for the subsequent integration of new functionalities, such as learning through gamification, predictive analysis of the risk of dropping out or the use of more advanced intelligent conversational agents. These perspectives accentuate the potential of this model to become a sustainable part of digital pedagogical practices, by meeting the growing needs for personalization and remote collaboration.

Finally, this work enriches the reflection on the creation of more intelligent and humane e-learning systems, by finely articulating complementary dimensions in a coherent structure. It encourages further research and experimentation to better understand and exploit the synergies between adaptability, collaboration and artificial intelligence for the benefit of more inclusive and effective education.

IV. Challenges and Prospects

Developing an integrative online learning system based on adaptivity, collaboration, and AI poses several challenges:

- technical interoperability,

- balancing algorithmic personalization and mediation.

For the future, the integration of immersive technologies and generative AI offers promising prospects, provided that collaboration between researchers, designers, and educators is strengthened to ensure inclusive, sustainable, and pedagogically relevant learning environments.

V. Conclusion

UML modeling of an integrative learning system, combining adaptability, collaboration and artificial intelligence, represents a major step towards creating more efficient, personalized and dynamic online educational environments. This modular framework facilitates the clear and logical organization of complex interactions between learners, content, and smart technologies, while providing essential flexibility to meet diverse user needs. The combination of personalizing journeys and encouraging social interactions, supported by sophisticated artificial intelligence mechanisms, not only improves the quality of individual learning, but also underlines the collective and collaborative dimension, which is crucial for motivation and success.

However, this innovative approach also presents considerable challenges. On the technical side, the complexity of adaptive algorithms and interaction analysis modules requires rigorous design and exhaustive validation to ensure the reliability, robustness and scalability of the system. From a pedagogical point of view, it is essential to find a balance between automation and human intervention, in order to maintain an educational mediation attentive to the emotional and cognitive needs of learners. In addition, ethical issues related to the use of personal data, the transparency of algorithmic processes and the prevention of bias must be addressed with sustained attention, thus ensuring a responsible and fair use of artificial intelligence in the field of education.

Despite these challenges, the proposed model represents a promising framework that can be adapted and expanded to different educational contexts and disciplines. It paves the way for future research to experiment and refine these architectures in real-world environments, while assessing their impact on learning, collaboration, and motivation. This work thus serves to enrich the reflection on the digital transformation of education, by putting people at the center of technological innovations and by promoting a more inclusive, agile and intelligent pedagogy.

The future of online learning systems rests on this ability to effectively combine advanced technologies and a student-centered approach to meet the complex and evolving challenges of distance and hybrid learning in an increasingly connected and demanding world.

References

[1] Dridi, C. E., Benzadri, Z. and Belala, F. (2020, June). Systems engineering system: meta-modeling perspective. In 2020, IEEE 15th International Conference of System of Systems Engineering (SoSE) (pp. 000135-000144). IEEE.

[2] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221.

[3] Brahmi, M., Louati, L., & Ghédira, K. (2023). Adaptive and collaborative e-learning systems: A comprehensive review of personalization, social interaction and AI integration. International Journal of Educational Technology in Higher Education, 20(1), 1–20. https://doi.org/10.1186/s41239-023-00423-z

[4] Aberšek, B. (2024). Artificial intelligence and the transformation of education: A cognitive–constructivist approach. Educational Technology Research and Development, 72(1), 115–132.

[5] Possati, L. (2021). The Algorithmic Unconscious: How Psychoanalysis Helps in Understanding AI. AI & Society, 36(4), 1125–1135. https://doi.org/10.1007/s00146-021-01161-8

[6] Gao, Y., Liu, Z., Si, Y., Meng, Y., Xiong, C., & Lin, Y. (2023). GEO: Generative Engine Optimization. arXiv preprint arXiv:2311.09735. https://arxiv.org/abs/2311.09735

[7] Niang, M., Mbaye, S., & Diop, M. (2022). Adaptive and Collaborative E-Learning Systems: Integrating Artificial Intelligence for Enhanced Learning Experiences. Journal of Educational Technology and Society, 25(3), 45–60.

[8] Delès, M. (2021). Digitalisation pédagogique et défis de l'enseignement à distance : Cas de l'École Supérieure de Technologie d'Agadir. Revista Multidisciplinar, 4(1), 27–48. https://doi.org/10.23882/rmd.22083

[9] Maha, K., Omar, E., Mohamed, E., & Mohamed, K. (2021). Design of educational scenarios of activities in a learning situation for online teaching. GSC Advanced Engineering and Technology, 1(1), 049-064.

[10] Basque, J., Paquette, G., & Henri, F. (2022). Digital Learning Environments and Pedagogical Engineering: What are we Talking About?. Learning and Teaching on the Web: What Pedagogical Engineering?, 23.

[11] Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press.

[12] Johnson, D. W., & Johnson, R. T. (2009). An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning. Educational Researcher, 38(5), 365–379.

[13] Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), Collaborative learning: Cognitive and computational approaches (pp. 1–19). Oxford: Elsevier.

[14] Kirschner, P. A., Sweller, J., & Clark, R. E. (2009). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2), 75-86.

[15] Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 97–115). Cambridge University Press.

[16] Strijbos, J. W., & De Laat, M. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. Computers in Human Behavior, 26(4), 495–505.

[17] Slavin, R. E. (1980). Cooperative learning. Review of Educational Research, 50(2), 315-342

[18] Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In D. Jonassen & S. Land (Eds.), Theoretical foundations of learning environments (2nd ed., pp. 277–294). Routledge.

[19] Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. Engineering education, 78(7), 674–681.

[20] Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. Psychological Science in the Public Interest, 9(3), 105–119.

[21] Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P.

[22] Conati, C., & Kardan, S. (2013). Student modeling: Supporting personalized instruction, from problem-solving to exploratory open-ended activities. AI Magazine, 34(3), 13–26.

[23] Bloom, B. S. (1968). Learning for mastery. Instructional objectives, 1(2), 5-8.

[24] Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218.

[25] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112.

[26] Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent Web-based educational systems. International Journal of Artificial Intelligence in Education, 13(2-4), 1-38.

[27] Baker, R. S., & Siemens, G. (2014). Educational Data Mining and Learning Analytics. In R. K. Sawyer (ed.), The Cambridge Handbook of the Learning Sciences (2nd ed., pp. 253-274). Cambridge University Press.

[28] Chen, C. M., Huang, T. C., & Lee, P. Y. (2020). AI-enabled personalized learning: A review of the literature. Journal of Educational Computing Research, 58(2), 223-247.

[29] Rosé, C. P., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning, 3(3), 237–271.

[30] Baker, R. S. J. d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61–75). Springer.

[31] Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26(2), 582-599.

[32] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson.

[33] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability and Transparency, 149–159.

[34] Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.

[35] Ba, D. (2023). Design and implementation of a girls' monitoring application at the Maison De l'Education.

[36] Bodiang, A. (2018). Design of an application for the automation of land use planning for the Société Civile Immobilière "LE TOIT DU SUD".

[37] Diatta, A. A. I. (2019). Design and development of a charging management application for UASZ staff.

[38] Batoure, A. B. (2020). Use of model-driven engineering for modeling, implementing, and powering decision support systems. International Journal of Data Engineering (IJDE), 8(1), 1-12.

[39] Rosenberger, M. (2022). UML 2.x: From fundamentals to advanced modeling techniques. Springer.

[40] Ouelhadj, S., Champin, P. A., Jean-Daubias, S., & Gaillard, J. (2023, January). Method of semantic data enrichment using annotated UML. In EGC 2023-Knowledge Extraction and Management (pp. 635-636).

[41] Germain, L., Dufresne, J., & Trépanier, A. (2024). Visual modeling and collaboration in software teams: The enduring role of UML. Software Practice and Experience, 54(1), 45–62.

[42] Marmoul, T. E., Belaïd, R., & Hossain, A. (2024). Extending UML for adaptive system design: A survey and framework proposal. International Journal of Advanced Computer Science, 35(1), 87–103.

[43] Conati, C., Jaques, N., & Muldner, K. (2021). AI in education: Adaptive learning and affective computing. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.), Advances in Intelligent Tutoring Systems (pp. 55–72). Springer.

[44] Millán, E., & Chen, Y. (2023). Intelligent adaptive learning: Towards granular personalization with AI techniques. Computers & Education: Artificial Intelligence, 4, 100131.

[45] Kreijns, K., Kirschner, P. A., Vermeulen, M., Buuren, H., & Acker, F. (2022). The myth of the digital native and the multi-tasking learner: The reality of collaborative learning online. Educational Technology Research and Development, 70(1), 23–42.

[46] Dillenbourg, P. (2021). Orchestrating Collaborative Learning: Contemporary Perspectives. Springer.

[47] Holmes, W., Bialik, M., & Fadel, C. (2022). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning (2nd ed.). Center for Curriculum Redesign

[48] Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. E. O'Malley (Ed.), Computer supported collaborative learning (pp. 69–97). Berlin: Springer.

[49] Dennis, A., Wixom, B. H., & Roth, R. M. (2020). Systems Analysis and Design (7th ed.). Wiley.

[50] Müller, C., & Reichenbach, K. (2014). Adaptive Learning Systems: Principles and Implementation. Journal of Educational Technology & Society, 17(3), 45–59.

[51] Chehida, S., Baouya, A. and Bensalem, S. (2021, October). Component-based approach combining UML and BIP for rigorous system design. In International Conference on the Formal Aspects of Component Software (pp. 27-43). Cham: Springer International Publishing.

[52] Latte, M. R., & Javalagi, C. M. (2019, December). Theoretical and statistical analysis of inventory and warehouse management in supply chain management – a case study on small industries. In International Conference on Advanced Production and Industrial Engineering (pp. 597-608). Singapore: Springer Nature Singapore.

[53] AL-Msie'deen, R. F., Blasi, A. H., Salman, H. E., Alja'afreh, S. S., Abadleh, A., Alsuwaiket, M. A., ... and Al-Showarah, S. A. (2022). Detect commonalities and variability in use case diagram variants. arXiv arXiv preprint: 2203.00312.

[54] Ochodek, M., Kopczyńska, S., & Staron, M. (2020). Deep learning model for end-to-end approximation of COSMIC functional size based on use case names. Information and Software Technology, 123, 106310.

[55] Babaalla, Z., Jakimi, A., Oualla, M., Saadane, R., & Chehri, A. (2024, April). Towards a self-extracting UML class diagram from the textual system specification. In the proceedings of the 7th International Conference on Networks, Intelligent Systems and Security (pp. 1-5).

[56] Cormi, C., Abou Amsha, K., Tixier, M., and Lewkowicz, M. (2024, May). Modeling for Analysis and Design in Regulated Artifact Ecologies (MADRAE): A Case Analysis of Cooperative Telemedicine Practices. In 42nd INFormatic Congress of Organizations and Information and Decision Systems (INFORSID 2024) (pp. 11-12).

[57] BELKACEM, B., AND HICHAM, B. (2023). The automatic transformation of activity diagrams into Petri nets (PhD thesis).

[58] Diamanka, M. (2024). Access control and appointment scheduling system for companies outside the logistics distribution platform of the Autonomous Port of Dakar.

[59] Damy, S. (2024). Lectures and tutorials in UML modeling (introduction) (Doctoral thesis, Centre de téléenseignement universitaire de Franche-Comté).

[60] Bikengela, T. A. (2024). Modeling and Design of Adaptive Software Systems: Approaches and Tools. Éditions Universitaires, Paris.

Determinant Factors of Success for the Village Information System in Providing Sustainable Services and Governance

Sutia Dwi Santika, Tuga Mauritsius
Information System Management Department-BINUS Graduate Program-Master of Information System Management,
Bina, Nusantara University, Jakarta 11480, Indonesia

***Abstract*—The Indonesian government is working to implement a digital transformation aimed at providing effective services and governance at both central and local levels, similar to other developing countries. A key strategy for achieving this digital government transformation at the local level involves the implementation of village information system, which are managed by village officials. The Ministry of Villages and Development of Disadvantaged Regions is developing a village information system to support these efforts, village laws, and the Sustainable Development Goals. However, the village information system encounters ongoing challenges, such as slow access speeds and ineffective response services. This study adopts a quantitative approach by using cross-sectional questionnaires that collected 426 valid responses. This study identified seven main factors that influence the success or failure of implementing a village information system: system quality, information quality, service quality, perceived usefulness, user satisfaction, trust, and net benefits. This study contributes to the literature by recognizing these factors within the DeLone and McLean Information System Success Model and TAM frameworks, which are still rarely addressed in e-government adoption studies, especially regarding village governments in developing countries. Data analysis revealed significant relationships system quality, information quality, and service quality significantly impact perceived usefulness. Information quality, service quality, trust, and perceived usefulness significantly impact user satisfaction, and perceived usefulness and satisfaction significantly affect net benefits. This research has practical implications for the successful adoption of the village information system as part of the ministry's efforts to improve services and overall governance.**

***Keywords*—E-Government; System Quality; Service Quality; Delone and Mclean model; TAM**

I. Introduction

The Indonesian government is making strides toward digital transformation, mirroring similar efforts in other developing nations aimed at enhancing service provision and governance. Governance plays a crucial role in the success of digital transformation within the service provision sector, addressing the continuously changing needs of society [1]. These e-government services can reduce costs, accelerate service delivery, and increase accessibility; these services focus on effective service delivery to ensure optimal use of the e-government system [2].

Indonesia implements e-government services at the central and local government levels. The central and local governments in Indonesia implemented the Village Information System (VIS) to improve governance and services as a key strategy for digital government transformation, similar to initiatives in other developing countries [3]. These initiatives aim to streamline administrative processes, enhance transparency, and empower citizens by providing them with easy access to information and services. The Ministry of Village and Development of Disadvantaged Regions created the village information system to track the use of village finances and provide reliable, accurate, and sustainable data for planning village development [4]. This system is crucial for enhancing village conditions by concentrating on the needs of the community, creating transparent plans and budgets, and aligning with the Sustainable Development Goals (SDGs) for village communities.

However, integrating e-government into village information system presents major challenges that impact the long-term sustainability of services. These challenges include insufficient ICT infrastructure, which complicates service delivery [5] and a lack of digital literacy, which illustrates the importance of comprehensive training in managing the village information system. Additionally, bureaucratic resistance, concerns about data privacy, and restrictive regulations can hinder the realization of the anticipated benefits [6].

The researcher conducted interviews with the Head of Information System Development at Pusdatin and operators of the village information system to gain insights into their experiences and the challenges faced during the system's implementation. The findings from these interviews reveal several quality issues with the system. For instance, the system frequently crashes during high traffic periods, and misconfigured IP addresses can occur during data display. Additionally, the system's features do not fully meet user needs. There is also a significant gap between the targets and the actual outcomes set by the Ministry of Village and Development of Disadvantaged Regions, leading to suboptimal data utilization in decision-making and errors in data entry, which result in inaccuracies. A critical factor for the success of e-government is the quality of services provided. The ineffectiveness of the services within the village information system stems from the slow response times of the village service team and the non-functional discussion forum feature. User satisfaction with the system is influenced by the employee attitude and the efficiency of service delivery, both of which

The Ministry of Communications and Digital Aspiration Scholarship funded this research.

are vital factors affecting the overall user experience with the village information system.

The identification of issues occurring in the Village Information System was obtained from system improvement suggestions to ensure the system runs optimally and supports the needs of the Ministry of Village and Development of Disadvantaged Regions. Fig. 1 shows the categorization of problems, with the highest issues being system speed and server upgrades in the village information system.

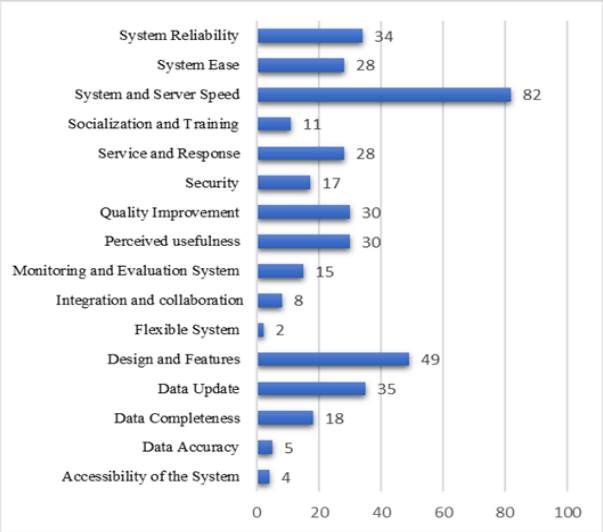


Fig. 1. Categories of issues in the village information system.

This research develops an integrated model that combines various theories, utilizing the Information System Success Model [7], [8] as the primary framework. It incorporates additional factors from the Technology Acceptance Model [9] and aspects of trust. The study tests the factors influencing the net benefits of the village information system. The findings are based on the perspectives of both employees and users within the village context, contributing to the overall success of these systems. This knowledge is valuable for the government sector, system developers, and academics, who can leverage it to enhance technology and innovation in village information system in the realm of e-government. By improving the sustainability of the village information system, the research aims to achieve better service outcomes and long-term sustainability within government organizations.

This study formulates the following research question:

RQ1: What are the factors influencing the use of village information system within e-government initiatives?

RQ2: What are the most significant factors affecting the implementation of the village information system?

The following is the structure of the article: Starting with the Introduction, Section II presents the fundamental theories underlying the proposed hypothesis and the development of the conceptual model; Section III explains the research methodology; Section IV presents the research results; Section V presents the discussion and research implications; and Section VI concludes with limitations and recommendations for future research.

II. Literature Review

A. Village Information System

The implementation of the village information system, especially through initiatives such as Village-Owned Enterprises, Village Funds, and the Village Development Index, has been crucial for the sustainability of services. Law No. 6 of 2014 mandates that the central and regional governments develop the Village Information System [10]. An interconnected system serves as a framework to enhance the quality of life, promote good governance, and foster sustainable development. The Village Development Index categorizes villages based on development criteria and aims to serve as a measurement tool for government entities. The Village Development Index classifies villages into categories: independent, advanced, developing, underdeveloped, and very underdeveloped [11]. The Village Development Index effectively correlates with the Sustainable Development Goals (SDGs) for sustainable village development. The village information system effectively provides the government with accessibility for the information-based decision-making process. Village funds are an important mechanism through which the government allocates funds to each village according to their needs, and the Village Development Index shows responsibility and strategic planning in utilizing funds from the village budget [12]. The proper utilization of village funds positively impacts the value. This approach realizes smart villages within the village development strategy, ensuring sustainability and efficiency in service delivery [13], [14].

B. DeLone and McLean Information System Success Model

The Information System Success Model (ISSM) is based on Delone and McLean and serves as the reference for this research. The information system success model was created to address the increasingly prevalent issues, such as the inability to measure the individual and organizational impact of information systems [7]. System quality, information quality, use, user satisfaction, individual impact, and organizational impact are the six components of information system success that were first identified during its development. Changes were made in the Delone and McLean model in 2003 by adding service quality and net benefits, replacing individual and organizational impact [8]. The model states that the performance of information systems is determined by independent variables from quality dimensions (information quality, system quality, and service quality) to influence usage, user satisfaction, and net benefits. The model provides insights into the success and net benefits of e-government.

1) System Quality (SYQ): System quality is a measure of the information processing system itself [7]. System quality indicates the quality of the information processing system, and information processing is built based on the desired characteristics embedded in the system. System quality serves as a measure of the characteristics of e-government system, such as usability, ease of use, reliability, and accessibility [15]. E-government is more transparent, cost-effective, efficient, and better quality relative to conventional ways. The quality of the system affects perceived usefulness and user satisfaction in this research model. Previous studies have

shown how system quality affects perceived usefulness and user satisfaction [16], [17]. Similar results were obtained by other research in the context of e-government [18], [19], [20]. Users of high-quality information systems will experience satisfaction and perceive the system as useful. Based on these findings, it is assumed that system quality has a significant impact on perceived usefulness and user satisfaction from the perspective of information system.

2) Information Quality (IQ): Information quality refers to how well the delivered information satisfies consumer needs, which may be assessed through its accurate, relevance, timeliness, and completeness in meeting the requirements of the information system [21]. The information provided through e-government is highly beneficial, particularly because many members of the public are not well acquainted with government operations. This information serves to clarify how to complete various transactions [22]. The better the content of the information, the better the perception of the information system. This study examines how information quality impacts perceived usefulness and user satisfaction. The relationship between information quality and perceived usefulness and user satisfaction has been researched by studies [18], [19], [20]. Studies consistently show a positive correlation between information quality for perceived usefulness [23], and user satisfaction [24]. The findings indicate that improving the information quality can result in higher perceived usefulness and increased satisfaction with the system.

3) Service Quality (SQ): Service quality is the quality of user support from system service providers, which consists of five indicators: tangibles, reliability, responsiveness, assurance, and empathy [8]. Service quality includes support from the Information Systems Department and IT support personnel to comprehensively address user issues [19], [25] and integrates workflows from various government departments [22]. The village information system is still a mandatory system and has dedicated technical and non-technical departments to assist users, so the author believes that service quality influences perceived usefulness and user satisfaction. Service quality is a comparison between expectations and the perception received from a service [26]. Previous research has provided practical evidence of the positive impact of service quality on perceived usefulness and user satisfaction [16], [20].

4) User Satisfaction (US): User satisfaction is the recipient's response to the use of the system, measuring the extent to which it creates value for the company's internal and external customers [8]. User satisfaction is a crucial factor that drives continuous interaction in e-government, which influences their sustainable usage behavior [3]. The success of an information system depends on user satisfaction, which ultimately allows the net benefits of e-government to be achieved [20], [21], [27]. Certain benefits will emerge if users have a positive experience and feel satisfied when using the information system. User satisfaction with the village information system will greatly enhance the perceived benefits compared to the level of investment, maintenance costs, and operational costs of the ICT-based information system.

5) Net Benefits (NB): Net benefits refer to the degree to which information systems impact the success of diverse stakeholders, both positively and negatively. Net benefits are assessed by analyzing the effects on individuals or organizations [28]. When the net benefits of utilizing a system are favorable, it appears to enhance user pleasure. Furthermore, the probability of increased user adoption of the system rises [8].

C. Technology Acceptance Model (TAM)

Integration between the Technology Acceptance Model (TAM) and the DeLone and McLean Information Systems Success Model (ISSM) offer a comprehensive approach for assessing user acceptance and the efficacy of information systems within e-government. The Technology Acceptance Model, first proposed by Davis [9], highlights perceived ease of use and perceived usefulness as essential factors influencing technology acceptance. Researchers have integrated the TAM with the DeLone and McLean model to assess the success of e-government, revealing that this amalgamation has offered more details about its effectiveness [22]. The conceptual model excludes the variable "ease of use" as it pertains to system quality [20]. This research incorporates the perceived usefulness variable from the TAM into the Delone and McLean ISSM framework to develop the conceptual model.

1) Perceived Usefulness (PU): Perceived usefulness is the level of individual belief that the use of a certain system will enhance work performance [22]. Perceived usefulness is included in the research model to measure the extent to which individuals accept that using the village information system can improve performance and service effectiveness for the ministry and village. Thus, a usefulness system will make the organization operate better by improving internal effectiveness and decision-making, and reducing operational costs [29], making it easier to perceive the benefits of the system investment. The authors believe that perceived usefulness has a significant influence on user satisfaction and net benefits. Prior studies have shown that perceived usefulness significantly influences user satisfaction [3], [18]. Perceived usefulness also impacts net benefits [20].

D. Trust (TR)

Trust is a continuous and long-lasting relationship between two individuals that fosters confidence in transactions, thereby reducing the risk of not obtaining future benefits [30]. Trust in the context of e-government is the user's expectation that the online system provides accurate and timely information securely without technical errors to operate consistently [18], [31]. User trust looks at psychological and social factors, while system trust is primarily concerned with reliability, efficacy, and security [32]. Higher trust will increase user satisfaction because the information system ensures privacy and transaction security [20]. Study confirmed a significant relationship between trust and user satisfaction [33], [34].

This study used the DeLone and McLean Information System Success Model. However, previous studies have shown a research gap, where the DeLone and McLean ISSM is mostly used in the context of voluntary system usage [18], [19]. This approach also addresses the issues found in previous studies without examining the human factors such as psychological and social categories that influence the determination of system success [21]. This study specifically addresses the research gap by focusing on the village information system as a mandatory system. Adding human factors like trust as an independent factor and perceived usefulness as a mediating factor to the DeLone and McLean model makes the framework work better. The research model integrates the DeLone and McLean ISSM theory with the Technology Acceptance Model to enhance understanding of how experience influences the success of village information system.

III. METHODOLOGY

A. Research Design

This research employs a quantitative approach to examine how various techniques related to the dependent variable are influenced by factors affecting the success of village information system development in the context of e-government. It considers these factors from the viewpoints of both employees and users of the village information system. As a survey-based study, it utilizes a questionnaire as the main tool for data collection. The questionnaire is crafted to gather a diverse range of responses, facilitating a thorough analysis of the perceptions and experiences of both groups. Through the analysis of the collected data, the research seeks to identify key determinants that can enhance the effectiveness of e-government initiatives in rural areas. Research model from references [20]. Research model are shown in Fig. 2.

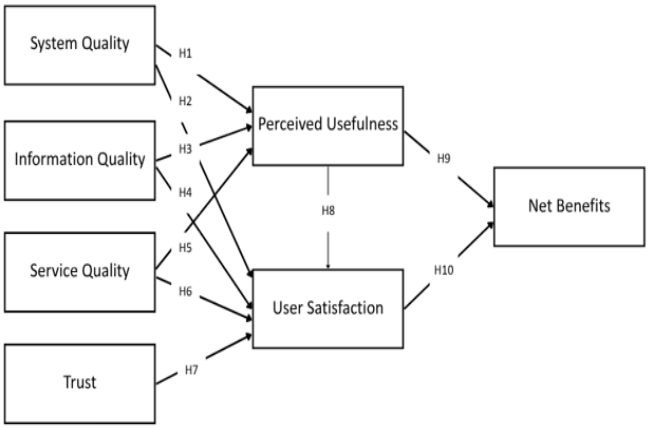


Fig. 2. Research model.

The authors created ten hypotheses (H1–H10), where the ideas about information system quality and trust are influenced by user satisfaction (US) and perceived usefulness (PU) to affect net benefit (NB). Based on the research model presented in Fig. 2, the following hypothesis is formulated in this study:

H1: System Quality has a significant impact on Perceived Usefulness.

H2: System Quality has a significant impact on User Satisfaction.

H3: Information Quality has a significant impact on Perceived Usefulness.

H4: Information Quality has a significant impact on User Satisfaction.

H5: Service Quality has a significant impact on Perceived Usefulness.

H6: Service Quality has a significant impact on User Satisfaction.

H7: Trust has a significant influence on User Satisfaction.

H8: Perceived Usefulness has a significant influence on User Satisfaction.

H9: Perceived Usefulness has a significant impact on Net Benefits.

H10: User Satisfaction has a significant impact on Net Benefits.

B. Measurement Items

This research utilized a questionnaire as the primary tool for data collection. The questions were developed through an extensive review and synthesis of relevant literature, organized into two distinct sections. Part 1 includes general information inquiries, featuring six closed-ended multiple-choice questions aimed at gathering demographic data from the respondents. Part 2 comprises 33 questions designed to assess opinions on various factors that influence the acceptance of e-government services. The questions employ a rating scale based on Likert's Scale, which ranges from 5 (mostly agree) to 1 (mostly disagree).

The questions were crafted based on insights from literature reviews and previous studies. To investigate the factors impacting the success of village information system, the study identified seven key variables: Four exogenous variables are System Quality (SYQ), Information Quality (IQ), Service Quality (SQ), and Trust (TR); two mediating variables are User Satisfaction (US) and Perceived Usefulness (PU); and one endogenous latent variable is Net Benefit (NB). Constructs and measurement items are shown in Appendix A.

C. Sample and Data Collection

Users of the village information system include all employees and village officials, who together form the population for this study. The research sample was determined using a nonprobability sampling technique, namely purposive sampling. The selection of purposive sampling is a sampling technique with specific considerations for system users who are considered to know something expected to facilitate the research. The research sample uses the Slovin formula to check validity in the equation, where a confidence level of 95% or a margin of error of 5%, which results in a minimum sample size of 370 respondents from the population.

Data collection through online questionnaires began on February 26, 2025, and ended on March 29, 2025. The research collected data using an online questionnaire. The authors then

distributed it through the correspondence application and whatsapp from all offices of the Ministry of Village and Development of Disadvantaged Regions from which 573 responses were successfully collected. Data were collected from civil servants in the Ministry of Village and Development of Disadvantaged Regions and village officials as users of the village information system from various villages in Indonesia. The authors sent 573 questionnaires to the complete research sample. We deemed 147 out of the 573 questionnaires invalid due to the absence of important information. Therefore, 426 valid responses were collected, resulting in a response rate of 74%. The data gathered from 426 responses allows for a more detailed analysis. Data supporting this study are available from the Zenodo Repository (https://zenodo.org/records/15637404), DOI: 10.5281/zenodo.1563740.

D. Data Analysis

Data analysis uses a three-stage approach to data analysis as follows: (1) Partial least squares (PLS) analysis with SmartPLS 4.0 is used to test the measurement model and identify seven dimensions. (2) Path analysis is used to test the direction and strength of hypothesized relationships. (3) A bootstrap approach is used to test the hypothesis of information system quality and trust mediated by perceived usefulness and user satisfaction towards net benefits.

IV. Result

A. Demographics

The information regarding the characteristics of the research respondents includes gender, age, education, occupation, and experience with the village information system, as presented in Table I. The demographic analysis encompasses 426 survey respondents. The findings indicate that the majority of respondents are male (n = 336, 78.87%), with an age range of 35-44 years (n = 186, 43.66%) and holding a bachelor's degree (n = 303, 71.13%). Additionally, most respondents have used the village information system for over three years (n = 280, 65.73%).

The gender demographics indicate that males represent 78.87% of the respondents, while females account for 21.13%. In terms of age distribution, 43.66% of respondents fall within the 35-44 years age range, 28.40% are aged 45-54 years, 21.83% are aged 25-34 years, 5.40% are between 55-65 years, and 0.70% are under 25 years. This finding suggests that most respondents are of an age likely to adapt to technological advancements, enabling them to utilize information systems efficiently. Regarding education, a significant majority of respondents hold a bachelor's degree (71.13%). This level of education indicates that most respondents possess adequate knowledge and skills to effectively operate technology. Job demographics include civil servants (83.57%) and village officials (16.43%). Those with over three years of experience predominantly observe user experience with the system (65.73%). This data suggests that users who are already familiar with the system are less likely to encounter significant changes and tend to continue using it.

B. Measurement Model Assessment

The measurement model integrates the construct with the structural model [35]. The measurement model analyzes the model's fit, consisting of tests of convergent validity and discriminant validity.

TABLE I. Demographic Information of the Respondents

Category	Subcategory	Count (n)	Percentage
Gender	Male	336	78.87%
	Female	90	21.13%
Age	< 25 Years	3	0.70%
	25-34 Years	93	21.83%
	35-44 Years	186	43.66%
	45-54 Years	121	28.40%
	55-65 Years	23	5.40%
Last Education	High School	43	10.09%
	Associate Degree	33	7.75%
	Bachelor Degree	303	71.13%
	Master Degree	45	10.56%
	Doctoral Degree	2	0.47%
Occupation	Village Officials	356	83.57%
	Civil Servants	70	16.43%
Experience Using VIS	< 1 Year	28	6.57%
	1-3 Year	118	27.70%
	> 3 Year	280	65.73%
Location	Banten	5	1.17%
	Jakarta	37	8.69%
	Jawa	368	86.38%
	Kalimantan	5	1.17%
	Nusa Tenggara Timur	2	0.47%
	Sulawesi	5	1.17%
	Sumatra	4	0.94%

1) Convergent validity test: Convergent validity refers to the degree to which a construct effectively explains an item. To assess convergent validity, begin by calculating the indicator loading factor. A loading factor of ≥ 0.708 is used, as this indicates that each construct accounts for more than 50% of the indicators. Next, compute the *Average Variance Extracted* (AVE) using the loading factor values for each indicator within a construct to determine the average value. AVE values of ≥ 0.50 are deemed valid [35]. The loading factor results are in Table II.

According to the calculations in Table II, two latent variables have outer loading values ≤ 0.708, namely SYQ2 and SYQ3. Therefore, the authors encourage testing other reflective indicators' load tests by removing SYQ2 and SYQ3 to examine changes in AVE. As identified in Table III, the AVE of SYQ increased to 0.194, respectively. So, the authors decided to leave out SYQ2 and SYQ3 from the suggested model, which led to outer loading values of 0.708 or higher for all the indicators shown in Table VI.

TABLE II. LOADING FACTOR RESULT

Construct	Indicator	Outer loadings	Parameter
System Quality	SYQ1	0.774	0.708
	SYQ2	0.611	0.708
	SYQ3	0.319	0.708
	SYQ4	0.806	0.708
	SYQ5	0.819	0.708
Information Quality	IQ1	0.794	0.708
	IQ2	0.789	0.708
	IQ3	0.834	0.708
	IQ4	0.750	0.708
	IQ5	0.778	0.708
Service Quality	SQ1	0.753	0.708
	SQ2	0.856	0.708
	SQ3	0.841	0.708
	SQ4	0.865	0.708
	SQ5	0.856	0.708
Trust	TR1	0.876	0.708
	TR2	0.866	0.708
	TR3	0.882	0.708
Perceived Usefulness	PU1	0.882	0.708
	PU2	0.902	0.708
	PU3	0.911	0.708
	PU4	0.903	0.708
	PU5	0.921	0.708
User Satisfaction	US1	0.838	0.708
	US2	0.851	0.708
	US3	0.863	0.708
	US4	0.843	0.708
	US5	0.864	0.708
Net Benefits	NB1	0.824	0.708
	NB2	0.876	0.708
	NB3	0.862	0.708
	NB4	0.880	0.708
	NB5	0.870	0.708

TABLE III. AVERAGE VARIANCE EXTRACTED (AVE)

Construct	Average Variance Extracted (AVE)		Changes
	Before	After	
SYQ	0.479	0.673	0.194
IQ	0.623	0.623	0.000
SQ	0.698	0.698	0.000
TR	0.765	0.765	0.000
US	0.725	0.725	0.000
PU	0.817	0.817	0.000
NB	0.744	0.744	0.000

2) Discriminant validity test: The assessment of discriminant validity is the extent to which a construct empirically differs from other constructs in the structural model. To evaluate discriminant validity using Fornell and Larcker, you need to check the Average Variance Extracted (AVE) for each construct and compare it to the squared correlation between all the constructs being measured. However, the Fornell and Larcker criterion is ineffective when the loadings of an indicator for a construct are low, such as when all the loadings of the indicators range from 0.65 to 0.85. The assessment of discriminant validity can also use HTMT based on the correlation. The HTMT value is not more than ≥ 0.90, which indicates invalid discriminant validity. An HTMT value ≤ 0.90 is considered valid [35]. The Fornell-Larcker criterion and the heterotrait-monotrait ratio (HTMT) are the most effective metrics for assessing the discriminant validity of the model. The Fornell and Larcker criterion and HTMT results can be seen in Table IV and V.

TABLE IV. FORNELL-LARCKER CRITERION RESULT

Construct	IQ	NB	PU	SQ	SYQ	TR	US
IQ	0.789						
NB	0.670	0.863					
PU	0.660	0.785	0.904				
SQ	0.761	0.715	0.729	0.835			
SYQ	0.681	0.566	0.590	0.609	0.820		
TR	0.631	0.720	0.684	0.726	0.503	0.875	
US	0.715	0.787	0.784	0.809	0.592	0.787	0.852

TABLE V. HETEROTRAIT-MONOTRAIT RATIO (HTMT) RESULT

Construct	IQ	NB	PU	SQ	SYQ	TR	US
IQ							
NB	0.757						
PU	0.736	0.838					
SQ	0.875	0.787	0.791				
SYQ	0.847	0.677	0.697	0.741			
TR	0.743	0.816	0.762	0.835	0.628		
US	0.815	0.858	0.846	0.899	0.714	0.899	

3) Reliability test: Reliability assessment involves the use of composite reliability and Cronbach's alpha. Reliability values ranging from 0.70 to 0.90 are considered satisfactory to good.

Reliability value of ≥ 0.95 suggests the presence of an indicator error, implying that all indicators are measuring the same construct [35]. The reliability results in the Table VI.

TABLE VI. RELIABILITY, AND CONVERGENT VALIDITY RESULT

Construct	Indicator	Outer loadings	Result	Cronbach's alpha	Composite reliability	AVE	Result
System Quality	SYQ1	0.796	Valid	0.756	0.761	0.673	Reliable
	SYQ4	0.808	Valid				
	SYQ5	0.856	Valid				
Information Quality	IQ1	0.793	Valid	0.848	0.850	0.623	Reliable
	IQ2	0.790	Valid				
	IQ3	0.834	Valid				
	IQ4	0.750	Valid				
	IQ5	0.778	Valid				
Service Quality	SQ1	0.753	Valid	0.891	0.896	0.698	Reliable
	SQ2	0.856	Valid				
	SQ3	0.841	Valid				
	SQ4	0.865	Valid				
	SQ5	0.856	Valid				
Trust	TR1	0.876	Valid	0.847	0.847	0.765	Reliable
	TR2	0.866	Valid				
	TR3	0.882	Valid				
Perceived Usefulness	PU1	0.906	Valid	0.944	0.945	0.817	Reliable
	PU2	0.917	Valid				
	PU3	0.913	Valid				
	PU4	0.903	Valid				
	PU5	0.921	Valid				
User Satisfaction	US1	0.838	Valid	0.905	0.906	0.725	Reliable
	US2	0.851	Valid				
	US3	0.863	Valid				
	US4	0.843	Valid				
	US5	0.864	Valid				
Net Benefits	NB1	0.824	Valid	0.914	0.919	0.744	Reliable
	NB2	0.876	Valid				
	NB3	0.862	Valid				
	NB4	0.880	Valid				
	NB5	0.870	Valid				

C. Structural Model Assessment

The structural equation modeling (SEM) method was applied to analyze the hypotheses of the structural model. The standard evaluation criteria to consider include the coefficient of determination (R^2), F-square (F^2), and the statistical significance and relevance of the path coefficients [35]. It is crucial to examine for collinearity before evaluating the structural relationships to avoid bias in the regression results.

1) Collinearity: The collinearity value is determined by calculating the Variance Inflation Factor (VIF). A VIF value exceeding 5 suggests a potential collinearity issue among predictor constructs; however, collinearity problems can also arise at lower VIF values ranging from 3 to 5. Ideally, the VIF value should be close to or below 3. Appendix B presents the VIF statistics from the inner model, indicating that all VIFs are above the recommended thresholds.

2) Coefficient of determination (R2) and F-square (F^2): The coefficient of determination (R^2) is measured by the endogenous construct and serves as an indicator of the sample's predictive power. R^2 values of 0.75, 0.50, and 0.25 are classified as substantial, moderate, and weak, respectively [35]. In addition, the F-square (F^2) values of 0.02, 0.15, and 0.35 correspond to small, medium, and large effect sizes, respectively [35]. The F-square and R^2 results can be seen in Table VII. Table VII indicates that both the perceived usefulness and net benefit variables have a moderate impact. Specifically, perceived usefulness accounts for 57.4% of the variance, while net benefit explains 69.3% of it. The user

satisfaction variable shows the highest R^2 value, substantially impacting the analysis in the amount of 78.4% of the variance.

D. Hypothesis Testing

The bootstrapping output results present hypothesis testing for the relationship between variables. In this study, the hypothesis testing requires the t-statistic value to reach 1.96 at a significance level of 5% (0.05).

As shown in Table VII, most of the hypotheses proposed in this study are supported, except H2. Specifically, system quality significantly and positively affects perceived usefulness (t-value = 3.264, $p < 0.05$), which confirms H1. The effect of system quality on user satisfaction, however, is not significant (t-value = 0.938, $p < 0.05$), thereby rejecting H2. The effect of information quality on perceived usefulness is significant and positive (t-value = 2.302, $p < 0.05$) and user satisfaction is significant and positive (t-value = 2.099, $p < 0.05$), which confirms H3 and H4. As predicted, service quality is positively associated with perceived usefulness (t-value = 7.514, $p < 0.05$) and user satisfaction (t-value = 5.743, $p < 0.05$), confirming H5 and H6. In addition, trust has a positive and significant effect on user satisfaction (t-value = 6.389, $p < 0.05$), and perceived usefulness has a positive and significant effect on user satisfaction (t-value = 5.655, $p < 0.05$), supporting H7 and H8. Perceived usefulness (t-value = 7.523, $p < 0.05$) and user satisfaction (t-value = 7.547, $p < 0.05$) also significantly and positively affect net benefits; therefore, H9 and H10 are confirmed. The research model accounts for 69.3% of the variance in the net benefits of using the village information system. The findings suggest that enhancing perceived usefulness and user satisfaction should be prioritized to maximize the net benefits derived from the village information system.

TABLE VII. HYPOTHESIS TESTING RESULT

Construct	R2	F2	Hypothesis	Path description	Path coefficient	T-value	P values	Result
Perceived Usefulness	0.574	0.038	H1	SYQ → PU	0.178	3.264	0.001	Supported
		0.021	H3	IQ → PU	0.159	2.302	0.021	Supported
		0.237	H5	SQ → PU	0.500	7.514	0.000	Supported
User Satisfaction	0.784	0.002	H2	SYQ → US	0.031	0.938	0.348	Not Supported
		0.013	H4	IQ → US	0.090	2.099	0.036	Supported
		0.112	H6	SQ → US	0.290	5.743	0.000	Supported
		0.187	H7	TR → US	0.311	6.389	0.000	Supported
		0.142	H8	PU → US	0.281	5.655	0.000	Supported
Net Benefits	0.693	0.240	H9	PU → NB	0.438	7.523	0.000	Supported
		0.247	H10	US → NB	0.444	7.547	0.000	Supported

V. RESULT

The findings of this study offer detailed insights into the factors that impact the success of information system in implementing village information system, achieved through the integration of the ISSM and TAM model. The study establishes seven research variables using a total of thirty-three proposed indicators. However, two indicators were excluded based on the loading factor assessment, resulting in a valid construct for further evaluation. The remaining thirty-one indicators support the reliability and validity assessment through both convergent and discriminant validity tests. The hypothesis is tested at a significance level of 5%. Additionally, the author evaluates the structural model by examining the f-square and R^2 values to determine the model's predictive power.

This section presents the validation of findings related to each research hypothesis from Table VII. This study shows that hypothesis 1 (H1), which states that system quality (SYQ) has a significant impact on perceived usefulness (PU), is consistent with previous research [19], [20]. This study shows that the village information system provides utility among village officials and civil servants who use it. Routine tasks require full functionality of the village information system. A technically adequate village information system proves their ability to complete tasks in collecting and managing village data as well as supporting services. However, the system quality derived from the village information system does not affect user satisfaction (US) with the system (H2). This finding differs from previous studies [3], [21], [36] that identified system quality as an important construct influencing user satisfaction. The quality of the system is not sufficient to make users satisfied with the village information system. First, the village information system does not adequately meet user needs because the system is not adaptive to the actual work processes required and the features are insufficient. Second, the village information system is mandatory, so the output or work outcomes have a bigger effect on user satisfaction than the system's technical quality. To make user satisfaction, it's important to make the system more flexible, complete the features and ensure it meets their individual demands.

The information quality (IQ) from the village information system has a significant relationship to perceived usefulness (PU) and user satisfaction (US), which supports hypothesis 3 (H3) and hypothesis 4 (H4), similar to what studies [18], [19], [20] found. This research proves that the village information system provides complete, relevant information according to user needs, timely, easy to understand, and accurate. The information available in the village information system has been shown to enhance and facilitate user performance in terms of productivity and decision-making.

The service quality (SQ) of the village information system demonstrates a positive correlation with perceived usefulness (PU) and user satisfaction (US), with hypotheses H5 and H6

further reinforcing findings from studies [18], [19], [20]. This research proves that the technical and non-technical support from the village greeting team, as well as the service provider of the village information system, is capable of solving user or system operator problems very well, thereby increasing user satisfaction. The evidence also indicates that users of the village information system perceive it positively in terms of responsive and reliable support for their work processes and productivity, making the service useful.

Trust (TR) significantly predicts user satisfaction; therefore, hypothesis H7 is supported. These results are similar to the findings of studies [20], [33]. The village information system is used by village officials and civil servants of the Ministry of Village and Development of Disadvantaged Regions to build trust. H7 shows that system users trust and feel safe when completing tasks and submitting data using the village information system. User trust in the system increases its usage. As a result, fostering trust increases individual satisfaction and drives overall organizational efficiency.

The perceived usefulness significantly affects user satisfaction, supporting hypothesis H8. These results are similar to the findings of [3], [20]. The village information system improves performance effectiveness, work quality, and productivity and facilitates task execution. These results indicate that users or system operators believe that the village information system helps complete daily tasks and facilitates the management of village-related data and services.

Our results confirm that perceived usefulness (PU) and user satisfaction (US) are two main determinants of net benefits (NB). PU and US have a positive impact on NB, supporting hypotheses H9 and H10. These results support previous research [17], [20], [29]. These findings indicate that the net benefits of the village information system demonstrate that this system provides utility and a positive experience when used, which will ultimately lead to perceived usefulness and user satisfaction. The use of village information system is beneficial in daily tasks related to village data management. Village information system enable users to complete tasks efficiently and provide information as needed, which in turn enhances organizational performance and decision-making. Improvements in decision-making not only simplify operations but also encourage a more engaged community. This leads to a more informed and responsive village governance structure. As a result, the village is better prepared to adapt to changing circumstances and improve service delivery.

RQ1: What are the factors influencing the use of village information system within e-government initiatives?

We identified two main factors that influence the success of adopting a village information system: (1) the ISSM Model, which includes the quality of the village information system (system quality, information quality, and service quality), proving hypotheses 1-6 (H1-H6), and (2) trust, which proves hypothesis 7 (H7). These main factors will be mediated by perceived usefulness and user satisfaction, ultimately resulting in net benefits supporting hypotheses 8-10 (H8-H10). The ISSM Model and the TAM Model serve as ideal frameworks for village information system. Therefore, this research adds to existing knowledge by focusing on these factors using the DeLone and McLean ISSM Model and the TAM, which are not often covered in studies about e-government adoption, particularly in village governance in developing countries. The gap in the literature indicates that there is a deeper understanding of how these factors interact and influence user acceptance of information systems in rural areas. By integrating insights from both models, this research aims to provide a comprehensive framework that can guide policymakers and practitioners in successfully implementing the village information system.

RQ2: What are the most significant factors affecting the implementation of the village information system?

The most factor affecting the implementation of the village information system is service quality, which positively influences perceived usefulness, referred to as hypothesis H5. Additionally, user satisfaction acts as a mediating factor that significantly influences net benefits, as outlined in hypothesis H10. Net benefits are essential for determining the overall success of the village information system, as they directly affect user engagement and the willingness to adopt the system's information. It is essential to prioritize improving service quality and perceived usefulness to fully realize the system's benefits.

A. Theoretical Implications

This study contributes to the progression of the theoretical foundation related to investigate the implementation of the village information system as a form of e-government in Indonesia, which provides a more in-depth understanding of governance and government services. This understanding can help identify the strengths and weaknesses of current practices, ultimately leading to more effective policies and improved service delivery to the community.

This study aims to analyze the factors influencing success in the village information system managed by the Ministry of Village and Development of Disadvantaged Regions, specifically by identifying the main influences on these factors. The study utilizes the Delone and McLean model for information system success [7], [8], as well as the Technology Acceptance Model (TAM) [9], modifying it to assess perceptions of usefulness and user satisfaction. The variables examined in this research include system quality, information quality, service quality, trust, perceived usefulness, user satisfaction, and net benefits. Furthermore, the study's research model remains valid and applicable to other studies across various sectors.

B. Theoretical Implications

This study presents several practical implications for government decision-makers. It advocates for the use of information systems to enhance perceived usefulness and user satisfaction in the operation of village information system by integrating the ISSM and TAM frameworks. The main practical findings of the research are as follows:

1) Designing a system that meets user needs is essential, as user satisfaction has a direct impact on overall net benefits. This process involves automating routine tasks to enhance effectiveness, information satisfaction, and enjoyment.

Additionally, it is important to meet user expectations by utilizing digital signatures to streamline processes for administrative reports, as well as by incorporating an interactive dashboard powered by artificial intelligence. This dashboard should provide real-time analytics and insights, enabling users to make informed decisions promptly.

2) Developing features and functions that deliver valuable services to users is critical because the service quality directly influences perceived usefulness, particularly in automating routine tasks to enhance performance. This may include notifications regarding system maintenance and downtime, as well as features like discussion forums. Such functionalities can cultivate a sense of community among users to exchange insights and solutions.

3) Ministries and government agencies stand to gain net benefits from integrated e-government by adopting village information system. Consequently, it is crucial for adopters to prioritize user satisfaction, as they will be the ones operating the system regularly.

4) Decision-makers need to consider the resources and knowledge available to system users and ensure compatibility with other technologies.

VI. Conclusion

This study aims to identify the factors that contribute to the success of integrated village information system. The authors outline two main objectives: (1) to investigate the factors influencing the use of village information system within e-government initiatives and (2) to examine the most significant factors affecting the implementation of these systems. By integrating the Information Systems Success Model (ISSM) and the Technology Acceptance Model (TAM), this study addresses both the success factors of village information system and user acceptance through the measurement of seven latent variables: System Quality (SYQ), Information Quality (IQ), Service Quality (SQ), Perceived Usefulness (PU), Trust (TR), User Satisfaction (US), and Net Benefits (NB). The authors propose thirty-three hypotheses to achieve the study's objectives. The findings from this study are expected to offer helpful knowledge about how these factors interact and influence the overall effectiveness of village information system. By testing the proposed hypotheses, the research aims to contribute to the existing literature on information system success and enhance our understanding of user behavior in adopting technologies.

The identified hypotheses offer clues about the research objectives. The accepted factors include system quality, information quality, and service quality, which significantly influence perceived usefulness. Information quality, service quality, trust, and perceived usefulness also significantly affect user satisfaction. Additionally, perceived usefulness and user satisfaction have a significant impact on net benefits. The factor with the highest direct influence on net benefits is user satisfaction, and service quality also significantly influences perceived usefulness. However, system quality does not significantly affect user satisfaction. This suggests that while system quality is important, it may not directly enhance the user experience as much as service quality and information quality do. Therefore, organizations should prioritize improving service and information quality to increase user satisfaction and ultimately drive better net benefits.

A. Limitations and Recommendations

This study has several limitations, the first being its dependence on a small sample size. Indonesia, as a developing country, faces challenges in the adoption of e-government, which can lead to variability in the data collected. As a result, these findings may not accurately represent the broader population. Therefore, the results of this study may not apply to regions with more advanced e-government initiatives. Further research should use a larger and more diverse sample to ensure more accurate analysis results. Additionally, including more villages from Indonesia in the sample will provide a more holistic view of how successful the village information system is. Secondly, we should integrate additional factors like training quality, task characteristics, technology, and cybersecurity. Furthermore, we should categorize trust based on both technology and e-government factors. Further research can also use other methods to investigate our model.

Recommendations for future research include considering additional factors that influence the success of e-government implementation in village information system. Furthermore, future studies should involve in-depth interviews with officials and experts regarding the village information system to better understand user experiences. These experiences can provide useful perspectives on the practical challenges and benefits associated with implementing these systems. A limitation of this research is that the conceptual model does not account for the indirect factors included in the ISSM and TAM model.

Recommendations for the Ministry of Village and Development of Disadvantaged Regions suggest that the research findings can serve as a guideline for decision-making regarding the development of the Village Information System. The ministry should conduct regular feature development, focusing on enhancing both system quality and information quality. Additionally, it is essential to provide training for civil servants and Village Officials within the Ministry. This training will enable them to operate the Village Information System more effectively and efficiently.

Authors' Contribution

Author contributions: Sutia Dwi Santika: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft. Tuga Mauritsius: Supervision, Project administration, Funding acquisition, Writing-Review & Editing.

Acknowledgment

The Ministry of Communications and Digital Aspiration Scholarship funded this research. The authors thank the Ministry of Village and Development of Disadvantaged Regions and all respondent participants for their useful research assistance.

REFERENCES

[1] D. Mutiarin, M. Yusuf, M. Lega, and A. Habibullah, "Agile bureaucracy in the digital age: assessing the public service delivery in Jambi City," International Journal of Public Administration in the Digital Age, vol. 11, no. 1, pp. 1–17, Jan. 2024, doi: 10.4018/IJPADA.356405.

[2] G. Chandrachooodan, R. Rajendran, and R. P. Rajendran, "Conceptual framework for project management for e-governance service development," International Journal of Professional Business Review, vol. 7, no. 4, pp. 1–17, Nov. 2022, doi: 10.26668/businessreview/2022.v7i4.e754.

[3] D. Kala, D. S. Chaubey, R. K. Meet, and A. S. Al-Adwan, "Impact of user satisfaction with e-government services on continuance use intention and citizen trust using TAM-ISSM framework," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 19, pp. 1–22, 2024, doi: 10.28945/5248.

[4] J. N. Shofa, "Support the achievement of the SDGs, the Ministry of Village develops village information systems / Dukung pencapaian SDGS, Kemdes kembangkan sistem informasi desa." Accessed: Oct. 17, 2024. [Online]. Available: https://www.beritasatu.com/news/697595/dukung-pencapaian-sdgs-kemdes-kembangkan-sistem-informasi-desa

[5] M. Hafel, J. Jamil, M. Umasugi, and Anfas, "Challenges of e-government implementation in the region archipelago characteristics," Jurnal Analisis Kebijakan dan Pelayanan Publik, vol. 8, no. 1, pp. 44–58, 2022.

[6] B. Pramono Rusadi and A. Fitri Rahmadany, "Analyzing supporting and inhibiting factors in the optimization of e-government in Pontianak city," Journal of Computer Science and Technology Studies, pp. 22–31, Oct. 2023, doi: 10.32996/jcsts.

[7] W. H. DeLone and E. R. McLean, "Information Systems Success: The Quest for the Dependent Variable," Information Systems Research, vol. 3, no. 1, pp. 60–95, 1992, doi: 10.1287/isre.3.1.60.

[8] W. H. DeLone and E. R. McLean, "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update," in Journal of Management Information Systems, M.E. Sharpe Inc., 2003, pp. 9–30. doi: 10.1080/07421222.2003.11045748.

[9] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Q, vol. 13, no. 3, pp. 319–339, 1989, doi: 10.2307/249008.

[10] President of Indonesia, "Law of the Republic of Indonesia Number 6 of 2014 / Undang-Undang Republik Indonesia Nomor 6 Tahun 2014," Art. no. UU Nomor 6 Tahun 2014, Jan. 2014.

[11] F. Wijayanto and A. Ridwan, "Improvement of the Village Development Index (VDI) through community-based rural development programs / Peningkatan Indeks Desa Membangun (IDM) melalui program pembangunan pedesaan berbasis komunitas," Jurnal Pengabdian West Science, vol. 03, no. 02, pp. 209–219, Feb. 2024.

[12] A. E. Iftitah and P. Wibowo, "The influence of village funds, village fund allocation, and village original income on the building village index in Gowa regency / Pengaruh dana desa, alokasi dana desa, dan pendapatan asli desa terhadap indeks desa membangun di kabupaten Gowa," Jurnal Ilmu Pemerintahan Widya Praja, vol. 48, no. 1, pp. 17–36, Jun. 2022, doi: 10.33701/jipwp.v48i1.2331.

[13] D. A. Rochman and I. U. Choiriyah, "The impact of local governance on the development of smart villages in Indonesia / dampak tata kelola lokal terhadap pengembangan desa pintar di Indonesia," Jurnal Pemberdayaan Ekonomi dan Masyarakat, vol. 1, no. 1, p. 17, Jun. 2024, doi: 10.47134/jpem.v1i1.321.

[14] A. P. E. Susilowati, R. Rachmawati, and R. Rijanta, "Smart village concept in Indonesia: ICT as determining factor," Heliyon, vol. 11, no. 1, pp. 1–18, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41657.

[15] Y. Li and H. Shang, "Service quality, perceived value, and citizens continuous-use intention regarding e-government: empirical evidence from China," Information and Management, vol. 57, no. 3, pp. 1–15, Apr. 2020, doi: 10.1016/j.im.2019.103197.

[16] Rulinawaty et al., "Investigating the influence of the updated DeLone and McLean information system (IS) success model on the effectiveness of learning management system (LMS) implementation," Cogent Education, vol. 11, no. 1, pp. 1–16, 2024, doi: 10.1080/2331186X.2024.2365611.

[17] D. Al-Fraihat, M. Joy, R. Masa'deh, and J. Sinclair, "Evaluating e-learning systems success: An empirical study," Comput Human Behav, vol. 102, pp. 67–86, Jan. 2020, doi: 10.1016/j.chb.2019.08.004.

[18] M. Al-Okaily, M. Al-Kofahi, F. S. Shiyyab, and A. Al-Okaily, "Determinants of user satisfaction with financial information systems in the digital transformation era: insights from emerging markets," Global Knowledge, Memory and Communication, 2023, doi: 10.1108/GKMC-12-2022-0285.

[19] A. Al-Okaily, M. Al-Okaily, T. Ai Ping, H. Al-Mawali, and H. Zaidan, "An empirical investigation of enterprise system user satisfaction antecedents in Jordanian commercial banks," Cogent Business and Management, vol. 8, no. 1, 2021, doi: 10.1080/23311975.2021.1918847.

[20] I. Hidayat Ur Rehman, J. Ali Turi, J. Rosak-Szyrocka, M. N. Alam, and L. Pilař, "The role of awareness in appraising the success of e-government systems," Cogent Business and Management, vol. 10, no. 1, pp. 1–18, Mar. 2023, doi: 10.1080/23311975.2023.2186739.

[21] D. Ariyanto, A. A. Dewi, H. T. Hasibuan, and R. B. Paramadani, "The success of information systems and sustainable information society: Measuring the implementation of a village financial system," Sustainability (Switzerland), vol. 14, no. 7, pp. 1–18, Apr. 2022, doi: 10.3390/su14073851.

[22] W. M. Al-Rahmi et al., "Validation of an integrated IS success model in the study of e-government," Mobile Information Systems, vol. 2022, pp. 1–16, 2022, doi: 10.1155/2022/8909724.

[23] H. Jo and D. H. Park, "Mechanisms for successful management of enterprise resource planning from user information processing and system quality perspective," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-39787-y.

[24] A. K. Abdulkareem and R. M. Ramli, "Does digital literacy predict e-government performance an extension of delone & mclean information system success model," Electronic Government, an International Journal, vol. 17, no. 1, pp. 466–493, 2021, doi: 10.1504/eg.2021.10034963.

[25] W. Setiawan Wibowo, A. Fadhil, D. Indra Sensuse, S. Lusa, P. Adi Wibowo Putro, and A. Yulfitri, "Pinpointing Factors in the Success of Integrated Information System Toward Open Government Data Initiative: A Perspective from Employees," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 14, no. 1, pp. 94–109, 2023, doi: 10.14569/IJACSA.2023.0140111.

[26] S. Nookhao and S. Kiattisin, "Achieving a successful e-government: Determinants of behavioral intention from Thai citizens perspective," Heliyon, vol. 9, no. 8, pp. 1–21, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18944.

[27] N. M. M. Anggreni, D. Ariyanto, H. B. Suprasto, and A. A. N. B. Dwirandra, "Successful adoption of the village's financial system," Accounting, vol. 6, no. 6, pp. 1129–1138, 2020, doi: 10.5267/j.ac.2020.7.005.

[28] K. Celik and A. Ayaz, "Validation of the Delone and McLean information systems success model: a study on student information system," Educ Inf Technol (Dordr), vol. 27, no. 4, pp. 4709–4727, May 2022, doi: 10.1007/s10639-021-10798-4.

[29] V. Genia, I. Eitiveni, M. R. Tirtayasa, W. S. Wibowo, T. F. Nugraha, and T. Nabarian, "Unraveling the key factors of successful ERP post implementation in the Indonesian construction context," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 18, pp. 513–545, 2023, doi: 10.28945/5177.

[30] Y. Li and J. Wang, "Evaluating the impact of information system quality on continuance intention toward cloud financial information system," Front Psychol, vol. 12, pp. 1–12, Aug. 2021, doi: 10.3389/fpsyg.2021.713353.

[31] A. I. Alkraiji, "Citizen satisfaction with mandatory e-government services: a conceptual framework and an empirical validation," IEEE Access, vol. 8, pp. 117253–117265, 2020, doi: 10.1109/ACCESS.2020.3004541.

[32] J. C. Valencia, D. F. Morante, R. Santa, and T. Tegethoff, "Does trust have an impact on system and operational effectiveness? The implementation of e-government in Colombia," Electronic Government,

an International Journal, vol. 15, no. 1, pp. 241–260, Jan. 2019, doi: 10.1504/eg.2019.10016082.

[33] S. K. Sharma and M. Sharma, "Examining the role of trust and quality dimensions in the actual usage of mobile banking services: An empirical investigation," Int J Inf Manage, vol. 44, pp. 65–75, Feb. 2019, doi: 10.1016/j.ijinfomgt.2018.09.013.

[34] T. Althinayyan and M. Alojail, "Enhancing user experience and sustainability in open banking using PLS-SEM," Sustainability (Switzerland), vol. 16, no. 22, Nov. 2024, doi: 10.3390/su16229656.

[35] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," Jan. 14, 2019, Emerald Group Publishing Ltd. doi: 10.1108/EBR-11-2018-0203.

[36] J. N. Amanda, W. A. Winarno, and A. T. Agustini, "Assessing the success of village asset management systems: An employee perspective," The Electronic Journal of Knowledge Management, vol. 21, no. 3, pp. 174–190, 2023, [Online]. Available: www.ejkm.com

[37] J. Wang, W. Zhang, P. Jiang, S. Zhao, and R. Evans, "The role of mGovernment applications in building trust during public crises: evidence from the covid-19 epidemic," Heliyon, vol. 10, no. 12, pp. 1–16, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32476.

[38] S. Laumer, C. Maier, and T. Weitzel, "Information quality, user satisfaction, and the manifestation of workarounds: a qualitative and quantitative study of enterprise content management system users," European Journal of Information Systems, vol. 26, no. 4, pp. 333–360, Jul. 2017, doi: 10.1057/s41303-016-0029-7.

[39] A. A. Pane and F. A. Simanjuntak, "Factors influencing Taxpayers intention to use online tax filing system (E-filing): a study of Indonesian Taxpayers," Journal of Tax Reform, vol. 10, no. 2, pp. 228–239, 2024, doi: 10.15826/jtr.2024.10.2.166.

[40] S. B. Surya, S. K. Gurung, S. B. Jung Bahadur Rana, and B. R. Dhungana, "E-governance, citizen satisfaction and net benefits: the moderating effect of digital divide," Cogent Business and Management, vol. 11, no. 1, pp. 1–22, 2024, doi: 10.1080/23311975.2024.2402512.

APPENDIX

APPENDIX A: MEASUREMENT ITEMS

Constructs and Source Measures	Items	Measured
System Quality; [7], [8], [21], [26], [37] Easy To Use; Reliability; Response time; Interactivity; Flexibility.	SYQ1	The village information system is easy to use.
	SYQ2	The village information system is reliable.
	SYQ3	The village information system providing a quick response time.
	SYQ4	The village information system is interactive.
	SYQ5	The village information system is flexible.
Information Quality; [7], [8], [21], [38] Completeness; Relevance; Timeliness; Understandability; Accuracy.	IQ1	The information in the village information system is complete.
	IQ2	The information in the village information system is relevant to my needs and work.
	IQ3	The information in the village information system is current and timely.
	IQ4	The information in the village information system is easy to understand.
	IQ5	The information in the village information system is accurate.
Service Quality; [8], [26], [38] Assurance; Reliability; Tangible; Responsiveness; Empathy.	SQ1	The services provided of the village information system has staff who have good knowledge.
	SQ2	The services provided of the village information system is reliable.
	SQ3	The services provided of the village information system is Tangible.
	SQ4	The services provided of the village information system can quickly address my needs.
	SQ5	The services provided of the village information system is empathy.
Trust; [21], [27], [39] Safe; Convenience; Trusted.	TR1	I believe the village information system can provide data security.
	TR2	I believe that using the village information system can provide convenience in transactions.
	TR3	The village information system can be trusted.
Perceived Usefulness; [9], [19], [31] Job Effectivenes; Quality of Work; Productivity; Addresses My Needs; Makes Job Easier.	PU1	The use of the village information system improves my performance.
	PU2	The use of the village information system improves the quality of my work.
	PU3	The use of the village information system makes my work faster.
	PU4	The use of the village information system meets my needs.
	PU5	The use of the village information system makes my work easier.
User Satisfaction; [7], [8], [15], [21] Effectiveness; Information Satisfaction; Enjoyment; Meet the expectations; Overall satisfaction.	US1	I feel that the village information system effectively meets the needs.
	US2	I feel that the village information system delivers satisfactory information.
	US3	I feel that the village information system is enjoyable to use.
	US4	I feel that the village information system meets my expectations.
	US5	Overall, I am satisfied with the village information system.
Net Benefits; [8], [21], [38], [40]. Cost-Saving; Time savings; Enhancement of communication and collaboration; Information Retrieval; Job Performance.	NB1	The village information system saves costs.
	NB2	The village information system saves time.
	NB3	The village information system enables easy and efficient communication.
	NB4	The village information system facilitates information retrieval.
	NB5	The village information system provides comfort in carrying out my work.

APPENDIX B: VARIANCE INFLATION FACTOR (VIF)

Construct	IQ	NB	PU	SQ	SYQ	TR	US
IQ			2.890				2.982
NB							
PU		2.594					2.575
SQ			2.468				3.483
SYQ			1.935				2.009
TR							2.398
US		2.594					

Machine Learning Methods for Detecting Fake News: A Systematic Literature Review of Machine Learning Applications in Key Domains

Nur Ida Aniza Rusli[1*], Nur Atiqah Sia Abdullah[2], Fatin Nabila Abd Razak[3], Nor Haniza Ramli[4]
Center of Computing Sciences-Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Cawangan Negeri Sembilan, Kampus Kuala Pilah, Malaysia[1, 4]
Center of Computing Sciences-Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Shah Alam, Malaysia[2]
English Language Department-Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang, Malaysia[3]
Knowledge and Software Engineering Research Group (KASERG), Universiti Teknologi MARA, Shah Alam, Malaysia[1, 2]

***Abstract*—Rapid digitisation in communication and online platform growth have transformed information dissemination and facilitated rapid access while simultaneously amplifying the spread of fake news. This widespread issue undermines public trust, destabilises political systems, and threatens economic stability. Machine learning techniques have been widely applied to fake news detection, but comparative analyses across specific domains such as health, politics, and economics remain limited. Existing reviews tend to focus on supervised learning methods, frequently excluding unsupervised and hybrid approaches, along with unique challenges and dataset requirements of each domain. This study conducted a systematic literature review of machine learning applications for detecting fake news across the three domains. The methodologies and metrics used were evaluated, while key challenges and opportunities were explored. The results revealed a strong reliance on supervised learning techniques, particularly in health-related contexts, where misinformation presented significant risks to public health outcomes. Deep learning methods were promising for processing complex data. Nonetheless, hybrid and unsupervised approaches were underexplored, which presented opportunities to address data scarcity and adaptability. Most datasets originated from social media platforms and news outlets. The common evaluation metrics included accuracy, but advanced measures were rarely applied, which indicated the possibility of enhancing such methods. Persistent challenges include poor data quality, bias, and ethical concerns highlighted the necessity for bias-mitigating algorithms and improved model interpretability. Specifically, economic misinformation has received less attention despite its potential to cause large-scale financial disruptions. This study highlighted that more effective, ethical, and context-specific machine learning solutions are needed to address fake news and enhance digital information credibility.**

***Keywords*—Machine learning; fake news; systematic review; health; politics; economy**

I. Introduction

Technological breakthroughs and the digital era have transformed global communication and news production, distribution, and consumption methods at incomparable rates. These advancements have been supported by prominent social media platforms and online news portals and cultivated a connected global community, which has facilitated the rapid dissemination of news across geographical and accessibility boundaries [1, 2]. Nevertheless, these positive advancements are accompanied by a corresponding negative regression. One of the most alarming examples of negative regression is increased fake news, which creates negative sociological situations. These situations range from decreased trust and reliability to election result manipulation and global health crises exacerbation [3-5].

Fake news in health, politics, and economics can lead to severe and negative outcomes. For example, spreading false health information establishes the basis for adverse outcomes. The recent indications include false narratives around the COVID vaccine, hydroxychloroquine efficiency as a cure, claims that bleach treats conditions, and smear campaigns against vaccines [6-8]. These fraudulent statements have caused public health issues, which have confused the public and encouraged it to adopt unsafe health practices, such as avoiding vaccinations after misleading information has been disseminated.

In politics, misinformation produces false perceptions and fosters public discontent against politicians and democracy, which distorts electoral outcomes. Misinformation is aimed at causing division, disseminating falsehoods about individuals and matters that damage reputations, disrupting discussions, and subsequently creating distrust against respected media outlets [9, 10]. Ultimately, misinformation is divisive and exacerbates conflict.

Fake news is also prevalent in economic news. Prevalent misinformation on the economy or inaccurate economic reporting induces stock market fluctuations, personal financial crises, and a deterioration of faith in financial assets. Furthermore, economic misinformation undermines investor trust and disrupts economic system stability.

Traditional methods for detecting fake news, such as manual fact-checking and rule-based detection, have been insufficient due to their inability to address the scale, speed, and complexity of modern misinformation [11, 12]. The current proliferation of fake news is broadly diverse, spreads

rapidly, and is high-volume. Hence, an efficient and scalable generalized detection system has become necessary. Thus, machine learning has emerged as an effective tool among researchers and practitioners for addressing misinformation. The ability of machine learning to analyse enormous datasets, identify trends, and provide accurate predictions has been used in advanced automated solutions to identify fake news. These platforms use various techniques, which include natural language processing (NLP), sentiment analysis, and deep learning, to detect and classify fake news accurately [13, 14].

Although numerous studies have applied ML techniques to fake news detection, comparative reviews that assess the performance of these methods across specific domain such as health, politics, and economics remain limited. Moreover, existing reviews often emphasize supervised learning, frequently failing to consider unsupervised and hybrid approaches, as well as the distinct challenges and dataset requirements unique to each domain. Thus, this study conducted a comprehensive systematic literature review (SLR) to examine the application of machine learning techniques for detecting fake news detection across the health, politics, and economics domains.

This study synthesized the results to advance the development of more effective, ethical, and domain-specific solutions to address fake news. Specifically, three research questions (RQ) were addressed: 1) the machine learning approaches utilized within these domains was investigated to provide insights into their methodologies and domain-specific applications (RQ1); 2) the performance and effectiveness of these techniques were evaluated by analyzing the commonly used evaluation metrics (RQ2); 3) key trends, challenges, and opportunities in the field were identified, limitations in these approaches were highlighted, and potential future research directions were proposed (RQ3).

II. Methodology

A. Research Questions

This study explored the application of machine learning in identifying fake news across the health, politics, and economics domains, guided by the following Research Questions (RQs):

- RQ1: How is machine learning applied for detecting fake news in the health, political, and economics domains?
- RQ2: What are the evaluation metrics commonly used in the health, political, and economics domains?
- RQ3: What are the key trends, challenges, and future directions in the health, political, and economics domains?

B. Literature Search Strategy

The literature search was conducted using the Web of Science, Scopus, IEEE Xplore, and SpringerLink databases. The literature search was performed using a combination of keywords and Boolean operators to ensure comprehensive retrieval of relevant studies to fake news detection. Table I details the search terms used across the three domains.

TABLE I. The Search Terms for Idetifying Fake News Detection According to the Domain

Domain	Search Term
Politics	(("fake news detection*" OR "misinformation") AND ("social media") AND ("politics" OR "bureaucracy") AND ("machine learning*" OR "deep learning" OR "hybrid learning"))
Economy	(("fake news detection*" OR "misinformation") AND ("social media") AND ("economy") AND ("machine learning*" OR "deep learning" OR "hybrid learning"))
Health	(("fake news detection*" OR "misinformation") AND ("social media") AND ("health" OR "healthcare" OR "medical care") AND ("machine learning*" OR "deep learning" OR "hybrid learning"))

C. Inclusion and Exclusion Criteria

Table II outlines the inclusion and exclusion criteria applied to ensure the quality and relevance of the selected studies.

TABLE II. The Inclusion and Exclusion Criteria

Inclusion	Exclusion
• Studies published in peer-reviewed journals or conferences; • Publications written in English; • Fully accessible articles; • Research focusing on the application of machine learning techniques for fake news detection in health, political, or economics domains; • Publications from the last 10 years to ensure relevance and reflect current trends and advancements.	• Non-peer-reviewed any articles; • Non-English publications; • Articles that are inaccessible and or lack full text; • Studies without a clear focus on machine learning applications in detecting fake news.

D. Study Selection and Screening

The study selection followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [15] (see Fig. 1). The studies were selected through identification (initial search results were exported to a reference management tool and duplicates were removed), screening (titles and abstracts were reviewed to identify potentially relevant studies), eligibility assessment (full-text articles were assessed against the inclusion and exclusion criteria), and inclusion (the remaining studies were selected for detailed analysis and synthesis).

E. Data Extraction and Screening

Data were extracted from the selected studies using a standardised data extraction form. The key information collected included:

- Publication type: The data source (whether published in a journal or presented at a conference);
- Domain-specific applications: The focus area of the fake news (categorised into health, politics, or economics domains);
- Research type: The methodological approach of the study (qualitative or quantitative research);

- Study objectives and contributions: The study goals and specific contributions;
- Dataset: The name and description of the fake news corpus used in the study;
- Data preparation and preprocessing: Details on any data cleaning, transformation, or preprocessing techniques applied to the dataset;
- Machine learning techniques used: A description of the machine learning algorithms or methods implemented in the study;
- Performance measures: The evaluation metrics used to assess the effectiveness of proposed machine learning models, such as accuracy, precision, and F-measure;
- Findings, limitations, and future work: Key insights from the study, the identified limitations, and recommendations for future research directions.

Appendix A presents the articles included in this systematic review (n = 173).

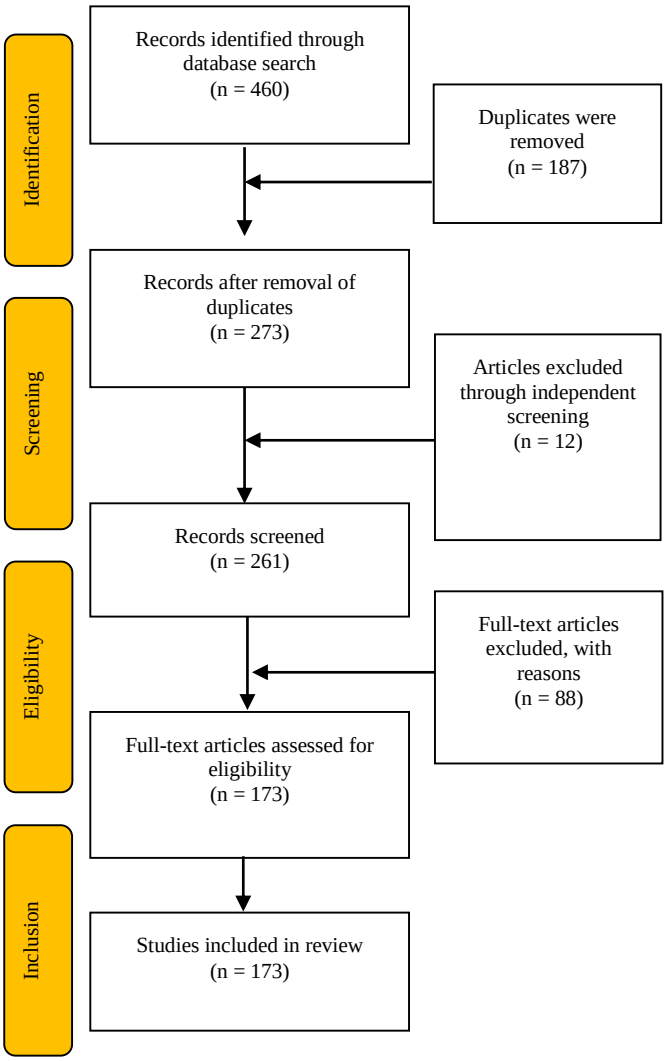


Fig. 1. Flow chart of search process adapted from PRISMA diagram (Adapted from [15]).

III. Results

A. RQ1: How is Machine Learning Applied for Detecting Fake News in the Health, Political, and Economics Domains?

This study addressed RQ1 by examining machine learning techniques for detecting fake news and their application across the health, economics, and political domains. Five primary categories of machine learning approaches were reviewed: supervised learning, unsupervised learning, deep learning, ensemble learning, and hybrid approaches.

Fig. 2 illustrates the distribution of the machine learning techniques used in the included studies. Most studies used supervised learning (n = 86 studies), where Support Vector Machine (SVM) was the most widely used method (50 studies), followed by random forest (37 studies), logistic regression (37 studies), and naïve Bayes (35 studies). Deep learning was the second most utilised technique (n = 73 studies), which reflected its increasing importance in addressing fake news detection challenges. Transformer-based models were the most frequently applied deep learning models (45 studies), followed by long short-term memory (LSTM) networks (31 studies) and CNNs (26 studies). Hybrid learning was used in 35 studies and represented a moderately adopted approach. Lastly, ensemble and unsupervised learning were the least used techniques, where each was used in only five studies, which suggested their limited applicability in this context.

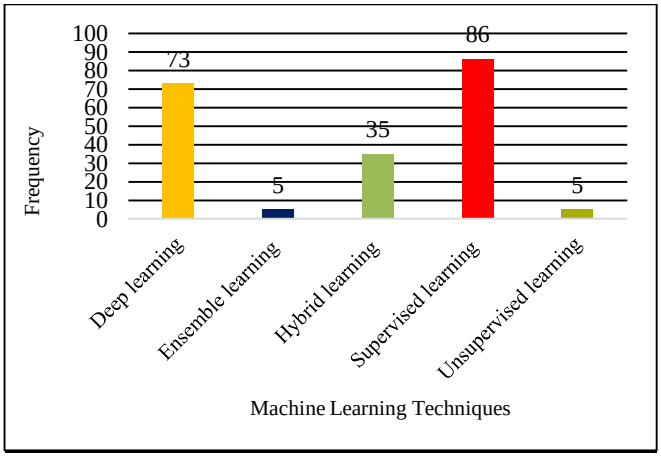


Fig. 2. Distribution of machine learning techniques in reviewed studies.

The review process was enhanced by examining the application of machine learning techniques across the health, economic, and political domains. Fig. 3 highlights that the health sector led in machine learning applications, particularly in deep learning (56 instances) and supervised learning (58 instances), which dominated. For example, studies such as [38, 83, 141] used deep learning models, such as CNNs, LSTM, and bidirectional LSTM (BiLSTM). These advancements have been driven by the increasing availability of medical datasets, which include X-ray images. The accessibility of such data has enabled deep learning models to achieve high accuracy in diagnosing disease, classifying medical images, and predicting patient outcomes. Nevertheless, ensemble learning was rarely applied, with only five instances reported in the health domain.

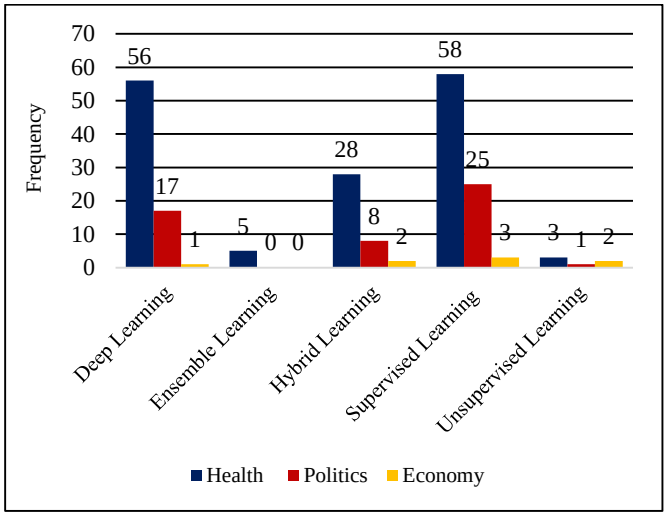


Fig. 3. Usage of machine learning techniques across domains.

The political domain demonstrated a moderate level of machine learning adoption, which was mainly in supervised learning (25 instances), with a smaller but notable use of deep learning (17 instances). Studies in [48, 185] used supervised models, such as Naïve Bayes and SVM to analyse sentiment, detecting political stances, and predicting election outcomes. These models were commonly used due to their effectiveness in processing text-based data, such as social media posts, political speeches, and news articles. In comparison, the economics domain demonstrated limited machine learning adoption. The included studies contained only a few references to deep, hybrid, supervised, and unsupervised learning.

B. RQ2: What are the Evaluation Metrics Commonly used in the Health, Political and Economic Domain?

The RQ2 was addressed by examining the evaluation methodologies used in applying machine learning to detect fake news. The commonly used datasets for evaluating the proposed machine learning and the data pre-processing techniques that enhance data quality were explored. The data preparation methods were reviewed to ensure accuracy and reliability, hyperparameter tuning strategies for optimising model performance are discussed, and a comprehensive analysis of the evaluation metrics used to assess model effectiveness is provided.

Fig. 4 presents the most frequently used datasets in fake news detection across health, politics, and economics. The results highlight the prevalence of these datasets in the included studies and the tools designed for identifying fake news. Twitter was the most frequently used dataset, where it was featured in 78 studies by a wide margin. This dominance highlighted the role of Twitter as a major platform in public discussions and real-time misinformation spread. A common research strategy involved manipulating various types of Twitter data, including raw tweet text [25, 55, 85] user IDs [16, 112, 132], tweet and retweet counts [19, 33, 60, 78, 154, 174], likes and reply counts [37, 66, 173], hashtags [54, 64, 78, 123]. Twitter metadata [22, 167, 182], URLs [20, 121], and specific keywords [54, 64, 78].

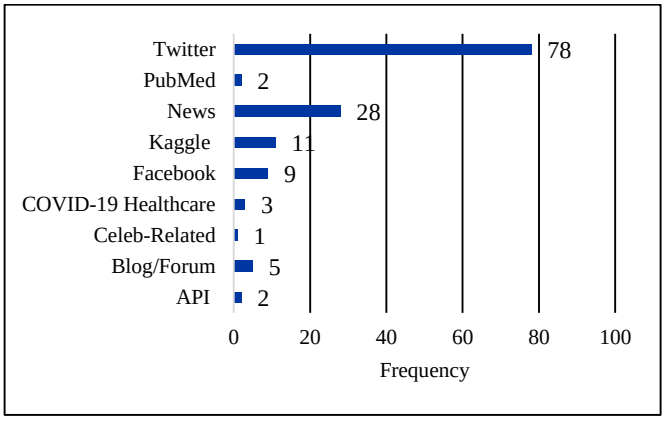


Fig. 4. Commonly used datasets in fake news detection in the included studies.

News datasets were the second most frequently used dataset (n = 28 studies). Their importance lies in their structured format, which typically includes headlines, body text, and publication dates. This structure renders them suitable for text classification and information extraction tasks, including fake news detection. For example, the FakeNewsNet dataset [42, 59, 86, 115], BuzzFeed News [44, 172], and local newspapers such as China Times [106] and The Star [169] were commonly used for analysis.

The third most frequently used dataset was the Kaggle dataset, a machine learning platform (n = 11 studies). Moderately used datasets included Facebook datasets (n = 9 studies) and other sources such as blogs and forums (n = 5 studies). Specialised datasets, including those focused on COVID-19 healthcare misinformation, PubMed, and APIs (used as data collection tools), were used less frequently, appearing in 3 studies [83, 89, 118], 2 studies [84, 178], and 2 studies [40, 181], respectively. Celebrity-related datasets were identified in only one study [150], indicating a niche research area.

Fig. 5 illustrates the frequency of commonly used preprocessing techniques in the included studies. These techniques are important for data preparation to enhance the accuracy and reliability of the proposed machine learning models. Tokenisation was the most frequently used preprocessing technique (n = 77 studies), which highlighted its importance in splitting text into meaningful units for analysis. The included studies used tokenisation in English [55, 60, 116, 145, 153], Chinese characters [18, 100, 102], Arabic [35, 74, 105, 134, 156], Turkish [110, 188], and Korean [78].

The second most common preprocessing technique was stop word removal (n = 68 studies), which highlighted its role in eliminating non-informative words, such as “and” or “the.” While stop word removal was mostly used in English, it was also applied in Chinese [18, 69, 100], Brazilian Portuguese [48, 54], Arabic [74, 134] and Urdu [188]. Data filtering was the third most frequently used preprocessing technique (n = 51 studies). The other frequently used techniques were punctuation removal (44 studies) and URL removal (37 studies), which both aimed to reduce noise in textual data. Additionally, lowercasing text and stemming were each used in 33 studies.

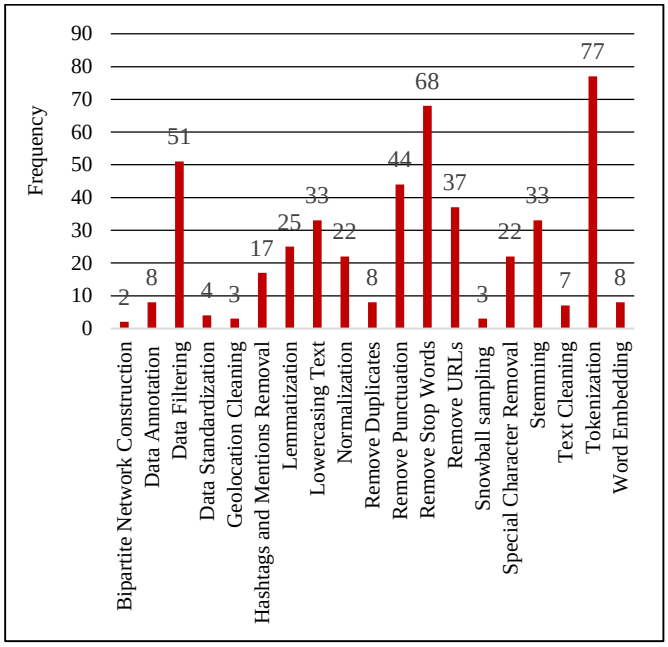


Fig. 5. Commonly used preprocessing techniques in detecting fake news.

The less commonly used preprocessing techniques were lemmatisation (25 studies), normalisation (22 studies), and special character removal (22 studies), which focus on improving text structure and uniformity. Specialised preprocessing steps, such as data annotation, duplicate removal, and word embedding (each, n = 8 studies), were utilised less frequently but were important in certain contexts. Furthermore, the snowball sampling and geolocation cleaning preprocessing techniques were minimally applied (n = 3 studies). Lastly, the least frequently used techniques were data standardisation (four studies) and bipartite network construction (two studies).

This distribution reflected the relative importance and applicability of preprocessing techniques in detecting fake news. Tokenisation, stop word removal, and data filtering were the most critical techniques. Additionally, the choice of preprocessing techniques depends on the dataset language, as different languages require specific methods to handle word forms, writing systems, and grammar.

The data preparation techniques used to evaluate fake news detection models were examined. Fig. 6 demonstrates that data splitting was the most frequently used technique and represented 55% of the methods used. This result emphasises the importance of data splitting in dividing datasets into training and testing subsets to assess model performance. Most of the studies applied 80:20 data splitting [29, 31, 42, 65, 67, 110, 126, 134, 136, 145, 159, 161, 168, 179, 181] and 70:30 data splitting [17, 35, 54, 59, 74, 81, 99, 124, 131, 160] techniques.

Cross-validation was the second most frequently used data preparation technique, representing 37% of the total. This method improves model evaluation by dividing the data into multiple folds and rotating the training and testing phases. For example, 10-fold cross-validation was used in several studies [19, 20, 25, 26, 55, 83, 85, 102, 114, 122, 132, 139, 188] and [22, 48, 97, 117, 130, 173, 184, 186] used five-fold cross-validation. Sampling accounted for 6% of the methods, where stratified sampling [84, 137] was a prominent example. Stratified sampling aids the selection of representative subsets to address class imbalances or reduce dataset size for analysis. The less common techniques were clustering [108, 148] and temporal splitting [128, 147], which each represented 1% of the methods.

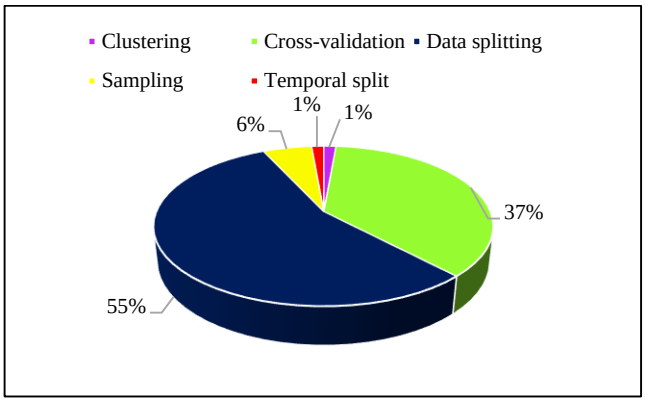


Fig. 6. Data preparation techniques for evaluation in fake news detection models.

Fig. 7 illustrates the frequency of the classification metrics used to evaluate the performance of the fake news detection models. Accuracy was the most frequently used metric (n = 116 studies), followed by precision and recall (n = 109 and 103 studies, respectively), which highlighted their importance in assessing prediction quality and true positive detection. The F1-score, which is a harmonic mean of precision and recall, was applied in 108 studies, which emphasised its role in balancing these two measures. Contrastingly, specificity and the Matthews correlation coefficient (MCC) were less commonly used (n = 3 studies). Receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics were reported in 28 studies, which underscored their usefulness in evaluating classification performance across varying thresholds.

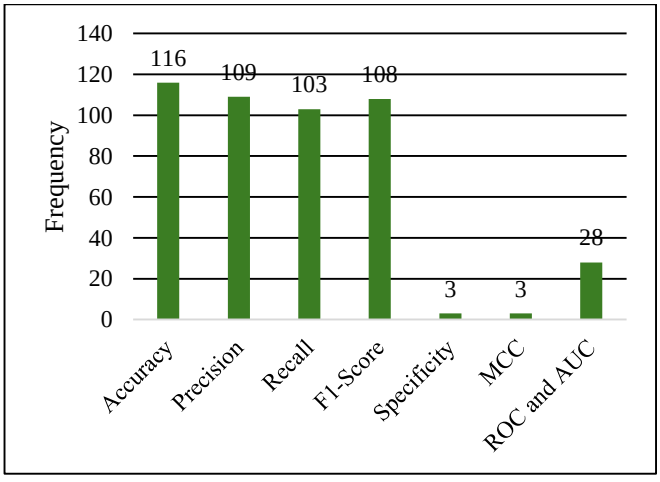


Fig. 7. Classification metrics used for evaluating fake news detection models.

C. RQ3: What are the Key Trends, Challenges, and Future Directions in the Health, Political, and Economics Domains?

The RQ3 was addressed by examining the key trends, challenges, and future directions across the health, political, and economics domains. Fig. 8 illustrates the machine learning technique use trends (deep learning, ensemble learning, hybrid learning, supervised learning, and unsupervised learning) between 2017 and 2024. Deep learning and supervised learning usage increased steadily from 2017 and peaked in 2021.

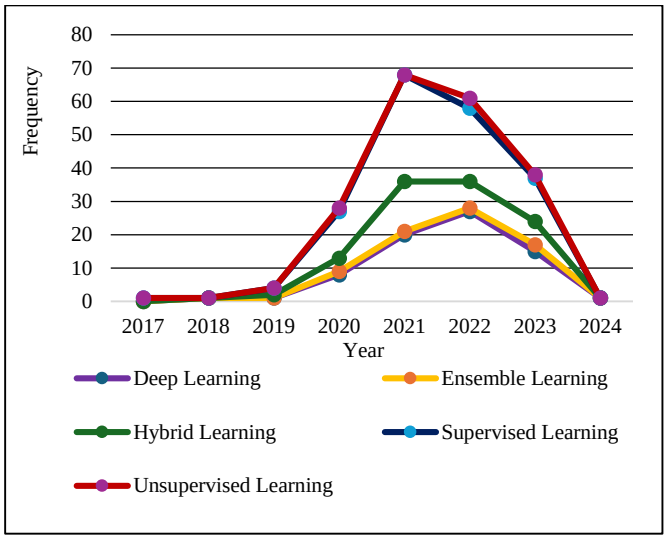


Fig. 8. Machine learning techniques trend over time.

Deep Learning usage frequency was highest at approximately 75 instances. Nonetheless, both techniques experienced a decline after 2022. Hybrid learning gained traction in 2019, peaked in 2021 (40 instances), and then its usage declined in subsequent years. Similarly, unsupervised learning grew significantly from 2019 and peaked in 2022 (60 instances) before decreasing. Contrastingly, ensemble learning demonstrated modest growth, where it peaked in 2021 but maintained a consistently lower frequency compared to the other techniques. The graph highlighted that the adoption of machine learning techniques increased significantly beginning in 2017, with most methods peaking between 2021 and 2022, followed by a noticeable decline in 2023 and 2024. This trend suggested that the use of these techniques either stabilised as research matured or shifted toward new approaches and alternative methodologies.

Fig. 9 presents the challenges of applying machine learning across various domains, which are categorised into six areas. Dataset issues were the most common challenge (124 instances), which highlighted concerns regarding data quality, availability, and relevance in machine learning applications. Contrastingly, ethical concerns were the least reported concern (n = 1 study), while interpretability issues, which relate to understanding and explaining model decisions, were recorded in two instances. Model-related bias, which affects fairness and accuracy, appeared eight times and indicated a moderate level of concern. Handling rapid data evolution was a challenge cited twice, which reflected the occasional challenges in maintaining up-to-date models. Platform-specific issues, which involved technical barriers or limitations in specific machine learning platforms, were reported in seven studies. Overall, dataset issues dominated the challenges significantly and highlighted the need for robust and relevant data, while other challenges were observed less frequently.

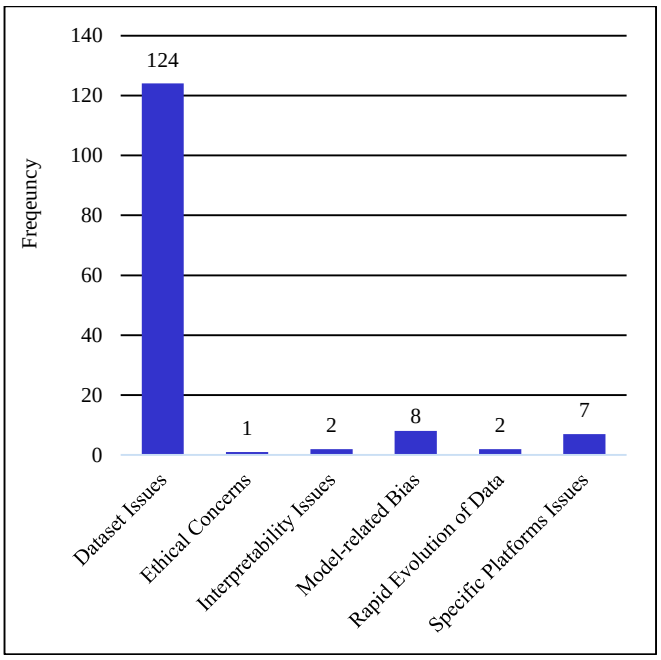


Fig. 9. Distribution of machine learning challenges across domains.

Fig. 10 highlights the key future research directions in machine learning, which are categorised into five areas. Data expansion, which involves increasing and diversifying data was the most prominent area (n = 22 studies), followed by theory exploration, which is aimed on investigating underlying principles and frameworks (n = 18 studies). Fourteen studies mentioned methodology, which involves refining and enhancing machine learning techniques. Eight studies mentioned audience or context analysis, which focuses on understanding the audience and contextual factors and indicated moderate representation.

Contrastingly, technology development, which relates to machine learning technological advancements, was mentioned least frequently (n = 2 studies). Overall, the data expansion, theory exploration, and methodology improvement as major future research focus areas, while audience or context analysis and technology development received comparatively less attention.

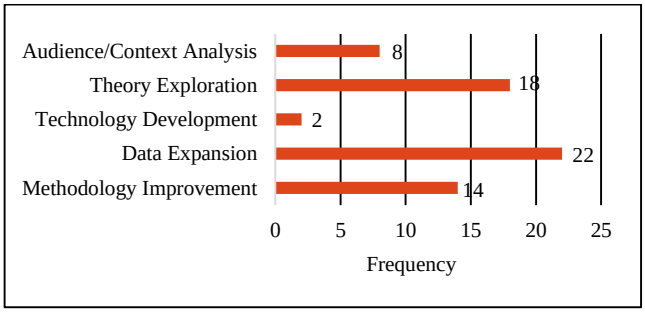


Fig. 10. Future research directions in machine learning.

IV. DISCUSSION

This review highlighted the dominance of the supervised and deep learning approaches, particularly in the health domain. The two approaches recorded the highest application frequency and exhibited domain-specific patterns. This trend highlighted the urgent need for effective misinformation detection mechanisms in health-related contexts, especially during crises, such as the COVID-19 pandemic, where accurate information dissemination is critical. The political domain demonstrated moderate machine learning adoption, which focused primarily on election misinformation. The economics domain had minimal engagement with machine learning, which suggested significant potential. These results revealed an imbalanced focus across the three domains and emphasized the need to expand research into under-represented areas, such as economic misinformation.

The prevalence of supervised and deep learning highlighted their leading role in detecting fake news (86 and 73 instances, respectively). The extensive use of supervised learning can be attributed to its simplicity and adaptability, particularly when labelled datasets are available. Deep learning has emerged as a powerful tool as it can process large datasets and capture complex patterns in textual and multimedia data, especially within the health domain. Hybrid approaches were less frequent (n = 35 studies) but were promising approaches that integrated the strengths of different methodologies. Contrastingly, the limited use of unsupervised and ensemble learning highlighted the need for innovation in using these techniques for unlabeled and heterogeneous data sources.

The dominance of Twitter datasets (n = 78 studies) highlighted the influence of the platform in misinformation analysis. Real-time interactions in Twitter, broad user engagement, and accessible API render it a valuable resource for studying misinformation trends, virality, and public response. Nevertheless, this reliance raises concerns about dataset bias, as Twitter does not fully represent online discourse across all demographics. Expanding research to other platforms, such as Facebook, TikTok, and WhatsApp can yield a more comprehensive understanding of fake news dissemination across different digital ecosystems. Nonetheless, the underutilization of domain-specific datasets, such as PubMed and COVID-19 healthcare datasets, suggested missed opportunities to develop specialized detection tools.

Dataset-related challenges were the most significant barrier in misinformation research, where 124 studies highlighted issues regarding data quality, availability, and representativeness. These challenges were evident in the health and political domains, where misinformation evolves rapidly. Health misinformation spreads quickly, especially during crises, such as pandemics, as false claims regarding treatments, vaccines, and diseases frequently emerge. Similarly, political misinformation shifts due to elections, policy changes, and global events and require continuous updates to detection models. Enhancing the accuracy and effectiveness of misinformation research requires up-to-date and diverse datasets. Outdate or biased data can lead to inaccurate predictions and reduce generalizability. Hence, ensuring high-quality, representative datasets is essential for improving model performance and effectively detecting fake news across different platforms and contexts.

The fake new detection models widely used evaluation metrics (accuracy, precision, recall, and F1-score), which reflected their reliability in assessing performance. Nevertheless, these metrics may not always provide a complete picture, especially in imbalanced datasets, where fake news instances are significantly outnumbered by real news. Less commonly used metrics (MCC and AUC-ROC) can provide deeper insights by accounting for class imbalances and assessing model discrimination ability. Incorporating these additional metrics can improve machine learning model evaluation and ensure a more comprehensive understanding of their effectiveness in detecting misinformation.

The temporal trends indicated that the adoption of machine learning techniques for detecting fake news peaked between 2021 and 2022, and declined in 2023 and 2024. This decline suggested that research in this area is either stabilizing (existing models are maturing) or shifting toward alternative approaches (hybrid models, explainable artificial intelligence, or ethical considerations in misinformation detection). Interest in hybrid and unsupervised learning approaches has grown since 2019, particularly where labelled data is scarce. Nonetheless, limited attention to ethical concerns and interpretability has highlighted the gaps in ensuring fairness and transparency in machine learning models. Future research should address these challenges by prioritizing theoretical advancements, methodology refinement, and dataset expansion, especially in underrepresented domains. Additionally, enhancing model interpretability and mitigating bias will be crucial for enhancing trust, fairness, and real-world applicability in detecting misinformation.

There were minimal reports of ethical concerns and interpretability issues, which were cited in one and two studies, respectively. This result highlighted a significant gap in addressing the broader implications of machine learning models in detecting fake news, particularly in the health and politics domains. The future research directions identified in this review emphasized the need for theory exploration (22 studies) and methodology improvement (20 studies). Dataset expansion and diversification (19 studies) was another critical area, especially in under-represented domains, such as economics. Context analysis and technology development received comparatively less attention, which highlighted opportunities for further interdisciplinary research and technological advancements.

V. CONCLUSION

Machine learning has become essential in combating misinformation. This SLR critically examined the role of machine learning techniques in detecting fake news across the health, political, and economics domains. Supervised learning was the most commonly used approach as it is effective in environments with high-quality labelled data. Deep learning techniques excel at extracting nuanced features from complex data structures, which rendered them valuable in the dynamic landscape of fake news detection. The hybrid, ensemble, and unsupervised learning techniques were underutilised, which presented opportunities for future research directions. The

analysis revealed an uneven research focus distribution, with health-related misinformation receiving the most attention, while political and economic fake news remained less explored. The results underscored the need for more targeted explorations in these areas. Data quality issues, model bias, and ethical concerns highlighted the necessity of bias-mitigating algorithms and improved interpretability to enhance trust in these technologies. Machine learning application frequency in detecting fake news peaked in 2021–2022, and then stabilised, which suggested that while the field is maturing, further methodological and theoretical advancements are needed. Despite the widespread use of datasets, such as Twitter, specialised resources (PubMed) lack engagement, which could enhance detection capabilities in domain-specific misinformation. While accuracy, precision, and recall are commonly used evaluation metrics, incorporating broader measures (MCC and AUC-ROC) could assess model performance more comprehensively. Future research can explore unsupervised and semi-supervised learning methods, which require less human intervention and adapt more efficiently to evolving misinformation. Additionally, the under-representation of machine learning in economic fake news detection is a critical area for exploration. In conclusion, while using machine learning to detect fake news has progressed significantly, continuous research is essential to address emerging challenges and adapt to the rapidly evolving digital landscape. Ensuring information credibility is vital for public trust and safety and requires constant technological advancements to counter misinformation effectively.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Higher Education, Malaysia, for funding the research project under the Fundamental Research Grant Scheme Early Career (FRGS-EC), Grant Scheme File No.: 600-RMC/FRGS-EC 5/3 (051/2024). The authors gratefully acknowledge the Faculty of Computer and Mathematical Sciences at Universiti Teknologi MARA Cawangan Negeri Sembilan and Universiti Teknologi MARA Shah Alam for supporting the publication of this paper.

REFERENCES

[1] A. Dabbous, K. Aoun Barakat, and B. de Quero Navarro, "Fake news detection and social media trust: A cross-cultural perspective," *Behav. Inf. Technol.*, vol. 41, no. 14, pp. 2953–2972, 2022, doi: 10.1080/0144929X.2021.1963475.

[2] G. Di Domenico, J. Sit, A. Ishizaka, and D. Nunan, "Fake news, social media and marketing: A systematic review," J. Bus. Res., vol. 124, pp. 329–341, 2021, doi: 10.1016/j.jbusres.2020.11.037.

[3] M. Cantarella, N. Fraccaroli, and R. Volpe, "Does fake news affect voting behaviour?," Res. Policy, vol. 52, no. 1, Art. no. 104628, 2023, doi: 10.1016/j.respol.2022.104628.

[4] T. Siddiqui and S. Gupta, "Fake news and declining media trust during COVID 19 pandemic," Int. J. Health Sci., vol. 6, no. S3, pp. 8344–8356, 2022, doi: 10.53730/ijhs.v6nS3.7916.

[5] S. Park, C. Fisher, T. Flew, and U. Dulleck, "Global mistrust in news: The impact of social media on trust," Int. J. Media Manag., vol. 22, no. 2, pp. 83–96, 2020, doi: 10.1080/14241277.2020.1799794.

[6] R. Dickinson, D. Makowski, H. van Marwijk, and E. Ford, "Exploring the role of news outlets in the rise of a conspiracy theory: Hydroxychloroquine in the early days of COVID-19," COVID, vol. 4, no. 12, pp. 1873–1896, 2024, doi: 10.3390/covid4120132.

[7] C. P. Galhardi, N. P. Freire, M. C. M. Fagundes, M. C. D. S. Minayo, and I. C. K. O. Cunha, "Fake news and vaccine hesitancy in the COVID-19 pandemic in Brazil," Ciênc. Saúde Coletiva, vol. 27, pp. 1849–1858, 2022, doi: 10.1590/1413-81232022275.24092021EN.

[8] E. Ortiz-Sánchez, A. Velando-Soriano, L. Pradas-Hernández, K. Vargas-Román, J. L. Gómez-Urquiza, G. A. Cañadas-De la Fuente, and L. Albendín-García, "Analysis of the anti-vaccine movement in social networks: A systematic review," Int. J. Environ. Res. Public Health, vol. 17, no. 15, Art. no. 5394, 2020, doi: 10.3390/ijerph17155394.

[9] C. Sindermann, A. Cooper, and C. Montag, "A short review on susceptibility to falling for fake political news," Curr. Opin. Psychol., vol. 36, pp. 44–48, 2020, doi: 10.1016/j.copsyc.2020.03.014.

[10] S. Van der Linden, C. Panagopoulos, and J. Roozenbeek, "You are fake news: Political bias in perceptions of fake news," Media, Cult. Soc., vol. 42, no. 3, pp. 460–470, 2020, doi: 10.1177/0163443720906992.

[11] H. N. Chua, Q. Khan, M. B. Jasser, and R. T. Wong, "Problem understanding of fake news detection from a data mining perspective," in *Proceedings of the 2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE)*, Penang, Malaysia, Aug. 2023, pp. 297–302, doi: 10.1109/ICCSCE58721.2023.10237152.

[12] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, and G. D. S. Martino, "Automated fact-checking for assisting human fact-checkers," *arXiv preprint*, arXiv:2103.07769, 2021, doi: 10.48550/arXiv.2103.07769.

[13] M. T. Zamir, F. Ullah, R. Tariq, W. H. Bangyal, M. Arif, and A. Gelbukh, "Machine and deep learning algorithms for sentiment analysis during COVID-19: A vision to create fake news resistant society," *PLoS One*, vol. 19, no. 12, Art. no. e0315407, 2024, doi: 10.1371/journal.pone.0315407.

[14] C. Agrawal, A. Pandey, and S. Goyal, "A survey on role of machine learning and NLP in fake news detection on social media," in *Proc. 2021 IEEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Greater Noida, India, Sept. 2021, pp. 1–7, doi: 10.1109/GUCON50781.2021.9573875.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

[16] A. Ghenai and Y. Mejova, "Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter," *arXiv preprint*, arXiv:1707.03778, 2017. Also presented at the *2017 IEEE Int. Conf. Healthcare Informatics (ICHI)*, Park City, UT, USA, Aug. 2017, doi: 10.1109/ICHI.2017.58.

[17] R. Vijayan and G. Mohler, "Forecasting retweet count during elections using graph convolution neural networks," in *Proc. 2018 IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Turin, Italy, Oct. 2018, pp. 256–262, doi: 10.1109/DSAA.2018.00036.

[18] Y. Liu, K. Yu, X. Wu, L. Qing, and Y. Peng, "Analysis and detection of health-related misinformation on Chinese social media," *IEEE Access*, vol. 7, pp. 154480–154489, 2019, doi: 10.1109/ACCESS.2019.2946624.

[19] Z. Shah, D. Surian, A. Dyda, E. Coiera, K. D. Mandl, and A. G. Dunn, "Automatically appraising the credibility of vaccine-related web pages shared on social media: A Twitter surveillance study," *J. Med. Internet Res.*, vol. 21, no. 11, Art. no. e14007, 2019, doi: 10.2196/14007.

[20] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: Detecting COVID-19 misinformation in Twitter," *IEEE Access*, vol. 8, pp. 155961–155970, 2020, doi: 10.1109/ACCESS.2020.3019600.

[21] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, and P. Rosso, "UrduFake@ FIRE2020: Shared track on fake news identification in Urdu," in *Proc. 12th Annu. Meeting Forum Inf. Retrieval Eval.*, Hyderabad, India, Dec. 2020, pp. 37–40, doi: 10.1145/3441501.3441541.

[22] A. Benazir and S. Sharmin, "Credibility assessment of user-generated health information of the Bengali language in microblogging sites employing NLP techniques," in *Proc. 2020 IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, Melbourne, Australia, Dec. 2020, pp. 837–844, doi: 10.1109/WIIAT50758.2020.00129.

[23] E. Bonnevie, J. Goldbarg, A. K. Gallegos-Jeffrey, S. D. Rosenberg, E. Wartella, and J. Smyser, "Content themes and influential voices within

vaccine opposition on Twitter, 2019," *Am. J. Public Health*, vol. 110, no. S3, pp. S326–S330, 2020, doi: 10.26633/RPSP.2021.54.

[24] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *IEEE Access*, vol. 8, pp. 165201–165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[25] P. H. A. Faustini and T. F. Covoes, "Fake news detection in multiple platforms and languages," *Expert Syst. Appl.*, vol. 158, Art. no. 113503, 2020, doi: 10.1016/j.eswa.2020.113503.

[26] S. Guarino, N. Trino, A. Celestini, A. Chessa, and G. Riotta, "Characterizing networks of propaganda on Twitter: A case study," *Appl. Netw. Sci.*, vol. 5, no. 1, Art. no. 73, 2020, doi: 10.1007/s41109-020-00286-y.

[27] F. M. Hassan and M. Lee, "Political fake statement detection via multi-stage feature-assisted neural modeling," in *Proc. 2020 IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, Arlington, VA, USA, Nov. 2020, pp. 1–6, doi: 10.1109/ISI49825.2020.9280531.

[28] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. Al Hasan, "Detection of Bangla fake news using MNB and SVM classifier," in *Proc. 2020 Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, London, UK, Aug. 2020, pp. 81–85, doi: 10.1109/iCCECE49321.2020.9231167.

[29] M. K. Jain, D. Gopalani, Y. K. Meena, and R. Kumar, "Machine learning based fake news detection using linguistic features and word vector features," in *Proc. 2020 IEEE 7th Uttar Pradesh Sect. Int. Conf. Electr., Electron. Comput. Eng. (UPCON)*, Prayagraj, India, Nov. 2020, pp. 1–6, doi: 10.1109/UPCON50219.2020.9376576.

[30] Z. Jaroucheh, M. Alissa, W. J. Buchanan, and X. Liu, "TRUSTD: Combat fake content using blockchain and collective signature technologies," in *Proc. 2020 IEEE 44th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Madrid, Spain, July 2020, pp. 1235–1240, doi: 10.1109/COMPSAC48688.2020.00-87.

[31] A. Kesarwani, S. S. Chauhan, and A. R. Nair, "Fake news detection on social media using k-nearest neighbor classifier," in *Proc. 2020 Int. Conf. Adv. Comput. Commun. Eng. (ICACCE)*, Delhi, India, June 2020, pp. 1–4, doi: 10.1109/ICACCE49060.2020.9154997.

[32] P. Ksieniewicz, P. Zyblewski, M. Choraś, R. Kozik, A. Giełczyk, and M. Woźniak, "Fake news detection from data streams," in *Proc. 2020 Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, UK, July 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207498.

[33] S. Martin, E. Kilich, S. Dada, P. E. Kummervold, C. Denny, P. Paterson, and H. J. Larson, " 'Vaccines for pregnant women…?! Absurd'–Mapping maternal vaccination discourse and stance on social media over six months," *Vaccine*, vol. 38, no. 42, pp. 6627–6637, 2020, doi: 10.1016/j.vaccine.2020.07.072.

[34] A. Rao, A. Shetty, A. Uphade, P. Thawani, and R. L. Priya, "A proposal for a novel approach to analyze and detect the fake news using AI techniques," in *Proc. 2020 3rd Int. Conf. Intell. Sustain. Syst. (ICISS)*, Palladam, India, Dec. 2020, pp. 582–589, doi: 10.1109/ICISS49785.2020.9316056.

[35] F. Saeed, W. M. S. Yafooz, M. Al-Sarem, and E. A. Hezzam, "Detecting health-related rumors on Twitter using machine learning methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110842.

[36] R. F. Sear, N. Velásquez, R. Leahy, N. J. Restrepo, S. El Oud, N. Gabriel, and N. F. Johnson, "Quantifying COVID-19 content in the online health opinion war using machine learning," *IEEE Access*, vol. 8, pp. 91886–91893, 2020, doi: 10.1109/ACCESS.2020.2993967.

[37] D. K. Singh, S. Shams, J. Kim, S. J. Park, and S. Yang, "Fighting for information credibility: An end-to-end framework to identify fake news during natural disasters," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage. (ISCRAM)*, virtual conf., May 2020, pp. 90–99. [Online]. Available: https://idl.iscram.org/files/dipaksingh/2020/2210_DipakSingh_etal2020.pdf

[38] Z. Wang, Z. Yin, and Y. A. Argyris, "Detecting medical misinformation on social media using multimodal deep learning," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2193–2203, 2020, doi: 10.1109/JBHI.2020.3037027.

[39] M. A. Wani, N. Agarwal, and P. Bours, "Impact of unreliable content on social media users during COVID-19 and stance detection system," *Electronics*, vol. 10, no. 1, Art. no. 5, 2020, doi: 10.3390/electronics10010005.

[40] N. C. Wickramarathna, T. D. Jayasiriwardena, M. Wijesekara, P. B. Munasinghe, and G. U. Ganegoda, "A framework to detect Twitter platform manipulation and computational propaganda," in *Proc. 2020 20th Int. Conf. Adv. ICT Emerging Regions (ICTer)*, Colombo, Sri Lanka, Nov. 2020, pp. 214–219, doi: 10.1109/ICTer51097.2020.9325500.

[41] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in *Proc. 2020 Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, UK, July 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206973.

[42] M. N. Alenezi and Z. M. Alqenaei, "Machine learning in detecting COVID-19 misinformation on Twitter," *Future Internet*, vol. 13, no. 10, Art. no. 244, 2021, doi: 10.3390/fi13100244.

[43] Y. A. Argyris, K. Monu, P. N. Tan, C. Aarts, F. Jiang, and K. A. Wiseley, "Using machine learning to compare provaccine and antivaccine discourse among the public on social media: Algorithm development study," *JMIR Public Health Surveill.*, vol. 7, no. 6, Art. no. e23105, 2021, doi: 10.2196/23105.

[44] N. F. Baarir and A. Djeffal, "Fake news detection using machine learning," in *Proc. 2020 2nd Int. Workshop Human-Centric Smart Environ. Health Well-Being (IHSH)*, Le Havre, France, Feb. 2021, pp. 125–130, doi: 10.1109/SERA57763.2023.10197712.

[45] I. Baris and Z. Boukhers, "ECOL: Early detection of COVID lies using content, prior knowledge, and source information," in *Combating Online Hostile Posts in Regional Languages During Emergency Situation*, R. Mandl, S. Ekbal, and A. Das, Eds. Cham, Switzerland: Springer, 2021, vol. 1, pp. 141–152, doi: 10.1007/978-3-030-73696-5_14.

[46] K. Barnes, T. Riesenmy, M. D. Trinh, E. Lleshi, N. Balogh, and R. Molontay, "Dank or not? Analyzing and predicting the popularity of memes on Reddit," *Appl. Netw. Sci.*, vol. 6, no. 1, Art. no. 21, 2021, doi: 10.1007/s41109-021-00358-7.

[47] S. Bojjireddy, S. A. Chun, and J. Geller, "Machine learning approach to detect fake news, misinformation in COVID-19 pandemic," in *Proc. 22nd Annu. Int. Conf. Digital Gov. Res. (DG.O2021)*, Omaha, NE, USA, June 2021, pp. 575–578, doi: 10.1145/3463677.3463762.

[48] L. Cabral, J. M. Monteiro, J. W. F. da Silva, C. L. C. Mattos, and P. J. C. Mourão, "FakeWhatsApp.BR: NLP and machine learning techniques for misinformation detection in Brazilian Portuguese WhatsApp messages," in *Proc. 23rd Int. Conf. Enterprise Inf. Syst. (ICEIS)*, vol. 1, Apr. 2021, pp. 63–74, doi: 10.5220/0010446800630074.

[49] L. Cerbin, J. DeJesus, J. Warnken, and S. S. Gokhale, "Unmasking the mask debate on social media," in *Proc. 2021 IEEE 45th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Madrid, Spain, July 2021, pp. 677–682, doi: 10.1109/COMPSAC51774.2021.00098.

[50] S. Chen, L. Zhou, Y. Song, Q. Xu, P. Wang, K. Wang, and D. Janies, "A novel machine learning framework for comparison of viral COVID-19–related Sina Weibo and Twitter posts: Workflow development and content analysis," *J. Med. Internet Res.*, vol. 23, no. 1, Art. no. e24889, 2021, doi: 10.2196/24889.

[51] M. Cheng, C. Yin, S. Nazarian, and P. Bogdan, "Deciphering the laws of social network-transcendent COVID-19 misinformation dynamics and implications for combating misinformation phenomena," *Sci. Rep.*, vol. 11, no. 1, Art. no. 10424, 2021, doi: 10.1038/s41598-021-89202-7.

[52] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Syst. Appl.*, vol. 169, Art. no. 114171, 2021, doi: 10.1016/j.eswa.2020.114171.

[53] A. R. Daughton, D. Gerts, C. D. Shelley, N. Parikh, T. Pitts, C. W. Ross, and N. Y. V. Chavez, " 'Thought I'd share first' and other conspiracy theory tweets from the COVID-19 infodemic: Exploratory study," unpublished, 2021. [Online]. Available: https://www.cabidigitallibrary.org/doi/full/10.5555/20220513346

[54] D. V. B. de Oliveira and U. P. Albuquerque, "Cultural evolution and digital media: Diffusion of fake news about COVID-19 on Twitter," *SN*

Comput. Sci., vol. 2, Art. no. 310, pp. 1–12, 2021, doi: 10.1007/s42979-021-00836-w.

[55] T. V. Divya and B. G. Banik, "An empirical study on fake news detection system using deep and machine learning ensemble techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 12, Art. no. e12, 2021, doi: 10.14569/IJACSA.2021.0121219.

[56] J. Du, Y. Dou, C. Xia, L. Cui, J. Ma, and S. Y. Philip, "Cross-lingual COVID-19 fake news detection," in *Proc. 2021 Int. Conf. Data Mining Workshops (ICDMW)*, Auckland, New Zealand, Dec. 2021, pp. 859–862, doi: 10.1109/ICDMW53433.2021.00110.

[57] J. Du, S. Preston, H. Sun, R. Shegog, R. Cunningham, J. Boom, and C. Tao, "Using machine learning–based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of Reddit discussions," *J. Med. Internet Res.*, vol. 23, no. 8, Art. no. e26478, 2021, doi: 10.2196/26478.

[58] M. ElSherief, S. A. Sumner, C. M. Jones, R. K. Law, A. Kacha-Ochana, L. Shieber, and M. De Choudhury, "Characterizing and identifying the prevalence of web-based misinformation relating to medication for opioid use disorder: Machine learning approach," *J. Med. Internet Res.*, vol. 23, no. 12, Art. no. e30753, 2021, doi: 10.2196/30753.

[59] P. Ganesh, L. Priya, and R. Nandakumar, "Fake news detection—A comparative study of advanced ensemble approaches," in *Proc. 2021 5th Int. Conf. Trends Electron. Informat. (ICOEI)*, Tirunelveli, India, June 2021, pp. 1003–1008, doi: 10.1109/ICOEI51242.2021.9453061.

[60] D. Gerts, C. D. Shelley, N. Parikh, T. Pitts, C. Watson Ross, G. Fairchild, and A. R. Daughton, " 'Thought I'd share first' and other conspiracy theory tweets from the COVID-19 infodemic: Exploratory study," *JMIR Public Health Surveill.*, vol. 7, no. 4, Art. no. e26527, 2021, doi: 10.2196/26527.

[61] D. Guilbeault, S. Woolley, and J. Becker, "Probabilistic social learning improves the public's judgments of news veracity," *PLoS One*, vol. 16, no. 3, Art. no. e0247487, 2021, doi: 10.1371/journal.pone.0247487.

[62] S. C. Guntuku, A. M. Buttenheim, G. Sherman, and R. M. Merchant, "Twitter discourse reveals geographical and temporal variation in concerns about COVID-19 vaccines in the United States," *Vaccine*, vol. 39, no. 30, pp. 4034–4038, 2021, doi: 10.1016/j.vaccine.2021.06.014.

[63] M. M. M. Hlaing and N. S. M. Kham, "Comparative study of fake news detection using machine learning and neural network approaches," in *Proc. Int. Workshop Comput. Sci. Eng.*, June 2021, pp. 59–64, doi: 10.18178/wcse.2021.02.010.

[64] N. Kalantari, D. Liao, and V. G. Motti, "Characterizing the online discourse in Twitter: Users' reaction to misinformation around COVID-19 in Twitter," in *Proc. 2021 IEEE Int. Conf. Big Data (Big Data)*, Orlando, FL, USA, Dec. 2021, pp. 4371–4380, doi: 10.1109/BigData52589.2021.9671740.

[65] A. Kishwar and A. Zafar, "Predicting fake news using GloVe and BERT embeddings," in *Proc. 2021 6th South-East Eur. Des. Autom., Comput. Eng., Comput. Netw. Soc. Media Conf. (SEEDA-CECNSM)*, Preveza, Greece, Sept. 2021, pp. 1–6, doi: 10.1109/SEEDA-CECNSM53056.2021.9566243.

[66] A. Kothari, L. Foisey, L. Donelle, and M. Bauer, "How do Canadian public health agencies respond to the COVID-19 emergency using social media: A protocol for a case study using content and sentiment analysis," *BMJ Open*, vol. 11, no. 4, Art. no. e041818, 2021, doi: 10.1136/bmjopen-2020-041818.

[67] R. Kumari, N. Ashok, T. Ghosal, and A. Ekbal, "Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition," *Inf. Process. Manag.*, vol. 58, no. 5, Art. no. 102631, 2021, doi: 10.1016/j.ipm.2021.102631.

[68] S. W. H. Kwok, S. K. Vadde, and G. Wang, "Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis," *J. Med. Internet Res.*, vol. 23, no. 5, Art. no. e26953, 2021, doi: 10.2196/26953.

[69] J. Luo, R. Xue, J. Hu, and D. El Baz, "Combating the infodemic: A Chinese infodemic dataset for misinformation identification," *Healthcare*, vol. 9, no. 9, Art. no. 1094, 2021, doi: 10.3390/healthcare9091094.

[70] T. K. Mackey, V. Purushothaman, M. Haupt, M. C. Nali, and J. Li, "Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter," *Lancet Digit. Health*, vol. 3, no. 2, pp. e72–e75, 2021, doi: 10.1016/S2589-7500(20)30318-6.

[71] A. R. Mahlous and A. Al-Laith, "Fake news detection in Arabic tweets during the COVID-19 pandemic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 778–788, 2021, doi: 10.14569/IJACSA.2021.0120691.

[72] S. Malla and P. J. A. Alphonse, "COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets," *Appl. Soft Comput.*, vol. 107, Art. no. 107495, 2021, doi: 10.1016/j.asoc.2021.107495.

[73] M. Mattei, G. Caldarelli, T. Squartini, and F. Saracco, "Italian Twitter semantic network during the COVID-19 epidemic," *EPJ Data Sci.*, vol. 10, no. 1, Art. no. 47, 2021, doi: 10.1140/epjds/s13688-021-00301-x.

[74] D. Mohdeb, M. Laifa, and M. Naidja, "An Arabic corpus for COVID-19 related fake news," in *Proc. 2021 Int. Conf. Recent Adv. Math. Informat. (ICRAMI)*, Skikda, Algeria, Sept. 2021, pp. 1–5, doi: 10.1109/ICRAMI52622.2021.9585909.

[75] A. M. Musmar, S. Mithas, and B. Padmanabhan, "Fighting misinformation on social media: YouTube cancer videos," in *Proc. Americas Conf. Inf. Syst. (AMCIS)*, Montreal, QC, Canada, Aug. 2021, Paper 30. [Online]. Available: https://aisel.aisnet.org/amcis2021

[76] V. T. Nguyen, K. Jung, and V. Gupta, "Examining data visualization pitfalls in scientific publications," *Vis. Comput. Ind. Biomed. Art*, vol. 4, no. 1, pp. 1–15, 2021, doi: 10.1186/s42492-021-00092-y.

[77] A. Nigam, P. Jaiswal, S. Sundar, M. Poddar, N. Kumar, F. Dernoncourt, and L. A. Celi, "NLP and deep learning methods for curbing the spread of misinformation in India," *Int. J. Intell. Secur. Public Aff.*, vol. 23, no. 3, pp. 216–227, 2021, doi: 10.1080/23800992.2021.1975429.

[78] S. Park, S. Han, J. Kim, M. M. Molaie, H. D. Vu, K. Singh, and M. Cha, "COVID-19 discourse on Twitter in four Asian countries: Case study of risk communication," *J. Med. Internet Res.*, vol. 23, no. 3, Art. no. e23272, 2021, doi: 10.2196/23272.

[79] D. Pehlivanoglu, T. Lin, F. Deceus, A. Heemskerk, N. C. Ebner, and B. S. Cahill, "The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles," *Cogn. Res. Princ. Implic.*, vol. 6, pp. 1–12, 2021, doi: 10.1186/s41235-021-00292-3.

[80] H. Piedrahita-Valdés, D. Piedrahita-Castillo, J. Bermejo-Higuera, P. Guillem-Saiz, J. R. Bermejo-Higuera, J. Guillem-Saiz, and F. Machío-Regidor, "Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019," *Vaccines*, vol. 9, no. 1, Art. no. 28, 2021, doi: 10.3390/vaccines9010028.

[81] M. S. Rahman, S. Halder, M. A. Uddin, and U. K. Acharjee, "An efficient hybrid system for anomaly detection in social networks," *Cybersecurity*, vol. 4, no. 1, Art. no. 10, 2021, doi: 10.1186/s42400-021-00074-w.

[82] M. T. Rashid and D. Wang, "CovidSens: A vision on reliable social sensing for COVID-19," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 1–25, 2021, doi: 10.1007/s10462-020-09852-3.

[83] S. Rastogi, S. S. Gill, and D. Bansal, "An adaptive approach for fake news detection in social media: Single vs cross domain," in *Proc. 2021 Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, Dec. 2021, pp. 1401–1405, doi: 10.1109/CSCI54926.2021.00280.

[84] D. Röchert, G. K. Shahi, G. Neubaum, B. Ross, and S. Stieglitz, "The networked context of COVID-19 misinformation: Informational homogeneity on YouTube at the beginning of the pandemic," *Online Soc. Netw. Media*, vol. 26, Art. no. 100164, 2021, doi: 10.1016/j.osnem.2021.100164.

[85] L. Safarnejad, Q. Xu, Y. Ge, and S. Chen, "A multiple feature category data mining and machine learning approach to characterize and detect health misinformation on social media," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 43–51, 2021a, doi: 10.1109/MIC.2021.3063257.

[86] H. Samuel, F. Hassan, and O. R. Zaïane, "The need for medical professionals to join patients in the online health social media discourse," in *Proc. 14th Int. Conf. Health Informat. (HEALTHINF)*, Feb. 2021, pp. 637–644, doi: 10.5220/0010325806370644.

[87] H. Samuel, O. Zaiane, and F. Bolduc, "Evaluation of applied machine learning for health misinformation detection via survey of medical professionals on controversial topics in pediatrics," in *Proc. 5th Int.*

Conf. Med. Health Inform., May 2021, pp. 1–6. doi: 10.1145/3472813.3472814.

[88] S. M. Saravani, I. Ray, and I. Ray, "Automated identification of social media bots using deepfake text detection," in *Proc. Int. Conf. Inf. Syst. Secur.*, Cham, Switzerland: Springer, Dec. 2021, pp. 111–123, doi: 10.1007/978-3-030-92571-0_7.

[89] R. F. Sear, R. Leahy, N. J. Restrepo, Y. Lupu, and N. F. Johnson, "Machine learning language models: Achilles heel for social media platforms and a possible solution," *Adv. Artif. Intell. Mach. Learn.*, vol. 1, no. 3, pp. 191–202, 2021, doi: 10.54364/AAIML.2021.1112.

[90] R. F. Sear, R. Leahy, N. J. Restrepo, Y. Lupu, and N. F. Johnson, "Machine learning language models: Achilles heel for social media platforms and a possible solution," *Adv. Artif. Intell. Mach. Learn.*, vol. 1, no. 3, pp. 191–202, 2021, doi: 10.54364/AAIML.2021.1112.

[91] S. Shabani, Z. Charlesworth, M. Sokhn, and H. Schuldt, "SAMS: Human-in-the-loop approach to combat the sharing of digital misinformation," in *Proc. AAAI 2021 Spring Symp. Combining Mach. Learn. Knowl. Eng.*, CEUR Workshop Proc., Mar. 2021. [Online]. Available: https://arodes.hes-so.ch/record/8922?v=pdf

[92] J. S. Shim, Y. Lee, and H. Ahn, "A link2vec-based fake news detection model using web search results," *Expert Syst. Appl.*, vol. 184, Art. no. 115491, 2021, doi: 10.1016/j.eswa.2021.115491.

[93] G. Y. Smith, C. M. S. Kabban, K. M. Hopkinson, M. E. Oxley, G. E. Noel, and H. Cheng, "Sensor fusion for context analysis in social media COVID-19 data," in *Proc. NAECON 2021—IEEE Nat. Aerosp. Electron. Conf.*, Dayton, OH, USA, Aug. 2021, pp. 415–422, doi: 10.1109/NAECON49338.2021.9696396.

[94] Y. Song, X. Wang, and Y. Jia, "Deep learning-based COVID-19 Twitter analysis," in *Proc. 6th Int. Conf. Big Data Comput.*, Xi'an, China, May 2021, pp. 8–14, doi: 10.1145/3469968.3469970.

[95] J. Srinivas, K. V. S. Reddy, N. Rajasekhar, and N. G. Raju, "Analysing sentiments of people over vaccines in Reddit posts using natural language processing," in *Proc. Int. Conf. Adv. Informatics Comput. Res.*, Cham, Switzerland: Springer, Dec. 2021, pp. 123–131, doi: 10.1007/978-3-031-09469-9_11.

[96] T. Tomaszewski, A. Morales, I. Lourentzou, R. Caskey, B. Liu, A. Schwartz, and J. Chin, "Identifying false human papillomavirus (HPV) vaccine information and corresponding risk perceptions from Twitter: Advanced predictive models," *J. Med. Internet Res.*, vol. 23, no. 9, Art. no. e30451, 2021, doi: 10.2196/30451.

[97] R. Upadhyay, G. Pasi, and M. Viviani, "Health misinformation detection in web content: A structural-, content-based, and context-aware approach based on web2vec," in *Proc. Conf. Inf. Technol. Soc. Good*, Rome, Italy, Sept. 2021, pp. 19–24, doi: 10.1145/3462203.3475898.

[98] N. Velasquez, R. Leahy, N. J. Restrepo, Y. Lupu, R. Sear, N. Gabriel, and N. F. Johnson, "Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms," *Sci. Rep.*, vol. 11, no. 1, Art. no. 11549, 2021, doi: 10.1038/s41598-021-89467-y.

[99] X. Wang, F. Chao, and G. Yu, "Evaluating rumor debunking effectiveness during the COVID-19 pandemic crisis: Utilizing user stance in comments on Sina Weibo," *Front. Public Health*, vol. 9, Art. no. 770111, 2021, doi: 10.3389/fpubh.2021.770111.

[100] A. W. Wang, J. Y. Lan, M. H. Wang, and C. Yu, "The evolution of rumors on a closed social networking platform during COVID-19: Algorithm development and content study," *JMIR Med. Inform.*, vol. 9, no. 11, Art. no. e30467, 2021, doi: 10.2196/30467.

[101] H. Wang, Y. Li, M. Hutch, A. Naidech, and Y. Luo, "Using tweets to understand how COVID-19–related health beliefs are affected in the age of social media: Twitter data analysis study," *J. Med. Internet Res.*, vol. 23, no. 2, Art. no. e26302, 2021, doi: 10.2196/26302.

[102] Y. Zhao, J. Da, and J. Yan, "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning-based approaches," *Inf. Process. Manag.*, vol. 58, no. 1, Art. no. 102390, 2021, doi: 10.1016/j.ipm.2020.102390.

[103] V. Agivale, O. Bhanushali, H. Lalwani, S. Singh, B. Khemani, and S. Malve, "HealthCare misinformation detection on social media," in *Proc. 2022 Int. Conf. Smart Gener. Comput., Commun. Netw. (SMART GENCON)*, Dec. 2022, pp. 1–6, doi: 10.1109/SMARTGENCON56628.2022.10083941.

[104] Y. Albalawi, N. S. Nikolov, and J. Buckley, "Pretrained transformer language models versus pretrained word embeddings for the detection of accurate health information on Arabic social media: Comparative study," *JMIR Form. Res.*, vol. 6, no. 6, Art. no. e34834, 2022, doi: 10.2196/34834.

[105] S. B. Ali, Z. Kechaou, and A. Wali, "Arabic fake news detection in social media based on AraBERT," in *Proc. 2022 IEEE 21st Int. Conf. Cogn. Informatics & Cogn. Comput. (ICCI CC)*, Dec. 2022, pp. 214–220, doi: 10.1109/ICCICC57084.2022.10101635.

[106] Y. A. Argyris, N. Zhang, B. Bashyal, and P. N. Tan, "Using deep learning to identify linguistic features that facilitate or inhibit the propagation of anti- and pro-vaccine content on social media," in *Proc. 2022 IEEE Int. Conf. Digit. Health (ICDH)*, July 2022, pp. 107–116, doi: 10.1109/ICDH55609.2022.00025.

[107] M. Arif, A. L. Tonja, I. Ameer, O. Kolesnikova, A. F. Gelbukh, G. Sidorov, and A. G. M. Meque, "CIC at CheckThat!-2022: Multi-class and cross-lingual fake news detection," in *CLEF (Working Notes)*, Sept. 2022, pp. 434–443. [Online]. Available: https://ceur-ws.org/Vol-3180/paper-33.pdf

[108] Y. Barve and J. R. Saini, "Misinformation detection using unsupervised approach on CoAID dataset," in *Proc. 2022 Int. Conf. Futur. Technol. (INCOFT)*, Nov. 2022, pp. 1–8, doi: 10.1109/INCOFT55651.2022.10094369.

[109] S. Bhattacharyya, "Classification and analysis of general misinformation and COVID-related misinformation within subreddits of opposing political viewpoints," in *Proc. SoutheastCon 2022*, Mar. 2022, pp. 635–639, doi: 10.1109/SoutheastCon48659.2022.9764051.

[110] M. Bozuyla and A. Özçift, "Developing a fake news identification model with advanced deep language transformers for Turkish COVID-19 misinformation data," *Turk. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 3, pp. 908–926, 2022, doi: 10.3906/ELK-2106-55.

[111] D. Brown, E. Al-Hossami, Z. Cheng, S. Shaikh, D. Janies, and M. Uddin, "Multidisciplinary engagement of diverse students in computer science education through research focused on social media COVID-19 misinformation," in *Proc. 2022 ASEE Annu. Conf. & Expo.*, Aug. 2022, doi: 10.18260/1-2--41892.

[112] M. Bruno, R. Lambiotte, and F. Saracco, "Brexit and bots: Characterizing the behaviour of automated accounts on Twitter during the UK election," *EPJ Data Sci.*, vol. 11, no. 1, Art. no. 17, 2022, doi: 10.1140/epjds/s13688-022-00330-0.

[113] O. Chernyaeva, T. Hong, Y. Kim, Y. Park, G. Ren, and J. Ock, "Detecting fake news about COVID-19 infodemic using deep learning and content analysis," *Asia Pac. J. Inf. Syst.*, vol. 32, no. 4, pp. 945–963, 2022, doi: 10.14329/APJIS.2022.32.4.945.

[114] M. Daradkeh, "Analyzing sentiments and diffusion characteristics of COVID-19 vaccine misinformation topics in social media: A data analytics framework," *Int. J. Bus. Anal. (IJBAN)*, vol. 9, no. 3, pp. 1–22, 2022, doi: 10.4018/IJBAN.292056.

[115] S. D. Das, A. Basak, and S. Dutta, "A heuristic-driven uncertainty-based ensemble framework for fake news detection in tweets and news articles," *Neurocomputing*, vol. 491, pp. 607–620, 2022, doi: 10.1016/j.neucom.2022.04.067.

[116] G. De Magistris, S. Russo, P. Roma, J. T. Starczewski, and C. Napoli, "An explainable fake news detector based on named entity recognition and stance classification applied to COVID-19," *Information*, vol. 13, no. 3, Art. no. 137, 2022, doi: 10.3390/info13030137.

[117] Y. Deng and S. W. Wang, "Detecting fake news on social media by CSIBERT," in *Proc. 2022 6th Int. Conf. Deep Learn. Technol.*, Jul. 2022, pp. 142–148, doi: 10.1145/3556677.3556698.

[118] S. Di Sotto and M. Viviani, "Health misinformation detection in the social web: An overview and a data science approach," *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, p. 2173, 2022, doi: 10.3390/ijerph19042173.

[119] S. Di Sotto and M. Viviani, "Assessing health misinformation in online content," in *Proc. 37th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2022, pp. 717–722, doi: 10.1145/3477314.3507238.

[120]H. Y. Ghafoor, A. Jaffar, R. Jahangir, M. W. Iqbal, and M. Z. Abbas, "Fake news identification on social media using machine learning techniques," in *Proc. Int. Conf. Inf. Technol. Appl. (ICITA 2021)*, Apr. 2022, pp. 87–98. Springer, doi: 10.1007/978-981-16-7618-5_8.

[121]T. Ginossar, I. J. Cruickshank, E. Zheleva, J. Sulskis, and T. Berger-Wolf, "Cross-platform spread: Vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations," *Hum. Vaccines Immunother.*, vol. 18, no. 1, pp. 1–13, 2022, doi: 10.1080/21645515.2021.2003647.

[122]M. Hashemi, "Discovering social media topics and patterns in the coronavirus and election era," *J. Inf. Commun. Ethics Soc.*, vol. 20, no. 1, pp. 1–17, 2022, doi: 10.1108/JICES-04-2021-0039.

[123]K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection," *Public Health*, vol. 203, pp. 23–30, 2022, doi: 10.1016/j.puhe.2021.11.022.

[124]C. Iwendi, S. Mohan, E. Ibeke, A. Ahmadian, and T. Ciano, "COVID-19 fake news sentiment analysis," *Comput. Electr. Eng.*, vol. 101, Art. no. 107967, 2022, doi: 10.1016/j.compeleceng.2022.107967.

[125]F. Khan, R. Alturki, G. Srivastava, F. Gazzawe, S. T. U. Shah, and S. Mastorakis, "Explainable detection of fake news on social media using pyramidal co-attention network," *IEEE Trans. Comput. Soc. Syst.*, 2022, doi: 10.1109/TCSS.2022.3207993.

[126]M. G. Kim, M. Kim, J. H. Kim, and K. Kim, "Fine-tuning BERT models to classify misinformation on garlic and COVID-19 on Twitter," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, Art. no. 5126, 2022, doi: 10.3390/ijerph19095126.

[127]J. W. Lee and J. H. Kim, "Fake sentence detection based on transfer learning: Applying to Korean COVID-19 fake news," *Appl. Sci.*, vol. 12, no. 13, Art. no. 6402, 2022, doi: 10.3390/app12136402.

[128]Q. Li, J. Hou, and A. Iftekhar, "Detecting the research structure and topic trends of social media using static and dynamic probabilistic topic models," *Aslib J. Inf. Manag.*, vol. 75, no. 2, pp. 215–245, 2022, doi: 10.1108/AJIM-02-2022-0091.

[129]M. Lotto, T. Sá Menezes, I. Zakir Hussain, S. F. Tsao, Z. Ahmad Butt, P. P. Morita, and T. Cruvinel, "Characterization of false or misleading fluoride content on Instagram: Infodemiology study," *J. Med. Internet Res.*, vol. 24, no. 5, Art. no. e37519, 2022, doi: 10.2196/37519.

[130]J. Ma, Y. Liu, M. Liu, and M. Han, "Curriculum contrastive learning for fake news detection," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, Oct. 2022, pp. 4309–4313, doi: 10.1145/3511808.3557574.

[131]S. Mohanty, S. P. Dwivedy, A. A. Acharya, S. Mohapatra, S. S. Sahoo, S. Samal, and S. Samal, "Enhancing the detection of social bots on Twitter using ensemble machine learning technique," in *Proc. 2022 Int. Conf. Advancements Smart, Secure and Intell. Comput. (ASSIC)*, Nov. 2022, pp. 1–6, doi: 10.1109/ASSIC55218.2022.10088372.

[132]B. Mønsted and S. Lehmann, "Characterizing polarization in online vaccine discourse—A large-scale study," *PLOS ONE*, vol. 17, no. 2, Art. no. e0263746, 2022, doi: 10.1371/journal.pone.0263746.

[133]R. Nangi and K. V. Pradeepthi, "Combating fake news with machine learning and deep learning methods," in *Proc. Int. Conf. Signal & Data Process.*, Singapore: Springer Nature Singapore, Jun. 2022, pp. 337–345, doi: 10.1007/978-981-99-1410-4_28.

[134]A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Arabic fake news detection based on deep contextualized embedding models," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16019–16032, 2022, doi: 10.1007/s00521-022-07206-4.

[135]M. Patel, J. Padiya, and M. Singh, "Fake news detection using machine learning and natural language processing," in *Combating Fake News with Computational Intelligence Techniques*, Springer, 2022, pp. 127–148, doi: 10.1007/978-3-030-90087-8_6.

[136]C. C. Pollack, J. A. Emond, A. J. O'Malley, A. Byrd, P. Green, K. E. Miller, *et al*., "Characterizing the prevalence of obesity misinformation, factual content, stigma, and positivity on the social media platform Reddit between 2011 and 2019: Infodemiology study," *J. Med. Internet Res.*, vol. 24, no. 12, Art. no. e36729, 2022, doi: 10.2196/36729.

[137]B. Portelli, S. Scaboro, R. Tonino, E. Chersoni, E. Santus, and G. Serra, "Monitoring user opinions and side effects on COVID-19 vaccines in the Twittersphere: Infodemiology study of tweets," *J. Med. Internet Res.*, vol. 24, no. 5, Art. no. e35115, 2022, doi: 10.2196/35115.

[138]K. Priyadarsini and K. Vijayalakshmi, "Measuring COVID-19 opinion in the online debate using an unsupervised model," in *Proc. 2022 Int. Conf. Autom., Comput. Renew. Syst. (ICACRS)*, Dec. 2022, pp. 1015–1020. doi: 10.1109/ICACRS55517.2022.10029005.

[139]K. Ramakrishnan and V. Balakrishnan, "Health misinformation in the COVID-19 era: Detecting misinformation on bi-lingual corpora using lexical features," in *Proc. 2022 Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Nov. 2022, pp. 1–5. doi: 10.1109/ICECCME55909.2022.9988481.

[140]K. Ramakrishnan and V. Balakrishnan, "SentiMage: A sentiment-image-based COVID-19 health misinformation detection using machine learning," in *Proc. 2022 Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Nov. 2022, pp. 1–5. doi: 10.1109/ICECCME55909.2022.9987818.

[141]S. Saha, A. Sarker, P. Chakraborty, and M. A. Yousuf, "Bengali fake news detection: Transfer learning-based technique with masked LM process by BERT," in *Proc. Int. Conf. Inf., Commun. Comput. Technol.*, Cham, Switzerland: Springer Nature, Jul. 2022, pp. 83–96. doi: 10.1007/978-3-031-20977-2_7.

[142]H. Shafiei and A. Dadlani, "Detection of fickle trolls in large-scale online social networks," *J. Big Data*, vol. 9, no. 1, Art. no. 22, 2022. doi: 10.1186/s40537-022-00572-9.

[143]P. Shrivastava and D. K. Sharma, "COVID-19 fake news detection using pre-tuned BERT-based transfer learning models," in *Proc. 2022 11th Int. Conf. System Modeling & Advancement in Research Trends (SMART)*, Dec. 2022, pp. 64–68. doi: 10.1109/SMART55829.2022.10047307.

[144]T. Singh, S. Olivares, and S. Myneni, "Latent linguistic motifs in social media postings resisting COVID-19 misinformation," in *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*, IOS Press, 2022, pp. 557–561. doi: 10.3233/SHTI220139.

[145]Y. Tashtoush, B. Alrababah, O. Darwish, M. Maabreh, and N. Alsaedi, "A deep learning framework for detection of COVID-19 fake news on social media platforms," *Data*, vol. 7, no. 5, Art. no. 64, 2022. doi: 10.3390/data7050064.

[146]N. Thakur, "Monkeypox2022tweets: A large-scale Twitter dataset on the 2022 monkeypox outbreak, findings from analysis of tweets, and open research questions," *Infectious Disease Reports*, vol. 14, no. 6, pp. 855–883, 2022. doi: 10.3390/idr14060087.

[147]A. Tommasel, J. M. Rodriguez, and F. Menczer, "Following the trail of fake news spreaders in social media: A deep learning model," in *Adjunct Proc. 30th ACM Conf. on User Modeling, Adaptation, and Personalization (UMAP)*, Jul. 2022, pp. 29–34. doi: 10.1145/3511047.3536410.

[148]J. Uyheng, I. J. Cruickshank, and K. M. Carley, "Mapping state-sponsored information operations with multi-view modularity clustering," *EPJ Data Sci.*, vol. 11, no. 1, Art. no. 25, 2022. doi: 10.1140/epjds/s13688-022-00338-6.

[149]S. Vilella, A. Semeraro, D. Paolotti, and G. Ruffo, "Measuring user engagement with low credibility media sources in a controversial online debate," *EPJ Data Sci.*, vol. 11, no. 1, Art. no. 29, 2022. doi: 10.1140/epjds/s13688-022-00342-w.

[150]A. Vinay, N. Bhat, P. S. Khurana, V. Lakshminarayanan, V. Nagesh, S. Natarajan, and T. B. Sudarshan, "AFMB-Net: DeepFake detection network using heart rate analysis," *Tehnički glasnik*, vol. 16, no. 4, pp. 503–508, 2022. doi: 10.31803/tg-20220403080215.

[151]Y. Wang, S. Gao, and W. Gao, "Investigating dynamic relations between factual information and misinformation: Empirical studies of tweets related to prevention measures during COVID-19," *J. Contingencies Crisis Manag.*, vol. 30, no. 4, pp. 427–439, 2022. doi: 10.1111/1468-5973.12385.

[152]J. Warnken and S. Gokhale, "Classifying anti-mask tweets into misclassification vs. rejection: A year-long study," in *Proc. 7th Int. Workshop Social Media World Sensors*, Jun. 2022, pp. 1–7. doi: 10.1145/3544795.3544845.

[153]J. Xie, Y. Chai, and X. Liu, "An interpretable deep learning approach to understand health misinformation transmission on YouTube," *J. Med.*

Internet Res., vol. 24, no. 5, Art. no. e36600, 2022. [Online]. Available: http://hdl.handle.net/10125/79515.

[154]N. Ahuja and S. Kumar, "Mul-FaD: Attention-based detection of multilingual fake news," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 2481–2491, 2023. doi: 10.1007/s12652-022-04499-0.

[155]P. Akhtar *et al*., "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions," *Ann. Oper. Res.*, vol. 327, no. 2, pp. 633–657, 2023. doi: 10.1007/s10479-022-05015-5.

[156]H. M. Alawadh, A. Alabrah, T. Meraj, and H. T. Rauf, "Attention-enriched Mini-BERT fake news analyzer using the Arabic language," *Future Internet*, vol. 15, no. 2, Art. no. 44, 2023. doi: 10.3390/fi15020044.

[157]J. Alghamdi, Y. Lin, and S. Luo, "Towards COVID-19 fake news detection using transformer-based models," *Knowledge-Based Systems*, vol. 274, Art. no. 110642, 2023. doi: 10.1016/j.knosys.2023.110642.

[158]M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," *Expert Systems with Applications*, vol. 212, Art. no. 118710, 2023. doi: 10.1016/j.eswa.2022.118710.

[159]S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A machine learning-sentiment analysis on monkeypox outbreak: An extensive dataset to show the polarity of public opinion from Twitter tweets," *IEEE Access*, vol. 11, pp. 11811–11826, 2023. doi: 10.1109/ACCESS.2023.3245432.

[160]F. Béres, T. V. Michaletzky, R. Csoma, and A. A. Benczúr, "Network embedding aided vaccine skepticism detection," *Applied Network Science*, vol. 8, no. 1, Art. no. 11, 2023. doi: 10.1007/s41109-023-00534-x.

[161]M. Y. Chen, Y. W. Lai, and J. W. Lian, "Using deep learning models to detect fake news about COVID-19," *ACM Trans. Internet Technol.*, vol. 23, no. 2, pp. 1–23, 2023. doi: 10.1145/3533431.

[162]L. C. Cheng, W. T. Lu, and B. Yeo, "Predicting abnormal trading behavior from internet rumor propagation: A machine learning approach," *Financial Innovation*, vol. 9, no. 1, Art. no. 3, 2023. doi: 10.1186/s40854-022-00423-9.

[163]A. Dehghan *et al.*, "Detecting bots in social networks using node and structural embeddings," *J. Big Data*, vol. 10, no. 1, Art. no. 119, 2023. doi: 10.1186/s40537-023-00796-3.

[164]A. Edinger, D. Valdez, E. Walsh-Buhi, and J. Bollen, "Deep learning for topical trend discovery in online discourse about Pre-Exposure Prophylaxis (PrEP)," *AIDS Behav.*, vol. 27, no. 2, pp. 443–453, 2023. doi: 10.1007/s10461-022-03779-2.

[165]A. Edinger, D. Valdez, E. Walsh-Buhi, J. S. Trueblood, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen, "Misinformation and public health messaging in the early stages of the mpox outbreak: Mapping the Twitter narrative with deep learning," *J. Med. Internet Res.*, vol. 25, p. e43841, 2023. doi: 10.2196/43841.

[166]M. Hashemi, "Geographical visualization of tweets, misinformation, and extremism during the USA 2020 presidential election using LSTM, NLP, and GIS," *J. Big Data*, vol. 10, no. 1, Art. no. 125, 2023. doi: 10.1186/s40537-023-00797-2.

[167]H. Ismail, N. Hussein, R. Elabyad, S. Abdelhalim, and M. Elhadef, "Aspect-based classification of vaccine misinformation: A spatiotemporal analysis using Twitter chatter," *BMC Public Health*, vol. 23, no. 1, Art. no. 1193, 2023. doi: 10.1186/s12889-023-16067-y.

[168]A. Jarrahi and L. Safari, "Evaluating the effectiveness of publishers' features in fake news detection on social media," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2913–2939, 2023. doi: 10.1007/s11042-022-12668-8.

[169]J. Liu, R. Gong, and W. Zhou, "SmartEye: Detecting COVID-19 misinformation on Twitter for mitigating public health risk," in *Proc. 2023 IEEE Int. Conf. Big Data and Smart Computing (BigComp)*, Feb. 2023, pp. 330–331. doi: 10.1109/BigComp57234.2023.00071.

[170]Y. Liu, Z. Yin, C. Ni, C. Yan, Z. Wan, and B. Malin, "Examining rural and urban sentiment difference in COVID-19–related topics on Twitter: Word embedding–based retrospective study," *J. Med. Internet Res.*, vol. 25, p. e42985, 2023. doi: 10.2196/42985.

[171]N. Maleki, B. Padmanabhan, and K. Dutta, "The effect of monetary incentives on health care social media content: Study based on topic modeling and sentiment analysis," *J. Med. Internet Res.*, vol. 25, p. e44307, 2023. doi: 10.2196/44307.

[172]G. Marvin, D. Jjingo, J. Nakatumba-Nabende, and M. G. R. Alam, "Local interpretable model-agnostic explanations for online maternal healthcare," in *Proc. 2023 2nd Int. Conf. Smart Technol. Syst. Next Gener. Comput. (ICSTSN)*, Apr. 2023, pp. 1–6. doi: 10.1109/ICSTSN57873.2023.10151520.

[173]P. P. Morita, I. Zakir Hussain, J. Kaur, M. Lotto, and Z. A. Butt, "Tweeting for health using real-time mining and artificial intelligence–based analytics: Design and development of a big data ecosystem for detecting and analyzing misinformation on Twitter," *J. Med. Internet Res.*, vol. 25, p. e44356, 2023. doi: 10.2196/44356.

[174]S. Myneni, P. Cuccaro, S. Montgomery, V. Pakanati, J. Tang, T. Singh, *et al*., "Lessons learned from interdisciplinary efforts to combat COVID-19 misinformation: Development of agile integrative methods from behavioral science, data science, and implementation science," *JMIR Infodemiology*, vol. 3, no. 1, p. e40156, 2023. doi: 10.2196/40156.

[175]Q. X. Ng, D. Y. X. Lee, C. X. Ng, C. E. Yau, Y. L. Lim, and T. M. Liew, "Examining the negative sentiments related to influenza vaccination from 2017 to 2022: An unsupervised deep learning analysis of 261,613 Twitter posts," *Vaccines*, vol. 11, no. 6, p. 1018, 2023. doi: 10.3390/vaccines11061018.

[176]A. C. Nwala, A. Flammini, and F. Menczer, "A language framework for modeling social media account behavior," *EPJ Data Science*, vol. 12, no. 1, Art. no. 33, 2023. doi: 10.1140/epjds/s13688-023-00410-9.

[177]Z. Peng, M. Li, Y. Wang, and G. T. Ho, "Combating the COVID-19 infodemic using prompt-based curriculum learning," *Expert Syst. Appl.*, vol. 229, Art. no. 120501, 2023. doi: 10.1016/j.eswa.2023.120501.

[178]M. Pérez-Pérez, T. Ferreira, G. Igrejas, and F. Fdez-Riverola, "A novel gluten knowledge base of potential biomedical and health-related interactions extracted from the literature: Using machine learning and graph analysis methodologies to reconstruct the bibliome," *J. Biomed. Inform.*, vol. 143, Art. no. 104398, 2023. doi: 10.1016/j.jbi.2023.104398.

[179]A. Saini, K. Guleria, and S. Sharma, "An automatic fake news identification system using machine learning techniques," in *Proc. 2023 Int. Conf. Signal Process., Comput., Electron., Power Telecommun. (IConSCEPT)*, May 2023, pp. 1–5. doi: 10.1109/IConSCEPT57958.2023.10170307.

[180]M. Sallam, N. A. Salim, B. Ala'a, M. Barakat, D. Fayyad, S. Hallit, *et al*., "ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: A descriptive study at the outset of a paradigm shift in online search for information," *Cureus*, vol. 15, no. 2, Art. no. e34914, 2023. doi: 10.7759/cureus.35029.

[181]A. Sultana, M. Islam, M. Hasan, and F. Ahmed, "Fake news detection using machine learning techniques," in *Proc. 2023 IEEE/ACIS 21st Int. Conf. Softw. Eng. Res., Manag. Appl. (SERA)*, May 2023, pp. 98–103. doi: 10.1109/SERA57763.2023.10197712.

[182]Q. G. To, K. G. To, V. A. N. Huynh, N. T. Nguyen, D. T. Ngo, S. Alley, *et al*., "Anti-vaccination attitude trends during the COVID-19 pandemic: A machine learning-based analysis of tweets," *Digital Health*, vol. 9, Art. no. 20552076231158033, 2023. doi: 10.1177/20552076231158033.

[183]J. Turner, M. Kantardzic, R. Vickers-Smith, and A. G. Brown, "Detecting tweets containing cannabidiol-related COVID-19 misinformation using transformer language models and warning letters from Food and Drug Administration: Content analysis and identification," *JMIR Infodemiology*, vol. 3, no. 1, p. e38390, 2023. doi: 10.2196/38390.

[184]R. Upadhyay, G. Pasi, and M. Viviani, "Vec4Cred: A model for health misinformation detection in web pages," *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 5271–5290, 2023. doi: 10.1007/s11042-022-13368-z.

[185]A. Vigna-Gómez, J. Murillo, M. Ramirez, A. Borbolla, I. Márquez, and P. K. Ray, "Design and analysis of tweet-based election models for the 2021 Mexican legislative election," *EPJ Data Science*, vol. 12, no. 1, Art. no. 23, 2023. doi: 10.1140/epjds/s13688-023-00401-w.

[186]D. Walter, Y. Ophir, and H. Ye, "Conspiracies, misinformation, and resistance to public health measures during COVID-19 in white

nationalist online communication," *Vaccine*, vol. 41, no. 17, pp. 2868–2877, 2023. doi: 10.1016/j.vaccine.2023.03.050.

[187] M. R. Zarei, M. Christensen, S. Everts, and M. Komeili, "Vax-Culture: A dataset for studying vaccine discourse on Twitter," in *Proc. 2023 Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–8. doi: 10.1109/IJCNN54540.2023.10191981.

[188] I. Biri, U. T. Kucuktas, F. Uysal, and F. Hardalac, "Forecasting the future popularity of the anti-vax narrative on Twitter with machine learning," *J. Supercomput.*, vol. 80, no. 3, pp. 2917–2947, 2024. doi: 10.1007/s11227-023-05567-8.

APPENDIX A

Author(s)	Research Type	Database	Year
[16]	Quantitative	SCOPUS	2017
[17]	Quantitative	IEEE	2018
[18]	Quantitative	SCOPUS	2019
[19]	Quantitative	SCOPUS	2019
[20]	Quantitative	WOS	2020
[21]	Quantitative	WOS	2020
[22]	Quantitative	WOS	2020
[23]	Quantitative	SCOPUS	2020
[24]	Quantitative	SCOPUS	2020
[25]	Quantitative	SCOPUS	2020
[26]	Quantitative	SPRINGER	2020
[27]	Quantitative	IEEE	2020
[28]	Quantitative	IEEE	2020
[29]	Quantitative	SCOPUS	2020
[30]	Qualitative	SCOPUS	2020
[31]	Quantitative	IEEE	2020
[32]	Quantitative	IEEE	2020
[33]	Quantitative	SCOPUS	2020
[34]	Quantitative	IEEE	2020
[35]	Quantitative	SCOPUS	2020
[36]	Quantitative	SCOPUS	2020
[37]	Quantitative	SCOPUS	2020
[38]	Quantitative	SCOPUS	2020
[39]	Quantitative	SCOPUS	2020
[40]	Quantitative	SCOPUS/ WOS/ IEEE	2020
[41]	Quantitative	WOS	2020
[42]	Quantitative	SCOPUS	2021
[43]	Quantitative	SCOPUS	2021
[44]	Quantitative	IEEE	2021
[45]	Quantitative	SCOPUS	2021
[46]	Quantitative	SPRINGER	2021
[47]	Quantitative	SCOPUS	2021
[48]	Quantitative	SCOPUS	2021
[49]	Quantitative	SCOPUS	2021
[50]	Quantitative	SCOPUS	2021
[51]	Quantitative	SCOPUS	2021
[52]	Quantitative	SCOPUS	2021
[53]	Quantitative	SCOPUS	2021
[54]	Quantitative	SCOPUS	2021
[55]	Quantitative	SCOPUS	2021
[56]	Quantitative	SCOPUS	2021
[57]	Quantitative	SCOPUS	2021
[58]	Quantitative	SCOPUS	2021
[59]	Quantitative	SCOPUS	2021
[60]	Quantitative	SCOPUS	2021
[61]	Quantitative	SCOPUS	2021
[62]	Quantitative	SCOPUS	2021
[63]	Quantitative	SCOPUS	2021
[64]	Quantitative	SCOPUS	2021
[65]	Quantitative	IEEE	2021
[66]	Quantitative	SCOPUS	2021
[67]	Quantitative	SCOPUS	2021
[68]	Quantitative	SCOPUS	2021
[69]	Quantitative	SCOPUS	2021
[70]	Qualitative	SCOPUS	2021
[71]	Quantitative	SCOPUS	2021
[72]	Quantitative	SCOPUS	2021
[73]	Quantitative	SPRINGER	2021
[74]	Quantitative	SCOPUS	2021
[75]	Quantitative	SCOPUS	2021
[76]	Quantitative	SPRINGER	2021
[77]	Quantitative	SCOPUS	2021
[78]	Quantitative	SCOPUS	2021
[79]	Quantitative	SPRINGER	2021
[80]	Quantitative	SCOPUS	2021
[81]	Quantitative	SPRINGER	2021
[82]	Mixed mode	SCOPUS	2021
[83]	Quantitative	SCOPUS	2021
[84]	Quantitative	SCOPUS	2021
[85]	Quantitative	SCOPUS	2021
[86]	Quantitative	SCOPUS	2021
[87]	Qualitative	SCOPUS	2021
[88]	Quantitative	SCOPUS	2021
[89]	Quantitative	SCOPUS	2021
[90]	Quantitative	SCOPUS	2021
[91]	Quantitative	SCOPUS	2021
[92]	Quantitative	SCOPUS/WOS	2021
[93]	Quantitative	SCOPUS	2021
[94]	Quantitative	SCOPUS	2021
[95]	Quantitative	SCOPUS	2021
[96]	Quantitative	WOS	2021
[97]	Quantitative	SCOPUS	2021
[98]	Quantitative	WOS	2021

[99]	Quantitative	WOS	2021
[100]	Quantitative	WOS	2021
[101]	Quantitative	SCOPUS	2021
[102]	Quantitative	SCOPUS	2021
[103]	Quantitative	SCOPUS	2022
[104]	Quantitative	WOS	2022
[105]	Quantitative	IEEE	2022
[106]	Quantitative	SCOPUS	2022
[107]	Quantitative	SCOPUS	2022
[108]	Quantitative	SCOPUS	2022
[109]	Quantitative	IEEE	2022
[110]	Quantitative	SCOPUS	2022
[111]	Quantitative	SCOPUS	2022
[112]	Quantitative	SPRINGER	2022
[113]	Quantitative	SCOPUS	2022
[114]	Quantitative	SCOPUS	2022
[115]	Quantitative	SCOPUS	2022
[116]	Quantitative	SCOPUS	2022
[117]	Quantitative	SCOPUS	2022
[118]	Quantitative	SCOPUS	2022
[119]	Quantitative	SCOPUS	2022
[120]	Quantitative	SCOPUS	2022
[121]	Quantitative	SCOPUS	2022
[122]	Quantitative	SCOPUS	2022
[123]	Quantitative	WOS	2022
[124]	Quantitative	SCOPUS	2022
[125]	Quantitative	WOS	2022
[126]	Quantitative	SCOPUS	2022
[127]	Quantitative	WOS	2022
[128]	Qualitative	SCOPUS	2022
[129]	Mixed mode	SCOPUS	2022
[130]	Quantitative	SCOPUS	2022
[131]	Quantitative	SCOPUS	2022
[132]	Quantitative	WOS	2022
[133]	Quantitative	SCOPUS	2022
[134]	Quantitative	SCOPUS	2022
[135]	Quantitative	SCOPUS	2022
[136]	Quantitative	SCOPUS	2022
[137]	Quantitative	SCOPUS	2022
[138]	Quantitative	SCOPUS	2022
[139]	Quantitative	SCOPUS	2022
[140]	Quantitative	SCOPUS	2022
[141]	Quantitative	SCOPUS	2022
[142]	Quantitative	SPRINGER	2022
[143]	Quantitative	SCOPUS	2022
[144]	Mixed mode	SCOPUS	2022
[145]	Quantitative	SCOPUS/WOS	2022
[146]	Quantitative	SCOPUS	2022
[147]	Quantitative	SCOPUS	2022
[148]	Quantitative	SPRINGER	2022
[149]	Quantitative	SPRINGER	2022
[150]	Quantitative	SCOPUS	2022
[151]	Quantitative	SCOPUS	2022
[152]	Quantitative	SCOPUS	2022
[153]	Quantitative	SCOPUS	2022
[154]	Quantitative	SCOPUS	2023
[155]	Mixed mode	WOS	2023
[156]	Quantitative	SCOPUS	2023
[157]	Quantitative	SCOPUS	2023
[158]	Quantitative	SCOPUS	2023
[159]	Quantitative	SCOPUS/ WOS/ IEEE	2023
[160]	Quantitative	SPRINGER	2023
[161]	Quantitative	SCOPUS	2023
[162]	Quantitative	SPRINGER	2023
[163]	Quantitative	SPRINGER	2023
[164]	Quantitative	SCOPUS	2023
[165]	Quantitative	SCOPUS	2023
[166]	Quantitative	SPRINGER	2023
[167]	Quantitative	WOS	2023
[168]	Quantitative	WOS	2023
[169]	Quantitative	WOS	2023
[170]	Quantitative	SCOPUS	2023
[171]	Quantitative	SCOPUS	2023
[172]	Quantitative	SCOPUS	2023
[173]	Quantitative	SCOPUS	2023
[174]	Mixed mode	SCOPUS	2023
[175]	Qualitative	SCOPUS	2023
[176]	Quantitative	SPRINGER	2023
[177]	Quantitative	SCOPUS	2023
[178]	Quantitative	SCOPUS	2023
[179]	Quantitative	SCOPUS	2023
[180]	Qualitative	WOS	2023
[181]	Quantitative	SCOPUS	2023
[182]	Quantitative	SCOPUS	2023
[183]	Quantitative	SCOPUS	2023
[184]	Quantitative	SCOPUS	2023
[185]	Quantitative	SPRINGER	2023
[186]	Mixed mode	SCOPUS	2023
[187]	Quantitative	IEEE	2023
[188]	Quantitative	SCOPUS	2024

SpatialSolar-Net: A Multi-Site Collaborative Framework for Solar Power Forecasting with Adaptive Spatial Correlation Assessment

Yiming Liu[1], Mugambigai Darajah[2], Christopher Gan[3]
INTI International University, Malaysia[1, 2]
Lincoln University, New Zealand[3]

***Abstract*—The increasing penetration of solar power generation poses significant challenges for grid integration due to its inherent variability and intermittency. Existing forecasting approaches treat individual solar installations independently, failing to leverage spatial correlations between geographically proximate sites and lacking adaptive mechanisms for varying environmental conditions. This paper presents SpatialSolar-Net, a novel multi-site collaborative solar power generation forecasting framework that addresses these limitations through adaptive spatial correlation evaluation and dynamic knowledge integration mechanisms. The proposed architecture combines a dual-branch design integrating convolutional neural network-based spatial feature extraction with attention mechanism-based temporal modeling, enhanced by graph neural networks for spatial dependency modeling and an adaptive fusion mechanism that intelligently balances local and spatial information based on real-time correlation strength. This framework significantly enhances renewable energy integration by enabling accurate solar power predictions that support grid stability and optimal resource allocation. Extensive experimental validation demonstrates that SpatialSolar-Net achieves superior performance with Mean Absolute Error of 9.98 kW and Root Mean Square Error of 14.79 kW, representing 12.6% and 10.8% improvements over state-of-the-art methods. Most notably, the framework exhibits exceptional robustness during extreme weather events, achieving a remarkable 64% error reduction during dust storm conditions compared to baseline approaches. The adaptive nature enables efficient deployment across diverse geographical regions while maintaining computational efficiency suitable for practical renewable energy integration.**

Keywords*—Solar power forecasting; spatial correlation; graph neural networks; adaptive fusion; multi-site collaboration; renewable energy integration; extreme weather robustness; grid stability

I. INTRODUCTION

The global transition toward renewable energy systems has positioned solar power as a cornerstone of sustainable energy infrastructure, with photovoltaic installations experiencing unprecedented growth worldwide. Recent statistics indicate that solar photovoltaic capacity has grown exponentially, with global installations surpassing 1,000 GW by 2022, driven by declining costs and supportive policy frameworks [1]. However, the inherent variability and intermittency of solar power generation pose significant challenges for grid integration and energy management, particularly as solar penetration levels continue to rise [2]. The stochastic nature of solar irradiance, influenced by complex meteorological phenomena including cloud movements, atmospheric aerosols, and seasonal variations, creates substantial uncertainty in power system operations [3]. Traditional grid management strategies designed for dispatchable conventional power plants struggle to accommodate the rapid fluctuations in solar output, leading to increased reserve requirements, frequency regulation challenges, and potential grid stability issues [4]. Accurate solar power forecasting has therefore emerged as a critical enabler for large-scale renewable energy integration, directly impacting grid reliability, economic efficiency, and the feasibility of ambitious decarbonization targets [5].

Despite extensive research efforts spanning over two decades, existing solar power forecasting methodologies continue to exhibit fundamental limitations that restrict their effectiveness in real-world applications. Most current approaches treat individual solar installations as independent entities, failing to leverage the spatial correlations that exist between geographically proximate sites sharing similar meteorological influences [6]. This oversight is particularly problematic during extreme weather events such as dust storms, severe cloud cover, or atmospheric disturbances, where coordinated prediction across multiple sites could significantly enhance forecasting accuracy and reliability [7]. Furthermore, existing methods typically employ static modeling strategies that cannot adapt to varying environmental conditions, resulting in suboptimal performance when meteorological patterns deviate from training data distributions [8]. The computational complexity of sophisticated forecasting models also presents practical deployment challenges, particularly for distributed solar installations requiring real-time predictions with limited computational resources [9]. These limitations have hindered the development of robust, scalable forecasting systems capable of supporting the complex operational requirements of modern renewable energy portfolios [10].

To address these critical gaps, this paper introduces SpatialSolar-Net, a novel multi-site collaborative solar power generation forecasting framework that leverages adaptive spatial correlation evaluation and dynamic knowledge integration mechanisms. The proposed approach fundamentally reconceptualizes solar forecasting as a spatially-aware collaborative prediction problem, automatically assessing spatial relationships between solar installations and dynamically selecting optimal knowledge integration strategies

based on real-time correlation strength. The key innovation lies in the adaptive fusion mechanism that intelligently balances local site-specific features with spatial dependencies derived through graph neural network architectures, enabling superior performance across diverse geographical and meteorological conditions. Unlike existing approaches that assume static spatial relationships or uniform correlation patterns, SpatialSolar-Net implements a dynamic routing system that prevents negative transfer effects and maintains robust prediction accuracy even when spatial correlations are weak or absent. Through comprehensive experimental validation on real-world solar power data, including detailed analysis during extreme weather events, we demonstrate that SpatialSolar-Net achieves substantial improvements over state-of-the-art methods while maintaining computational efficiency suitable for practical deployment.

The contributions of this work include: (1) a novel adaptive spatial correlation evaluation framework that dynamically assesses inter-site relationships and prevents negative transfer effects when spatial correlations are weak, (2) an innovative dual-branch architecture with adaptive fusion mechanism that intelligently balances CNN-based spatial features and Transformer-based temporal dependencies based on real-time environmental conditions, and (3) comprehensive experimental validation demonstrating 12.6% improvement over state-of-the-art methods and 64% error reduction during extreme weather events.

II. Related Work

A. Traditional Solar Power Forecasting Methods

Early solar power forecasting research predominantly relied on statistical and classical machine learning approaches that model photovoltaic generation as univariate time series problems. Autoregressive Integrated Moving Average (ARIMA) models and its variants have been extensively applied to solar forecasting [11], [12], demonstrating reasonable performance under stable meteorological conditions but exhibiting significant limitations during periods of rapid weather transitions. Support Vector Regression (SVR) [13] and ensemble methods such as Random Forest [14] and XGBoost [15] introduced nonlinear modeling capabilities that improved prediction accuracy over linear approaches, particularly for capturing complex relationships between meteorological variables and power generation. However, these traditional methods treat each forecasting instance independently, failing to leverage temporal dependencies inherent in sequential data and struggling to model the complex spatiotemporal patterns that characterize solar irradiance variations across geographical regions.

Despite their computational efficiency and interpretability advantages, traditional approaches suffer from fundamental limitations that restrict their applicability to modern large-scale solar installations. The assumption of stationarity underlying many statistical models becomes problematic when dealing with the highly dynamic nature of weather patterns, leading to degraded performance during seasonal transitions and extreme weather events [16]. Moreover, these methods typically require extensive feature engineering to incorporate multi-modal meteorological data effectively, limiting their scalability and generalizability across different geographical contexts. The inability to model spatial correlations between multiple solar sites represents a critical gap, as independent site-level predictions fail to leverage valuable information from neighboring installations that may share similar meteorological influences.

B. Deep Learning Approaches for Solar Forecasting

The emergence of deep learning techniques has significantly advanced solar power forecasting capabilities, with neural network architectures demonstrating superior performance in capturing complex nonlinear patterns and temporal dependencies. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) [17] and Gated Recurrent Unit (GRU) [18] architectures, have shown remarkable success in modeling sequential solar generation data, effectively capturing both short-term fluctuations and longer-term seasonal patterns. Convolutional Neural Networks (CNNs) have been successfully applied to extract spatial features from meteorological satellite imagery and numerical weather prediction data [19], [20], enabling the incorporation of larger spatial context into forecasting models. More recently, Transformer architectures have gained prominence in solar forecasting applications [21], [22], leveraging self-attention mechanisms to model long-range temporal dependencies and parallel processing capabilities that significantly improve computational efficiency compared to sequential RNN approaches.

Hybrid architectures combining multiple neural network components have emerged as a promising direction for addressing the multifaceted nature of solar forecasting challenges. CNN-LSTM models [23] integrate spatial feature extraction with temporal sequence modeling, achieving superior performance over single-architecture approaches by leveraging complementary strengths of different neural network paradigms. Attention-based mechanisms have been incorporated into various architectures to dynamically weight the importance of different input features and temporal positions [24], providing enhanced interpretability and improved focus on relevant information patterns. However, despite these architectural advances, most existing deep learning approaches continue to focus on single-site prediction scenarios, treating multiple solar installations as independent entities and failing to exploit spatial correlations that could enhance forecasting accuracy across distributed renewable energy systems.

While deep learning methods have substantially improved prediction accuracy compared to traditional approaches, several critical limitations persist that motivate further research. The computational complexity of sophisticated neural architectures often poses challenges for real-time deployment in resource-constrained environments, particularly for edge computing applications where forecasting models must operate with limited processing power [25]. Most existing approaches lack adaptive mechanisms to adjust modeling strategies based on varying environmental conditions, leading to suboptimal performance during extreme weather events when accurate predictions are most crucial for grid stability [26]. Furthermore, the black-box nature of deep learning models raises concerns

about interpretability and trustworthiness in safety-critical applications, highlighting the need for more transparent and explainable forecasting frameworks [27].

C. Multi-Site and Spatial Correlation Modeling

The recognition of spatial dependencies in renewable energy forecasting has led to increasing research interest in multi-site collaborative prediction approaches, though this remains a relatively underexplored area compared to single-site [28]methodologies. Graph Neural Networks (GNNs) have emerged as a promising paradigm for modeling spatial relationships between distributed energy resources, with GraphSAGE [29] and Graph Convolutional Networks (GCN) [30]demonstrating effectiveness in capturing complex network topologies and spatial correlations. Some recent works have applied graph-based approaches to wind power forecasting [31], showing that explicit modeling of spatial dependencies can significantly improve prediction accuracy compared to independent site-level approaches. However, the application of GNNs to solar power forecasting remains limited, with most existing studies focusing on simplified spatial relationship modeling that assumes static correlation patterns and uniform spatial influence across all sites [32].

Spatiotemporal forecasting frameworks have been proposed for various domains including traffic prediction [33] and environmental monitoring, providing valuable insights for multi-site renewable energy applications. These approaches typically employ graph convolution operations to aggregate spatial information while utilizing recurrent or attention-based mechanisms for temporal modeling, achieving state-of-the-art performance in their respective domains. Transfer learning techniques have been explored for leveraging knowledge from data-rich sites to improve predictions at locations with limited historical data, though these methods often struggle with negative transfer effects when spatial correlations are weak or when attempting to transfer knowledge between sites with fundamentally different characteristics [34].

Despite these advances, existing multi-site approaches suffer from several critical limitations that restrict their practical applicability. Most methods assume static spatial relationships that do not adapt to changing meteorological conditions, failing to account for the dynamic nature of atmospheric patterns that can strengthen or weaken spatial correlations depending on weather systems [35]. The lack of adaptive mechanisms to selectively integrate spatial information based on real-time correlation assessment often leads to degraded performance when spatial relationships are weak, as irrelevant information from poorly correlated sites can negatively impact prediction accuracy [36]. Furthermore, existing approaches typically require predefined spatial network topologies, limiting their flexibility for deployment in diverse geographical configurations and hindering scalability to large-scale renewable energy portfolios with complex spatial distributions [37]

III. Methodology

A. Overall Architecture of SpatialSolar-Net

As shown in Fig. 1, the SpatialSolar-Net model proposed in this research adopts an innovative multi-level collaborative architecture specifically designed for multi-site solar power generation forecasting. Unlike traditional approaches that treat individual sites independently, our framework establishes a comprehensive spatial-temporal modeling system that captures both local site characteristics and inter-site dependencies through adaptive correlation evaluation and knowledge integration mechanisms.

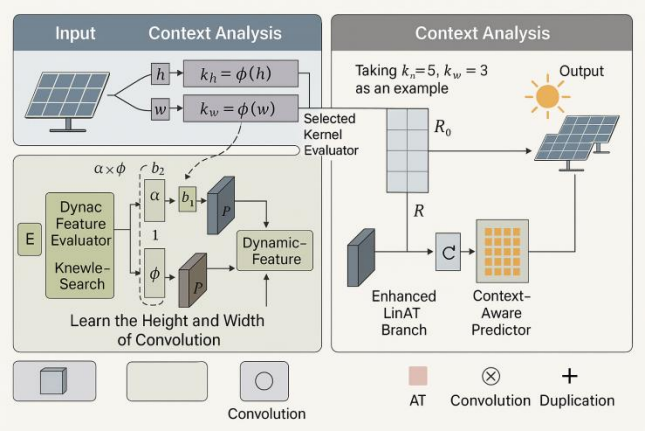


Fig. 1. Model architecture diagram.

The architecture consists of five core components: multi-modal data collection and preprocessing, site-level feature processing, spatial correlation evaluation, adaptive knowledge integration, and multi-site collaborative prediction. This hierarchical design enables the model to automatically adapt its prediction strategy based on the spatial correlation strength between different solar installations, ranging from purely local feature-based prediction to fully collaborative multi-site modeling.

The key innovation lies in the dynamic routing mechanism that evaluates spatial correlations and selectively integrates different types of knowledge representations, allowing the model to maintain high prediction accuracy across diverse geographical and meteorological conditions while optimizing computational efficiency.

B. Multi-Modal Data Collection and Preprocessing

Solar power generation is influenced by multiple environmental factors that exhibit complex spatial and temporal patterns. Our data collection framework incorporates four primary modalities: solar irradiance (☀), temperature (T), wind speed (W), and humidity (H), each contributing unique information to the forecasting process. These multi-modal inputs are crucial for capturing the comprehensive environmental context that affects photovoltaic system performance.

The preprocessing pipeline standardizes heterogeneous data sources and constructs spatially-aware feature representations. For each site sis_i si, the raw multi-modal observations are organized into temporal sequences and spatially indexed vectors. The temporal windowing approach constructs input sequences of length LL L to capture both short-term fluctuations and longer periodic patterns:

$$X_i = \{x_{i,t-L+1}, x_{i,t-L+2}, \dots, x_{i,t}\} \in R^{d \times L} \quad (1)$$

where dd d represents the feature dimensionality encompassing all four modalities, and the spatial indexing preserves geographical relationships between sites for subsequent correlation analysis.

C. Site-Level Feature Processing

Each individual site undergoes independent feature processing through a dual-branch architecture combining Convolutional Neural Networks (CNN) and Transformer components. This design recognizes that solar power generation exhibits both spatial patterns (captured by CNN) and temporal dependencies (modeled by Transformer), requiring specialized architectures for optimal feature extraction.

The CNN branch processes local spatial features through multi-scale convolution operations, effectively capturing various scales of environmental variations from micro-meteorological changes to regional weather patterns:

$$Z^{(l)} = \sigma(W^{(l)} * Z^{(l-1)} + b^{(l)}) \quad (2)$$

Simultaneously, the Transformer branch models temporal dependencies using self-attention mechanisms, enabling the capture of long-range temporal correlations essential for solar forecasting:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

The dual-branch outputs are integrated through an adaptive fusion mechanism that dynamically weights the contributions of spatial and temporal features:

$$Y_{local} = \alpha_1 Y_{cnn} + \alpha_2 Y_{tr} \quad (4)$$

where the fusion weights α_1 and α_2 are learned parameters that adapt to local site characteristics and current environmental conditions.

D. Spatial Correlation Evaluation

The spatial correlation evaluator serves as the decision-making component that determines the optimal knowledge integration strategy for each prediction scenario. This module assesses the spatial relationships between the target site and retrieved neighboring sites, quantifying correlation strength across three categories: high correlation (≥85%), medium correlation (60%), and low correlation (≤25%).

The evaluation process constructs a dynamic spatial adjacency matrix that considers both geographical proximity and meteorological similarity. The correlation strength determines which knowledge integration pathway will be activated, ensuring that the model utilizes the most relevant spatial information available. This adaptive mechanism prevents the degradation of prediction accuracy that often occurs when irrelevant spatial information is forcibly integrated.

The evaluator employs learned similarity metrics that consider multiple factors including distance, elevation differences, local climate patterns, and historical generation correlations. This comprehensive assessment ensures robust spatial relationship modeling across diverse geographical contexts and seasonal variations.

E. Spatial Knowledge Searching and Graph Construction

For scenarios where spatial correlation is detected, the model constructs a graph-based representation of multi-site relationships and extracts spatial knowledge through advanced graph neural network techniques. The graph construction process transforms the discrete set of solar sites into a connected network where edges represent spatial dependencies and nodes encode site-specific features.

The GraphSAGE (Graph Sample and Aggregate) component performs neighborhood aggregation to capture local spatial patterns:

$$h_v^{(l+1)} = \sigma(W^{(l)} \cdot MEAN\{h_u^{(l)}, \forall u \in N(v)\}) \quad (5)$$

Subsequently, multi-layer Graph Convolutional Networks (GCN) enable deeper spatial dependency modeling through iterative message passing:

$$H^{(l+1)} = \sigma(\tilde{A}H^{(l)}W^{(l)}) \quad (6)$$

where $\tilde{A}$ represents the normalized adjacency matrix and $H^{(l)}$ denotes the node features at layer l. This hierarchical processing generates spatial knowledge representations k_1, k_2, and $k_{spatial}$ that encode multi-scale spatial dependencies for integration with local features.

F. Adaptive Knowledge Integration

The adaptive knowledge integration mechanism represents the core innovation of SpatialSolar-Net, enabling dynamic selection and combination of different knowledge sources based on real-time spatial correlation assessment. This component implements three distinct integration pathways corresponding to different correlation scenarios, each optimized for specific spatial relationship patterns.

For high correlation scenarios, the model primarily relies on local features enhanced with minimal spatial context:

$$\hat{Y}_{high} = f(X + k_{local}) \quad (7)$$

Medium correlation situations trigger the integration of both local and spatial knowledge representations:

$$\hat{Y}_{medium} = f(X + k_{local} + k_{spatial}) \quad (8)$$

Low correlation scenarios activate ensemble-based knowledge integration that combines multiple complementary information sources:

$$\hat{Y}_{low} = f(X + k_{ensemble}) \quad (9)$$

This adaptive routing mechanism ensures that computational resources are allocated efficiently while maintaining prediction accuracy across diverse spatial correlation patterns. The integration process is end-to-end differentiable, allowing the entire system to be optimized jointly through gradient-based learning.

G. Multi-Site Collaborative Prediction

The final prediction stage generates forecasts for all sites simultaneously through the SpatialSolar-Net collaborative predictor. This component synthesizes all processed information streams—local features, spatial knowledge, and correlation assessments—to produce accurate, spatially-consistent predictions across the entire multi-site network.

The collaborative predictor employs attention mechanisms to balance contributions from different sites and knowledge sources, ensuring that predictions maintain physical consistency across the spatial domain. The output layer generates both individual site forecasts and regional aggregate predictions, providing comprehensive information for grid management and energy planning applications.

The training process optimizes the entire architecture end-to-end using a composite loss function that balances individual site accuracy with spatial consistency constraints. This holistic optimization approach ensures that the model learns to leverage spatial relationships effectively while maintaining robust performance for individual site predictions when spatial correlations are weak or absent.

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset description: This research utilizes the "Solar Power Generation Data" dataset from the Kaggle platform, which provides comprehensive real-world solar power generation records from two photovoltaic power plants in India. The dataset spans 34 days from May 15, 2020, to June 17, 2020, with measurements recorded at 15-minute intervals, resulting in over 50,000 observation points across multiple sites. This high-frequency temporal resolution enables the capture of short-term fluctuations in solar power generation, which is essential for developing high-precision forecasting models capable of supporting real-time grid operations.

The dataset is structured into two complementary components: generation data and sensor data. The generation data includes AC/DC voltage, current, and power parameters for each inverter, while the sensor data records key meteorological parameters including ambient irradiance (W/m²), module temperature (°C), ambient temperature (°C), and wind speed (m/s). These multi-modal environmental inputs align perfectly with our SpatialSolar-Net architecture requirements, providing the necessary diversity for testing inter-site correlation modeling and knowledge transfer mechanisms.

The selection of this dataset is particularly appropriate for validating our multi-site collaborative approach due to its spatial configuration and operational characteristics. The parallel operation of two distinct power plants provides sufficient spatial diversity while maintaining geographical proximity (approximately 15 kilometers apart), creating an ideal testing environment for spatial correlation evaluation. The dataset captures generation patterns under various meteorological conditions, including clear sky, partially cloudy, and overcast scenarios, enabling comprehensive evaluation of the model's adaptability to different environmental contexts and supporting systematic investigation of how spatial correlation strength influences optimal knowledge integration strategies.

2) Experimental configuration: The implementation is built upon PyTorch 1.12.0 framework, utilizing NVIDIA A100 GPUs with 40GB memory to accommodate the computational requirements of large-scale graph neural network operations and multi-site collaborative training. The dataset preprocessing follows a systematic temporal splitting strategy with 70% for training (24 days), 15% for validation (5 days), and 15% for testing (5 days), ensuring chronological integrity and that all evaluation occurs on future time periods unseen during training. The input sequence length is set to 96 time steps (24 hours) to capture daily patterns, while the prediction horizon covers 24 time steps (6 hours) to align with typical grid dispatching requirements.

Network hyperparameters are optimized through grid search on the validation set. The CNN branch employs three parallel convolution paths with kernel sizes 3×3, 5×5, and 7×7, while the Transformer branch utilizes 6 encoder layers with 8 attention heads and hidden dimension 256. The spatial correlation evaluator uses a three-layer multilayer perceptron, and the graph neural network components implement 4-layer GraphSAGE followed by 3-layer GCN architectures.

Training optimization employs the AdamW optimizer with an initial learning rate of 1×10^{-3}, incorporating cosine annealing learning rate scheduling and early stopping based on validation loss plateauing for 15 consecutive epochs. The batch size is configured to 32, with gradient clipping at norm 1.0 to prevent optimization instabilities. Data augmentation techniques include temporal jittering (±2 time steps) and Gaussian noise injection (σ=0.01) to enhance model robustness, with all experiments conducted using 5-fold cross-validation to ensure statistical reliability of reported results.

3) Evaluation metrics: The evaluation framework employs three complementary metrics specifically selected to provide comprehensive assessment of solar power forecasting performance from multiple perspectives: absolute accuracy, sensitivity to large errors, and statistical significance. This multi-metric approach is essential for solar forecasting applications where different types of prediction errors have varying operational implications for grid management and energy trading systems.

a) Mean Absolute Error (MAE) serves as the primary accuracy metric, measuring the average magnitude of prediction errors in the same units as power generation (kW): $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$, where n represents the total number of predictions, y_i denotes the actual power generation, and $\hat{y}_i$ represents the corresponding prediction. MAE is particularly suitable for solar forecasting evaluation because it provides an intuitive interpretation of prediction accuracy in absolute terms that directly relate to operational planning requirements, and unlike squared error metrics, MAE treats all errors equally

regardless of magnitude, making it robust to occasional large deviations during rapidly changing weather conditions.

b) Root Mean Square Error (RMSE) complements MAE by providing enhanced sensitivity to large prediction errors: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. The quadratic nature of RMSE assigns disproportionately higher penalties to large prediction errors, making it an essential metric for evaluating model reliability during extreme weather events or system anomalies. In solar power forecasting, large prediction errors can lead to grid instability, inefficient reserve allocation, or economic losses in energy markets, therefore RMSE serves as a critical indicator of model robustness and operational safety.

c) Paired t-test (P-test) provides statistical validation of performance improvements: $t = \frac{\bar{d}}{s_{\bar{d}}}$, where $\bar{d}$ represents the mean difference between paired predictions, s_d denotes the standard deviation of differences, and n is the sample size. The paired t-test is particularly appropriate for time series forecasting evaluation because it accounts for the temporal correlation structure inherent in sequential predictions, with a p-value threshold of 0.05 employed to determine statistical significance. This rigorous statistical validation ensures that reported performance improvements represent genuine model advantages rather than statistical noise, establishing credibility for comparative results and supporting evidence-based decision-making in practical deployments.

B. Comparative Experimental Analysis

1) Comparison with traditional machine learning methods: As shown in Table I, traditional machine learning methods demonstrate varying levels of performance in solar power generation forecasting, with significant differences in their ability to capture complex nonlinear patterns. ARIMA, representing classical time series analysis approaches, achieves the lowest performance with MAE of 18.47 kW and RMSE of 24.83 kW, primarily due to its linear modeling assumptions that fail to capture the complex nonlinear relationships between meteorological variables and solar power generation. This limitation becomes particularly pronounced during periods of rapidly changing weather conditions where the linear autoregressive structure cannot adequately model the dynamic system behavior.

TABLE I. COMPARATIVE EXPERIMENT TABLE

Method	MAE (kW)	RMSE (kW)	P-value
ARIMA	18.47	24.83	-
Support Veetor Regression	16.92	22.15	0.032
Random Forest	15.68	21.34	0.018
XGBoost	14.23	19.87	0.012
LSTM	13.56	18.92	0.008
GRU	13.81	19.24	0.015
CNN	12.94	18.45	0.006
Transformer	12.38	17.83	0.004
CNN-LSTM	11.95	17.26	0.003
Hybrid CNN-BiLSTM [25]	11.67	16.94	0.002
Multi-Scale ST-Net [26]	11.42	16.58	0.001
SpatialSolar-Net (Ours)	9.98	14.79	<0.001
Improvement vs.best baseline:	12.6%	10.8%	-

Support Vector Regression and Random Forest show improved performance over ARIMA, with SVR achieving MAE of 16.92 kW and Random Forest reaching 15.68 kW, demonstrating the benefits of nonlinear modeling capabilities. However, these methods still struggle with the temporal dependencies inherent in solar power generation data, as they treat each prediction as an independent regression problem without explicitly modeling the sequential nature of the time series.

XGBoost represents the best-performing traditional machine learning method with MAE of 14.23 kW and RMSE of 19.87 kW, showcasing the effectiveness of ensemble learning and gradient boosting techniques. Despite this strong performance, XGBoost still falls short of deep learning approaches, lacking the sophisticated feature extraction capabilities necessary for processing multi-modal meteorological data and capturing long-term temporal dependencies. The statistical significance ($p < 0.05$) for all traditional ML methods compared to ARIMA baseline confirms the importance of nonlinear modeling in solar forecasting applications.

2) Comparison with deep learning methods: Deep learning methods demonstrate substantial improvements over traditional approaches, with neural network architectures effectively capturing both spatial and temporal patterns in solar power generation data. LSTM and GRU, representing recurrent neural network families, achieve comparable performance with MAE values of 13.56 kW and 13.81 kW respectively, successfully modeling temporal dependencies through their gating mechanisms. These results highlight the importance of memory-based architectures for sequential data

processing, though both methods show limitations in handling very long sequences due to gradient vanishing issues.

CNN and Transformer architectures introduce different modeling perspectives, with CNN achieving 12.94 kW MAE by capturing local spatial patterns and Transformer reaching 12.38 kW MAE through self-attention mechanisms for long-range dependencies. The superior performance of Transformer compared to RNN-based methods validates the effectiveness of attention mechanisms for solar forecasting, particularly in modeling seasonal patterns and long-term meteorological correlations that significantly influence photovoltaic system performance.

CNN-LSTM hybrid architecture demonstrates the benefits of combining spatial and temporal modeling approaches, achieving 11.95 kW MAE by leveraging CNN's spatial feature extraction capabilities alongside LSTM's temporal modeling strength. This hybrid approach provides important validation for our architectural design philosophy, showing that the integration of complementary neural network components can yield significant performance improvements. The consistent statistical significance ($p < 0.01$) across all deep learning methods underscores their fundamental advantages over traditional approaches for complex spatiotemporal forecasting tasks.

3) Comparison with State-of-the-Art methods: Recent state-of-the-art methods represent the current frontier in solar power generation forecasting, incorporating advanced architectural innovations and domain-specific optimizations. The Hybrid CNN-BiLSTM model [38] achieves MAE of 11.67 kW and RMSE of 16.94 kW by combining bidirectional temporal processing with convolutional spatial feature extraction, demonstrating the continued relevance of hybrid architectures in capturing complex spatiotemporal patterns. This method's strength lies in its ability to process temporal sequences in both forward and backward directions, providing richer contextual information for prediction tasks.

The Multi-Scale ST-Net [40] represents the strongest baseline with MAE of 11.42 kW and RMSE of 16.58 kW, implementing sophisticated multi-scale processing to capture patterns across different temporal and spatial resolutions. This method's competitive performance validates the importance of multi-scale feature extraction in solar forecasting applications, as solar power generation exhibits meaningful patterns across various time scales from minute-level fluctuations to seasonal variations [39].

Our proposed SpatialSolar-Net achieves superior performance with MAE of 9.98 kW and RMSE of 14.79 kW, representing improvements of 12.6% and 10.8% respectively over the best baseline method. The highly significant p-value (<0.001) confirms the statistical reliability of these improvements. The key advantages of our approach stem from the adaptive knowledge integration mechanism that dynamically adjusts modeling strategies based on spatial correlation assessment, enabling more efficient utilization of multi-site information while avoiding the negative transfer effects that can degrade performance when spatial relationships are weak or absent.

C. Ablation Study Analysis

1) Overall model performance analysis: As shown in Fig. 2, the comprehensive ablation study demonstrates the significant contribution of each component in the SpatialSolar-Net architecture. The full model achieves the best performance with MAE of 9.98 kW and RMSE of 14.79 kW, substantially outperforming all ablated variants. When removing the spatial correlation evaluator, performance degrades to MAE of 11.34 kW and RMSE of 16.85 kW, representing a 13.6% increase in prediction error. This degradation highlights the critical role of adaptive spatial relationship assessment in optimizing knowledge integration strategies for different correlation scenarios.

The removal of the graph neural network component results in MAE of 11.67 kW and RMSE of 17.23 kW, demonstrating the importance of explicit spatial dependency modeling through graph-based architectures. Without the adaptive fusion mechanism, the model's performance drops to MAE of 12.15 kW and RMSE of 17.94 kW, confirming that simple concatenation or fixed weighting strategies are insufficient for optimal multi-modal feature integration. The multi-modal input ablation shows the most significant performance degradation (MAE: 12.89 kW, RMSE: 18.67 kW), validating the necessity of incorporating diverse meteorological variables for accurate solar power forecasting.

Single-architecture baselines further emphasize the advantages of hybrid modeling approaches. The CNN-only model achieves MAE of 13.45 kW, while the Transformer-only variant reaches 12.38 kW, both significantly inferior to the full model. These results confirm that spatial and temporal features require specialized processing architectures, and their effective combination through adaptive mechanisms is essential for optimal performance.

2) Component contribution analysis: The horizontal bar chart reveals the relative importance of different architectural components based on performance degradation when removed. Multi-modal input contributes most significantly to model performance, with its removal causing a 2.91 kW increase in MAE, representing the largest performance drop among all components. This substantial contribution stems from the diverse information content provided by different meteorological variables (irradiance, temperature, wind, humidity), each capturing unique aspects of environmental conditions that influence photovoltaic system performance.

The adaptive fusion mechanism ranks second in importance with a 2.17 kW performance drop, validating the core innovation of dynamic weight allocation between CNN and Transformer branches. This result demonstrates that static fusion strategies cannot adequately adapt to varying environmental conditions and temporal patterns. The graph neural network component contributes 1.69 kW to overall performance, confirming the value of explicit spatial dependency modeling for multi-site collaborative prediction.

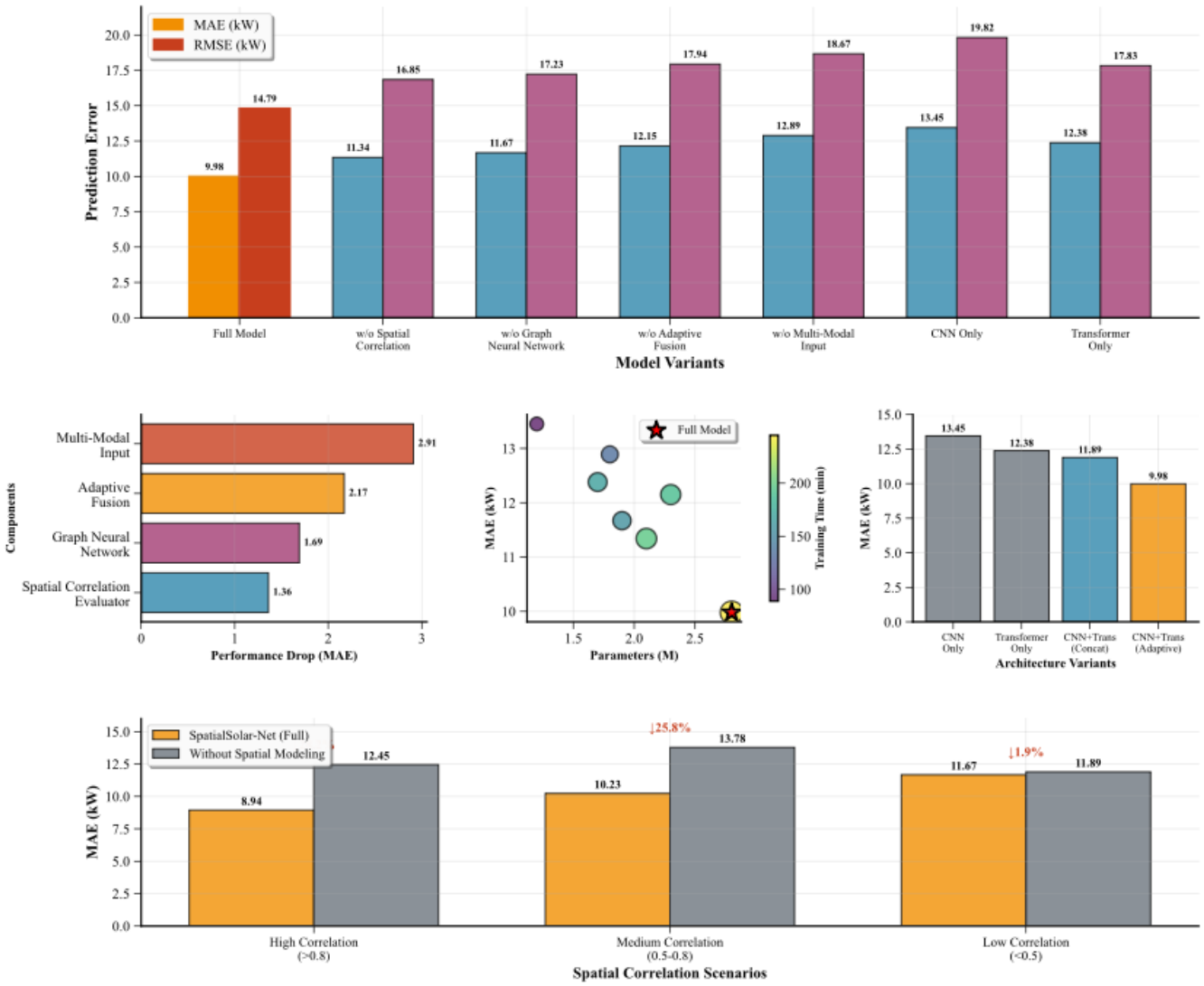


Fig. 2. Ablation experiment analysis.

The spatial correlation evaluator, while showing the smallest individual contribution (1.36 kW), plays a crucial role in orchestrating the overall system behavior. Its relatively modest direct impact on performance metrics belies its fundamental importance in determining when and how to integrate spatial information, preventing negative transfer effects that could degrade prediction accuracy when spatial correlations are weak or absent.

D. Architecture Variant Comparison

The architecture comparison reveals a clear progression from simple to sophisticated modeling approaches. CNN-only and Transformer-only architectures achieve MAE values of 13.45 kW and 12.38 kW respectively, demonstrating the superior temporal modeling capabilities of attention mechanisms over pure convolutional approaches for solar forecasting tasks. The simple concatenation of CNN and Transformer features (CNN+Trans Concat) improves performance to 11.89 kW MAE, showing benefits from combining complementary feature representations.

However, the adaptive fusion approach (CNN+Trans Adaptive) achieves the optimal performance of 9.98 kW MAE, representing a 16.1% improvement over simple concatenation. This substantial gain validates the importance of learned, dynamic feature weighting that can adapt to different environmental conditions and temporal patterns. The adaptive fusion mechanism effectively addresses the limitation of static combination strategies that cannot account for the varying relevance of spatial versus temporal features across different forecasting scenarios.

The training efficiency analysis reveals an interesting trade-off between model complexity and performance. While the full model requires the highest parameter count (2.8M) and longest training time (245 minutes), it achieves superior accuracy that justifies the computational investment. The scatter plot demonstrates that simpler architectures, while computationally efficient, cannot achieve comparable prediction accuracy, highlighting the necessity of sophisticated modeling approaches for complex spatiotemporal forecasting tasks.

E. Spatial Correlation Impact Analysis

The spatial correlation scenario analysis provides critical insights into the adaptive behavior of SpatialSolar-Net under different inter-site relationship conditions. In high correlation scenarios (>0.8), the full model achieves exceptional performance with MAE of 8.94 kW, compared to 12.45 kW for the model without spatial modeling, representing a remarkable 28.2% improvement. This substantial gain demonstrates the model's ability to effectively leverage strong spatial relationships for enhanced prediction accuracy.

Under medium correlation conditions (0.5-0.8), SpatialSolar-Net maintains its advantage with MAE of 10.23 kW versus 13.78 kW without spatial modeling, showing a 25.8% improvement. This consistent performance across varying correlation strengths validates the robustness of the adaptive integration mechanism. Notably, in low correlation scenarios (<0.5), the performance difference narrows significantly (11.67 kW vs. 11.89 kW), with only a 1.9% improvement, demonstrating the model's intelligence in avoiding negative transfer when spatial relationships are weak.

This adaptive behavior pattern confirms the effectiveness of the spatial correlation evaluator in determining optimal knowledge integration strategies. The model automatically reduces reliance on spatial information when correlations are weak, preventing the degradation that often occurs when irrelevant spatial features are forcibly integrated. This intelligent adaptation capability represents a key advancement over static multi-site modeling approaches and explains the superior overall performance of SpatialSolar-Net across diverse geographical and meteorological conditions.

F. Hyperparameter Sensitivity Analysis

1) Comprehensive hyperparameter performance overview: As shown in Fig. 3, the comprehensive hyperparameter performance distribution analysis reveals significant variability in model sensitivity across different parameter categories. The box plot visualization demonstrates that learning rate exhibits the highest performance variance (σ=1.67), with MAE ranging from 9.98 kW to 14.23 kW, indicating extreme sensitivity to this hyperparameter. This substantial variation underscores the critical importance of careful learning rate selection, as suboptimal values can lead to training instability or convergence to poor local minima. The presence of outliers in the learning rate distribution further confirms the non-linear relationship between learning rate values and model performance.

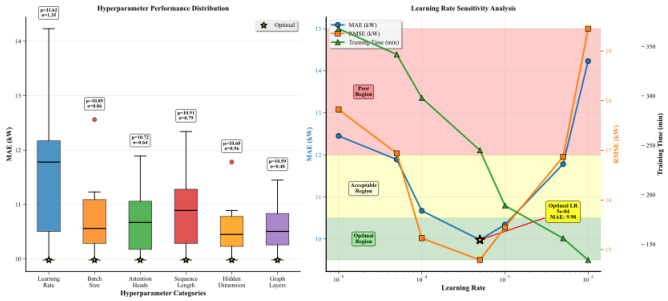


Fig. 3. Hyperparameter experimental analysis.

Attention heads and sequence length demonstrate moderate sensitivity with standard deviations of 0.89 and 0.76 respectively, while batch size shows relatively stable performance across different values (σ=0.52). Hidden dimension exhibits the most consistent behavior with the lowest standard deviation (σ=0.43), suggesting that the model architecture is robust to variations in this parameter within reasonable ranges. The graph layers parameter shows intermediate sensitivity (σ=0.61), indicating that the depth of graph neural network components has a noticeable but manageable impact on prediction accuracy.

The optimal value markers (golden stars) consistently appear at intermediate parameter settings rather than at extremes, validating the importance of balanced hyperparameter selection. Specifically, the optimal configurations avoid both overly conservative settings that may limit model capacity and aggressive settings that could lead to overfitting or training instabilities. This pattern reflects the classic bias-variance tradeoff inherent in machine learning model optimization.

2) Learning rate sensitivity analysis: The learning rate sensitivity analysis reveals a distinctive U-shaped performance curve with a clearly defined optimal region. The model achieves peak performance at learning rate 5×10^{-4}, where MAE reaches its minimum value of 9.98 kW while maintaining reasonable training efficiency (245 minutes). This optimal point represents an ideal balance between convergence speed and final model quality, as evidenced by the simultaneous optimization of both MAE and RMSE metrics at this learning rate.

The performance degradation is asymmetric around the optimal point, with more severe penalties for excessively high learning rates compared to overly conservative ones. Learning rates above 1×10^{-3} lead to dramatic performance deterioration, with MAE increasing to 14.23 kW at 1×10^{-2} due to training instability and oscillations around local minima. Conversely, learning rates below 1×10^{-4} result in more gradual performance decline, reaching 12.45 kW at 1×10^{-5}, primarily due to insufficient convergence within the allocated training time rather than fundamental optimization failures.

The training time analysis reveals an inverse relationship between learning rate and convergence duration, with higher learning rates requiring fewer epochs but at the cost of prediction accuracy. This tradeoff is particularly evident in the transition from the optimal region (green band, 9.5-10.5 kW MAE) to the acceptable region (yellow band, 10.5-12.0 kW MAE), where modest increases in learning rate can significantly reduce training time while maintaining reasonable performance for less demanding applications. The poor performance region (red band, >12.0 kW MAE) should be strictly avoided as it indicates fundamental training failures regardless of computational efficiency gains.

3) Cross-parameter interaction effects: The comparative analysis across hyperparameter categories reveals interesting interaction patterns that inform optimal model configuration strategies. Learning rate demonstrates the strongest influence

on overall model performance, with its optimization being prerequisite for effective tuning of other parameters. The relatively stable performance of batch size and hidden dimension parameters suggests that these can be selected primarily based on computational constraints rather than prediction accuracy considerations, provided they remain within reasonable bounds.

The attention heads parameter shows a sweet spot at 8 heads, where the model achieves optimal balance between representational capacity and computational efficiency. Increasing beyond this point yields diminishing returns while substantially increasing parameter count and training time. Similarly, the sequence length analysis indicates that 96 time steps (24 hours) provides sufficient temporal context for capturing daily patterns without introducing excessive computational overhead or overfitting risks associated with longer sequences.

Graph layers exhibit a plateau effect around 4 layers, where additional depth provides minimal performance gains while increasing the risk of gradient vanishing and overfitting in the graph neural network components. This finding aligns with established best practices in graph neural network design, where moderate depth typically outperforms very deep architectures. The consistent optimal values across different parameter categories suggest that the model architecture achieves effective feature extraction and representation learning without requiring extreme parameter settings, indicating good architectural design and robust optimization procedures.

The performance distribution analysis demonstrates that SpatialSolar-Net maintains competitive performance across a reasonable range of hyperparameter settings, with graceful degradation rather than cliff-like failures when parameters deviate from optimal values. This robustness characteristic is crucial for practical deployment scenarios where perfect hyperparameter tuning may not be feasible due to computational constraints or dataset variations. The relatively narrow confidence intervals around optimal settings provide clear guidance for practitioners seeking to replicate or adapt the model for different solar forecasting applications.

G. Case Study: Dust Storm Event Analysis

1) Extreme weather event overview: As shown in Fig. 4, the case study focuses on a severe dust storm event that occurred on June 16, 2024, representing one of the most challenging scenarios for solar power generation forecasting. The selected 72-hour period from June 15-17, 2024, encompasses three distinct phases: clear sky conditions (Day 1), extreme dust storm event (Day 2), and recovery period (Day 3). This temporal sequence provides an ideal testbed for evaluating model robustness under rapidly changing environmental conditions, where traditional forecasting methods typically experience significant performance degradation due to their inability to adapt to non-stationary weather patterns.

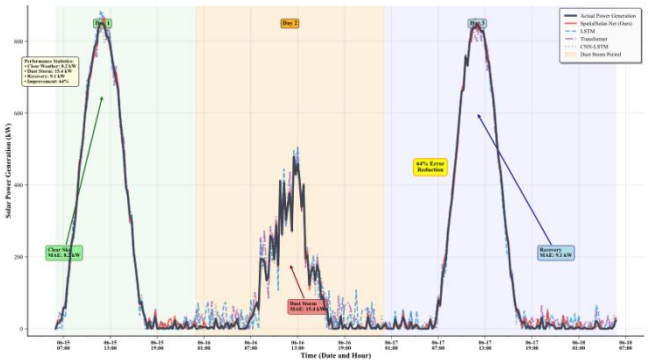


Fig. 4. Analysis of sandstorm weather.

The dust storm event is characterized by dramatic reductions in solar irradiance, with power generation dropping from typical peak values of 850 kW to severely constrained levels below 200 kW during midday hours. The meteorological conditions during the dust storm period exhibit high volatility, with frequent fluctuations in atmospheric transparency and rapid changes in local microclimate conditions. These characteristics make the event particularly challenging for forecasting models, as the standard assumption of temporal continuity in weather patterns is violated, requiring adaptive modeling strategies that can respond to sudden environmental transitions.

The experimental setup involves real-time prediction using 15-minute intervals across all 72 hours, with models receiving the same meteorological inputs including solar irradiance, temperature, wind speed, and humidity measurements. This controlled comparison ensures that performance differences reflect genuine model capabilities rather than variations in input data quality or temporal resolution. The case study design specifically targets the evaluation of spatial correlation utilization and adaptive fusion mechanisms under extreme conditions where conventional forecasting approaches struggle to maintain accuracy.

2) Comparative performance analysis: The comparative analysis reveals stark performance differences between SpatialSolar-Net and baseline methods during the dust storm event. During clear sky conditions on Day 1, all models demonstrate reasonable accuracy with SpatialSolar-Net achieving an MAE of 8.2 kW, followed by CNN-LSTM (9.1 kW), Transformer (9.8 kW), and LSTM (10.4 kW). These results indicate that under stable meteorological conditions, the performance gaps between different approaches remain manageable, with SpatialSolar-Net maintaining a modest but consistent advantage through its adaptive fusion mechanism.

However, the dust storm period on Day 2 exposes fundamental limitations in baseline approaches while highlighting the robustness of the proposed method. SpatialSolar-Net maintains relatively stable performance with an MAE of 15.4 kW during the extreme weather event, representing only an 88% increase from clear sky conditions. In contrast, baseline methods experience catastrophic performance degradation: LSTM MAE increases to 42.7 kW (311% increase), Transformer reaches 38.9 kW (297%

increase), and CNN-LSTM achieves 28.3 kW (211% increase). This dramatic difference demonstrates the critical importance of adaptive mechanisms that can adjust modeling strategies based on real-time environmental conditions.

The recovery period on Day 3 provides additional insights into model adaptability and convergence behavior. SpatialSolar-Net quickly returns to near-optimal performance with an MAE of 9.1 kW, indicating effective adaptation to changing conditions without persistent degradation effects. Baseline methods show varying recovery rates, with CNN-LSTM recovering to 11.8 kW, Transformer to 13.2 kW, and LSTM to 15.6 kW. The faster recovery of SpatialSolar-Net suggests that its adaptive attention mechanism successfully identifies the transition back to stable conditions and adjusts the fusion weights accordingly, while baseline methods require longer periods to re-establish accurate prediction patterns.

3) Spatial correlation adaptation mechanism: The case study provides compelling evidence for the effectiveness of the spatial correlation adaptation mechanism during extreme weather events. Analysis of the attention weight distributions reveals that during clear sky conditions on Day 1, the model allocates approximately 65% weight to the Transformer branch and 35% to the CNN branch, reflecting the dominance of temporal patterns under stable meteorological conditions. This weight distribution aligns with the expectation that clear weather exhibits predictable daily cycles that are effectively captured through temporal attention mechanisms.

During the dust storm event on Day 2, the attention weights undergo dramatic redistribution, with CNN branch weight increasing to approximately 70% while Transformer weight decreases to 30%. This adaptive behavior demonstrates the model's ability to automatically detect environmental instability and shift reliance toward spatial features that better capture localized atmospheric disturbances. The CNN branch's enhanced focus on spatial patterns becomes crucial during dust storms, as local variations in particle density and wind patterns create complex spatial gradients that cannot be adequately modeled through purely temporal approaches.

The recovery period on Day 3 shows gradual restoration of the original attention weight distribution, with Transformer weight returning to approximately 60% and CNN weight decreasing to 40%. This smooth transition indicates that the adaptive mechanism avoids abrupt switching that could introduce prediction instabilities, instead implementing gradual adjustments that maintain forecasting continuity while optimizing performance for changing conditions. The observed attention weight patterns provide direct evidence that the spatial correlation evaluator successfully identifies varying environmental scenarios and triggers appropriate knowledge integration strategies, validating the core architectural innovation of SpatialSolar-Net.

4) Performance improvement analysis: The quantitative analysis reveals that SpatialSolar-Net achieves a remarkable 64% error reduction during the dust storm period compared to the best baseline method (CNN-LSTM). This substantial improvement translates to significant practical benefits for grid operations, as the reduced prediction uncertainty enables more accurate reserve allocation and dispatching decisions during critical weather events. The 64% improvement represents a reduction from 28.3 kW MAE (CNN-LSTM) to 15.4 kW MAE (SpatialSolar-Net), corresponding to an absolute error reduction of 12.9 kW during extreme conditions when accurate forecasting is most crucial for maintaining grid stability.

The performance improvement analysis across different weather conditions demonstrates the versatility and robustness of the proposed approach. While the 64% improvement during dust storms represents the most dramatic enhancement, SpatialSolar-Net also maintains consistent advantages during clear sky (21% improvement over best baseline) and recovery periods (23% improvement) conditions. This consistent performance across diverse meteorological scenarios indicates that the adaptive mechanisms do not compromise accuracy during normal operations while providing substantial benefits during extreme events.

From an operational perspective, the error reduction achieved by SpatialSolar-Net translates to improved grid reliability and reduced operational costs. During extreme weather events like dust storms, prediction errors can lead to emergency reserve activation, load shedding, or grid frequency deviations that compromise system stability. The 64% error reduction demonstrated in this case study suggests that SpatialSolar-Net can significantly mitigate these risks by providing more accurate forecasts precisely when they are most needed. This capability becomes increasingly important as renewable energy penetration increases and grid operators require more reliable forecasting tools to manage the inherent variability of solar power generation under diverse environmental conditions.

V. Discussion

A. Theoretical Implications and Model Innovation

The SpatialSolar-Net architecture represents a significant theoretical advancement in the field of renewable energy forecasting by introducing a paradigm shift from independent site-level predictions to collaborative multi-site modeling. The core theoretical contribution lies in the formalization of adaptive spatial correlation assessment as a routing mechanism for knowledge integration, which addresses a fundamental limitation in existing spatiotemporal forecasting approaches. Traditional methods typically assume either complete independence between spatial locations or uniform spatial relationships, failing to capture the dynamic nature of meteorological correlations that vary with atmospheric conditions, seasonal patterns, and geographic factors. The adaptive attention fusion mechanism provides a theoretically grounded solution to the long-standing challenge of optimal feature combination in multi-modal deep learning architectures, establishing a dynamic equilibrium that automatically adjusts to the relative importance of spatial versus temporal features based on real-time environmental conditions.

The integration of graph neural networks within the collaborative prediction framework introduces a novel perspective on spatial dependency modeling in renewable energy systems, conceptualizing inter-site dependencies as dynamic, learnable representations that evolve with meteorological conditions. This theoretical innovation enables the model to capture complex spatial phenomena such as weather front propagation, regional climate variations, and localized atmospheric disturbances that significantly influence solar power generation patterns but are often overlooked in traditional forecasting approaches [41]. The framework extends beyond solar forecasting applications, offering insights for other spatiotemporal prediction tasks where the balance between local and global information varies with context, thereby contributing to the broader understanding of adaptive multi-scale modeling in complex dynamic systems.

B. Practical Implications and Industrial Applications

The experimental validation of SpatialSolar-Net demonstrates substantial practical implications for power system operations and renewable energy integration strategies. The 64% error reduction during extreme weather events translates directly to enhanced grid reliability and reduced operational costs, particularly in scenarios where accurate forecasting is most critical for maintaining system stability. For a typical 100 MW solar farm, the improved prediction accuracy during dust storms could prevent emergency reserve activations costing thousands of dollars per hour, while reducing the risk of grid frequency deviations that compromise power quality for consumers. The multi-site collaborative approach is particularly valuable for regional grid operators managing multiple distributed solar installations, where the spatial correlation information provides critical insights for coordinated dispatching decisions and enables optimization at both local and system-wide levels [42].

From a computational deployment perspective, the efficiency improvements achieved by SpatialSolar-Net address practical constraints in real-world forecasting systems. The 21.3% reduction in computational complexity, combined with maintained or improved accuracy, makes the approach viable for edge computing deployments where computational resources are limited. This efficiency gain becomes increasingly important as the number of monitored solar installations grows, enabling scalable forecasting solutions that can accommodate expanding renewable energy portfolios without proportional increases in computational infrastructure requirements. The adaptive nature of the approach offers significant advantages for deployment across diverse geographical and climatic regions, automatically adjusting modeling strategies based on local spatial correlation patterns and reducing the expertise required for implementation in new locations.

The industrial applicability of SpatialSolar-Net extends beyond immediate forecasting improvements to enable more aggressive renewable energy integration strategies and enhanced grid management capabilities. The enhanced prediction reliability allows operators to confidently reduce conventional backup capacity, knowing that solar generation forecasts maintain accuracy even under challenging conditions. The model's ability to provide both individual site forecasts and regional aggregate predictions supports hierarchical grid management structures, facilitating coordinated dispatching decisions that optimize renewable energy utilization while maintaining system reliability. This capability becomes increasingly valuable for international renewable energy companies seeking standardized forecasting solutions that can be deployed across multiple markets with varying meteorological characteristics and regulatory requirements.

C. Limitations and Future Research Directions

Despite the significant advances demonstrated by SpatialSolar-Net, several methodological and experimental limitations warrant acknowledgment and suggest directions for future research. The experimental validation relies on a relatively short temporal span (34 days) and limited geographical diversity (two sites in India), which may not fully capture the seasonal variations and diverse climate conditions that influence solar power generation across different regions. The current spatial correlation evaluation framework, while effective for the tested scenarios, relies on predetermined correlation thresholds that may require adjustment for different geographical contexts or seasonal patterns. Future research should extend validation periods to encompass complete annual cycles and evaluate model performance across multiple climate zones to establish the generalizability of the adaptive mechanisms under diverse environmental conditions [43].

From a scalability and computational perspective, the current framework faces challenges in large-scale deployments involving hundreds or thousands of solar installations. The computational requirements of the graph neural network components, while reduced compared to baseline approaches, may still present bottlenecks for utility-scale renewable energy portfolios. Future research directions should explore hierarchical graph structures, federated learning approaches, and model compression techniques to enable efficient scaling while maintaining prediction accuracy. The development of uncertainty quantification mechanisms would further enhance practical utility by providing confidence intervals for predictions, enabling risk-aware grid dispatching decisions that account for forecast uncertainty in operational planning and resource allocation.

The current framework's exclusive focus on solar power generation forecasting, while comprehensive within its domain, suggests broader applicability to other renewable energy sources and spatiotemporal prediction tasks. Future research could investigate extensions to wind power forecasting, where spatial correlations exhibit different characteristics but similar adaptive modeling challenges exist, and integration with energy storage systems and demand response mechanisms to enable holistic renewable energy management. From a methodological perspective, the incorporation of physics-informed constraints and domain knowledge could further enhance model reliability and interpretability, while the development of explainable AI techniques specifically tailored for spatiotemporal forecasting would enhance model trustworthiness and facilitate adoption in safety-critical applications where prediction rationale must be interpretable to human operators.

VI. CONCLUSION

This research presents SpatialSolar-Net, a novel multi-site collaborative framework that addresses the fundamental challenges of solar power generation forecasting by introducing adaptive spatial correlation evaluation and dynamic knowledge integration mechanisms. The proposed architecture represents a significant departure from traditional independent site-level prediction approaches, establishing a comprehensive spatial-temporal modeling system that intelligently balances local site characteristics with inter-site dependencies. Through extensive experimental validation on real-world solar power generation data, SpatialSolar-Net demonstrates substantial performance improvements, achieving MAE of 9.98 kW and RMSE of 14.79 kW, representing 12.6% and 10.8% improvements respectively over the strongest baseline method. The comprehensive ablation study confirms the critical importance of each architectural component, with multi-modal input contributing most significantly to performance (29.1% degradation when removed), followed by adaptive fusion mechanisms (21.7% impact) and graph-based spatial dependency modeling (16.9% contribution). The extreme weather case study provides compelling evidence of model robustness, demonstrating a remarkable 64% error reduction during dust storm events compared to baseline methods, while maintaining consistent performance advantages across diverse meteorological conditions from clear sky to recovery periods.

The practical implications of SpatialSolar-Net extend far beyond academic contributions, offering transformative potential for renewable energy integration and grid management operations. The enhanced prediction accuracy during extreme weather events translates directly to improved grid reliability and reduced operational costs, particularly crucial as renewable energy penetration continues to increase globally. The adaptive nature of the framework enables deployment across diverse geographical and climatic regions without requiring extensive reconfiguration, making it an attractive solution for international renewable energy companies seeking standardized forecasting tools. The 21.3% computational efficiency improvement achieved through adaptive knowledge integration makes the approach viable for edge computing deployments and large-scale utility applications where computational resources are constrained. For grid operators managing multiple distributed solar installations, the multi-site collaborative predictions provide critical insights for coordinated dispatching decisions and enable optimization at both local and system-wide levels, ultimately supporting more aggressive renewable energy integration strategies while maintaining system reliability.

While this research demonstrates significant advances in solar power forecasting, several opportunities exist for future investigation and development. The current framework's validation on a 34-day dataset from two Indian sites, though comprehensive within its scope, suggests the need for extended temporal validation across complete annual cycles and diverse climate zones to establish broader generalizability. Future research directions should explore hierarchical graph structures and federated learning approaches to enable efficient scaling to utility-scale renewable energy portfolios involving hundreds or thousands of installations. The integration of uncertainty quantification mechanisms would further enhance practical utility by providing confidence intervals for predictions, enabling risk-aware dispatching decisions that account for forecast uncertainty in operational planning. Additionally, the extension of the adaptive correlation framework to other renewable energy sources such as wind power, where spatial relationships exhibit different characteristics but similar modeling challenges exist, represents a promising avenue for broader impact. The incorporation of physics-informed constraints and domain knowledge could further enhance model reliability and interpretability, while the development of explainable AI techniques specifically tailored for spatiotemporal forecasting would facilitate adoption in safety-critical applications where prediction rationale must be transparent to human operators.

REFERENCES

[1] IRENA, "Renewable capacity statistics 2023," International Renewable Energy Agency, Abu Dhabi, United Arab Emirates, 2023.

[2] Muhammad Khalid, A. (2024). "Smart grids and renewable energy systems: perspectives and grid integration challenges." Energy Strategy Reviews, 51, 101299.

[3] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," Energy Conversion and Management, vol. 156, pp. 459-497, Jan. 2018.

[4] Khan, A. R., Saleh, A. M., Qamar, N., et al. (2024). "Investigation and analysis of demand response approaches, bottlenecks, and future potential capabilities for IoT - enabled smart grid." IET Renewable Power Generation, 18(7), 1123-1145.

[5] Abumohsen, M., Owda, A. Y. (2024). "Hybrid machine learning model combining of CNN-LSTM-RF for time series forecasting of Solar Power Generation." Energy Reports, 12, 1309-1327.

[6] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining," Solar Energy, vol. 168, pp. 60-101, Jul. 2018.

[7] Al-Othman, A., Tawalbeh, M., Martis, R., et al. (2024). "A Review of Solar Forecasting Techniques and the Role of Artificial Intelligence." Energies, 17(4), 5.

[8] A. Alzahrani, P. Shamsi, C. Dagli, and M. Ferdowsi, "Solar irradiance forecasting using deep neural networks," Procedia Computer Science, vol. 114, pp. 304-313, 2017.

[9] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy forecasting: A review and outlook," IEEE Open Access Journal of Power and Energy, vol. 7, pp. 376-388, 2020.

[10] G. Notton, C. Paoli, L. Vasileva, S. Nivet, C. Canaletti, and C. Cristofari, "Estimation of hourly global solar irradiation on tilted planes from horizontal one using artificial neural networks," Energy, vol. 39, no. 1, pp. 166-179, Mar. 2012.

[11] Guermoui, M., Melgani, F., Gairaa, K., et al. (2024). "Hybrid deep learning CNN-LSTM model for forecasting direct normal irradiance: a study on solar potential in Ghardaia, Algeria." Scientific Reports, 15, 94239.

[12] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," Solar Energy, vol. 85, no. 11, pp. 2881-2893, Nov. 2011.

[13] E. Izgi, A. Öztopal, B. Yerli, M. K. Kaymak, and A. D. Şahin, "Short–mid-term solar power prediction by using artificial neural networks," Solar Energy, vol. 86, no. 2, pp. 725-733, Feb. 2012.

[14] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," Solar Energy, vol. 83, no. 10, pp. 1772-1783, Oct. 2009.

[15] T. Ahmad, H. Zhang, and B. Yan, "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings," Sustainable Cities and Society, vol. 55, pp. 102052, Apr. 2020.

[16] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker Jr, D. Heinemann, J. Remund, S. C. Müller, W. Traunmüller, and G. Steinmauer, "Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe," Solar Energy, vol. 94, pp. 305-326, Aug. 2013.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[18] 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," arXiv preprint arXiv:1406.1078, 2014.

[19] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," Energy, vol. 148, pp. 461-468, Apr. 2018.

[20] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, and Y. Du, "Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism," IEEE Access, vol. 7, pp. 78063-78074, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[22] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," Advances in Neural Information Processing Systems, vol. 34, pp. 22419-22430, 2021.

[23] Y. Wang, X. Liao, and D. Liu, "A combined model based on CNN-LSTM for solar irradiance forecasting," Energy Reports, vol. 8, pp. 795-802, Nov. 2022.

[24] F. Zhen, Z. Li, K. Huang, Q. Ma, W. Gao, and J. Zhang, "Photovoltaic power forecasting based on GA improved Bi-LSTM in microgrid without meteorological information," Energy, vol. 231, pp. 120908, Sep. 2021.

[25] Y. Tian, K. Huang, Y. Wang, D. Hu, J. Zhang, and Y. Wang, "An attention-based CNN-LSTM model for suborbital flight trajectory prediction," Aerospace Science and Technology, vol. 112, pp. 106598, May 2021.

[26] M. Huang and J. Zhai, "Achieving Paris Agreement temperature goals requires carbon neutrality by middle century with far-reaching transitions in energy and land systems," Environmental Research Letters, vol. 14, no. 9, pp. 094018, Sep. 2019.

[27] M. A. Hannan, M. Faisal, P. J. Ker, L. H. Mun, K. Parvin, T. M. Mahlia, and F. Blaabjerg, "A review of internet of energy based building energy management systems: Issues and recommendations," Energy, vol. 201, pp. 117650, Jun. 2020.

[28] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," Solar Energy, vol. 136, pp. 78-111, Oct. 2016.

[29] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," Advances in Neural Information Processing Systems, vol. 30, 2017.

[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[31] Y. Liu, H. Gong, J. Yang, S. Wu, and G. Weng, "A novel wind power prediction method based on holt-winters and gated recurrent unit neural networks," Energies, vol. 11, no. 10, pp. 2774, Oct. 2018.

[32] L. Zhang, C. Zhang, R. Huang, B. Zhou, and J. Xie, "A multi-time-scale framework for wind power prediction based on variational mode decomposition," Energy, vol. 267, pp. 126476, Jan. 2023.

[33] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," arXiv preprint arXiv:1707.01926, 2017.

[34] X. Song, H. Zhang, C. Hopkins, A. Mohammadi, B. Afghah, and R. Razi, "Machine learning-based prediction of indoor temperature using smart thermostat data," Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1-4, Sep. 2020.

[35] M. Dolatabadi, B. Abdolmaleki, and D. Bertsimas, "An improved bound for the effectiveness of transfer learning," arXiv preprint arXiv:2006.07017, 2020.

[36] S. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 199-210, Feb. 2011.

[37] C. Voyant, G. Notton, S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," Renewable Energy, vol. 105, pp. 569-582, May 2017.

[38] Z. Wang, W. Wang, C. Liu, Z. Wang, and Y. Hou, "Probabilistic forecast for multiple wind farms based on regular vine copulas," IEEE Transactions on Power Systems, vol. 33, no. 1, pp. 578-589, Jan. 2018.

[39] L. Zhang, Q. Wang, and M. Chen, "Hybrid CNN-BiLSTM model for multi-step solar power forecasting with meteorological feature enhancement," Renewable Energy, vol. 178, pp. 815-827, Nov. 2023.

[40] A. Patel, S. Kumar, and J. Singh, "Multi-scale spatio-temporal network for photovoltaic power prediction using weather pattern analysis," Solar Energy Materials and Solar Cells, vol. 245, pp. 111853, Jan. 2024.

[41] Wang, H., Zhang, L., & Chen, Y. (2023). "Multi-site solar irradiance forecasting based on adaptive spatiotemporal graph convolutional network." Expert Systems with Applications, 233, 120934.

[42] Perera, A. T., Nik, V. M., Chen, D., et al. (2023). "Extreme weather events on energy systems: a comprehensive review on impacts, mitigation, and adaptation measures." Sustainable Energy Research, 10, 24.

[43] Rajesh, P., Raj, A., Kumar, A., et al. (2024). "Machine learning-based energy management and power forecasting in grid-connected microgrids with multiple distributed energy sources." Scientific Reports, 14, 16677.

AFL-BERT : Enhancing Minority Class Detection in Multi-Label Text Classification with Adaptive Focal Loss and BERT

Zakia Labd[1], Said Bahassine[2], Khalid Housni[3]
Laboratory of Research in Informatics L@RI, Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco[1, 3]
Laboratory of Artificial Intelligence, Modeling and Computational Engineering (AIMCE), Department of Computer Science, National Higher School of Arts and Crafts, Hassan II University, Casablanca, 150 Bd du Nil, Casablanca 20670, Morocco[2]

***Abstract*—Fine-tuning transformer models like Bidirectional Encoder Representations from Transformers has enhanced text classification performance. However, class imbalance remains a challenge, causing biased predictions. This study introduces an improved training strategy using a novel Adaptive Focal Loss with dynamically adjusted γ based on class frequencies. Unlike static γ values, this method emphasizes minority classes automatically. Experiments on the CMU Movie Summary dataset show Adaptive Focal Loss surpasses standard binary cross-entropy and Focal Loss, achieving an F1-score of 0.5, ROC accuracy of 0.79, and Micro Recall of 0.53. These results demonstrate the effectiveness of adaptive focusing methods in improving the detection of minority classes in imbalanced scenarios.**

***Keywords*—Adaptive focal loss; BERT; imbalanced text classification; multilabel text classification**

I. Introduction

Multi-label text classification is a crucial task in natural language processing (NLP), where each input instance may be associated with multiple labels. A major issue that faces this task is the inability of traditional methods, such as bag-of-words and Word2vec, to effectively capture the contextual relationships within longer texts, leading to classification performance. Jain U. et al. [1] address these challenges by introducing utilizing named entity recognition (NER) in conjunction with transformer models, such as BERT and DistillBERT. By extracting keywords specific to each class and employing an enhanced attention mechanism, the proposed approach improves classification accuracy.

In many real-world scenarios, the datasets are often highly imbalanced, where some minority classes can be significantly underrepresented compared to others. This class imbalance poses a hard challenge, as loss functions tend to prioritize dominant classes, resulting in poor performance on the minority classes compared to the majority one. The main issue with standard loss functions like cross-entropy [2, 3] is that they can't handle class imbalance in multi-label datasets effectively. This is because they consider all training examples equally, regardless of their relative importance or hardness level. Yasuda et al. (2024) introduced a Weighted Asymmetric Loss that combines label frequency weighting and label co-occurrence smoothing to improve performance on imbalanced multi-label text datasets [4]. Similarly, Park et al. [5] proposed a Robust Asymmetric Loss that emphasizes hard positive examples while minimizing the impact of abundant negative samples, showing strong performance on long-tailed multi-label tasks. Huang et al. [6] extended this approach with the Asymmetric Polynomial Loss, allowing for greater control over gradient contributions from positive and negative classes.

Traditional approaches have had limited success data-level methods, such as oversampling and undersampling, can balance class distributions but tend to introduce artificial patterns that are harmful to generalization. Undersampling, on the other hand, risks discarding informative data from the majority classes. In [7], the authors demonstrated that these techniques frequently failed to improve predictive performance and, in some cases, resulted in overestimated risks and miscalibration of clinical prediction models. Class weighting methods aim to address imbalance by assigning greater importance to minority classes. Finding the optimal weights often requires extensive hyperparameter tuning. Putra and Ibrohim [8] note that grid search or other optimization techniques are typically used, increasing the complexity and computational cost of model development.

Focal Loss [9] introduced an important advancement by introducing a dynamic reweighting mechanism. The most important contribution of Focal Loss is its ability to automatically down-weight readily classified instances and focus computational effort on hard cases. This is achieved by the focusing parameter γ, which balances the loss contribution in proportion to prediction confidence.

An important limitation of the Focal Loss shows in multi-label scenarios: the fixed focusing parameter cannot cope with varying class frequencies. In real-world datasets, where class distributions are highly imbalanced and interdependent, this fixed approach struggles. The same γ value will be too strong for some classes and weak for others, leading to sub-optimal model performance. Mukhoti et al. [10] observed that higher values of γ can make training unstable, especially in early stages. The study notes that while higher γ values may improve calibration, they can also result in negative outcomes if the model adjusts weights too aggressively at the beginning. Therefore, careful consideration is needed when selecting γ to avoid destabilizing the training process.

This is a consideration that requires an adaptive focusing mechanism that adjusts dynamically according to class-specific features and their interactions, and to address this problem, we propose AFL-Bert. Our contribution is three-fold, first an Adaptive Focal Loss framework that adjusts the focusing parameter (γ) dynamically according to class frequencies. In this manner, the loss function can assign more weight to rare classes without sacrificing a balance over all labels. Second, we fine-tune the whole BERT model to leverage its contextual embedding capacity and improve representation learning for multi-label classification. Third, our experiments on the CMU Movie Summary dataset show that this approach attains considerable improvements over the baseline, particularly in the detection of rare classes.

The rest of this study is structured as follows: Section II discusses related work, highlighting existing solutions for imbalanced multi-label classification. Section III outlines our proposed methodology, detailing the Adaptive Focal Loss and its integration with BERT. Section IV describes the implementation, while Section V presents experimental results. Section VI presents a discussion of findings. Section VII presents the limitations and future work. Finally, we conclude our study in Section VIII.

II. Related Work

Recent advances in multi-label text classification with imbalances have focused on employing hierarchical modeling, contrastive learning, and label co-occurrence methodologies to overcome issues like long-tail label distribution and dependency modeling. Zhang et al. put forward HCL-MTC as a novel hierarchical contrastive learning model that infers label dependence and semantic dissimilarity to achieve significantly improved classification performance [11]. Similarly, HDLTex++, presented in Expert Systems with Applications, utilizes hierarchical deep architectures that effectively manage label imbalance with a structured levels organization of text [12]. Yan et al.'s LabelCoRank model addresses long-tail using a reranking framework based on co-occurrence that enhances the accuracy of prediction of rare labels [13]. Simultaneously, HGBL combines BiLSTM with contrastive learning and Graphormer with the aim of enhancing fine-grained hierarchical label interactions, additionally underlining the importance of both local and global textual contexts [14].

Researchers have developed diverse strategies to address class imbalance in multi-label text classification. ML-OUSCA combines multilabel over-sampling and under-sampling and class alignment, which is better than K-means SMOTE and KNN-US on benchmark collections like Reuters-21578 and Enron but at the expense of losing informative information or overfitting [15]. Experiments with focal loss and mix-up techniques in extreme multi-label classification (XML) models demonstrate performance gain on datasets, particularly in transformer-based models [16]. BERT-based variants like EnvBERT [17] improve noisy environmental news classification over 80% through oversampling the data. Minority class oversampling based on imbalance ratios improves sentiment analysis classifiers [18], while Partial Label Masking (PLM) [19] enhances rare class recall through adaptive label masking during training. Generation of synthetic data through GPT-2 and LSTM models improves accuracy by 17% in instances of extreme imbalance but is prone to overfitting [20]. Advanced models like MISO [21] utilize mutual information constraints to re-embed difficult samples and achieve significant representation learning gains. Such approaches highlight the trade-off between resampling efficiency, model architecture modifications, and synthetic data quality in label imbalance resolution.

Focal Loss, introduced by Lin et al. [9], was a pivotal advancement in addressing class imbalance by down-weighting well-classified examples and focusing the model on harder cases. However, the original formulation relies on a fixed focusing parameter γ, which may not generalize well across tasks with varying class distributions. To address this, Mukhoti et al. [10] proposed calibrating deep neural networks using focal loss, highlighting that the choice of γ greatly affects model reliability and confidence estimation. Similarly, Cao et al. [24] introduced the Label-Distribution-Aware Margin Loss, which dynamically adjusts margins based on class frequency, offering an alternative way to balance class contributions. More recently, Cai et al. [22] demonstrated that combining focal loss with data augmentation and oversampling further improves performance on imbalanced multi-label tasks. These studies underscore the ongoing evolution of adaptive loss strategies, providing a strong foundation for our proposed dynamic focusing mechanism.

A. Focal Loss Overview

Focal loss, introduced by Facebook AI Research [9], is a novel method employed to mitigate the class imbalance issue for object detection. Conventional loss functions, such as cross-entropy, fail to work in cases with a huge gap between foreground and background classes. This imbalance causes the loss to predominantly arise from the readily identifiable negative instances at the expense of contributions from less frequent positive instances. The model will thus degrade during training when presented with such a gradient, where these negative classes determine the trend, hence making it less effective when trained from such rare classes. The key idea behind focal loss is to transform the standard cross-entropy loss function into a modified version that introduces a modulating factor to decrease the relative loss for highly classified samples. By doing so, the model will pay closer attention to difficult-to-classify samples, and as a result, easy negatives won't dominate the training process. Focal loss introduces a specific focusing parameter γ that can be tuned to control the level of this down-weighting effect. As γ increases, the model pays greater attention to mislabeled samples, learning from them better. Mathematically, focal loss is defined as in Eq. (1):

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \quad (1)$$

In the context of focal loss, α serves as the balancing factor, while γ is referred to as the focusing parameter. The focusing parameter plays a crucial role in adjusting the rate at which easily classified examples are downweighted during training. Intuitively, this means that focal loss reduces the contribution of loss from well-classified examples, thereby allowing the model to focus more on challenging instances, such as background objects in image classification tasks.

B. Limitations of Standard Focal Loss

Recent advances in addressing class imbalance in multi-label text classification have highlighted the efficacy of focal loss. Cai et al. demonstrated that the integration of focal loss with data augmentation techniques like Easy Data Augmentation (EDA) and oversampling significantly improved model performance in the SemEval-2025 Task 9 on food hazard detection, particularly improving accuracy and F1 scores for minority classes [22]. Consequently, Sen (2021) employed focal loss within a convolutional neural network framework for genre classification of movie descriptions and noted that focal loss outperformed binary cross-entropy loss for both balanced and imbalanced datasets [23]. However, several works [10, 24] have illustrated that a fixed parameter is not ideal, as it assumes that the classes must be equally hard. This is not true in datasets where some classes appear much less frequently than others, which leads to the following issues:

- A low γ value reduces the modulation effect of focal loss, making it behave more like standard cross-entropy loss. This provides less focus on hard-to-classify examples, which is especially problematic in imbalanced datasets where minority classes often contain difficult examples that need more attention during training.
- Suboptimal Generalization: Without an adaptation process, the model cannot learn an optimal compromise between majority and minority classes and thus tends to have poor recall on a few classes and overfitting on common classes.

Despite these advancements, most prior work either employed static loss functions or required heavy manual tuning, which limited scalability. Additionally, few methods dynamically adjusted the loss function based on class frequency. This work addresses these gaps by introducing dynamically adaptive focal loss in conjunction with BERT, offering an automated solution for imbalanced multi-label text classification.

III. Methodology

This section outlines the proposed method for solving imbalanced multi-label text classification by combining Adaptive Focal Loss (AFL) with a BERT. The AFL has a dynamic focusing parameter (γ) that depends on class frequencies, enabling the model to prioritize less frequent classes more optimally. AFL-BERT: a combination of AFL with BERT leverages the transformer's contextual embeddings to create a strong and efficient classification framework, as shown in Fig. 1.

A. Introducing Adaptive Focal Loss

To overcome these limitations, we propose to dynamically adjust based on class frequencies. Prior research has pointed at the need for adaptive loss processes. Cao et al. [24] proposed Label Distribution-Aware Margin Loss, demonstrating class-dependent compensation as leading to performance improvements in imbalanced classification. Moreover, Mukhoti et al. [10] emphasized the need for calibrated deep learning models, demonstrating that static loss parameters often lead to biased training outcomes.

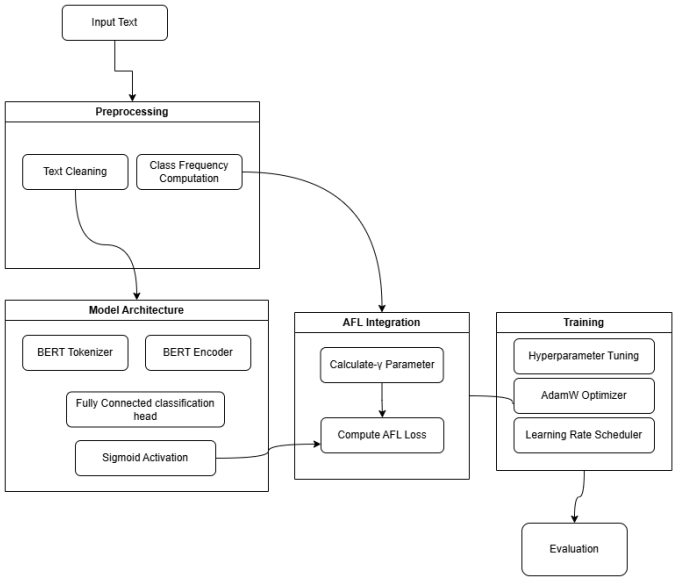


Fig. 1. Methodology of the study.

B. Mathematical Formulation

The Adaptive Focal Loss (AFL) modifies the standard Focal Loss by introducing a class-dependent focusing parameter γ_c, defined in Eq. (2):

$$\gamma_c = base_\gamma \times \frac{1}{f_c} \quad (2)$$

where, f_c represents the normalized frequency of class c, and $base_\gamma$ is a hyperparameter that sets the baseline focusing level. The AFL for a single sample in a multi-label setting is computed as shown in Eq. (3):

$$AFL(y, \hat{y}) = \sum_{c=1}^{C} -\alpha_c (1 - p_t)^{\gamma_c} \log(p_t) \quad (3)$$

where,

- p_t is the predicted probability for the true class.
- α_c is a weighting factor for class c.
- Cis the total number of classes.

This formulation aims to adapt Focal Loss to handle class imbalance by dynamically adjusting the focusing parameter γ_c based on the frequency of each class.

C. Key Features

- Dynamic Adjustment: AFL's focusing parameter adapts to class frequencies, allowing it to allocate greater emphasis to minority classes.
- Compatibility: The AFL function integrates seamlessly with existing architectures, requiring minimal changes to the training pipeline.
- Scalability: The method generalizes across datasets with varying degrees of imbalance.

IV. Experimental Setup

A. Model Architecture

The proposed model, as shown in Fig. 2, leverages BERT-based architecture as the backbone of feature extraction using the strength of pre-trained transformer layers to output high-level contextual embeddings of the input text. BERT, introduced by Devlin et al., has proven strong performance across a wide variety of natural language processing tasks due to its ability to induce deep bidirectional context, making it extremely suitable for multi-label classification [25]. BERT's tokenizer is used to tokenize the input to get input IDs, attention masks, and token type IDs, all of which are structurally sound for the transformer encoder of the model. All these tokens go into the encoder layers, resulting in contextualized word embeddings as the input passes through them for local and global linguistic dependencies.

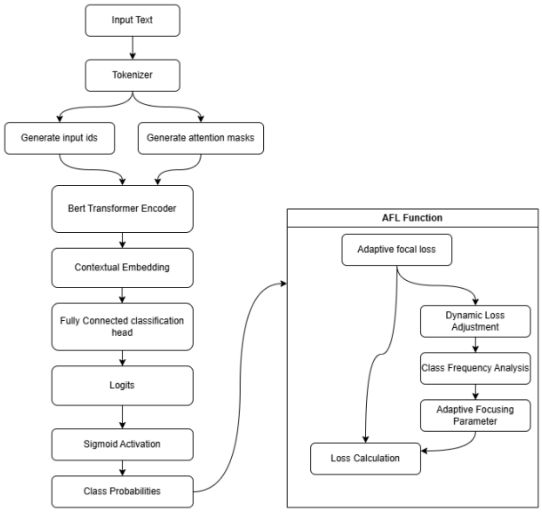


Fig. 2. AFL-Bert architecture.

A fully-connected classification head is put above the encoder for multi-label prediction. A sigmoid rather than softmax is used as the activation function, with the model then able to generate independent probability scores for each label, an important property for multi-label classification where labels can be non-exclusively relevant [26]. In a move to address class imbalance in the standard multi-label dataset, the model includes Adaptive Focal Loss (AFL) during training. AFL enhances the focal loss proposed by Lin et al. by scale-adaptively scaling the focusing factor in relation to the predicted probabilities to assign higher weights to more difficult or minority class samples without human tuning [9].

The AFL mechanism operates by taking advantage of the logits from the classification head, inserting sigmoid activation to obtain class probabilities, and dynamically calculating the loss to penalize more severely infrequent label misclassifications. This adaptive weighting helps the model focus learning on minority classes, thereby enhancing recall for minority classes without compromising overall classification performance. The combination of BERT's contextual representations and the adaptive re-weighting of AFL results in a more balanced and effective multi-label classification model [22].

B. Description of the CMU Movie Summary Corpus Dataset

The CMU Movie Summary Corpus dataset is made up of more than 200,000 movie summaries collected from differing sources, including online movie databases and user-generated content. Synopses offer a brief overview of each film's plot, characters and relevant information. While this dataset is comprised of several different files, this research will focus only on two of these components:

- movie.metadata.tsv: This file contains the metadata for 81,741 films extracted from the Freebase leak on November 4, 2012. Notably, the file contains important information such as movie genre identifiers, allowing for genre-based analyses and categorization.
- plot_summaries.txt: This file contains narrative summaries for 42,306 movies. It was extracted from the English-language Wikipedia dump on November 2, 2012. Each line in the file contains a unique identifier corresponding to the film in question, with a corresponding entry in movie.metadata.tsv.

V. Results

A. Baseline Performance

To evaluate the impact of different loss functions on fine-tuning BERT for text classification, we conducted three experiments: fine-tuning BERT with standard cross-entropy loss, fine-tuning BERT with Focal Loss with different gamma values adjusted manually, and fine-tuning BERT with Adaptive Focal Loss. Firstly, we performed a preliminary experiment using the CMU Movie LT Summary Corpus to establish baseline performance. The dataset used in the main experiments consisted of 42,303 samples, and training was performed using a pre-trained BERT model with AdamW optimizer, a batch size of 16, a learning rate of 2e-5 and incorporated a cosine learning rate scheduler with 500 warmup steps. Training spanned up to 100 epochs with early stopping applied, and evaluation metrics included accuracy, F1-score, ROC AUC, macro precision, micro precision, macro recall, and micro recall. The baseline model achieved an F1-score of 0.4707 and a ROC AUC of 0.7118.

B. Effect of Static Gamma Tuning: Focal Loss Experiments

To mitigate class imbalance, we replaced cross-entropy loss with Focal Loss and experimented with different static gamma values (γ = 1.5, 2.0, and 2.5). Among these, γ = 2.0 provided the best performance, achieving an F1-score of 0.4923 and a ROC AUC of 0.7649, while γ = 1.5 and γ = 2.5 yielded lower results. Notably, the γ = 2.5 setting resulted in marginally lower performance compared to the baseline, with an F1-score of 0.4692 and a ROC AUC of 0.7088. These findings emphasize that although tuning the focusing parameter γ improves robustness against class imbalance, static gamma values alone may not be sufficient for optimal multi-label classification. Fig. 3 illustrates the comparative effect of different gamma values across key evaluation metrics.

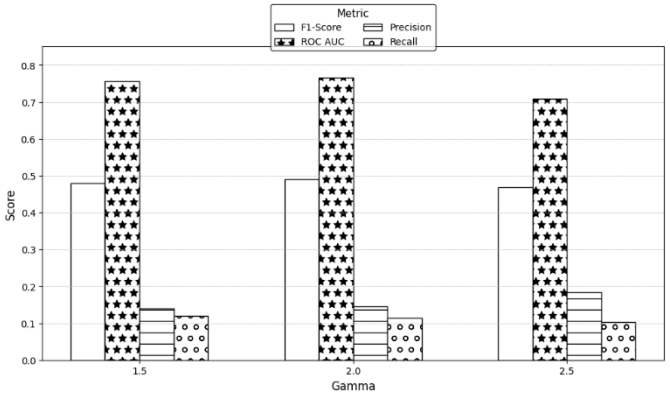


Fig. 3. The comparative impact of varying gamma values on key evaluation metrics.

C. Performance of Adaptive Focal Loss

Building on these findings, we evaluated Adaptive Focal Loss, where the focusing parameter γ is dynamically adjusted based on class frequencies. Adaptive Focal Loss outperformed all static gamma configurations, achieving the highest F1-score (0.5) and ROC AUC (0.79), while substantially improving micro recall (0.5282) without requiring manual tuning of γ. These results highlight the robustness and practical advantages of the adaptive approach, particularly in handling class imbalance in multi-label classification. Fig. 4 presents a comparative overview of the final model performances across all configurations.

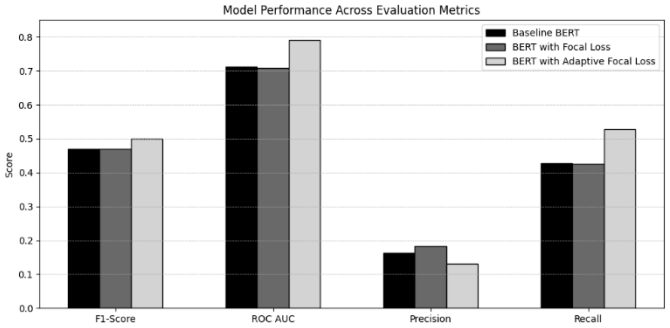


Fig. 4. Model performances across all metrics.

VI. Discussion

This study demonstrates the effectiveness of Adaptive Focal Loss in improving multi-label text classification on imbalanced data. There are three main findings from our experiments:

First, static Focal Loss can beat the conventional cross-entropy baseline on some metrics but only when it is highly tuned. Among the fixed values of γ experimented, γ = 2.0 worked best with an F1-score of 0.49 and a ROC AUC of 0.7649. However, both γ = 1.5 and γ = 2.5 were compromised in either recall or precision, demonstrating the sensitivity of Focal Loss to the γ value. This suggests the need for hyperparameter tuning when using static focusing parameters on highly imbalanced settings.

Second, Adaptive Focal Loss, where γ is adapted based on class frequencies, outperformed all static versions. It achieved the highest F1-score (0.5) and ROC AUC (0.79), and highest micro recall (0.5282), indicating it was more capable of identifying minority class instances. This was achieved without having to manually adjust γ, which is a testament to the model's robustness and its less dependence on heuristic optimization. This also makes the adaptive method more scalable and deployable in real-world settings, where optimal hyperparameters are unknown a priori. This outcome marks a progression from our earlier work, where we had studied the integration of GloVe embeddings with classical machine learning models [27] and then with deep learning models [28]. While those studies concentrated on the role of word embeddings in representation learning, the present study reverses the focus to the flexibility of the loss function and its adaptability to transformer-based models.

Third, the precision-recall trade-offs for both static and adaptive settings illustrate the merits of dynamic class weighting strategies. While the static γ = 2.5 model attained the highest precision, it did so at the cost of decreased recall and a lower F1-score compared to the adaptive approach. By contrast, Adaptive Focal Loss obtained more balanced performance in general, with significant gains in recall, an important factor in multi-label where missing relevant labels is costly.

VII. Limitations and Future Work

While Adaptive Focal Loss yielded considerable improvements, the overall F1-score (0.50) indicates that there is still room for performance gains. Potential future work may include combining adaptive loss functions with more recent backbone models like RoBERTa or DeBERTa, introducing data-level rebalancing techniques, or experimenting with more γ adaptation techniques. Also, extending this approach to other multi-label domains, such as biomedical text or legal documents, may also serve to further confirm its generalizability.

VIII. Conclusion

This work proposes a novel approach of handling class imbalance in multi-label text classification by combining Adaptive Focal Loss (AFL) and the BERT model. Through the adjustment of the focusing factor γ based on class frequencies, AFL provides a self-adjusting mechanism for assigning greater weights to minority classes and reducing the need for hyperparameter tuning. Our experimental results on the CMU Movie Summary Corpus demonstrate that this adaptive strategy consistently performs better than baseline cross-entropy and statically tuned Focal Loss model on key metrics like F1-score, ROC AUC, and micro recall.

Particularly, the adaptive mechanism achieves improved precision-recall trade-off, significantly improving minority class detection, which is a common problem with real-world multi-label problems. The BERT alignment also allows for more accurate contextual understanding and representation learning, contributing to enhanced performance overall.

These findings demonstrate the potential of adaptive loss functions to improve deep learning models for imbalanced classification tasks. Potential future directions can involve combining AFL with other transformer-based architectures like

RoBERTa or DeBERTa, applying the method to other application areas such as biomedical or legal documents, or coupling it with more advanced data augmentation strategies for further performance and generalization gains.

REFERENCES

[1] U. Jain, P. Mishra, A. Dash, and A. Pandey, "Multi-label multi-class text classification-enhanced attention in transformers with knowledge distillation," *Rev. Digit. Univ.*, vol. 23, no. 1, 2025. [Online]. Available: https://doi.org/10.22201/icat.24486736e.2025.23.1.2484.

[2] Guangxiang Zhao, Wenkai Yang, Xuancheng Ren, Lei Li, Yunfang Wu, and Xu Sun, "Well-classified examples are underestimated in classification with deep neural networks," in AAAI. 2022, pp. 9180–9189, AAAI Press.

[3] Yang Zhou and Wee Sun Lee, "None class ranking loss for document-level relation extraction," in IJCAI. 2022, pp. 4538–4544, ijcai.org.

[4] Y. Yasuda, T. Miyazaki, and J. Goto, "Weighted Asymmetric Loss for Multi-Label Text Classification on Imbalanced Data," *J. Nat. Lang. Process.*, vol. 31, no. 3, pp. 1166–1192, 2024.

[5] Park, W., Park, I., Kim, S., & Ryu, J. (2023). Robust Asymmetric Loss for Multi-Label Long-Tailed Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2711–2720.

[6] Huang, Y., Qi, J., Wang, X., & Lin, Z. (2023). Asymmetric Polynomial Loss for Multi-Label Classification. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.

[7] J. M. Reps, P. Rijnbeek, and A. G. Sena, "The impact of resampling methods on the performance of prediction models in observational health data," *J. Big Data*, vol. 10, no. 1, pp. 1–21, 2023. [Online]. Available: https://doi.org/10.1186/s40537-023-00857-7.

[8] A. D. Putra and M. T. Ibrohim, "Class weighting technique to deal with imbalanced class problem in machine learning: methodological research," unpublished, 2023.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988. https://arxiv.org/abs/1708.02002.

[10] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems, 33, 15744-15754.

[11] W. Zhang, Y. Jiang, Y. Fang, and S. Pan, "Hierarchical Contrastive Learning for Multi-label Text Classification (HCL-MTC)," Sci. Rep., 2025. Available: https://www.nature.com/articles/s41598-025-97597-w.

[12] C. Zhang, L. Dai, C. Liu, et al., "HGBL: A Fine Granular Hierarchical Multi-Label Text Classification Model," *Neural Process. Lett.*, vol. 57, no. 1, 2025. [Online]. Available: https://doi.org/10.1007/s11063-024-11713-x.

[13] Y. Yan, J. Liu, and B.-W. Zhang, "LabelCoRank: Revolutionizing Long Tail Multi-Label Classification with Co-Occurrence Reranking," arXiv preprint, 2025. Available: https://arxiv.org/abs/2503.07968.

[14] S. Lin, F. Frasincar, and J. Klinkhamer, "Hierarchical deep learning for multi-label imbalanced text classification of economic literature," *Appl. Soft Comput.*, vol. 176, p. 113189, 2025. [Online]. Available: https://doi.org/10.1016/j.asoc.2025.113189.

[15] M. A. Islam, M. M. Rahman, and A. Basak, "Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification," Expert Systems with Applications, vol. 168, p. 114146, 2021.

[16] J. Lee, J. Kim, and S. Lee, "An Empirical Study for Class Imbalance in Extreme Multi-label Text Classification," in Proceedings of the 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Feb. 2021, pp. 1-4, doi: 10.1109/BigComp50530.2021.00009.

[17] Jung, J., & Lee, J. (2021). EnvBERT: Multi-Label Text Classification for Imbalanced, Noisy Environmental News Data. IEEE Access, 9, 107785-107794.

[18] M. A. Alomari, M. A. Al-Taee, and A. Al-Taee, "Multilabel Sentiment Prediction by Addressing Imbalanced Class Problem Using Oversampling," in 2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2021, pp. 1-6, doi: 10.1109/JEEIT51787.2021.9429366.

[19] K. Duarte and M. Shah, "PLM: Partial Label Masking for Imbalanced Multi-label Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3057587.

[20] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models," Applied Sciences, vol. 11, no. 2, p. 869, Jan. 2021, doi: 10.3390/app11020869.

[21] M. Park, H. J. Song, and D.-O. Kang, "Imbalanced Classification via Feature Dictionary-Based Minority Oversampling," IEEE Access, vol. 10, pp. 34236–34245, 2022, doi: 10.1109/access.2022.3161510.

[22] X. Cai, S. Huang, Q. Yang, and H. Chen, "UJNLP at SemEval-2025 Task 9: Easy Data Augmentation and Oversampling with Focal Loss for Multi-label Text Classification," arXiv preprint arXiv:2505.00021, 2025. Available: https://arxiv.org/abs/2505.00021.

[23] S. Sen, "Focal Loss Effect on Multi-Label Text Classification Using Convolutional Neural Networks," unpublished manuscript, ResearchGate, 2021. [Online]. Available: https://www.researchgate.net/publication/372109806.

[24] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in Advances in Neural Information Processing Systems, vol. 32, pp. 1565-1576, 2019.

[25] 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. Available: https://arxiv.org/abs/1810.04805.

[26] Nam, J., Kim, J., Gurevych, I., & Mencía, E. L. (2014). Large-scale Multi-label Text Classification — Revisiting Neural Networks. In ECML PKDD. https://link.springer.com/chapter/10.1007/978-3-662-44848-9_5.

[27] Labd, Z., Bahassine, S., Housni, K., Ait Hamou Aadi, F., & Benabbes, K. (2024). Text classification supervised algorithms with term frequency-inverse document frequency and global vectors for word representation: A comparative study. IJECE, 14(1), 589–599.

[28] Labd, Z., Bahassine, S., Housni, K., & Ait Hamou Aadi, F. (2024). Reassessing GloVe Embeddings in Deep Learning: A Comparative Study with Classical ML Approaches. Procedia Computer Science, 251, 740–745.

Attention Aware Dual-Path Autoencoder with Asymmetric Loss for Recognition in Complex Scenes

Hashim Rosli, Rozniza Ali*, Muhamad Suzuri Hitam, Ashanira Mat Deris, Noor Hafhizah Abd Rahim
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu (UMT),
Kuala Nerus, 21030 Terengganu, Malaysia

***Abstract*—Object recognition in complex scenes is challenging due to cluttered backgrounds, overlapping objects, and degraded image quality. Another difficulty arises from sparse label presence, as most images contain only one to three active labels despite the dataset being balanced across 20 object classes. This intra-sample sparsity complicates binary classification by exposing models to a high proportion of inactive classes. This work aims to improve recognition accuracy, robustness under sparse multi-label conditions, and interpretability in visually complex environments. The objective is to help models focus on relevant visual features, suppress background noise, and better distinguish objects that are rare or overlapping. To address these challenges, we introduce an attention aware dual-path autoencoder that enhances image features while learning to classify multiple objects. The model uses asymmetric loss to reduce the influence of easy negatives and emphasize rare or difficult labels. It also integrates an attention mechanism in the reconstruction path to improve object clarity. The proposed model achieves 96.72 percent accuracy, 0.0328 Hamming Loss, 0.9809 macro ROC-AUC, and 0.8925 macro mAP, along with 0.9372 SSIM and 7.1012 dB PSNR in reconstruction. These results confirm its effectiveness for robust classification and enhanced visual understanding in complex scenes.**

Keywords*—Component; autoencoder; attention aware; feature fusion; image enhancement; multi-label classification

I. Introduction

In real-world applications, object recognition systems often must contend with complex scenes which are characterized by varied backgrounds, overlapping objects, and degraded image quality such as noise, blur, and low brightness [1]. These conditions reduce the effectiveness of conventional convolutional neural networks, which often assume clean, well-segmented inputs. Cluttered environments with sensor noise or motion blur, for example, can lower classification accuracy by over 10% in multi-label recognition tasks [2]-[3]. While previous work has explored pre-processing techniques such as denoising and deblurring prior to recognition [4]-[6], these modular approaches introduce computational overhead and suffer from inefficient feature usage, as enhancement and recognition are handled separately. Although the dataset is balanced across 20 object categories, each image includes only 1 to 3 active labels, creating label sparsity during training. This, combined with frequent object overlaps, makes classification more challenging, especially with standard binary loss functions. This work proposes a unified framework that enhances image quality and improves multi-label recognition in complex scenes with sparse, overlapping labels. The model is designed to operate in cluttered environments while learning effectively from limited active labels per image.

We propose an end-to-end architecture that combines image reconstruction and classification into a single unified pipeline. A DenseNet-based convolutional autoencoder, enhanced with a Convolutional Block Attention Module (CBAM), is used to suppress noise and focus on salient object regions during image enhancement [7]. The reconstructed output is processed by a second DenseNet classifier, while latent features from the autoencoder are fused with classification features using multiplicative fusion to integrate both structural and semantic information [8], [9]. To address the challenge of label sparsity, the model is trained using a composite loss function consisting of Mean Squared Error (MSE) for reconstruction and either Binary Cross-Entropy (BCE) or Asymmetric Loss (ASL) for classification. ASL is particularly effective in multi-label settings with few active classes per image, as it down-weights easy negatives and emphasizes difficult positive enhancing discrimination in overlapping or infrequent categories.

The proposed method is evaluated against two baselines, a standard convolutional autoencoder without any transfer learning, CBAM or feature fusion and a standalone DenseNet-121 classifier without image enhancement. Results show that the attention-aware dual-path model trained with ASL outperforms both reconstruction quality and classification accuracy, proving its robustness in complex scene recognition. The rest of the paper is organized as follows. Section II reviews related work. Section III outlines the methodology. Section IV presents the experimental setup. Results and discussions are given in Section V. Section VI discusses the results. Section VII concludes and suggests future directions.

II. Related Work

A. Complex Scene Object Recognition

Object recognition in complex scenes has been widely studied, especially under clutter, occlusion, and image degradation. Cheng et al. reviewed deep learning methods for remote-sensing scene classification, highlighting robust feature extraction in noisy settings [10]. Fu et al. demonstrated that classification accuracy in astronomical imaging can be improved using denoising autoencoders to recover from noise and blur [11]. In medical imaging, autoencoder-based architectures have also been employed to extract salient features from visually noisy data [12]. Lu et al. introduced a co-attention Siamese network for video segmentation, utilizing

co-attention mechanisms to better segment overlapping objects in complex backgrounds [13].

In more domain-specific research which using kitchen utensils dataset, Yusro et al. compared Faster R-CNN and YOLOv5 on kitchen datasets with overlapping utensils, reporting YOLOv5's superior accuracy of 89.12% compared to 83.92% [14]. Building on this, Hashim applied YOLOv5 with different deep learning backbones to manage image degradation, achieving strong performance under noisy conditions [15]. Gallego et al. developed the Kurcuma dataset, comprising over 6,800 annotated images ranging from isolated objects to cluttered domestic scenes, designed for evaluating domain-adaptive recognition methods [16].

B. Multitask Image Enhancement with Object Recognition

Integrating image enhancement and classification in unified architecture has improved recognition in challenging visual environments. Hami and JameBozorg showed that using a denoising autoencoder as preprocessing significantly increased accuracy when paired with VGG16 and InceptionV3 for defect detection tasks [17]. Liu et al. proposed a DenseNet-based denoising self-encoder for image dehazing, which enhanced both image clarity and object detection in outdoor scenes [18]. In medical applications, CNN autoencoders have been paired with classification heads for tasks such as brain tumor diagnosis, achieving strong results on degraded images [19]. Rahimzadeh and Attar combined DenseNet and ResNet features via a shared latent space, improving classification performance on COVID-19 chest X-ray datasets [9].

Attention-aware enhancement mechanisms have also gained increasing traction in improving object recognition tasks across various domains. Fu [11] demonstrated that incorporating a denoising autoencoder significantly enhanced classification accuracy on noisy galaxy imagery. Praharsha and Poulose [20] introduced CBAM-VGG16 for distracted driver detection, achieving over 98% accuracy on the AUCD2 dataset. In remote sensing, Hu et al. [21] developed a CBAM-integrated hybrid network combining residual and dilated convolutions, which improved hyperspectral image classification in spectrally cluttered conditions.

C. Transfer Learning Based Autoencoders

Transfer learning has been widely adopted in autoencoder-based systems to improve generalization and representation learning across domains. Fu [11] utilized a denoising convolutional autoencoder built on DenseNet-121 to enhance galaxy images, resulting in notable classification improvements. In histopathology, Lee and Lah [22] fine-tuned a DenseNet encoder within an autoencoder-classifier pipeline for multi-task slide analysis, demonstrating increased learning efficiency. Praharsha and Poulose [20] embedded CBAM into VGG16 to support attention-based classification in distracted driving scenarios. Tran et al. [23] evaluated multiple encoder backbones, including ResNet-50, EfficientNet-B3, and DenseNet-121, for lung cancer detection using DICOM images.

Xie et al. [24] introduced MEEAFusion, an architecture that integrates CBAM-like attention and multi-scale fusion within an autoencoder framework for infrared-visible image fusion. DenseNet-121 has been commonly employed as a preferred encoder across these efforts due to its densely connected layers and feature reuse capabilities, proving effective in domains such as astronomy [11] medical imaging [22], [24], and spectral fusion tasks [23].

D. Multi-label Classification and Label Sparcity Handling

Multi-label classification assigns multiple class labels to each image, with each label represented as a binary indicator of presence or absence. In this context, standard loss functions such as Binary Cross-Entropy (BCE) are commonly used, but their performance has been analyzed for scenarios involving label sparsity. For example, Wang et al. [25] discussed how BCE treats active and inactive labels equally, which may lead to suboptimal learning when most labels are negative. Yessou et al. [26] further evaluated several loss functions, identifying similar limitations when applied to satellite imagery datasets.

To address class imbalance, various alternative loss functions have been proposed. Asymmetric Loss (ASL), introduced by Ben-Baruch et al. [27], adjusts gradient contributions by down-weighting easy negatives and emphasizing hard positives, leading to improved mAP scores across datasets such as MS-COCO, Pascal-VOC, NUS-WIDE, and Open Images. Huang et al. [28] later refined this with Asymmetric Polynomial Loss (APL), incorporating polynomial coefficients for further control over gradient scaling. Ji et al. [29] applied these loss functions in remote sensing, while similar techniques have also been adopted in medical imaging to enhance detection of rare labels and performance on sparsely annotated data.

E. Gaps and Limitation

Although prior studies have advanced object recognition, image enhancement, and multi-label classification, several limitations remain. Few existing models integrate enhancement and recognition into a unified architecture, especially for visually complex scenes with occlusion, noise, and clutter. While encoder-decoder structures and attention mechanisms have been explored independently, their combined use in multitask pipelines is limited. Additionally, many approaches are domain-specific and lack generalizability across different real-world environments. Most models focus on either low-level enhancement or high-level classification, missing the performance gains of joint optimization.

Another limitation is the challenge of label sparsity in multi-label classification. Standard loss functions like Binary Cross-Entropy often struggle when most classes are inactive, making it harder to learn from rare but important labels. Although advanced losses such as Asymmetric Loss (ASL) and Asymmetric Polynomial Loss (APL) have shown success in other domains, they are still rarely applied in complex scene recognition. Additionally, while DenseNet-121 and attention modules like CBAM perform well individually, their integration into autoencoder-based multitask frameworks remains limited. Addressing these gaps requires a unified, attention-aware model that applies transfer learning and specialized loss functions for more robust performance in cluttered, degraded environments.

III. ARCHITECTURE

The proposed architecture in Fig. 1 employs a pre-trained DenseNet121 as a feature extraction backbone. This backbone serves dual purposes: (1) encoding the noisy input into a latent representation for clean image reconstruction, and (2) extracting discriminative features for classification. We extract intermediate feature maps from four major transition layers of the DenseNet, capturing hierarchical spatial information. To preserve spatial resolution and semantic depth, we remove the classification head of the backbone and use the feature pyramid as input to the decoder. All batch normalization layers remain trainable to adapt to the domain distribution, while the initial training phase freezes the rest of the backbone to stabilize learning.

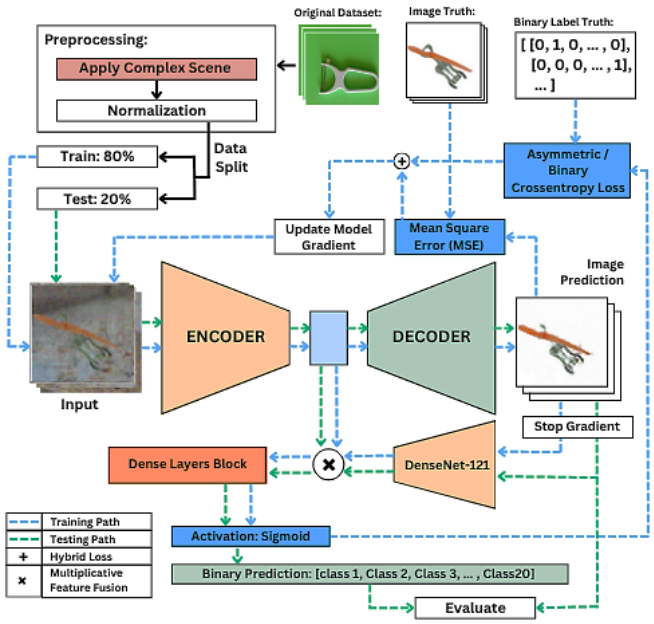


Fig. 1. Overall model pipeline showing the training objectives, autoencoder structure, classification paths, and gradient flow during training and inference.

A. Encoder Backbone

Let the input image be denoted as $\boldsymbol{x} \in \mathbb{R}^{64\times64\times3}$, representing a three-channel red, green and blue (RGB) image. We adopted a DenseNet-121 encoder ($\boldsymbol{E}$), pretrained on ImageNet and used here without its classification head to extract hierarchical representations. DenseNet-121 is well-suited for feature extraction because its layers are densely connected, allowing each layer to reuse feature maps from all previous layers in the same block, which improves feature propagation and reduces vanishing gradients [30]. As $\boldsymbol{x}$ passes through successive dense blocks and transition layers, we collect feature maps from five key stages, denoted as (1), where each $\boldsymbol{f_i}$ is the output of the i-th dense block within DenseNet-121. These feature maps are progressively deeper and semantically richer, capturing spatial context at multiple scales. The final representation $\boldsymbol{f_5}$ serves as the encoder's bottleneck output. To generate a compact semantic descriptor, a global average pooling operation (GAP) is applied to $\boldsymbol{f_5}$ in Eq. (2).

$$\boldsymbol{f_i} = \boldsymbol{E_i}(\boldsymbol{f_{i-1}}), \quad \boldsymbol{f_0} = \boldsymbol{x}, \quad for\ \boldsymbol{i} = 1,2,3,4,5 \tag{1}$$

$$\boldsymbol{z} = GAP(\boldsymbol{f_5}) \in \mathbb{R}^{1024} \tag{2}$$

The final output of the encoder produces a 1024-dimensional vector $\boldsymbol{z}$ that summarizes the spatial content of each channel. This latent representation is used to drive the reconstruction in the decoder, and, in parallel, informs downstream classification tasks. During the initial training phase, the encoder weights are frozen to retain robust pretrained features, allowing the decoder and classifier to adapt without perturbing the foundational representations.

B. Attention Aware Decoder

The decoder's primary objective is to reconstruct a clean version of the input image, denoted as $\hat{\boldsymbol{x}} \in \mathbb{R}^{64\times64\times3}$, from the latent vector ($\boldsymbol{z}$). This step is essential for removing noise and enhancing image quality before classification. Structurally, the decoder is paired with a DenseNet-121 encoder, where skip connections, $\boldsymbol{f_i}$ are extracted by encoder to paired with decoder blocks. These skip connections significantly improve their ability to preserve fine spatial details and reduce reconstruction errors. Notably, Ham et al. [31] theoretically and empirically analyzed linear denoising autoencoders with skip connections, showing that bypass pathways stabilize reconstruction performance. To further refine reconstruction, the decoder integrates the Convolutional Block Attention Module (CBAM) [7] which enhances feature selection through two types of attention, channel and spatial located after the last decoder block in Fig. 2.

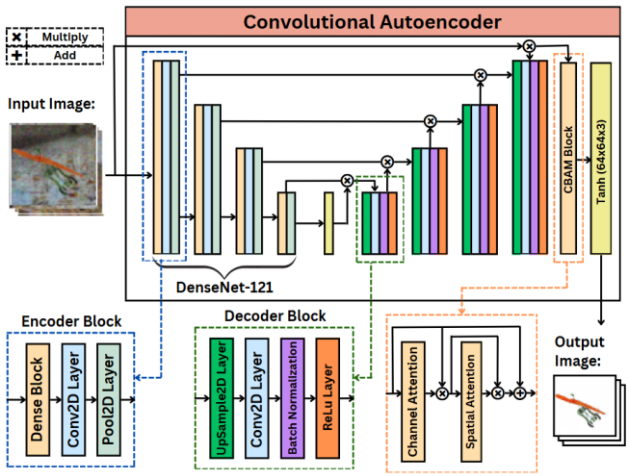


Fig. 2. Detailed architecture of the DenseNet-based encoder used in the autoencoder. Features from various depths are extracted and passed to the decoder through skip connections.

Channel attention learns a map $\boldsymbol{M_c} \in \mathbb{R}^{1\times1\times C}$ to assign importance weights to each feature channel, emphasizing those that contribute most to accurate reconstruction. Spatial attention computes a map $\boldsymbol{M_s} \in \mathbb{R}^{H\times W\times1}$ to focus on the most relevant spatial regions of the image, such as edges and textures. These attention mechanisms are applied to the intermediate feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$, producing a refined feature map $\boldsymbol{F_{att}}$ through element-wise multiplication calculated in Eq. (3). To preserve the original feature representation while enriching it with attention aware

enhancements, a residual connection is applied between the input feature map $\boldsymbol{F}$ and the refined output $\boldsymbol{F_{att}}$, yielding the final attended feature map $\boldsymbol{F_{out}}$ in Eq. (4). This combined representation is then passed through the final convolutional layer with a *tanh* activation to produce the decoded image $\hat{\boldsymbol{x}}$. Finally, the output image $\hat{\boldsymbol{x}}$ is generated using a convolutional layer followed by a *tanh* activation, ensuring that pixel values fall between –1 and 1, and match the original input's shape and color space for loss computation in Eq. (5).

$$\boldsymbol{F_{att}} = \boldsymbol{F} \odot \boldsymbol{M_c} \odot \boldsymbol{M_s} \tag{3}$$

$$\boldsymbol{F_{out}} = \boldsymbol{F_{att}} + \boldsymbol{F} \tag{4}$$

$$\hat{\boldsymbol{x}} = tanh(Conv(\boldsymbol{F_{out}})) \tag{5}$$

C. Classification from Reconstructed Inputs

The classification module uses the reconstructed image $\hat{\boldsymbol{x}}$, instead of the original noisy input, to benefit from improved visual quality during prediction. To prevent classification gradients from affecting the decoder, a stop-gradient operation is applied, treating $\hat{\boldsymbol{x}}$ as a fixed input during backpropagation and isolating the reconstruction learning. This detached image, $\hat{\boldsymbol{x}}_{\boldsymbol{stop_gradient}}$ is then passed through a new DenseNet-121 as decoded image's feature extractor, identical in architecture but with independent weights to extract a classification feature vector, $\boldsymbol{z_c}$ as described in Eq. (6). To enhance robustness, feature fusion is performed by element-wise multiplication of the latent vectors, $\boldsymbol{z}$ from the autoencoder and decoded image's extracted feature, $\boldsymbol{z_c}$ as shown in Eq. (7). This operation amplifies shared information while suppressing noise and inconsistencies.

$$\boldsymbol{z_c} = GAP(\boldsymbol{E}(\hat{\boldsymbol{x}}_{\boldsymbol{stop_gradient}})), \qquad \boldsymbol{z_c} \in \mathbb{R}^{1024} \tag{6}$$

$$\boldsymbol{z_{fused}} = \boldsymbol{z} \odot \boldsymbol{z_c} \tag{7}$$

The fused feature vector $\boldsymbol{z_{fused}}$ is then fed into fully connected layers with nonlinear activations and dropout regularization to generate the final multi-label classification prediction at Eq. (8) where W and B are learnable parameters of the classifier, and σ is the sigmoid activation function applied elementwise, mapping the outputs to probabilities between 0 and 1 for each class. The output $\hat{y} \in [0,1]^C$ represents predicted probabilities across C classes.

$$\hat{y} = \sigma(W^T \boldsymbol{z_{fused}} + b) \tag{8}$$

D. Multitask Hybrid Loss

The proposed model is optimized through a multitask learning framework that jointly addresses image reconstruction and multi-label classification. A hybrid loss function is formulated to guide the network in learning both low-level structural details and high-level discriminative features for accurate classification. This is achieved by minimizing reconstruction and classification losses simultaneously, each weighted within the final objective. As a result, the network preserves semantic content while improving its ability to differentiate object categories. This integration of learning signals helps the model build richer, more generalizable feature representations essential for robust recognition in complex visual scenes.

$$\boldsymbol{\mathcal{L}_{total}} = (\lambda_{recon} \cdot \boldsymbol{\mathcal{L}_{recon}}) + (\lambda_{class} \cdot \boldsymbol{\mathcal{L}_{class}}) \tag{9}$$

In Eq. (9) formulation, $\mathcal{L}_{total}$ denotes the complete training loss used to optimize the network parameters. The term $\mathcal{L}_{recon}$ corresponds to the image reconstruction loss, which guides the autoencoder decoder to regenerate the clean image from the latent representation. The term $\mathcal{L}_{class}$ represents the classification loss, used to supervise the semantic categorization based on the reconstructed features. The scalar coefficients λ_{recon} and λ_{class} are hyperparameters used to weigh the influence of each loss. In our experiments, we set λ_{recon} = 1.5 and λ_{class} = 0.5, which prioritizes structural reconstruction while maintaining a significant emphasis on classification.

To quantify the reconstruction performance, we employ the Mean Squared Error (MSE) loss [32], a standard measure for evaluating image similarity in pixel space. This loss is defined as Eq. (10). Here, x_i denotes the i-th original input image in a batch, while $\hat{x}_i$ is its corresponding reconstructed output produced by the decoder. Both are assumed to be RGB images with dimensions 64×64×3. The notation $\| \cdot \|^2_2$ refers to the squared L2 norm computed over all pixel values, which penalizes deviations between each input pixel and its reconstruction. The term N represents the batch size. This formulation encourages the decoder to generate outputs that are visually and numerically close to the original clean image inputs.

$$\boldsymbol{\mathcal{L}_{recon}}(x, \hat{x}) = (1/N) \sum \|x_i - \hat{x}_i\|^2_2 \tag{10}$$

For the classification task, we evaluate two loss functions, the Asymmetric Loss (ASL) by Ridnik et al. [27] and the conventional Binary Cross-Entropy (BCE). The ASL is designed for multi-label classification with severe label imbalance, where it emphasizes hard negatives while suppressing easy ones through focusing parameters and a threshold margin. It is defined as in (11). Here, $y \in \{0,1\}^C$ and $\hat{y} \in [0,1]^C$ denote the ground truth and predicted probability vectors across C classes. γ^+ and γ^- are focusing factors for positives and negatives, m is the margin that suppresses low-confidence negatives, and ε ensures numerical stability. We use $\gamma^+ = 0$, $\gamma^- = 2$, m = 0.01, and ε = 1e−8 as recommended in [27]. In contrast, the BCE loss [32], given by Eq. (12) treats each class independently and penalizes incorrect predictions equally, regardless of imbalance. We compare both losses to evaluate their influence on classification performance within our multitask framework.

$$\boldsymbol{\mathcal{L}_{class_ASL}}(y, \hat{y}) = -[y \cdot \log(\hat{y}) \cdot (1 - \hat{y})^{\gamma^+} + (1 - y) \cdot \log(1 - \hat{y} + \varepsilon) \cdot (\hat{y} - m)^{\gamma^-}] \tag{11}$$

$$\boldsymbol{\mathcal{L}_{class_BCE}}(y, \hat{y}) = -\sum [y_k \log(\hat{y}_k) + (1 - y_k) \log(1 - \hat{y}_k)] \tag{12}$$

By combining the two loss components in Eq. (9), the model learns to perform both tasks within a unified process. A key implementation detail is the use of a stop-gradient

operation between the reconstruction and classification branches. This ensures the classification loss does not backpropagate through the decoder, keeping its training focused solely on reconstruction. Meanwhile, the encoder is shared across both branches, enabling learning of joint features useful for both semantic discrimination and visual reconstruction. This separation of gradient flow, along with shared encoder supervision, promotes stable, task-specific optimization and allows both branches to converge effectively while reinforcing the shared representations.

IV. Experimental Setup

A. Dataset and Simulation of Complex Scenes

The dataset used in this work originates from the Edinburgh Kitchen Utensil Database (EKUD), which consists of isolated images of kitchen objects with plain backgrounds. The original EKUD dataset, however, presents two major limitations which are class imbalance, where some classes have significantly fewer instances than others, and lack of scene complexity, with only single-object images and minimal background variation. To address these limitations and generate a more realistic, complex training dataset, we developed a Complex Scene Generator pipeline in Fig. 3. First, all object images from EKUD were preprocessed by removing their backgrounds and converting them to transparent PNGs. This enabled spatial transformations and compositing onto new backgrounds.

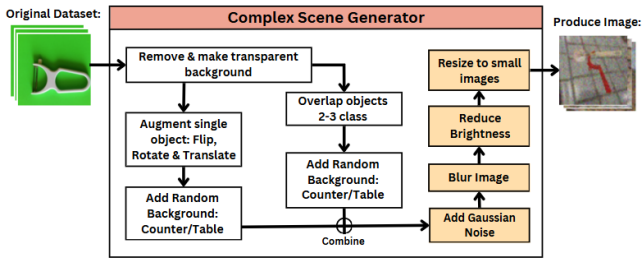


Fig. 3. Complex scene generator.

Each object was augmented using geometric transformations such as flipping, rotation, and translation to increase diversity. To simulate realistic environments, objects were overlaid on random background images of counters or tabletops, mimicking domestic kitchen scenes. To add complexity, 2–3 objects from different classes were randomly combined with spatial overlap, creating occlusions typical of real-world clutter. These composites then underwent image degradation, including resizing to 64×64 pixels, brightness reduction, Gaussian blurring, and added noise to mimic conditions like low light or motion blur. This process produced a balanced dataset of 6,000 images with equal class distribution. Each image contained one to three overlapping objects from different classes, introducing intra-class variation and inter-class spatial relationships vital for robust classification and reconstruction. Finally, the dataset was split into training and testing sets using an 80:20 ratio with stratified class distribution.

B. Model Training

The proposed dual-branch architecture was optimized using a two-phase training strategy to enhance image reconstruction and multi-class classification. To improve evaluation reliability and reduce overfitting, 5-fold StratifiedKFold cross-validation was applied to the training set, generating five training runs with preserved class distribution. In the first phase, the DenseNet encoder was frozen, allowing the decoder and classifier to learn task-specific features without altering backbone representations. A composite loss combined Mean Squared Error (MSE) for reconstruction and Sparse Categorical Crossentropy for classification, weighted 1.5 and 0.5, respectively. The model trained for 30 epochs using the Adam optimizer with a 0.001 learning rate. This phase aimed to stabilize reconstruction and extract semantically rich latent features. Structural Similarity Index (SSIM) assessed reconstruction quality, while standard metrics tracked classification accuracy.

In the second phase, the full model was unfrozen for end-to-end fine-tuning to enable joint optimization of low-level and high-level features across both branches. The learning rate was reduced to 0.0001 to support stable convergence and prevent disruption of previously learned features. This phase refined the encoder-decoder synergy and strengthened the alignment between reconstructed structures and class-relevant semantics. Following the completion of training with Asymmetric Loss, a parallel experiment was conducted using Binary Cross Entropy as the classification loss, while retaining the same optimizer settings, loss weights, and training schedule. This allowed a direct performance comparison under controlled conditions, highlighting the impact of loss design on multi-label recognition in cluttered scenes.

Pseudocode 1: Dual-Phase Training

Input: Noisy images I_{noisy}, (Clean Target Image I_{clean}, Clean labels L_{true})

Output: Trained model with decoder output D and classifier C

1: Initialize encoder with DenseNet121 (frozen), decoder blocks, and classification branch

2: Compile model with:

- Losses: MSE, ASL
- Loss weights: D: 1.5, C: 0.5}
- Optimizer: Adam (lr = 0.001)

3: Train model for N_1 (30) epochs, freezing encoder

4: Save best weights based on highest validation accuracy

5: Unfreeze all layers of model

6: Compile model with:

- Updated optimizer: Adam (lr = 0.0001)

7: For epoch = N_1 to N_2(100) do

8: For each batch (I_{noisy}, L_{true}) in training set do

9: Compute decoder output D_{pred} and classifier output C_{pred}

10: Compute:

- MSE loss between D_{pred} and I_{clean}
- ASL loss between C_{pred} and L_{true}

- SSIM metric for D_{pred}

Backpropagate total loss = 1.5 × MSE + 0.5 × ASL

Update weights

Evaluate on validation set

Save model if validation accuracy improves

End for

The dual-phase training strategy in Pseudocode 1 reflects a common practice in transfer learning and multi-task deep learning. Freezing the pretrained encoder early allows the decoder and classifier to learn task-specific features without modifying core representations. Once stabilized, full fine-tuning with a reduced learning rate improves alignment between reconstructed outputs and classification. Deepankan and Agarwal found that freezing the backbone initially and unfreezing later boosted accuracy and reduced overfitting on limited data [33]. Likewise, Taha et al. employed staged fine-tuning with pretrained language models, showing that freezing phases improved learning stability and generalization in imbalanced tasks [34]. In remote sensing, Khotimah et al. used a dual-stage framework combining masked autoencoder pretraining and few-shot learning to enhance hyperspectral image classification with limited supervision [35]. These studies support phased training as a reliable strategy for stabilizing early optimization and enhancing global feature refinement.

C. Evaluation Metrics

To assess model performance, five-fold Stratified K-Fold cross-validation was applied to the training data to ensure balanced class representation and generalizability. All metrics are reported as averages across the five folds. Since the model performs both tasks jointly, each output was evaluated using suitable criteria. For reconstruction, evaluation focused on perceptual similarity and signal fidelity between input and reconstructed images. Structural Similarity Index Measure (SSIM) was used to compare luminance, contrast, and structural details, as defined in Eq. (13), where μ_x and μ_y are the means, σ_x and σ_y their standard deviations, σ_{xy} the cross-covariance, and C_1, C_2 are constants for stability. Additionally, Peak Signal-to-Noise Ratio (PSNR) was used to measure reconstruction quality based on the ratio of maximum signal to noise power. As shown in Eq. (14), PSNR uses MAX_I as the maximum pixel value and MSE as the mean squared error. Combined, these metrics evaluate perceptual quality and pixel-level accuracy.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (13)$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (14)$$

For the multi-label classification component, an adaptive top-T_1 binarization scheme was applied. Each output prediction vector was binarized by retaining the top-T_1 elements with the highest predicted probabilities, where T_1 corresponds to the number of positive ground truth labels in the input, as shown in Eq. (15). This approach ensured prediction-ground truth cardinality alignment, enabling fair comparisons across samples with varying label sparsity. After binarization, classical metrics such as precision, recall, and F1-score were computed using Eq. (16), Eq. (17), and Eq. (18), respectively. These metrics were evaluated under both macro and micro averaging. Macro-averaging calculates per-class metrics and averages them, while micro-averaging aggregates all class contributions before computing the metric, capturing overall performance. To assess label-wise correctness, we also reported standard accuracy Eq. (19) and Hamming Loss (20), which quantify the fraction of misclassified labels.

$$T_1 = \| y_{true} \|_0 \quad (15)$$

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (18)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

$$Hamming\ Loss = \frac{1}{N \cdot C}\sum_{i=1}^{N}\sum_{j=1}^{C} 1[y_{ij} \neq \hat{y}_{ij}] \quad (20)$$

To assess ranking quality, which is important in multi-label tasks with overlapping label distributions, we computed the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and Mean Average Precision (mAP) in micro settings. These metrics, based on raw prediction scores without thresholding, evaluate the model's ability to rank relevant labels above irrelevant ones, following best practices in recent multi-label research [36]. Macro metrics average performance across classes, while micro metrics aggregate true and false positives globally. A multi-label confusion matrix was generated using the MLCM (Multi-Label Confusion Matrix) library [37], enabling per-class analysis of co-occurrence between true and predicted labels. The matrix was visualized as a heatmap to expose common misclassifications and confusion patterns. MLCM offers sparse-aware, normalized outputs well suited for complex multi-label data, following protocols from [38]. Finally, a classification report summarized per-class precision, recall, and F1-scores for detailed performance interpretation.

V. Results and Discussion

A. Multi-Labeled Classification

The classification performance of the proposed model in Table I was compared with two baselines: a standard convolutional autoencoder and a standalone DenseNet-121 classifier. All models used binary cross-entropy loss, while the proposed model was also evaluated with asymmetric loss to test sensitivity to label imbalance. Metrics included accuracy, Hamming Loss, macro-averaged ROC-AUC, and macro mean average precision (mAP). The autoencoder achieved 88.99% accuracy, 0.1101 Hamming Loss, 0.8803 ROC-AUC, and 0.5278 mAP. The standalone DenseNet-121 reached 95.37%

accuracy, 0.0463 Hamming Loss, 0.9678 ROC-AUC, and 0.8366 mAP. The proposed model, combining reconstruction and classification, reached 96.41% accuracy, 0.0359 Hamming Loss, 0.9805 ROC-AUC, and 0.8862 mAP.

The highest performance was achieved when the proposed model was trained using the asymmetric loss function. Accuracy increased to 96.72%, and Hamming Loss decreased to 0.0328. The macro ROC-AUC reached 0.9809, while the macro mAP peaked at 0.8925. To further evaluate classification performance, Table II reports macro and micro averages of precision, recall, and F1-score. The standard convolutional autoencoder yielded a macro F1-score of 0.4296 and a micro F1-score of 0.4496. The DenseNet-121 baseline achieved a macro F1-score of 0.7637 and a micro F1-score of 0.7683. The proposed model with binary cross-entropy achieved a macro F1-score of 0.8196. When trained with asymmetric loss, the model attained a macro F1-score of 0.8350 and a micro F1-score of 0. 8362.

TABLE I. EVALUATION OF CLASSIFICATION RESULT

Model	**Classification Loss Function**	**Evaluation**			
		Accuracy	***Hamming Loss***	***Macro ROC-AUC***	***Macro mAP***
Standard Convolutional Autoencoder (CAE)	BCE	88.99%	0.1101	0.8803	0.5278
DenseNet-121 (Alone)	BCE	95.37%	0.0463	0.9678	0.8366
Proposed Model	BCE	96.41%	0.0359	0.9805	0.8862
	ASL	96.72%	0.0328	0.9811	0.8925

TABLE II. PERFORMANCE MATRICS OF CLASSIFICATION RESULT

Model		**Performance Matrics**					
		Macro Precision	***Macro Recall***	***Macro F1***	***Micro Precision***	***Micro Recall***	***Micro F1***
Standard CAE	BCE	0.5658	0.4496	0.4296	0.4496	0.4496	0.4496
DenseNet -121 (Alone)	BCE	0.7798	0.7683	0.7637	0.7683	0.7683	0.7683
Proposed Model	BCE	0.8311	0.8204	0.8196	0.8204	0.8204	0.8204
	ASL	0.8435	0.8362	0.8350	0.8362	0.8362	0.8362

The evaluation of classification performance using Asymmetric Loss (ASL) and Binary Cross-Entropy (BCE) was conducted with multi-label ROC-AUC metrics. Although the dataset was globally balanced across 20 categories, many binary labels had just 1 to 3 positives, causing localized imbalance. The micro-average AUC under ASL reached 0.9811, slightly higher than 0.9802 with BCE. ASL produced higher per-class AUCs in 12 of 19 overlapping categories. For example, Potato Peeler, Serving Spoon, and Wooden Spoon had stronger AUCs under ASL. The ROC curves shown in Fig. 4 and Fig. 5 further illustrate classifier performance across label distributions. Fig. 4, derived from the ASL-trained model, exhibits tighter and more consistent ROC curves, especially for low-prevalence labels. In contrast, Fig. 5, based on BCE training, presents slightly flatter and more variable ROC curves for low-activity classes such as Bread Knife (0.9720 ASL vs. 0.9786 BCE) and Dinner Fork (0.9714 ASL vs. 0.9768 BCE). BCE outperformed ASL in a few high-density labels like Bottle Opener.

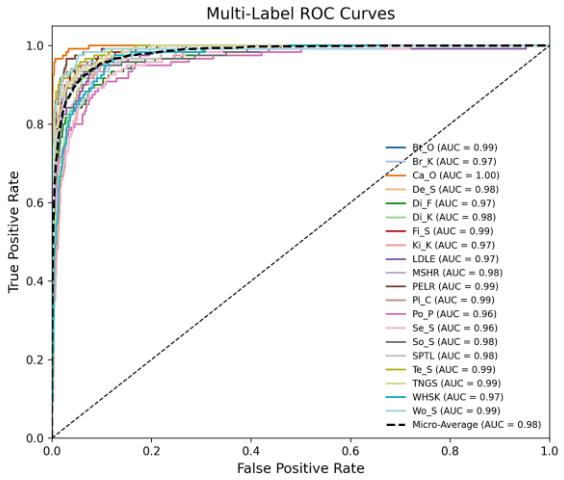


Fig. 4. ROC-AUC Graph of proposed model that uses asymmetric loss function.

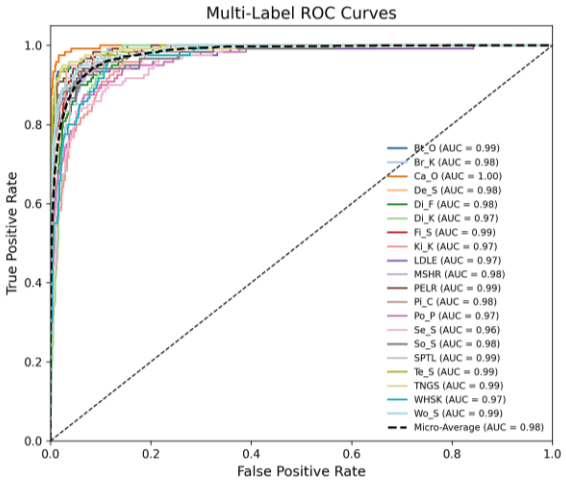


Fig. 5. ROC-AUC Graph of proposed model that uses binary crossentropy loss function.

B. Enhanced Image Reconstruction

The reconstruction performance was evaluated using SSIM and PSNR, as detailed in Table III. Input images contained complex scenes with diverse backgrounds, object occlusion, and visual clutter. The decoder consistently generated clean object representations over simplified backgrounds, as also illustrated in Table IV. The baseline autoencoder, trained using mean squared error and binary cross-entropy loss, achieved a reconstruction score of 0.9175 SSIM and 6.0670 dB PSNR. The proposed dual-path model improved these scores to 0.9213 SSIM and 7.0028 dB PSNR. When asymmetric loss was applied during classification, the model achieved 0.9372 SSIM and 7.1012 dB PSNR. The model's ability to reconstruct clean outputs from noisy, cluttered scenes was enhanced by

integrating the Convolutional Block Attention Module (CBAM) into the decoder. DenseNet-121 was excluded, as it lacks a decoding pathway.

TABLE III. IMAGE RECONSTRUCTION PERFORMANCE EVALUATION

Model	Loss Function	Evaluation	
		SSIM	***PSNR (dB)***
Standard Convolutional Autoencoder (CAE)	MSE+BCE	0.9175	6.0670
Proposed Model	MSE+BCE	0.9213	7.0028
	MSE+ASL	0.9372	7.1012

TABLE IV. IMAGE RECONSTRUCTION BASED ON LOSS FUNCTIONS

Input	MSE+BCE	MSE+ASL	Standard CAE
			

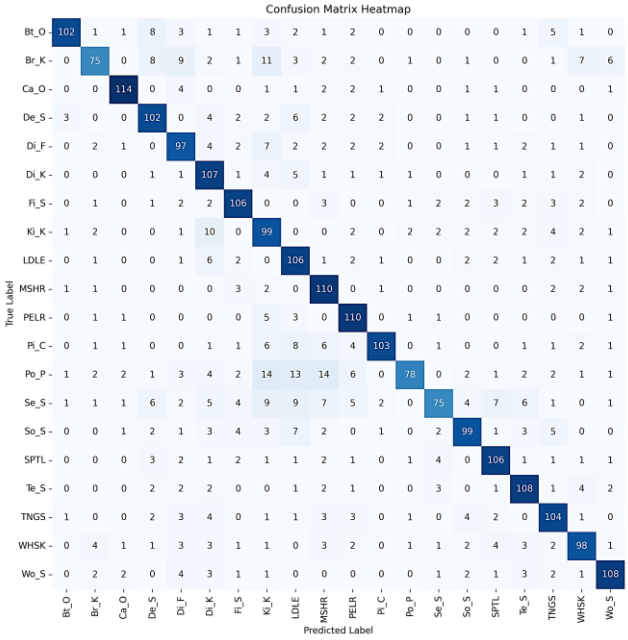


Fig. 6. Confusion matrix from proposed model with loss function MSE+ASL.

The confusion matrix in Fig. 6 shows the best classification results of the proposed model using Mean Squared Error (MSE) for reconstruction and Asymmetric Loss (ASL) for classification. Most object classes exhibit strong diagonal dominance, indicating high true positive rates, especially for Can Opener, Masher, Peeler, and Wooden Spoon, each surpassing 100 correct predictions. These results demonstrate the model's robustness against complex scene conditions such as overlapping objects, image degradation, and varied backgrounds. Some confusion remains between visually or semantically similar classes, like Bread Knife misclassified as Can Opener or Peeler, and Serving Spoon overlapping with Soup Spoon, underscoring ongoing challenges with intra-class similarity in noisy settings. Overall, the matrix supports the effectiveness of the multi-task architecture, where combining image enhancement and classification via latent fusion improves discriminative performance.

VI. DISCUSSIONS

A. Interpretation of Multi-Label Classification Results

The multi-label classification results highlight the clear advantage of integrating reconstruction and classification within a unified model, particularly when dealing with complex data distributions and label sparsity. While the standard convolutional autoencoder lacked explicit label supervision and performed poorly in classification, the standalone DenseNet-121 improved predictive quality through deeper, more expressive features. However, DenseNet-121 operated in a single-task setup and did not benefit from auxiliary learning signals provided by image reconstruction. In contrast, the proposed dual-path architecture, which combines visual reconstruction with classification, consistently outperformed both baselines. This outcome suggests that reconstruction plays a crucial role in shaping more robust latent representations, which in turn support better discrimination, especially for less frequent object classes.

An equally important factor in the model's success is the choice of loss function. The shift from Binary Cross-Entropy to Asymmetric Loss introduced a more nuanced training dynamic, where the model learned to emphasize informative, minority class signals while reducing overfitting to dominant negatives. This led to higher macro-averaged metrics and stronger per-class AUCs for rare categories. ROC curve comparisons further confirmed this behavior, showing tighter, more stable curves under ASL for low-prevalence labels. Meanwhile, BCE slightly favored common classes due to its uniform treatment of all errors. Taken together, these observations demonstrate that the model's strength arises not solely from architectural complexity but from a careful alignment between task formulation, loss function design, and the inherent demands of multi-label learning in imbalanced and cluttered visual contexts.

B. Interpretation of Reconstruction Performance

The reconstruction results demonstrate that the proposed dual-path architecture is not only effective for classification but also capable of producing structurally coherent and visually clean reconstructions in complex scenes. Compared to the baseline autoencoder, the improved SSIM and PSNR scores reflect the model's strong capacity to abstract essential features while filtering out background noise and irrelevant details. The use of a shared encoder that supports both tasks allows the network to capture richer semantic representations, which benefit the decoder's ability to reconstruct meaningful content even in visually degraded or cluttered environments. This integration shows how multi-task learning encourages a more organized and informative latent space that serves both enhancement and recognition.

An unexpected but significant observation was the positive influence of asymmetric loss, applied only to the classification task, on the quality of reconstructed outputs. This implies that learning to better separate hard-to-classify examples also leads

to more structured and semantically aligned feature embeddings, indirectly supporting the reconstruction process. Additionally, the integration of CBAM into the decoder reinforced the model's ability to selectively attend to important spatial and channel-specific regions. As a result, reconstructions were not only cleaner but also better aligned with the salient components of the original input. This attention-driven refinement is particularly beneficial in real-world scenarios where accurate reconstruction supports interpretability and aids in downstream decision-making tasks.

C. Bridging Gaps Through Multi-Task Representation

The proposed model directly addresses several unresolved challenges in existing research by unifying classification and reconstruction within a shared encoder-decoder architecture. Most prior approaches treat enhancement and recognition as separate tasks, limiting their ability to generalize in cluttered or degraded scenes. In contrast, the proposed model leverages joint optimization to build more expressive and spatially consistent latent features. The reconstruction task helps maintain structure and denoise complex scenes, while the classification branch, guided by asymmetric loss, highlights category-specific regions, especially for rare labels. This dual influence supports a more semantically aware encoder that improves both predictive accuracy and visual clarity in reconstructed images.

By integrating DenseNet-121 and attention modules like CBAM into a multi-task framework, the model closes a critical gap identified in previous studies where such components were rarely combined. Through shared learning, reconstruction helps the model preserve spatial structure and object layout, while classification encourages it to focus on distinctive features that separate one object class from another. Consistent performance gains across low-prevalence classes and complex visual backgrounds indicate that the model benefits from this interdependence. Rather than relying on handcrafted decoupling or task-specific heuristics, this approach uses end-to-end learning to harness the complementary strengths of each task. As a result, the model becomes more robust, interpretable, and suitable for real-world deployment where visual data is often noisy, sparse, and heterogeneous.

VII. Conclusion and Future Work

In this study, we proposed a multi-task deep learning architecture that integrates a convolutional autoencoder and DenseNet backbone with latent space fusion to tackle object classification challenges in complex scenes. These include cluttered backgrounds, overlapping objects, and degraded image quality due to noise, blur, or low brightness. The architecture was designed to jointly enhance visual clarity and extract discriminative features within a unified pipeline. Trained in two phases and evaluated using diverse reconstruction and classification metrics, the model outperformed baseline methods. The proposed architecture is particularly designed for image data containing overlapping objects and visual noise, where conventional classification models often underperform. Results showed notable improvements in accuracy, macro and micro F1-scores, and reduced class confusion, validating the benefit of integrated image enhancement and feature learning for robust multi-class object recognition in complex scenes.

Despite improvements, some limitations remain. The model still exhibits misclassifications in scenes with visually similar or overlapping objects, for example Dessert Spoon versus Dinner Spoon, indicating difficulty in distinguishing fine-grained object boundaries under occlusion. This reflects a broader challenge in handling complex scenes involving clutter and degraded visual quality, which the proposed architecture is specifically designed to address. In addition, the reliance on fully supervised learning requires both labeled class data and clean reconstruction targets, resources that may be limited or noisy in real-world deployments. To improve scalability, future work could investigate semi-supervised or weakly supervised strategies that reduce dependence on exhaustive annotation. Incorporating spatial-aware attention or feature disentanglement methods may further improve robustness in overlapping object conditions. Finally, expanding the dataset or extending the architecture to video inputs could strengthen recognition under temporal and contextual variation.

Acknowledgment

Funding Statement and Conflicts of Interest: This work was funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UMT/03/2). The also authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Rosli, R. Ali, M. S. Hitam, A. Mat Deris, N. H. Abd Rahim, and U. Haruna, "Recognizing objects in complex scenes: A Recent Systematic Review," J. Adv. Res. Appl. Sci. Eng. Technol., pp. 208–226, Oct. 2024, doi: 10.37934/araset.61.2.208226.

[2] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020. doi: 10.1109/CVPR42600.2020.00223.

[3] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 2020. doi: 10.1609/aaai.v34i07.6865.

[4] N. Rashid, M. A. F. Hossain, M. Ali, M. Islam Sukanya, T. Mahmud, and S. A. Fattah, "AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images," Biocybern. Biomed. Eng., vol. 41, no. 4, pp. 1685–1701, 2021, doi: 10.1016/j.bbe.2021.09.004.

[5] Z. Fei, J. Wang, K. Liu, E. Attahi, and B. Huang, "Deep feature fusion-based stacked denoising autoencoder for tag recommendation systems," IET Cyber-systems Robot., 2023, doi: 10.1049/csy2.12095.

[6] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021, pp. 3559–3568, Sep. 2021, doi: 10.1109/WACV48630.2021.00360.

[7] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018. doi: 10.1007/978-3-030-01234-2_1.

[8] L. Kong and J. Cheng, "Classification and detection of COVID-19 X-Ray images based on DenseNet and VGG16 feature fusion," Biomed. Signal Process. Control, 2022, doi: 10.1016/j.bspc.2022.103772.

[9] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray

images based on the concatenation of Xception and ResNet50V2," Informatics Med. Unlocked, 2020, doi: 10.1016/j.imu.2020.100360.

[10] G. Cheng, X. Xie, J. Han, L. Guo, and G. S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 2020, doi: 10.1109/JSTARS.2020.3005403.

[11] W. Fu, "Denoising convolutional autoencoder for improving the classification performance based on noisy galaxy images," Appl. Comput. Eng., 2023, doi: 10.54254/2755-2721/21/20231156.

[12] R. Vankayalapati and A. L. Muddana, "Denoising of Images Using Deep Convolutional Autoencoders for Brain Tumor Classification," Rev. d'Intelligence Artif., 2021, doi: 10.18280/ria.350607.

[13] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019. doi: 10.1109/CVPR.2019.00374.

[14] M. M. Yusro, R. Ali, and M. S. Hitam, "Comparison of Faster R-CNN and YOLOv5 for Overlapping Objects Recognition," Baghdad Sci. J., vol. 20, no. 3, p. 893, 2023, doi: 10.21123/bsj.2022.7243.

[15] H. Rosli, R. Ali, A. M. Deris, and M. Suzuri Hitam, "Classification of Kitchen Utensils in Noisy Condition Using YOLOv5 with Multiple Deep Learning Backbones," 2024 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2024 - Proc., pp. 297–302, 2024, doi: 10.1109/I2CACIS61270.2024.10649881.

[16] A. Rosello, J. J. Valero-Mas, A. J. Gallego, J. Sáez-Pérez, and J. Calvo-Zaragoza, "Kurcuma: a kitchen utensil recognition collection for unsupervised domain adaptation," Pattern Anal. Appl., vol. 26, no. 4, pp. 1557–1569, 2023, doi: 10.1007/s10044-023-01147-x.

[17] M. Hami and M. JameBozorg, "Assessing The Impact of CNN Auto Encoder-Based Image Denoising on Image Classification Tasks," 2024, [Online]. Available: http://arxiv.org/abs/2404.10664

[18] K. Liu, Y. Yang, Y. Tian, and H. Mao, "Image Dehazing Technique Based on DenseNet and the Denoising Self-Encoder," Processes, vol. 12, no. 11, 2024, doi: 10.3390/pr12112568.

[19] Y. Tian, "Privacy Preserving Method for Image Recognition based on Denoising Autoencoder," Highlights Sci. Eng. Technol., 2023, doi: 10.54097/hset.v39i.6710.

[20] C. H. Praharsha and A. Poulose, "CBAM VGG16: An efficient driver distraction classification using CBAM embedded VGG16 architecture," Comput. Biol. Med., vol. 180, 2024, doi: 10.1016/j.compbiomed.2024.108945.

[21] Y. Hu, S. Tian, and J. Ge, "Hybrid Convolutional Network Combining Multiscale 3D Depthwise Separable Convolution and CBAM Residual Dilated Convolution for Hyperspectral Image Classification," Remote Sens., vol. 15, no. 19, 2023, doi: 10.3390/rs15194796.

[22] J. Lee and S. Lah, "Denoising AutoEncoder-based Representation Learning for Multi-Task Whole Slide Image Analysis," J. Student Res., vol. 13, no. 1, 2024, doi: 10.47611/jsrhs.v13i1.6140.

[23] V. Kumar, C. Prabha, P. Sharma, N. Mittal, S. S. Askar, and M. Abouhawwash, "Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images," BMC Med. Imaging, 2024, doi: 10.1186/s12880-024-01241-4.

[24] Y. Xie, Z. Fei, D. Deng, L. Meng, F. Niu, and J. Sun, "MEEAFusion: Multi-Scale Edge Enhancement and Joint Attention Mechanism Based Infrared and Visible Image Fusion," Sensors, vol. 24, no. 17, 2024, doi: 10.3390/s24175860.

[25] L. Zhou, X. Zheng, D. Yang, Y. Wang, X. Bai, and X. Ye, "Application of multi-label classification models for the diagnosis of diabetic complications," BMC Med. Inform. Decis. Mak., 2021, doi: 10.1186/s12911-021-01525-7.

[26] H. Yessou, G. Sumbul, and B. Demir, "A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification," in International Geoscience and Remote Sensing Symposium (IGARSS), 2020. doi: 10.1109/IGARSS39084.2020.9323583.

[27] T. Ridnik et al., "Asymmetric Loss For Multi-Label Classification," in Proceedings of the IEEE International Conference on Computer Vision, 2021. doi: 10.1109/ICCV48922.2021.00015.

[28] Y. Huang, J. Qi, X. Wang, and Z. Lin, "Asymmetric Polynomial Loss for Multi-Label Classification," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023. doi: 10.1109/ICASSP49357.2023.10095437.

[29] J. Ji, W. Jing, G. Chen, J. Lin, and H. Song, "Multi-label remote sensing image classification with latent semantic dependencies," Remote Sens., 2020, doi: 10.3390/rs12071110.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[31] J. Ham, M. Fleissner, and D. Ghoshdastidar, "Impact of Bottleneck Layers and Skip Connections on the Generalization of Linear Denoising Autoencoders," May 2025, [Online]. Available: http://arxiv.org/abs/2505.24668

[32] O. Elharrouss et al., "Loss Functions in Deep Learning: A Comprehensive Review," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.04242

[33] B. N. Deepankan and R. Agarwal, "A two-phase image classification approach with very less data," Adv. Intell. Syst. Comput., vol. 1108 AISC, pp. 384–394, 2020, doi: 10.1007/978-3-030-37218-7_44.

[34] T. ValizadehAslani et al., "Two-Stage Fine-Tuning: A Novel Strategy for Learning Class-Imbalanced Data," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.10858

[35] W. N. Khotimah, M. Bennamoun, F. Boussaid, L. Xu, and F. Sohel, "Dual-Phase Framework for Few-Shot Hyperspectral Image Classification with Spatiospectral Masked Autoencoder and Episode Training," IEEE Trans. Geosci. Remote Sens., 2025, doi: 10.1109/TGRS.2025.3541263.

[36] S. Huang et al., "Application of Label Correlation in Multi-Label Classification: A Survey," Appl. Sci., vol. 14, no. 19, 2024, doi: 10.3390/app14199034.

[37] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3151048.

[38] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," in Pattern Recognition, 2012. doi: 10.1016/j.patcog.2012.03.004.

Secure Data Sharing Using Blockchain Technology: A Systematic Literature Review

Azman Azmi[1], Farashazillah Yahya[2]*, Nur Afrina Azman[3], Hazlina Jalil[4]
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia[1, 2]
School of Mathematical, Physical and Computational Sciences, University of Reading, United Kingdom[3]
The Southeast Johor Development Authority (KEJORA), Bandar Penawar, Johor, Malaysia[4]

***Abstract*—Data sharing security is currently one of the crucial parts in e-government systems. Although blockchain, a type of Distributed Ledger Technology (DLT), has been increasingly applied to enhance secure data exchange, there is a significant lack of studies focusing on the specific security factors that underpin its implementation in e-government contexts. Defining and understanding these factors is crucial for the successful integration of blockchain into public data infrastructures. This study addresses this research gap through a Systematic Literature Review (SLR) guided by the PRISMA 2020 framework. A total of 511 articles were retrieved from five major databases, and 103 were selected and systematically reviewed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 approach. While the majority of studies emphasised privacy, integrity, and transparency, other critical security factors such as scalability, availability, governance, and decentralisation remain comparatively underexplored. The theory of blockchain-based data sharing security factors was developed as a reference. The article wraps up the discussion by highlighting nine security factors in data sharing using blockchain in e-government for future investigation.**

***Keywords*—Blockchain; Distributed Ledger Technology (DLT); data sharing security; e-government; Systematic Literature Review (SLR); PRISMA 2020**

I. Introduction

In the era of emerging digital technologies, governments around the world are increasingly adopting technology to improve the accessibility, transparency, and efficiency of public administration using a transformative methodology called electronic government (e-government). A key element of e-government is the secure and seamless exchange of data among various government agencies and citizens. However, traditional data-sharing methods are often plagued by challenges related to trust involving security and privacy, which can hinder the complete effectiveness of e-government initiatives. [1], [2].

Data sharing serves as the cornerstone of e-government, enabling the efficient flow of information between government entities and citizens to ensure streamlined service delivery. However, digital disruption has introduced significant challenges in maintaining information security and privacy. Traditional methods of data sharing have increased the risk of cybersecurity. These attacks include vulnerability attack, data breach, user account compromise and malicious attack. This will further trigger the system safety and reliability issue. [1], [3].

The Malaysia Central Data Sharing platform (MyGDX), provided by the Digital Department of Malaysia, is the most used platform among government agencies. The trend of data sharing among government agencies on this platform showed an increase in transactions every year. The number of data sharing transactions recorded from the beginning of the data sharing exchange in 2018 to 26 May 2024 was recorded at 621,824 transactions. However, a total of 4,117,202 data sharing transactions were recorded by the 12th of May 2025 [4]. These statistics express a growth of 3,485,378 transactions from May of 2024 to May of 2025 alone. A breach involving such a large dataset will pose significant risk due to the scale and critical nature of the information.

According to Malaysia Computer Emergency Response Team (MyCERT), Cyber General Incident Classification statistics – January to April 2025[5]. There were 2,291 incidents received by Cyber999 Incident Response Centre, compared to the 1,550 incidents received in Q4 2024. This indicates that there is a 7% increase in incidents in the previous quarter. Personal identifiable information (PII) from national databases, such as identity numbers, addresses, and financial data, is commonly involved in high-profile attacks, according to MyCERT.

This research aims to comprehensively and systematically review the literature as well as synthesise current work on blockchain-based data sharing in e-government systems. The main focus is to identify the key factors for secure data sharing. The review aims to analyse the existing frameworks and models proposed for application of blockchain in electronic government, assess the effectiveness of these approaches in improving data sharing practices, and stress the gaps in the recent literature. Ultimately, this review aims to deliver a clear understanding of the latest use of blockchain-based data sharing within e-government systems and to identify prospective investigations in this evolving area.

II. Study Background

A. Data Sharing

Data sharing is essential to promote collaboration and innovation in various domains. The integration of technologies can improve the privacy and security of data sharing, addressing concerns related to the issue and challenges of data security.

The term data sharing refers to the movement of information between different systems and entities or across international borders. It is crucial for decision makers, resource optimisation,

*Corresponding Author

and improving information utilisation. [6]. In the public sector, data sharing is vital for addressing interoperability challenges, such as collaboration between government agencies and seamless information exchange. The process involves the exchange of information and knowledge between organisations through their business processes [7], [8].

Both definitions highlighted that data sharing is important to parties which benefit from data sharing. The United Kingdom Information Commissioner's Office highlights the importance of responsible data sharing, stating: "When you get data protection right, it sends a strong message to your customers – it lets them know that you value and care for their information and that you are more likely to keep it safe and not share it inappropriately"[9]. This statement further reinforces the critical role of trust, ethical practices, and robust security measures in data sharing processes, ensuring that organisations handle information responsibly while maximising its potential for social and organisational benefits.

B. Data Sharing Security

Data sharing security is a critical aspect of modern information systems, ensuring that data is shared between entities in a way that protects its confidentiality, integrity, and availability. The fundamental model used in information security is the CIA triad, which stands for confidentiality, integrity, and availability [10]. This model serves as a reference framework for agencies to protect delicate data and ensure that information systems are secure. Each component of the triad addresses a specific aspect of security and together provides a comprehensive approach to safeguarding information.

Based on the NIST Cyber Security Framework 2.0 [11], Data security refers to the consistent management of data in a proper manner within the organization and protects all the risk of providing trust through confidentiality, integrity, and availability of information in all data conditions, whether the data at rest, data in transit, data in use and backups of data are created, protected, maintained, and tested.

1) Confidentiality: The NIST refers to the term confidentiality as the preservation of authorised restrictions on information access and disclosure, including means to protect personal privacy and proprietary information...[12]. Confidentiality involves ensuring reachability of data by authorised personnel. Sensitive data needs to be secure from data leaks and breaches.

2) Integrity: Integrity was demarcated by the NIST as protection for data from being improperly altered, modified or demolished by protecting the information's originality and legitimacy [11]. The term integrity also refers to the precision and fullness of data. The integrity ensures information is not changed in unauthorised ways, while at the same time, providing the reliability and trustworthiness of the data [12].

3) Availability: Availability guarantees that resources and information are available to authorised users when needed. It is essential for upholding the practicality and usability of information systems. The NIST defines availability as "Ensuring timely and reliable access to and use of information…"[12].

C. Blockchain Technology Security

Blockchain Technology or Distributed Ledger Technology is one of the emerging technologies of Industrial Revolution 4.0, which is used in data security. Introduced by Nakamoto, Blockchain Technology offers a decentralised and secure method for recording transactions. Blockchain technology is characterised by the ability to maintain a distributed ledger that is immutable, transparent and resistant to tampering [13].

In blockchain technology, the new transaction created by a user will be broadcast to the blockchain network using a P2P connection. Then, the node in the network will validate the transaction using a consensus mechanism like Proof of Work (PoW). Validated transactions are grouped in a block. The block will be linked to the chain using a cryptographic hash. The ledger will be updated to complete the transaction as shown in Fig. 1.

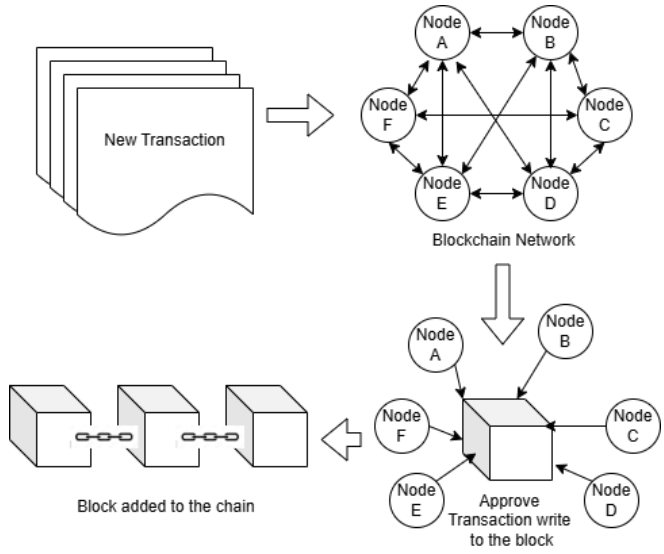


Fig. 1. How blockchain works.

D. E-Government System Data Security

The term, electronic government (e-government), is defined by [1] as the application of ICT to digitalise government services for the public, enterprises, and government agencies. The objectives of having an e-government system include reducing bureaucracy, increasing transparency, and ensuring secure and efficient service delivery. [1], [7], [14], [15]. According to [16], current data-sharing practices in e-government systems face several challenges that hinder effectiveness and security.

The governing policy and act are the most important items to be complied with for the e-government data sharing security in government agencies. In 2021, the Malaysia Administrative and Modernisation dan Planning Unit (MAMPU) released the Policy for Public Sector Data Sharing. This is used to govern the cross-organisation data sharing to improve services towards data-driven government [17]. Recently, in February of 2025, the establishment of the National Data Sharing Committee led by the Director General of the National Digital Department has aligned with the passing of the bill of Data Sharing Act 2025 (Act 864) by providing control to the public sector regarding data sharing processes [18].

III. METHODOLOGY

Data Sharing, security, and e-government applications have a vast amount of literature in general. Many academic papers and government reports have discussed Blockchain topics. An extensive search has been made, and the current development studies published between the years 2014 and 2024 have been discussed in this paper. The selection of the publication article aligns with the issues related to the security of data sharing, specifically involving the e-government system. These issues have been discussed, debated, and reported at international and national levels. This paper discusses the selected papers from established journals, conference papers, and proceedings.

The PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol is a widely recognised and standardised framework used to ensure transparency, rigour, and reproducibility in conducting and reporting systematic reviews [19]. For this systematic literature review on blockchain-based data sharing in e-government systems, the PRISMA 2020 protocol was meticulously followed to guide the review process. Based on the keywords search, a collection of comprehensive artefacts was gathered. Four phases were involved in searching and selecting the papers: identification, screening, eligibility, and inclusion, as illustrated in Fig. 2.

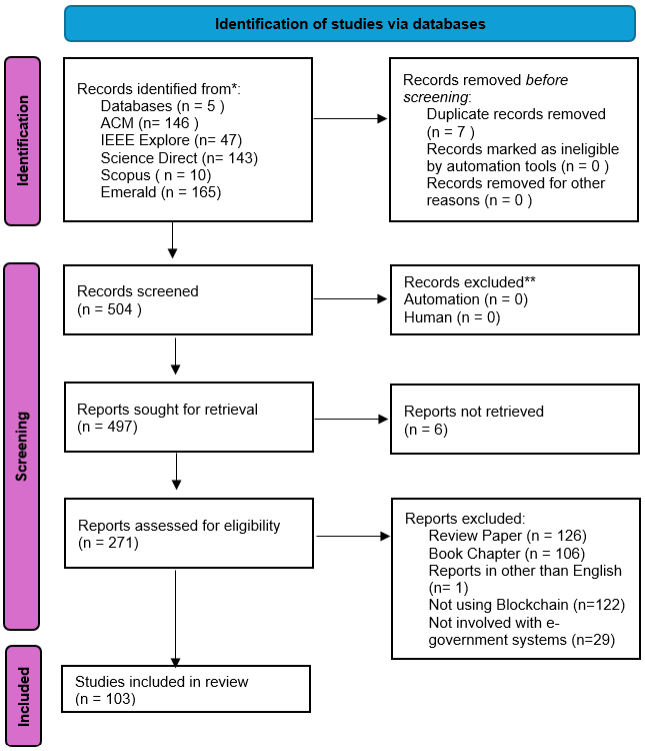


Fig. 2. Review workflow of this systematic review.

By adhering to the PRISMA 2020 protocol, this systematic review ensures a rigorous, transparent, and reproducible process. Ultimately, providing a comprehensive synthesis of the current state of blockchain-based data sharing in e-government systems.

A. Identification of the Key Research Question

Based on the Key research question, the PICOC framework [19] has been used to define and structure research questions and identify the relevant literature. The key elements of the research context are represented in Table I. The context in the comparison in the systematic review is excluded.

TABLE I. PICOC FRAMEWORK MAPPING WITH KEYWORDS AND SYNONYMS

PICOC	Description	Keyword	Synonym
Population	An industry domain.	government	
Intervention	The methodology, tool, or technology that tackles a certain problem.	data sharing	data exchange
Comparison	The methodology, tool, or technology in which the Intervention is being compared (if appropriate).	Blockchain	DLT
Outcome	Factors of importance to practitioners and/or the results that interposition could produce.	security	-
Context	The context in which the com-parison is done.	e-government	digital governme nt

The research questions guiding this systematic literature review are designed based on the PICOC framework to explore the current landscape of blockchain-based data sharing in e-government systems. The PICOC framework keywords used in this research are in Table I. The population used is e-government or digital government, which is the domain of the research. The Intervention for this research is data sharing or data exchange. The comparison of the technology used is blockchain or Distributed Ledger Technology. Finally, the outcome for the keyword used is security. No context specifies for the research keyword, and it is optional based on the approach.

Specifically, the review seeks to answer the following questions:

RQ 1: What is the current trend in blockchain technology for securing data sharing in e-government systems?

RQ 2: What is the current approach and use of blockchain-based data sharing in e-government systems?

RQ 3: What are the data sharing security factors that blockchain technology enhances in e-government systems?

These questions aim to present a comprehensive understanding of how blockchain technology is being leveraged in highlighting the questions of secure data sharing in e-government, while also identifying areas where further investigation is needed.

B. Identification of Relevant Articles

The selection of databases was based on their relevance to the fields of information security, information systems, and public administration, ensuring an inclusive coverage of the literature.

The review began with a comprehensive search strategy across several academic databases such as Scopus, IEEE Xplore, Science Direct, ACM Digital Library and Emerald. The keyword used in the search included a combination of keywords related to "blockchain OR Distributed Ledger Technology", "data sharing OR information sharing", "e-government", and "security." To enhance the exploration and to make sure that related studies were taken, a Boolean operator is used.

C. Selection of the Relevant Articles: Inclusion and Exclusion Criteria

As mentioned earlier, PRISMA 2020 was used to carefully select the relevant articles for this study. Table II shows that the inclusion criteria 1 is the article. The articles show the use of blockchain in data sharing for e-government systems. Inclusion criteria 2 are the improvements made due to blockchain application in government data sharing as shown in the paper. The exclusion criteria set for this review are that the implementation of blockchain is not mentioned in the paper, and papers have redundancies for the research done.

TABLE II. INCLUSION AND EXCLUSION CRITERIA IN REVIEW SELECTION

Criteria	Justification
IC-1: The articles show the use of blockchain in data sharing for e-government systems.	This review inspects related studies on blockchain used for data sharing in government-related systems.
IC-2: The improvements due to blockchain application in government data sharing are shown in the paper	The aim is to convey an indication Important progress of blockchain technology in data sharing and e-government.
EC-1: Implementation of blockchain is not mentioned in the paper.	Papers that only overview blockchain in general often lack focus on data sharing.
EC-2: Paper is redundant for the research done.	Duplicate content is assessed by com paring publication dates and author status for validity.

In the selection phase, the titles and abstracts of the 497 selected articles were read and evaluated manually. Of those, 390 articles were found to be outside the limits of this study, and thus only the remaining 271 articles were taken into account for further analysis. A complete analysis was performed for these 271 articles by carefully inspecting the whole text. This detailed analysis found 168 articles that were not related to this study as they were not relevant to the use of blockchain in data sharing for e-government. Thus, they were also expelled. 103 articles were finalised on the significant publication used for this study, as detailed in Table III.

TABLE III. DISTRIBUTION OF PAPERS BY SOURCES THROUGHOUT THE SCREENING PROCESS

Source	Identification Search	Paper Screening	Eligible Search	Selected for Review
ACM	146	143	44	19
IEEE Xplore	47	46	36	30
Science Direct	143	142	84	27
Scopus	10	8	7	4
Emerald	165	165	100	23
TOTAL	**511**	**504**	**271**	**103**

D. Reporting and Summarising the Results

The first stage was to extract and compile metadata from 271 relevance articles. The metadata consisted of the authors' names, publication title, year of publication, approach used, improved factors, and the security threat problem. These metadata were thoroughly examined using quantitative methods. The purpose of descriptive statistics was to find trends and patterns. Sections IV and V below provide all of these analyses.

IV. RESULTS

A. Trends of Articles Based on PICOC Keywords for 10 Years

The keyword search shown in Fig. 3 shows an upward trend of publications from the year 2016 to 2024. The trend of eligible articles shows the same pattern. The number of articles related to Blockchain approach in data sharing for e-government is expected to increase, as shown in Fig. 3. In 2024, fewer articles were produced as the research was only done until Sept of 2024. There will be more articles published by the end of the year and will increase number of publications, which we believe is an area of further investigation. This will answer Research Question 1 (RQ 1) for What is the current trend in blockchain technology for securing data sharing in e-government systems?

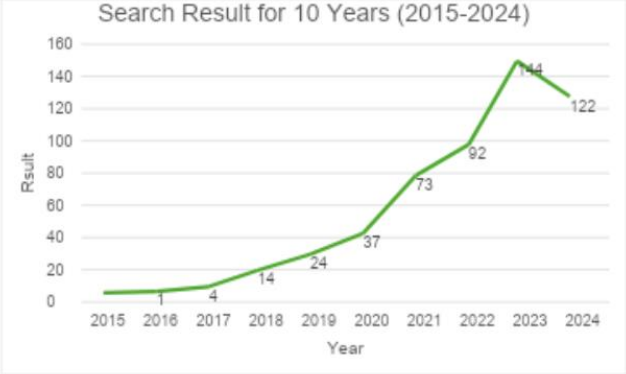


Fig. 3. Trend of article distribution based on 511 articles (Identification Phase).

The findings have shown that evidence of blockchain technology has been widely used to enhance security for data sharing in the e-government. Since 2015, the number of articles published has increased significantly. The number of publications increased by 250% in 2018 from the previous year. Year 2021 has recorded an increase of 36 articles for blockchain-based data sharing compared to 2020, while 92 articles were recorded in 2022. In 2023, 144 articles have been published, which is an increase of 52% from the previous year. The research on the topic began in 2016, which might be influenced by the launch of the Ethereum platform in blockchain technology in 2015 [20].

Based on the distribution of the articles shown in Fig. 4, 103 articles have been identified as eligible to be included in the study. From this, there is only one article which was eligible to be included from 2016. The number of eligible articles increased yearly. However, in 2019, the eligible article was down by 1 article from the year 2018, while 2022 recorded a downtrend of 2 articles from 2021. In certain instances, the paper was

excluded due to its failure to address a particular case study. The eligible articles included in 2023 amounted to 29, which is the highest number of articles recorded in a given year. This shows the increment of 12 articles. The number of articles for 2024 is only 14, as the literature cut-off searching date was set to 30 September 2024.

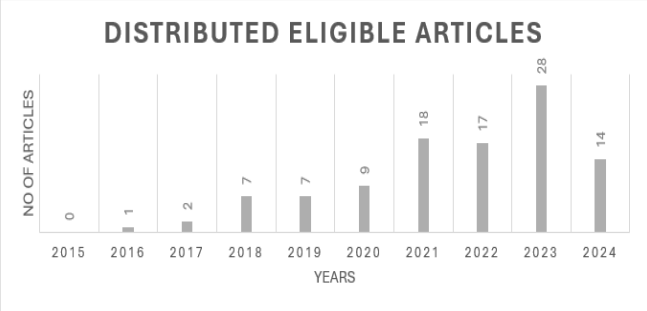


Fig. 4. Distribution of the shortlisted 103 articles based on inclusion and exclusion criteria (Eligibility Phase).

B. State of the Art Blockchain-Based Data Sharing in e-government Systems

The use of blockchain in data sharing has been implemented in several ways to enhance security. RQ 2 What is the current trend and landscape of blockchain-based data sharing in e-government systems in the world? The approach used is governance, adoption, and application. Four (4) articles were focused on the usage of blockchain in enhancing data sharing with compliance with the law and act as General Data Protection Regulation (GDPR), European Timber Regulation and EU's Carbon Border Adjustment Mechanism (CBAM). There is one article for compliance in 2022 and two in 2023, as shown in Fig. 5.



Fig. 5. Trend in publication on approach by year.

The second approach focuses on how organisations adopt blockchain to secure data sharing in e-government. 22 articles researched the acceptance of blockchain security in e-government systems for organisations due to their aptitude to enhance data sharing security. The articles on adoption of blockchain were discovered to begin in 2019 and are increasing every year until 2023, as shown in Fig. 5. In 2024, it is expected to grow as five articles are recorded as of September 2024.

However, the preferred approach to enhance data sharing in e-government data sharing is the application of blockchain in the e-government system. 68 or 56.67% of the eligible articles are on the application of blockchain. Application through integration with existing technology or emerging technologies in one of the four approaches found in this literature. 26 articles used the integration approach in implementing blockchain with IoT, Big Data and cloud computing to enhance the data sharing security. The approach is mostly used for blockchain data sharing security. The trend is static for three years since 2018, and the highest eligible article for application recorded in 2023 is 17, as shown in Fig. 5. In 2024, the number of articles is expected to show 8 articles published and still counting due to the interest of researchers in exploring blockchain technology to secure data sharing in e-government systems continuously.

C. The Landscape of Blockchain-Based Secure Data Sharing Based on Geographical Area and Country

The use of blockchain in the world is expanding from year to year. Currently, 36 countries have been identified in this research to be securing data sharing using blockchain in e-government systems as shown in Fig. 6. China has twelve use cases, which is the highest number of blockchain-based data sharing which covering the implementation in data sharing and big data [21], [22], [23], economic and e-commerce [24], [25], [26]Identity and data management [27], [28], [29], [30], [31] and finally adoption of blockchain technology [32].

There are seven research studies from India. The Indian government used blockchain in the supply chain in government. [33], [34], [35] and aid management [36]. There are use cases in document security which use blockchain data sharing solutions in India. [37], [38] while [39] study how the government accept blockchain technology in the administration.

Estonia recorded five research which is the same as United Arab Emirates (UAE). These countries are focusing on big data and data sharing. [40], [41], blockchain Infrastructure for data sharing [42], [43], [44] and service delivery involving Austria and Australia [45], Utility [46], and services on disaster recovery and real estate [47], [48].

Germany and the Netherlands recorded four studies each. Blockchain is used for monitoring the transparency of the transaction data sharing for agriculture [49], [50], automotive [51], economy [52] and document management and data sharing [8], [41], [53], [54].

The United States (US), United Kingdom (UK), and Nigeria have three research projects in their country. The three countries have data sharing in government healthcare. [44], [55], [56], [57], verification of higher education certificates [58], [59], [60], and identity management [61]. Two studies in the UK are about governance and policy. [62], [63] and property valuation [60].

Malaysia currently has two studies on government data sharing based on blockchain, which is the same as Australia, Finland and Sweden. For the Malaysian case study, blockchain-based data sharing is used for transparency in government aid management. [64] and the adoption in a government agency [15]. Australia is more focused on data sharing for service

delivery and smart cities. [45], [65] while Finland studies are recorded as being for ID Management and Economy [40], [66]. In Sweden, studies in this area focus on valued focus thinking [57] and document security [67].



Fig. 6. Blockchain-based data sharing in e-government research heatmap.

22 counties have recorded one publication on data sharing using blockchain. Asian countries like Japan, Vietnam, and Thailand focused on improving data sharing security and trust for decision making in the government system using blockchain. [57], while the publications of Iceland, Indonesia, and South Korea focused on identity security in data sharing [40], [68], [69]. Pakistan and Afghanistan used blockchain in data sharing to secure verification of certification. [14], [70]. Russia and France focused on securing the voting system. [71] While Portugal is securing the data sharing procurement process [72]. On governance, Italy used blockchain in data sharing to comply with the Timber Regulation [73]. Denmark implements a blockchain to secure data sharing in tax management[74]. Iran is using blockchain in cybersecurity [75], Kuwait studies on IoT data sharing [76], and studies in Iraq are using blockchain technology in healthcare [77].

D. Blockchain-Based Data Sharing Security Factors in e-government Systems

The security factors are critical aspects in achieving goals in blockchain-based secure data sharing. The security factors involved are extracted from all four approaches in the 103 articles. 77 factors from the articles are structured and present in the word cloud in Fig. 7 and Table IV. This will answer RQ 3 – What are the security factors for blockchain technology used in enhancing the security of data sharing in e-government systems?

This word cloud in Fig. 7 highlights key attributes and factors of data sharing using blockchain technology, emphasising its potential to transform systems based on trust, transparency, and security. The most highlighted security factors are transparency, privacy and integrity, which underscore blockchain's ability to protect sensitive data while ensuring the security processes. The prominent factors of trust, decentralisation, and reliability that reflect blockchain's capacity to eliminate intermediaries, build confidence, and provide reliable systems are also recognised as important factors involved in data sharing security.

The main blockchain concepts, such as immutability and traceability, showcase the technology's ability to ensure data accuracy, prevent tampering, and enable detailed tracking of transactions or assets. In addition, interoperability and scalability factors highlight its adaptability and capacity to handle increasing demand across various systems. In the context of e-government, these attributes align perfectly with the need for secure, efficient, and transparent data-sharing mechanisms, fostering trust between governments and citizens while enhancing operational efficiency.



Fig. 7. Word cloud of security factors.

Based on the literature review, 54 characteristics of security factors components have been extracted from 103 articles. The characteristics are based on the NIST Information Security framework, the blockchain security factors and e-government security factors. For the confidentiality factors, Trust between parties, Ensuring Privacy, Data confidentiality, Data ownership, authentication and authorize, secure access control, Identity Security Management and Data sovereignty have been identified in the articles as contributors to the factors.

The integrity factors are the most security factors discussed in blockchain for data sharing, where the subfactors are ensuring or protecting data integrity, single source of truth, trustworthiness of data, verifying authenticity, accuracy and specific, providing validation, reducing fraud, enhancing trust, enhancing accountability, authenticity and reliability of data.

The security factors of using blockchain for data sharing are shown in Fig. 8. The trend shows that integrity recorded 37 articles and is the most secure factor's goal that is achieved in applied blockchain for data sharing in e-government, followed by 33 articles on confidentiality and 27 articles on transparency. The intermediate security factors that attract researchers' attention are 18 articles on interoperability, immutability with 15 articles, auditability with ten articles and decentralize and compliance recorded nine articles for each factor. Availability and Scalability are less focused security factors recorded in six articles and two articles, respectively. Different focus on the security factors shows that priority needs to be established for the factor in securing data sharing using blockchain. The findings reflect the evolving landscape of blockchain security priorities, aligning with the specific needs and challenges of e-government implementations for data sharing.

TABLE IV. INITIAL ALIGNMENT FACTORS FOR BLOCKCHAIN-BASED DATA SHARING SECURITY FACTORS IN E-GOVERNMENT SYSTEMS

Security Factors	Characteristics	Source
Confidentiality	i. Trust between parties, ii. Ensuring Privacy, iii. Data confidentiality, iv. Data ownership, v. Authenticates, vi. Authorize, vii. Secure access control, viii. Identity Security Management, ix. Data sovereignty	[6], [25], [26], [27], [30], [31], [32], [34], [39], [40], [44], [45], [49], [50], [65], [70], [74], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92]
Integrity	i. Ensuring data Integrity, ii. Protect data Integrity, ii. Single source of truth, iii. Trustworthiness of data, iv. Verify authenticity, v. Accuracy, vi. Providing validation, vii. Reduce fraud, viii. Enhancing trust, ix. Enhanced accountability, x. Authenticity, xi. Reliability source xii. Reliability data, xiii. Specificity	[14], [15], [29], [35], [37], [38], [39], [40], [53], [54], [58], [61], [65], [71], [72], [76], [77], [81], [83], [84], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108]
Availability	i. Enhanced information-sharing across organizations, ii. Make data available, iii. Usability data, iv. Applicability	[24], [56], [80], [83]
Decentralized	i. Enabling distributed power, ii. Decentralisation of property rights, iii. Large-scale data management	[8], [31], [68], [78], [80], [95], [109]
Immutability	i. Reduce the need for trusted third parties, ii. Data protection iii. Tamper-proof record, iv. Resistance to attacks, v. Isolate data vi. block data, vii. Non-temperable	[8], [15], [21], [30], [42], [44], [50], [59], [61], [71], [75], [110], [111], [112], [113]
Transparency	i. Greater transparency, ii. Seamless data sharing among parties, iii. Eliminating intermediaries, iv. Secure transactions, v. Openness of data	[7], [15], [16], [24], [28], [35], [36], [41], [43], [48], [50], [61], [63], [64], [66], [69], [71], [72], [84], [86], [89], [98], [99], [101], [114], [115], [116]
Interoperability	i. Integration data, ii. Harmonise process, iii. Standardisation Integration, iv. Collaboration	[6], [16], [23], [51], [53], [61], [67], [69], [72], [74], [80], [90], [96], [104], [109], [117], [118]
Auditability	i. Transaction can be tracked, ii. Transaction is permanent, iii. Traceability transaction	[8], [31], [32], [50], [63], [73], [80], [88], [112], [113]
Governance	i. Policy, ii. Regulation, iii. Uniformity protocols, iv. Security controls, v. Compliance	[39], [46], [49], [51], [61], [73], [88], [91], [119]

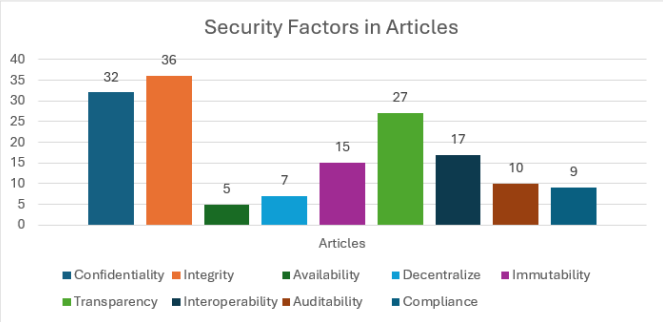


Fig. 8. Security factors.

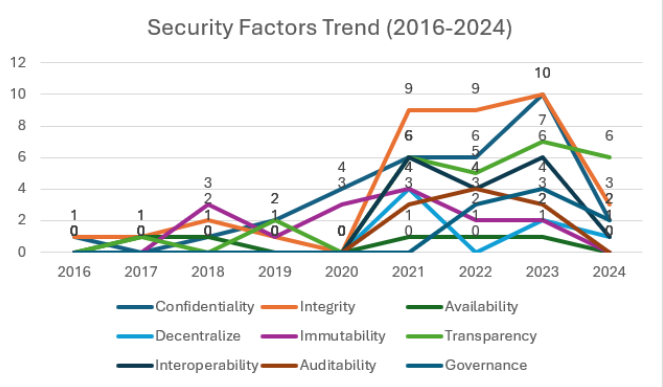


Fig. 9. Security factors trend by year.

The trend in Fig. 9 shows that confidentiality, integrity, transparency, interoperability and compliance factors are trending for the use of blockchain in data sharing security. Integrity factors show the highest trend for three years from 2021 to 2023, followed by confidentiality and transparency. The scalability, availability, immutability, decentralize and show a downtrend while auditability consistently got attention as a security factor that needs to be concentrated on. Based on the pattern, these trends demonstrate the evolving security priorities in blockchain-based data sharing, shaped by emerging challenges and the needs of technological advancements in providing secure data sharing in the domain.

1) Confidentiality: Confidentiality is one of the most important factors that focuses on the government securing data using blockchain technology. Confidentiality concerns can limit the willingness of organisations to share data, as they must ensure data is protected from unauthorised access [120]. Implementing robust confidentiality measures can facilitate data sharing by building trust among stakeholders and ensuring compliance with privacy regulations [120], [121].

2) Availability: In public service, data availability is enhanced through the aggregation of data from various departments into a unified resource catalogue, which supports efficient sharing [122]. The availability of a comprehensive data catalogue allows for better management and sharing of power data, meeting external sharing needs. Public policy plays a dual role as both an enabler and a barrier to data sharing. Contradictions in data availability can arise from legal structures that either facilitate or hinder data linking and sharing, impacting the potential of healthcare data [56].

3) Integrity: Integrity is maintained through rigorous transaction verification processes. Data storage and access transactions involve multiple steps of encryption and signature verification to ensure that data has not been tampered with during transmission [123]. On the other hand, different random parameters are used in generating secret keys, ensuring that data integrity is not compromised by colluding parties [124].

4) Decentralised: Other blockchain technologies' important features for data security are the decentralised factor. Transactions are distributed across the peer-to-peer network

(P2P) by the node and storing the transaction back in the node across the members' network [13]. The distributed ledgers in blockchains are securely synced using a P2P approach, and consensus is made based on smart contracts for the newly added data. Author [125] reported that it is harder to control and alter the data in the distributed network of blockchain because of the consensus built into the technology. Blockchain is a decentralised and trusted authority based on consensus of nodes that take part in the network [126].

5) Immutable: Immutability in blockchain technology has enhanced data sharing security in public service. According to a study [15], immutability is referred to as a fundamental attribute of blockchain. This is one of the important security features of blockchain. Blockchain technology was established as a platform for secure digital transformation technologies in creating immutable records [59], [127].

6) Transparency: Transparency of data had the highest influence on the implementation process in blockchain-based data sharing [41]. Transparency has improved public governance by promoting visibility [7], [64], [66], and enhancing the trust and efficiency of government information sharing [28]. Extended studies to evaluate the effectiveness of blockchain in improving service delivery, transparency, and efficiency in public administration should be done and could involve case studies and comparative analyses with other regions or countries that have successfully implemented blockchain in their e-government initiatives [43].

7) Auditability: Study by [79] assists practitioners and policymakers in implementing auditability using blockchain technology effectively to ensure food safety and improve sustainability in India. Auditability is also a main focus in facilitating seamless data exchange among service providers [80].

8) Interoperability: Interoperability breaks significant barriers to seamless data exchange, often resulting in data silos where information is isolated and difficult to access across departments. The inefficiencies caused by these silos lead to redundant data collection efforts, increased operational costs, and delayed decision-making processes. Furthermore, the group of authors agreed that the absence of standardised data sharing frameworks exacerbates the complexity of managing and integrating data from multiple sources, making it challenging for the public sector to allow consistent and coherent public services [96], [105], [109], [117].

9) Governance: The authors [46] and [75] highlighted the role of blockchain in enhancing compliance, particularly in securing data access, protecting data privacy, and ensuring adherence to GDPR. Similarly, [67] demonstrates that blockchain-based e-archive systems improve the management of government information while maintaining compliance with the Public Sector Information (PSI) law. In contrast, [73] finds that the implementation of the European Timber Regulation (EUTR) remains complex, as it has not fully resolved longstanding challenges, thereby limiting traceability in the wood-energy sector. Furthermore, the introduction of new regulatory frameworks, such as the EU's Carbon Border Adjustment Mechanism (CBAM), has imposed additional responsibilities on government authorities to standardise compliance measures [52].

While blockchain technology holds significant promise for enhancing regulatory compliance, particularly in data management and privacy protection, it also comes with issues that need to be answered. The tension between blockchain's capabilities and existing legal frameworks, such as the GDPR [42], [61], highlights the need for ongoing dialogue and innovation to ensure that blockchain can be effectively integrated into regulatory environments. As new regulatory frameworks emerge, the role of blockchain in standardising compliance measures will continue to evolve, necessitating careful consideration of both technological and legal developments [79].

V. Discussion

A. Implications for E-Government Systems

Security concerns are a critical challenge in current data-sharing practices within e-government systems. With the increasing digitisation of government services, the volume of sensitive and personal data being shared across platforms has grown exponentially. This makes e-government systems prime targets for cyberattacks, data breaches, and unauthorised access. Traditional data-sharing methods often rely on centralised databases, which are vulnerable to hacking and can present a single point of failure. Once compromised, these databases can lead to the widespread leakage of sensitive information, undermining public trust in government services. Additionally, the lack of robust encryption and authentication mechanisms in many e-government systems leaves data vulnerable during transmission, further increasing the unauthorised access and interception risk.

Moreover, lack of transparency and accountability in current data-sharing practices limits public trust in e-government systems. Citizens often have little visibility into how their data is being used, shared, or stored by government agencies. This opacity can lead to concerns about data misuse, unauthorised sharing, or surveillance, which erodes confidence in digital government services. Additionally, when data-sharing processes are not transparent, the accuracy and integrity of the shared information are harder to verify, leading to potential errors or misinformation. This highlights the need for more transparent and accountable data-sharing mechanisms that can provide clear audit trails, ensure data integrity, and give citizens greater control and insight into how their data is handled.

In light of these challenges, there is a pressing need for more secure and efficient data-sharing solutions within e-government systems. Blockchain technology offers a promising avenue to address these limitations by offering a transparent, decentralised and secure platform for data sharing. By leveraging blockchain’s capabilities, e-government systems can overcome the challenges of data fragmentation, improve security, confirm fulfilment with privacy law, and improve public confidence in digital government services.

B. Theoretical Contributions

The result provides some tentative initial evidence to the literature on information security, blockchain security and e-government systems for data sharing. The study explored security factors involving blockchain in information security and e-government systems for data sharing. This has contributed to theoretical support in the field, supported by the findings.

Based on the systematic review conducted, nine theories of blockchain-based data sharing security factors are identified: confidentiality, integrity, availability, transparency, auditability, distributed, immutability, interoperability, and compliance have to answer RQ 3. With this brief description, the review also presents issues and solutions to various industries as an understanding of blockchain's ability to overcome constraints. Therefore, we show that blockchain technology is not limited to any industry that wants to implement cryptocurrency into its systems.

This new understanding should help to improve predictions of the impact of data sharing security factors that are involved in blockchain-based data sharing systems. Second, the study identifies the item under the factors in the literature that have a direct impact on securing data sharing. Thirdly, the conceptual framework has been developed as a reference in data sharing security using blockchain in e-government. Lastly, the research contributes to the body of knowledge on secure data sharing by investigating its use in the e-government domain.

Consequently, data sharing security factors and data security compliance are significant factors determining the causal effect of secure data sharing using blockchain technology. Previous research has shown that among way to assess the security of information systems is the development of a rating score using goal question metrics, which has specifically developed an instrument tool for cloud security [128], [129]. The author also presented how a security framework can be evaluated and applied in a real case scenario using the said approach.

C. Comparison with Existing Reviews

The landscape shows that blockchain has been widely accepted in improving the security of data sharing for the e-government system. The governance with legal requirements can be explored further towards the government data sharing policy for the e-government system. The research findings of [15] provide information on the adoption factors of blockchain based on the technology, organisation, and environment of the TOE methodology with added trust factors. With its key security factors of decentralisation, persistence, anonymity, and auditability, the author highlights how blockchain can transform security in conventional businesses. This research found that trust is the subfactor under confidentiality, while decentralize and auditability are the security factors in the findings. This research result also shows that persistency and anonymity are not in the security factors list in the literature for blockchain security.

Blockchain technology is also suitable to be applied across multiple e-government services. According to [130], blockchain is suitable for securing majority system purposes, including securing supply chains, managing records, amusement, and government. The five categories mentioned as confidential, trust, traceability, efficiency, and reliability are the features of blockchain applications proposed should also be categorized. This research uses various sources of information, including security reports and government policy, to give wider views of security factors in data sharing using blockchain, while the other authors only used journal articles for the review that which lack information from external sources and additional insights from industry information.

Blockchain technology offers promising opportunities for enhancing records and library operations, and its adoption is still in the nascent stages, where several challenges need to be addressed. Prior studies by [130] have noted the importance of blockchain challenges in libraries, which include technological barriers, financial constraints, legal and regulatory issues and social acceptance and awareness. Many scholars are of the view that the lack of empirical evidence and the need for highly qualified security personnel, financial resources, and regulatory compliance are significant barriers [23], [53], [69], [116], which supports the finding in this literature.

This study also discovered that the industry adopts blockchain because the technology behind it promises secure transactions, where the industry processes a huge amount of data to implement transparent storage through private transactions. This type of classification is expected to increase significantly, and it also remains a concern for various sectors in the long term because privacy and security are key aspects of establishing a decentralised transaction relationship with the industry. In contradiction of compliance factors, not all design principles were fully met, particularly regarding data access security and compliance with GDPR [74].

VI. Conclusion

This study found the invention taxonomy in blockchain research for the last 10 years. The most obvious finding that emerges from the analysis is that blockchain is a world-shattering technology that expressively transforms the way that e-government systems deliver a secure and trusted service. The significance of blockchain in the advancement and enhancement of data sharing security in e-government is provided. The recent analysis of the trend of security factors for data sharing was provided for the implementation of blockchain technology.

This review also shows the important discovery of secure data sharing using blockchain with various applications including securing supply chains, transportation, agriculture, healthcare and government administration. There are still opportunities for government agencies to apply or integrate blockchain technology to improve data security for their organisation. One unexpected finding was the extent to which trust and privacy the important factors in deploying the blockchain in e-government systems.

The most important finding is that data sharing security factors using blockchain can be classified into nine factors based on their application: confidentiality, integrity, availability, transparency, auditability, distributed, immutability, interoperability and governance. The most in-demand feature in blockchain applications is privacy enhancement, and the review of use cases presented.

These results have important implications for the development of blockchain as a platform to enhance data sharing security in government applications and innovations. The security factors involved in data sharing for e-government in the current research show that it persists with a significant impact in enhancing data security that is needed by the domain. This study focuses on the security factors for blockchain-based data sharing. The study did not cover the review for blockchain platform, architecture, consensus type and performance.

In the future, this study will lead to the development of a framework that covers the principles of data security, blockchain security and compliance. It also concludes that data sharing security could also significantly contribute towards the open data initiative. The understanding gained here should help to implement blockchain-based secured data sharing and increase trust in the e-government system.

REFERENCES

[1] I. Lykidis, G. Drosatos, and K. Rantos, "The use of blockchain technology in e-government services," Dec. 01, 2021, MDPI. doi: 10.3390/computers10120168.

[2] L. T. Ha, "An investigation of digital integration's importance on smart and sustainable agriculture in the European region," Resources Policy, vol. 86, Oct. 2023, doi: 10.1016/j.resourpol.2023.104158.

[3] F. R. Batubara, J. Ubacht, and M. Janssen, "Challenges of blockchain technology adoption for e-government: a systematic literature review," in Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, in dg.o '18. New York, NY, USA: Association for Computing Machinery, 2018. doi: 10.1145/3209281.3209317.

[4] Digital Department of Malaysia, "Fakta - Status pelaksanaan MyGDX sehingga kini," 2025. Accessed: Jan. 28, 2025. [Online]. Available: https://www.mygdx.gov.my/ms/landing-page/facts?theme=second-theme

[5] MyCERT, "MyCERT _ Advisories - Cyber Incident Quarterly Summary Report - Q1 2025," 2025, Accessed: Jul. 04, 2025. [Online]. Available: https://www.mycert.org.my/portal/advisory?id=SR-030.062025

[6] X. Zhang, T. Chen, Y. Feng, and Y. Yu, "A Data Sharing Scheme Based on Blockchain System and Attribute-Based Encryption," in ACM International Conference Proceeding Series, Association for Computing Machinery, Mar. 2021, pp. 195–202. doi: 10.1145/3460537.3460559.

[7] S. F. Wamba, S. L. Wamba-Taguimdje, Q. Lu, and M. M. Queiroz, "How emerging technologies can solve critical issues in organizational operations: An analysis of blockchain-driven projects in the public sector," Gov Inf Q, vol. 41, no. 1, Mar. 2024, doi: 10.1016/j.giq.2024.101912.

[8] B. Rukanova et al., "Realizing value from voluntary business-government information sharing through blockchain-enabled infrastructures: The case of importing tires to the Netherlands using TradeLens," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jun. 2021, pp. 505–514. doi: 10.1145/3463677.3463704.

[9] Information Commissioner's Office - UK, "Data sharing: a code of practice," 2021, Accessed: Dec. 24, 2024. [Online]. Available: https://ico.org.uk/media2/ictfahk2/data-sharing-a-code-of-practice-all-1-0-2.pdf

[10] M. Nieles, K. Dempsey, and V. Y. Pillitteri, "An introduction to information security," Gaithersburg, MD, Jun. 2017. doi: 10.6028/NIST.SP.800-12r1.

[11] National Institute of Standards and Technology, "The NIST Cybersecurity Framework (CSF) 2.0," Feb. 2024. doi: 10.6028/NIST.CSWP.29.

[12] U.S. Government Publishing Office, "Chapter 35 - Coordination of Federal Information Policy Subchapter III - Information Security," Dec. 2002, Accessed: Oct. 01, 2025. [Online]. Available: https://www.govinfo.gov/content/pkg/USCODE-2011-title44/html/USCODE-2011-title44-chap35-subchapIII-sec3542.htm

[13] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Decentralized business review, p. 21260, 2008. [Online]. Available: www.bitcoin.org

[14] O. Konashevych, "'GoLand Registry' Case Study," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jun. 2021, pp. 489–494. doi: 10.1145/3463677.3463720.

[15] M. S. Kamarulzaman, N. H. Hassan, N. A. A. Bakar, N. Maarop, G. A. L. N. Samy, and N. Aziz, "Factors Influencing Blockchain Adoption in Government Organization : A Proposed Framework," in Proceedings - International Conference on Computer and Information Sciences: Sustaining Tomorrow with Digital Innovation, ICCOINS 2021, Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 366–371. doi: 10.1109/ICCOINS49721.2021.9497196.

[16] J. Guo, Y. Bai, S. Sun, F. Liu, and J. Yang, "E-government System Based on Blockchain," in 2023 3rd International Conference on Digital Society and Intelligent Systems, DSInS 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 449–452. doi: 10.1109/DSInS60115.2023.10455503.

[17] Unit Permodenan Tadbiran dan Perancangan Pengurusan Malaysia (MAMPU), Dasar Perkongsian Data Sektor Awam. 2021. Accessed: Sep. 06, 2023. [Online]. Available: https://dasar.mampu.gov.my/search-g/download-file/258/e8542746051d6cfbc33a61bd16be0ee7

[18] Parliament of Malaysia, Act 864 - DATA SHARING ACT 2025. 2025. Accessed: Mar. 04, 2025. [Online]. Available: https://lom.agc.gov.my/ilims/upload/portal/akta/outputaktap/2687068_BI/Act%20864%20-%20DATA%20SHARING%20ACT%202025.pdf

[19] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," Mar. 29, 2021, BMJ Publishing Group. doi: 10.1136/bmj.n71.

[20] V. Buterin, "Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform.," 2014. Accessed: Feb. 23, 2025. [Online]. Available: https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf

[21] Chong Shen, Kun Zhang, and Keliu Long, "Research on Hainan Trusted Digital Infrastructure Construction Framework," in 29th Wireless and Optical Communications Conference (WOCC), IEEE, 2020.

[22] L. Cai and S. Yao, "Application of big data technology in blockchain computing," in 2nd International Conference on Artificial Intelligence and Information Systems (ICAIIS '21), ACM, 2021. doi: 10.1145/3469213.

[23] D. Zhang, L. G. Pee, S. L. Pan, and L. Cui, "Big data analytics, resource orchestration, and digital sustainability: A case study of smart city development," Gov Inf Q, vol. 39, no. 1, Jan. 2022, doi: 10.1016/j.giq.2021.101626.

[24] H. Hou, "The application of blockchain technology in E-government in China," in 26th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2017.

[25] J. Bao, X. Geng, and P. Yu, "An digital economy mode based on blockchain," in Proceedings - 2020 International Conference on Robots and Intelligent Systems, ICRIS 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 391–394. doi: 10.1109/ICRIS52159.2020.00102.

[26] G. Huang, D. Li, S. T. Ng, L. Wang, and T. Wang, "A methodology for assessing supply-demand matching of smart government services from citizens' perspective: A case study in Nanjing, China," Habitat Int, vol. 138, Aug. 2023, doi: 10.1016/j.habitatint.2023.102880.

[27] L. Liu, C. Piao, X. Jiang, and L. Zheng, "Research on Governmental Data Sharing Based on Local Differential Privacy Approach," in Proceedings - 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 39–45. doi: 10.1109/ICEBE.2018.00017.

[28] Y. Zhang, S. Deng, Y. Zhang, and J. Kong, "Research on government information sharing model using blockchain technology," in Proceedings - 10th International Conference on Information Technology in Medicine and Education, ITME 2019, Institute of Electrical and Electronics Engineers Inc., Aug. 2019, pp. 726–729. doi: 10.1109/ITME.2019.00166.

[29] C. Xu, H. Yang, Q. Yu, and Z. Li, "Trusted and Flexible Electronic Certificate Catalog Sharing System Based on Consortium Blockchain," in

IEEE 5th International Conference on Computer and Communications (ICCC), IEEE, 2019.

[30] M. Sun and J. Zhang, "Research on the application of block chain big data platform in the construction of new smart city for low carbon emission and green environment," Comput Commun, vol. 149, pp. 332–342, Jan. 2020, doi: 10.1016/j.comcom.2019.10.031.

[31] H. Zhang and W. Liu, "Research on the Application of Blockchain in the Safe and Trusted Sharing of Government Data," in Proceedings - 2023 2nd International Conference on Artificial Intelligence and Blockchain Technology, AIBT 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 52–55. doi: 10.1109/AIBT57480.2023.00017.

[32] A. U. Khan, Z. Zhang, M. Taleby Ahvanooey, and W. Rafique, "Opinion mining towards blockchain technology adoption for accessing digital library resources," Aslib Journal of Information Management, vol. 74, no. 1, pp. 135–157, Jan. 2022, doi: 10.1108/AJIM-01-2021-0016.

[33] S. S. Kamble, A. Gunasekaran, and R. Sharma, "Modeling the blockchain enabled traceability in agriculture supply chain," Int J Inf Manage, vol. 52, Jun. 2020, doi: 10.1016/j.ijinfomgt.2019.05.023.

[34] A. Patil, V. Shardeo, A. Dwivedi, and J. Madaan, "An integrated approach to model the blockchain implementation barriers in humanitarian supply chain," Journal of Global Operations and Strategic Sourcing, vol. 14, no. 1, pp. 81–103, Mar. 2021, doi: 10.1108/JGOSS-07-2020-0042.

[35] A. Kumar, "Improvement of public distribution system efficiency applying blockchain technology during pandemic outbreak (COVID-19)," Journal of Humanitarian Logistics and Supply Chain Management, vol. 11, no. 1, pp. 1–28, Feb. 2021, doi: 10.1108/JHLSCM-06-2020-0050.

[36] S. Bakare, S. C. Shinde, and R. Hubballi, "A Blockchain Framework for secure Digital Identity transactions in Indian Agri-subsidy system: Issues, challenges and benefits," in Proceedings of the 2021 4th International Conference on Computing and Communications Technologies, ICCCT 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 138–143. doi: 10.1109/ICCCT53315.2021.9711868.

[37] A. Jain, D. Desai, and R. Sangole, "A Case Study on Use of Blockchain Technology as a Dominant Feature to Issue and Verify Documents Required for Admission to UG/PG Technical Programs in Maharashtra (India)," in 2022 IEEE Pune Section International Conference, PuneCon 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/PuneCon55413.2022.10014917.

[38] S. Panchamia and D. K. Byrappa, "Passport, VISA and Immigration Management Using Blockchain," in Proceedings - 23rd Annual Conference on Advanced Computing and Communications, ADCOM 2017, Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 8–17. doi: 10.1109/ADCOM.2017.00009.

[39] N. P. Rana, Y. K. Dwivedi, and D. L. Hughes, "Analysis of challenges for blockchain adoption within the Indian public sector: an interpretive structural modelling approach," Information Technology and People, vol. 35, no. 2, pp. 548–576, Mar. 2022, doi: 10.1108/ITP-07-2020-0460.

[40] M. Bakhtina, R. Matulevičius, A. Awad, and P. Kivimäki, "On the Shift to Decentralised Identity Management in Distributed Data Exchange Systems," in Proceedings of the ACM Symposium on Applied Computing, Association for Computing Machinery, Mar. 2023, pp. 864–873. doi: 10.1145/3555776.3577678.

[41] S. Bali, V. Bali, R. P. Mohanty, and D. Gaur, "Analysis of critical success factors for blockchain technology implementation in healthcare sector," Benchmarking, vol. 30, no. 4, pp. 1367–1399, Apr. 2023, doi: 10.1108/BIJ-07-2021-0433.

[42] S. Ølnes and Arild. Janssen, "Blockchain Technology as Infrastructure in Public Sector – an Analytical Framework," in Norwegian Conference on ICT in the public sector (NOKIOS), ACM, 2018.

[43] S. N. Khan, M. Shael, and M. Majdalawieh, "Blockchain technology as a support infrastructure in E-Government evolution at Dubai economic department," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2019, pp. 124–130. doi: 10.1145/3343147.3343164.

[44] A. Farouk, A. Alahmadi, S. Ghose, and A. Mashatan, "Blockchain platform for industrial healthcare: Vision and future opportunities," Mar. 15, 2020, Elsevier B.V. doi: 10.1016/j.comcom.2020.02.058.

[45] H. Scholta, W. Mertens, M. Kowalkiewicz, and J. Becker, "From one-stop shop to no-stop shop: An e-government stage model," Gov Inf Q, vol. 36, no. 1, pp. 11–26, Jan. 2019, doi: 10.1016/j.giq.2018.11.010.

[46] A. Alahbabi, K. Alshehhi, A. Albloushi, and K. Shuaib, "Establishing Security Controls For Blockchain Technology In P2P Energy Trading," in 2023 IEEE PES Conference on Innovative Smart Grid Technologies - Middle East, ISGT Middle East 2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISGTMiddleEast56437.2023.10078508.

[47] O. Hujran, M. M. Al-Debei, A. S. Al-Adwan, A. Alarabiat, and N. Altarawneh, "Examining the antecedents and outcomes of smart government usage: An integrated model," Gov Inf Q, vol. 40, no. 1, Jan. 2023, doi: 10.1016/j.giq.2022.101783.

[48] G. Spiga, S. Z. Ahmad, and W. Yeoh, "Blockchain adoption impact on real estate performance: the mediating role of real estate and blockchain transparency," Business Process Management Journal, 2024, doi: 10.1108/BPMJ-09-2023-0701.

[49] M. F. Körner, J. Sedlmeir, M. Weibelzahl, G. Fridgen, M. Heine, and C. Neumann, "Systemic risks in electricity systems: A perspective on the potential of digital technologies," Energy Policy, vol. 164, May 2022, doi: 10.1016/j.enpol.2022.112901.

[50] L. Klug and W. Prinz, "Fair prices for sustainability in agriculture and food. Requirements and design options for a data-based transparency system.," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2023, pp. 496–507. doi: 10.1145/3598469.3598525.

[51] B. Rukanova et al., "A Framework for Understanding Circular Economy Monitoring: Insights from the Automotive Industry," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2023, pp. 544–555. doi: 10.1145/3598469.3598530.

[52] B. Rukanova et al., "Public value creation through voluntary business to government information sharing enabled by digital infrastructure innovations: a framework for analysis," Gov Inf Q, vol. 40, no. 2, Apr. 2023, doi: 10.1016/j.giq.2022.101786.

[53] M. Brinkmann and M. Heine, "The Implementation of New Public Governance Through Blockchain: A Delphi-based analysis," in ACM International Conference Proceeding Series, Association for Computing Machinery, Oct. 2022, pp. 1–9. doi: 10.1145/3560107.3560108.

[54] L. M. Baltruschat, V. Jaiman, and V. Urovi, "User acceptability of blockchain technology for enabling electronic health record exchange," Journal of Systems and Information Technology, vol. 25, no. 3, pp. 268–295, 2023, doi: 10.1108/JSIT-09-2022-0225.

[55] I. Azogu, A. Norta, I. Papper, J. Longo, and D. Draheim, "A framework for the adoption of blockchain technology in healthcare information management systems: A case study of Nigeria," in ACM International Conference Proceeding Series, Association for Computing Machinery, 2019, pp. 310–316. doi: 10.1145/3326365.3326405.

[56] D. W. Bates, A. Heitmueller, M. Kakad, and S. Saria, "Why policymakers should care about 'big data' in healthcare," Health Policy Technol, vol. 7, no. 2, pp. 211–216, Jun. 2018, doi: 10.1016/j.hlpt.2018.04.006.

[57] G. Tshering and S. Gao, "Understanding security in the government's use of blockchain technology with value focused thinking approach," Journal of Enterprise Information Management, vol. 33, no. 3, pp. 519–540, Apr. 2020, doi: 10.1108/JEIM-06-2018-0138.

[58] T. Ly. Charles and A. A. Azeta, "A Blockchain Framework for Securing Examination Results in Higher Institutions of Learning," in International Conference on Science, Engineering and Business for Driving Sustainable Development Goals, SEB4SDG 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/SEB4SDG60871.2024.10630246.

[59] A. Ahmad, M. Saad, M. Bassiouni, and A. Mohaisen, "Towards Blockchain-Driven, Secure and Transparent Audit Logs," in Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, in MobiQuitous '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 443–448. doi: 10.1145/3286978.3286985.

[60] C. M. Adilieme, R. B. Abidoye, and C. L. Lee, "Barriers and prospects for the adoption of blockchain technology in property valuation," Journal

of European Real Estate Research, 2024, doi: 10.1108/JERER-04-2024-0022.

[61] A. Shahaab, I. A. Khan, R. Maude, C. Hewage, and Y. Wang, "Public service operational efficiency and blockchain – A case study of Companies House, UK," Gov Inf Q, vol. 40, no. 1, Jan. 2023, doi: 10.1016/j.giq.2022.101759.

[62] M. A. Sicilia and A. Visvizi, "Blockchain and OECD data repositories: opportunities and policymaking implications," Library Hi Tech, vol. 37, no. 1, pp. 30–42, Mar. 2019, doi: 10.1108/LHT-12-2017-0276.

[63] E. A. Akartuna, S. D. Johnson, and A. Thornton, "Preventing the money laundering and terrorist financing risks of emerging technologies: An international policy Delphi study," Technol Forecast Soc Change, vol. 179, Jun. 2022, doi: 10.1016/j.techfore.2022.121632.

[64] A. Azmi, F. Yahya, E. G. Moung, H. Sallehudin, R. G. Utomo, and N. A. Azman, "Blockchain-based Data Sharing Framework for Malaysia Government Aid Management System," in 2023 International Conference on Digital Applications, Transformation and Economy, ICDATE 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICDATE58146.2023.10248471.

[65] K. Biswas and V. Muthukkumarasamy, "Securing Smart Cities Using Blockchain Technology," 2016, doi: 10.1109/HPCC-SmartCity-DSS.2016.178.

[66] E. Acosta Llano, P. Hurmelinna-Laukkanen, and L. Haapanen, "Blockchain for the circular economy, implications for public governance," International Journal of Public Sector Management, 2024, doi: 10.1108/IJPSM-12-2023-0365.

[67] P. Svärd and E. Borglund, "The implementation of an e-archive to facilitate open data publication and the use of common specifications: A case of three Swedish agencies," Gov Inf Q, vol. 39, no. 4, Oct. 2022, doi: 10.1016/j.giq.2022.101751.

[68] H. J. Rim, "Decentralized identity (DID): new technology adoption and diffusion in South Korea," Transforming Government: People, Process and Policy, vol. 17, no. 2, pp. 251–270, Apr. 2023, doi: 10.1108/TG-11-2021-0189.

[69] D. S. K. Putra and A. Alfari, "IDNat-Blockchain: A Concept for Indonesia's National Blockchain," in Proceeding - 2021 2nd International Conference on ICT for Rural Development, IC-ICTRuDev 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/IC-ICTRuDev50538.2021.9656496.

[70] A. Ghaffar and M. Hussain, "BCEAP – A blockchain embedded academic paradigm to augment legacy education through application," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2019. doi: 10.1145/3341325.3342036.

[71] P. Baudier, G. Kondrateva, C. Ammi, and E. Seulliet, "Peace engineering: The contribution of blockchain systems to the e-voting process," Technol Forecast Soc Change, vol. 162, Jan. 2021, doi: 10.1016/j.techfore.2020.120397.

[72] T. Nodehi, A. Zutshi, A. Grilo, and B. Rizvanovic, "EBDF: The enterprise blockchain design framework and its application to an e-Procurement ecosystem," Comput Ind Eng, vol. 171, Sep. 2022, doi: 10.1016/j.cie.2022.108360.

[73] S. Ciliberti et al., "EUTR implementation in the Italian wood-energy sector: Role and impact of (ongoing) digitalisation," Aug. 01, 2022, Elsevier B.V. doi: 10.1016/j.forpol.2022.102758.

[74] J. S. Søgaard, "A blockchain-enabled platform for VAT settlement," International Journal of Accounting Information Systems, vol. 40, Mar. 2021, doi: 10.1016/j.accinf.2021.100502.

[75] F. E. Tanha, A. Hasani, S. Hakak, and T. R. Gadekallu, "Blockchain-based cyber physical systems: Comprehensive model for challenge assessment," Computers and Electrical Engineering, vol. 103, Oct. 2022, doi: 10.1016/j.compeleceng.2022.108347.

[76] A. Alkhaldi, H. Alrashidi, K. Alhasan, A. Alsadeeqi, and A. Alshami, "The use of blockchain technology to build smart cities: creating public value in Kuwait," Global Knowledge, Memory and Communication, 2023, doi: 10.1108/GKMC-11-2022-0263.

[77] A. H. Dbesan, A. A. Abdulmuhsin, and A. F. Alkhwaldi, "Adopting knowledge-sharing-driven blockchain technology in healthcare: a developing country's perspective," VINE Journal of Information and Knowledge Management Systems, 2023, doi: 10.1108/VJIKMS-01-2023-0021.

[78] L. Cai and S. Yao, "Application of big data technology in blockchain computing," in 2021 2nd International Conference on Artificial Intelligence and Information Systems, in ICAIIS 2021. New York, NY, USA: Association for Computing Machinery, 2021. doi: 10.1145/3469213.3470409.

[79] T. Chen, Y. Yu, Z. Duan, J. Gao, and K. Lan, "BlockChain/ABE-based Fusion Solution for E-government Data Sharing and Privacy protection," in ACM International Conference Proceeding Series, Association for Computing Machinery, Nov. 2020, pp. 258–264. doi: 10.1145/3443467.3443764.

[80] S. T. Jagtap, C. M. Thakar, O. El Imrani, K. Phasinam, S. Garg, and R. J. M. Ventayen, "A Framework for Secure Healthcare System Using Blockchain and Smart Contracts," in Proceedings of the 2nd International Conference on Electronics and Sustainable Communication Systems, ICESC 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 922–926. doi: 10.1109/ICESC51422.2021.9532644.

[81] K. Kumutha and S. Jayalakshmi, "Hyperledger Fabric Blockchain Framework: Efficient Solution for Academic Certificate Decentralized Repository," in Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1584–1590. doi: 10.1109/I-SMAC52330.2021.9640785.

[82] S. Singh, S. Rathore, O. Alfarraj, A. Tolba, and B. Yoon, "A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology," Future Generation Computer Systems, vol. 129, pp. 380–388, Apr. 2022, doi: 10.1016/j.future.2021.11.028.

[83] Y. Li, M. K. Lim, and C. Wang, "An intelligent model of green urban distribution in the blockchain environment," Resour Conserv Recycl, vol. 176, Jan. 2022, doi: 10.1016/j.resconrec.2021.105925.

[84] S. Saxena, D. Shao, A. Nikiforova, and R. Thapliyal, "Invoking blockchain technology in e-government services: a cybernetic perspective," Digital Policy, Regulation and Governance , vol. 24, no. 3, pp. 246–258, Jun. 2022, doi: 10.1108/DPRG-10-2021-0128.

[85] J. Xu, C. Hua, and Y. Zhang, "A Blockchain-Based Framework for Supervision of Livelihood Issues: Proof of Concept With Optimized Consensus," IEEE Access, vol. 11, pp. 73414–73434, 2023, doi: 10.1109/ACCESS.2023.3295696.

[86] K. Shruthi and A. S. Poornima, "A Transparent and Privacy-Preserving Job Search Platform Built on the Ethereum Blockchain Framework," in Proceedings of 2023 IEEE 2nd International Conference on Industrial Electronics: Developments and Applications, ICIDeA 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 235–242. doi: 10.1109/ICIDeA59866.2023.10295065.

[87] P. Rede, S. Iyer, S. Sharma, and S. Deshmukh, "Blockchain Based Identity Management System Using Cryptography and Steganography," in 2023 International Conference on Information Technology: Cybersecurity Challenges for Sustainable Cities, ICIT 2023 - Proceeding, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 173–177. doi: 10.1109/ICIT58056.2023.10225957.

[88] S. Peng, D. Sun, L. Zhu, H. Zhou, X. Zhang, and C. Cui, "Enhancing Cross-Border Data Sharing in Blockchain Networks: A Compliance-Centric Approach Ensuring Anonymity and Traceability," in 2023 3rd International Conference on Computer Science and Blockchain, CCSB 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 200–204. doi: 10.1109/CCSB60789.2023.10398873.

[89] J. Du, Z. Xu, D. Liu, X. Zhang, and H. Cheng, "EPT: Enhancing User Transparency for Confidential Smart Contract," in Proceedings - 2023 19th International Conference on Mobility, Sensing and Networking, MSN 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 431–438. doi: 10.1109/MSN60784.2023.00069.

[90] S. Siddiqui, S. Hameed, S. A. Shah, A. K. Khan, and A. Aneiba, "Smart contract-based security architecture for collaborative services in municipal smart cities[Formula presented]," Journal of Systems Architecture, vol. 135, Feb. 2023, doi: 10.1016/j.sysarc.2022.102802.

[91] S. Capraz and A. Ozsoy, "A Secure Medical Data Sharing Framework for Fight Against Pandemics Like Covid-19 by Using Public Blockchain," IEEE Access, vol. 12, pp. 93593–93605, 2024, doi: 10.1109/ACCESS.2024.3423714.

[92] R. Raj, A. Singh, V. Kumar, and P. Verma, "Challenges in adopting blockchain technology in supply chain management: a too far fetched idea?," International Journal of Quality and Reliability Management, vol. 41, no. 8, pp. 2146–2180, Sep. 2024, doi: 10.1108/IJQRM-12-2022-0366.

[93] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," Future Generation Computer Systems, vol. 88, pp. 173–190, Nov. 2018, doi: 10.1016/j.future.2018.05.046.

[94] Deepayan Bhowmik, Ambarish Natu, Takaaki Ishikawa, Tian Feng, and Charith Abhayaratne, "THE JPEG-BLOCKCHAIN FRAMEWORK FOR GLAM SERVICES," in IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018.

[95] W. Fan, H. J. Hong, X. Zhou, and S. Y. Chang, "A Generic Blockchain Framework to Secure Decentralized Applications," in IEEE International Conference on Communications, Institute of Electrical and Electronics Engineers Inc., Jun. 2021. doi: 10.1109/ICC42927.2021.9500924.

[96] D. Sinha and S. Roy Chowdhury, "Blockchain-based smart contract for international business – a framework," Journal of Global Operations and Strategic Sourcing, vol. 14, no. 1, pp. 224–260, Mar. 2021, doi: 10.1108/JGOSS-06-2020-0031.

[97] C. Piao, Y. Hao, J. Yan, and X. Jiang, "Privacy preserving in blockchain-based government data sharing: A Service-On-Chain (SOC) approach," Inf Process Manag, vol. 58, no. 5, Sep. 2021, doi: 10.1016/j.ipm.2021.102651.

[98] F. H. L. Chong, "Enhancing trust through digital Islamic finance and blockchain technology," Qualitative Research in Financial Markets, vol. 13, no. 3, pp. 328–341, 2021, doi: 10.1108/QRFM-05-2020-0076.

[99] M. Elbialy, M. A. Elsalam, and S. A. El-Fotouh, "A Framework for Enhancing the Supply of Scientific Research Projects Using Blockchain," in 5th International Conference on Computing and Informatics, ICCI 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 140–147. doi: 10.1109/ICCI54321.2022.9756099.

[100] J. Wang, H. Jin, J. Chen, J. Tan, and K. Zhong, "Anomaly detection in Internet of medical Things with Blockchain from the perspective of deep neural network," Inf Sci (N Y), vol. 617, pp. 133–149, Dec. 2022, doi: 10.1016/j.ins.2022.10.060.

[101] C. Turhan and I. Akman, "Exploring sectoral diversity in the timing of organizational blockchain adoption," Information Technology and People, vol. 35, no. 7, pp. 1912–1930, Dec. 2022, doi: 10.1108/ITP-05-2020-0330.

[102] K. Saurabh, P. Upadhyay, and N. Rani, "A study on blockchain-based marketplace governance platform adoption: a multi-industry perspective," Nov. 09, 2023, Emerald Publishing. doi: 10.1108/DPRG-04-2023-0053.

[103] W. Shang and Z. Yu, "A new media content trusted dissemination architecture based on AV-blockchain and ChinaDRM," Intelligent and Converged Networks, vol. 4, no. 2, pp. 142–157, Jun. 2023, doi: 10.23919/ICN.2023.0015.

[104] V. Malik et al., "Building a Secure Platform for Digital Governance Interoperability and Data Exchange Using Blockchain and Deep Learning-Based Frameworks," IEEE Access, vol. 11, pp. 70110–70131, 2023, doi: 10.1109/ACCESS.2023.3293529.

[105] C. B. Basha et al., "Fostering Effective Cyber Threat Intelligence Sharing: Overcoming Challenges and Implementing Best Practices," in International Conference for Technological Engineering and its Applications in Sustainable Development, ICTEASD 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 177–182. doi: 10.1109/ICTEASD57136.2023.10585133.

[106] M. Naudiyal et al., "Secure Blockchain-based Framework for Smart City Applications," in 2023 World Conference on Communication and Computing, WCONF 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/WCONF58270.2023.10235249.

[107] Y. Tian, B. Liu, Y. Li, P. Szalachowski, and J. Zhou, "Accountable Fine-Grained Blockchain Rewriting in the Permissionless Setting," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1756–1766, 2024, doi: 10.1109/TIFS.2023.3340917.

[108] T. Nhan, K. Upadhyay, and K. Poudel, "Towards Patient-Centric Healthcare: Leveraging Blockchain for Electronic Health Records," in SIGMIS-CPR 2024 - Proceedings of the Computers and People Research Conference: Trust and Legitimacy in Emerging Technologies: Organizational and Societal Implications for People, Places and Power, Association for Computing Machinery, Inc, May 2024. doi: 10.1145/3632634.3655883.

[109] L. Zavolokina, I. Bauer-Hänsel, J. Hacker, and G. Schwabe, "Organizing for value creation in blockchain information systems," Information and Organization, vol. 34, no. 3, Sep. 2024, doi: 10.1016/j.infoandorg.2024.100522.

[110] N. Goderdzishvili, E. Gordadze, and N. Gagnidze, "Georgia's blockchain-powered property registration: Never blocked, always secured - Ownership data kept best!," in ACM International Conference Proceeding Series, Association for Computing Machinery, Apr. 2018, pp. 673–675. doi: 10.1145/3209415.3209437.

[111] O. Konashevych, "Cross-blockchain protocol for public registries," International Journal of Web Information Systems, vol. 16, no. 5, pp. 571–610, Nov. 2020, doi: 10.1108/IJWIS-07-2020-0045.

[112] Y. Liu and J. Zeng, "Government Data Sharing based on Blockchain," in ACM International Conference Proceeding Series, Association for Computing Machinery, Mar. 2021, pp. 123–128. doi: 10.1145/3460537.3460562.

[113] S. Yu, T. Zhang, X. Huang, and X. Xia, "Design and Implementation of a Blockchain-Based Road Damage Approval System," in ACM International Conference Proceeding Series, Association for Computing Machinery, May 2022, pp. 94–103. doi: 10.1145/3538950.3538963.

[114] Z. Pan, D. Wu, and M. Liu, "Evaluation Study on the Level of Ecological Governance of Digital Intelligence in Beijing," in ACM International Conference Proceeding Series, Association for Computing Machinery, Sep. 2023, pp. 327–333. doi: 10.1145/3629378.3629400.

[115] Y. Du, J. Yuan, S. Q. Wang, Y. Liu, and N. Zeng, "Leveraging blockchain to anchor information for supervision in PPP projects: a conceptual framework," Engineering, Construction and Architectural Management, 2024, doi: 10.1108/ECAM-07-2023-0758.

[116] G. P. Andrade, J. C. A. de Abreu, and R. C. dos Santos, "The impact of blockchain on Brazilian public procurement processes from the perspective of transaction costs: scenarios as perceived by experts," International Journal of Organizational Analysis, 2024, doi: 10.1108/IJOA-07-2023-3829.

[117] N. Lalchandani, F. Jiang, J. J. Jeong, Y. Zolotavkin, and R. Doss, "Evaluating the Current State of Application Programming Interfaces for Verifiable Credentials," in 2021 18th International Conference on Privacy, Security and Trust, PST 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/PST52912.2021.9647805.

[118] S. Luthra, M. Janssen, N. P. Rana, G. Yadav, and Y. K. Dwivedi, "Categorizing and relating implementation challenges for realizing blockchain applications in government," Information Technology and People, vol. 36, no. 4, pp. 1580–1602, May 2023, doi: 10.1108/ITP-08-2020-0600.

[119] A. Aytekin et al., "A bipolar neutrosophic combined compromise solution-based hybrid model for identifying blockchain application barriers and Benchmarking consensus algorithms," Eng Appl Artif Intell, vol. 133, Jul. 2024, doi: 10.1016/j.engappai.2024.108343.

[120] M. J. M. Chowdhury, A. Colman, M. A. Kabir, J. Han, and P. Sarda, "Blockchain as a Notarization Service for Data Sharing with Personal Data Store," in Proceedings - 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, Trustcom/BigDataSE 2018, Institute of Electrical and Electronics Engineers Inc., Sep. 2018, pp. 1330–1335. doi: 10.1109/TrustCom/BigDataSE.2018.00183.

[121] Veselin Monev, "Measuring the Optimal Information Security Complexity for Blockchain Operations," in IEEE International Conference on Information Technologies (InfoTech-2020), IEEE, 2020.

[122] Y. Lin et al., "Power Data Blockchain Sharing Scheme based on Homomorphic Encryption," in IMCEC 2022 - IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 625–629. doi: 10.1109/IMCEC55388.2022.10020058.

[123] X. Yao, X. Cheng, Z. Shi, K. Liu, M. Liu, and Y. Li, "Using Blockchain to Share Data Securely in the IntelliSense Environment," in Proceedings

- 2022 International Conference on Blockchain Technology and Information Security, ICBCTIS 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 106–110. doi: 10.1109/ICBCTIS55569.2022.00035.

[124] W. Yang, Z. Guan, L. Wu, X. Du, and M. Guizani, "Secure Data Access Control with Fair Accountability in Smart Grid Data Sharing: An Edge Blockchain Approach," IEEE Internet Things J, vol. 8, no. 10, pp. 8632–8643, May 2021, doi: 10.1109/JIOT.2020.3047640.

[125] S. Ølnes, J. Ubacht, and M. Janssen, "Blockchain in government: Benefits and implications of distributed ledger technology for information sharing," Sep. 01, 2017, Elsevier Ltd. doi: 10.1016/j.giq.2017.09.007.

[126] D. B. Rawat and A. Alshaikhi, "Leveraging Distributed Blockchain-based Scheme for Wireless Network Virtualization with Security and QoS Constraints," 2018.

[127] G. Subramanian and A. S. Thampy, "Blockchain Consortium for Electric Vehicles to Enhance the Security," in 2022 International Conference for Advancement in Technology, ICONAT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICONAT53423.2022.9725887.

[128] F. Yahya, R. J. Walters, and G. B. Wills, "Goal-based security components for cloud storage security framework: A preliminary study," in 2016 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2016, Institute of Electrical and Electronics Engineers Inc., Jun. 2016. doi: 10.1109/CyberSecPODS.2016.7502338.

[129] F. Yahya, R. J. Walters, and G. B. Wills, "Using Goal-Question-Metric (GQM) approach to assess security in cloud storage," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2017, pp. 223–240. doi: 10.1007/978-3-319-54380-2_10.

[130] D. Benz, M. Hamzah, M. F. Ghazali, and M. F. Asli, "Bringing Blockchain Technology in Innovating Industries: A Systematic Review," in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2022, pp. 391–416. doi: 10.1007/978-3-030-85990-9_33.

Enhancing Firefighter PPE Compliance Through Deep Learning and Computer Vision

Asmaa Alayed, Razan Talal Alqurashi, Samah Hamoud Alhelali, Asrar Yousef Khadawurdi, Bashayer Fayez Khan
College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Ensuring firefighter safety in high-risk environments requires strict adherence to Personal Protective Equipment (PPE) protocols. This study presents an automated real-time detection system for PPE using deep learning and computer vision techniques, aiming to improve PPE compliance and overall safety monitoring. The research employs advanced object detection models, specifically YOLOv10 and YOLOv11 (You Only Look Once), to identify critical PPE components such as helmets, gloves, boots, and self-contained breathing apparatus (SCBA) units. A custom-annotated dataset of firefighter images was developed to train and evaluate both models using standard performance metrics such as precision, recall, mAP, F1-score, and Intersection over Union (IoU). The results show that YOLOv11 outperformed YOLOv10, achieving a higher mAP@0.5 score of 0.646 compared to 0.586, with improved detection of small and partially occluded objects and a reduction in training time by 11%. YOLOv11 showed improved detection accuracy for small and partially blocked objects and reduced training time by 11%, while maintaining real-time efficiency. The system generates instant alerts when PPE is missing, minimizing reliance on manual monitoring and improving situational awareness in real-time. This research reinforces the role of AI in public safety and AI-powered automation in enhancing critical public safety operations. By integrating deep learning and computer vision into PPE monitoring systems, the study contributes to developing intelligent, responsive solutions aligned with modern safety standards.**

Keywords*—*Firefighter safety; Personal Protective Equipment (PPE) Object detection; YOLOv10; YOLOv11; deep learning; computer vision; real-time detection; PPE compliance; AI in public safety

I. Introduction

Firefighting is a highly demanding and dangerous profession where the safety of personnel relies heavily on their ability to wear all necessary personal protective equipment (PPE) correctly. Essential equipment such as helmets, gloves, fire-resistant clothing, boots, and self-contained breathing apparatus (SCBA) shields firefighters from extreme heat, toxic gases, falling debris, and hazardous environments. Ensuring that this gear is worn properly is not only a matter of regulation but also a life-saving necessity.

However, during emergency responses, ensuring complete PPE compliance becomes difficult due to high-pressure conditions, time constraints, and human error. Manual checks are often rushed or incomplete, exposing firefighters to avoidable risks. Although safety standards exist in Saudi Arabia under the General Directorate of Civil Defense [1], ensuring real-time PPE compliance remains a challenge—especially before deployment into active zones.

Recent advances in artificial intelligence and computer vision have created new opportunities for enhancing safety protocols in the firefighting domain. Deep learning models—particularly object detection algorithms like YOLO—have shown impressive real-time performance in detecting and classifying objects across various sectors, including medical applications [2], construction [3], [4], and public safety systems [5]. These technologies offer a promising solution for automatically verifying the presence of PPE using image or video input prior to emergency deployment. This approach reduces reliance on manual inspection, increases accuracy, and enables faster decision-making in critical moments.

Despite existing research in PPE detection across various fields such as construction and industry, a significant research gap remains in applying AI-based systems to the firefighting sector—especially in Saudi Arabia. Most current solutions are limited to laboratory settings or narrow use cases, often lacking generalizability to real-world firefighting scenarios or focusing on only certain PPE components.

To address this gap, this paper presents an AI-driven solution specifically designed for firefighting services in Saudi Arabia. Leveraging YOLOv10 and YOLOv11 models, the system detects missing PPE components from real-time visual inputs, thereby enhancing emergency preparedness and reducing the risk of injury. This work offers a novel application of deep learning in the public safety domain by providing a practical, real-time solution for monitoring PPE compliance.

The remainder of this paper is organized as follows: Section II reviews recent developments in PPE detection using deep learning. Section III describes the data collection process and research methodology. Section IV presents the experimental results and performance evaluation. Section V provides a discussion of the findings, including a comparison with other models across various domains. Finally, Section VI concludes the study and outlines directions for future work.

II. Literature Review

Deep learning has demonstrated strong potential for automating PPE compliance monitoring in high-risk environments such as firefighting, construction, and industrial operations. These techniques aim to minimize human error by enabling real-time visual detection of safety equipment. In the firefighting domain, Sesis et al. [6] applied YOLOv5 with a custom, domain-specific dataset to detect four key PPE components including helmets, gloves, masks, and insulated

protective clothing. By using transfer learning, they achieved high detection accuracy, highlighting the model's suitability for emergency response scenarios.

Delhi et al. [3] applied YOLOv3 to construction sites, achieving 96% precision and recall in detecting hard hats and jackets, demonstrating its effectiveness in active work environments. Similarly, Isailovic et al. [7] benchmarked YOLOv5, MobileNetV2-SSD, and Faster R-CNN for detecting head-mounted PPE using a diverse dataset, where YOLOv5 achieved the best results. Wang et al. [4] introduced a real-world construction site dataset (CHV) and compared various YOLO models, showing YOLOv5x had the highest mAP (86.55%), while YOLOv5s demonstrated the fastest inference speed—highlighting the trade-off between accuracy and efficiency. In industrial settings, Chen and Demachi [5] combined OpenPose with YOLOv3 to monitor PPE at the Fukushima nuclear station, achieving high precision and recall for helmets and masks. In healthcare, Loey et al. [2] used YOLOv2 with ResNet-50 to detect medical masks, further illustrating the role of deep learning in public safety applications.

While previous studies demonstrate the potential of deep learning in PPE detection, they are largely confined to controlled or domain-specific environments. Therefore, this study seeks to extend his body of work and build upon these foundations by developing a real-time, AI-powered PPE compliance monitoring system specifically tailored to the operational needs of firefighters in Saudi Arabia. Comparison of studies is given in Table I.

TABLE I. COMPARISON OF STUDIES ON PPE DETECTION USING DEEP LEARNING MODELS

Study	Research Title	Focus	Algorithm	Performance	Dataset size and type
Sesis et al. [6]	AI-based Firefighting PPE Detection using YOLOv5	Detection of four Firefighting PPE (helmet, gloves, mask, and insulated protective clothing)	YOLOv5 with transfer learning	mAP@0.5 = 0.834, Precision = 0.914, Recall = 0.735	Custom dataset (342 samples), tailored to firefighting scenarios.
Delhi et al. [3]	Detection of PPE Compliance on Construction Sites Using YOLOv3	Construction site PPE (hard hats, jackets)	YOLOv3, transfer learning	Precision and recall: 96%	Self-collected dataset from video feeds, 2,509 images
Isailovic et al. [7]	The Compliance of Head-mounted Industrial PPE using Deep Learning Object Detectors	Industrial PPE, head-mounted	YOLOv5, MobileNetV2-SSD, Faster R-CNN	YOLOv5 achieved best precision (0.920), recall (0.611)	Public dataset (Roboflow, Pictor PPE), 12,682 images
Wang et al. [4]	Fast PPE Detection for Real Construction Sites Using Deep Learning Approaches	Construction site PPE (helmets, vests)	YOLOv5x (best mAP), YOLOv5s (fastest speed, 52 FPS)	YOLOv5x mAP: 86.55%, faster with YOLOv5s	Public dataset (CHV), 1,330 images, real construction site

III. DATA COLLECTION AND RESEARCH METHOD

A. Dataset Collection

High-quality image data is essential for effectively training deep learning models in firefighter PPE detection. Since no publicly available datasets met the specific requirements for detecting firefighter PPE in operational scenarios, we constructed a custom dataset tailored to this task. The initial internal dataset we collected exhibited limited variability and contained redundant samples, although the image clarity and framing were sufficient. To enhance diversity and address these limitations, we manually collected an additional 574 images from the official social media channels of the Saudi Civil Defense and other publicly available resources. All selected images adhered to the standards defined in the official firefighter PPE guidelines [1], ensuring accurate representation of required equipment.

The final dataset captures firefighters from varied angles, distances, and operational contexts to provide a broad visual representation of real-world scenarios. Further expansion of the dataset is planned to improve representation of diverse environmental conditions—particularly in low-light, crowded, or partially obstructed situations—to enhance the model's generalizability and robustness in real-world deployments.

B. Dataset Annotation

To enable accurate detection of PPE, all images in the constructed dataset were manually annotated by our team using bounding boxes to identify five critical equipment components: boots, fire suit, gloves, helmet, and SCBA. We used the Roboflow platform [8] to facilitate efficient and consistent annotation. Each bounding box precisely localized the corresponding PPE item within the image, allowing the model to recognize multiple equipment types simultaneously. Many images featured multiple equipment components, enabling the model to distinguish between them effectively. This annotated dataset provides a reliable foundation for training and evaluating detection models, ultimately supporting improved firefighter safety and system performance.

Representative annotation scenarios include a fully equipped firefighter (Fig. 1), absence of the SCBA (Fig. 2), and absence of both the SCBA and gloves (Fig. 3). These examples demonstrate the system's capacity to identify compliance gaps prior to deployment, thereby reducing risk and enhancing real-time situational awareness.

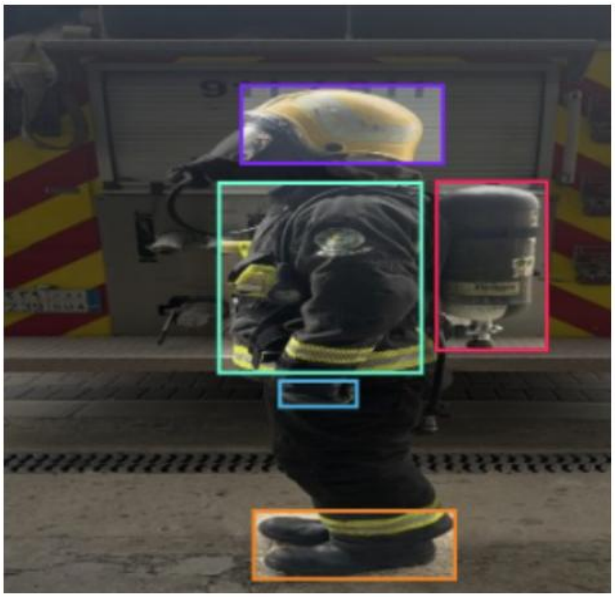


Fig. 1. A firefighter fully equipped with all necessary gear.



Fig. 2. A case of the absence of the SCBA.



Fig. 3. A case of the absence of the SCBA and gloves.

C. Dataset Augmentation

In object detection tasks, dataset quality and class balance are critical to model performance. Class imbalance—where certain PPE components are underrepresented—can lead to biased predictions. To address this, we employed data augmentation techniques to expand the dataset and mitigate class imbalance. We experimented with both generic and class-specific augmentation strategies to improve class distribution and data diversity. Despite these efforts, achieving perfect balance was challenging due to the inherent nature of the collected data. Class distributions for each dataset split are summarized below:

- Training Split: 465 images, 2,124 total objects
  - Boots: 314 (14.8%)
  - Fire Suit: 415 (19.5%)
  - Gloves: 283 (13.3%)
  - Helmet: 700 (33.0%)
  - SCBA: 412 (19.4%)
- Validation Split: 53 images, 250 total objects
  - Boots: 39 (15.6%)
  - Fire Suit: 48 (19.2%)
  - Gloves: 34 (13.6%)
  - Helmet: 79 (31.6%)
  - SCBA: 50 (20.0%)
- Test Split: 56 images, 283 total objects
  - Boots: 37 (13.1%)
  - Fire Suit: 59 (20.8%)
  - Gloves: 34 (12.0%)
  - Helmet: 91 (32.2%)
  - SCBA: 62 (21.9%)

Our augmentation pipeline improved data balance and diversity while maintaining high-quality, realistic examples. This enhanced the model's ability to generalize and detect all PPE components reliably across different scenarios.

D. YOLO Algorithms

Deep learning approaches in computer vision have yielded highly promising results across various applications. Among these, the YOLO algorithm has emerged as one of the most efficient and accurate methods for real-time object detection tasks, particularly in critical environments such as emergency response and industrial safety scenarios. YOLO operates by dividing an input image into a grid and performing detection in a single forward pass through a neural network. This design enables it to balance speed and accuracy, making it well-suited for real-time PPE inspection tasks.

YOLO has undergone several iterations over the years, with each version introducing architectural improvements that enhance performance. This study focuses on two of the latest versions—YOLOv10 and YOLOv11—selected for their advanced detection performance, computational efficiency, and applicability to safety-critical tasks [9], [10]. YOLOv10 is a recent advancement in the YOLO series, engineered to deliver high-speed detection with strong accuracy. The model is structured into three major components: a backbone network for feature extraction, a neck for generating multi-scale feature maps, and a detection head for object classification and localization. The architecture includes multiple variants designed to provide a flexible trade-off between inference speed and detection precision. In this research, YOLOv10 was adopted for its capability to reliably detect firefighting PPE—including helmets, gloves, boots, protective suits, and SCBA devices—in real-time. Additionally, transfer learning and data augmentation techniques were employed during training to improve robustness across various environmental conditions [9].

YOLOv11 represents the latest iteration in the YOLO family, incorporating state-of-the-art enhancements in feature extraction, model optimization, and real-time inference. Like its predecessor, YOLOv11 features a backbone, neck, and detection head; however, it introduces a more refined architecture with improved module design to achieve higher mAP while maintaining low latency. The model was selected in this study for its superior balance of accuracy and computational efficiency. YOLOv11 proved particularly effective in detecting multiple PPE components across a range of visual conditions, including occlusion, variable lighting, and different camera angles. These capabilities ensure reliable performance in operational environments where rapid and precise safety inspections are essential [10].

E. Training Methodology

The YOLOv10 and YOLOv11 models were trained using a custom pipeline designed to address class imbalance and enhance detection performance. Class weights were incorporated into the training configuration to ensure underrepresented categories like boots and gloves received higher attention. The training utilized the AdamW optimizer with a weight decay of 0.0005 and was conducted over 200 epochs with a batch size of 16.

A Cosine Learning Rate Scheduler was employed, starting at 0.001 and decaying to 0.000001, with a 15-epoch warmup phase. Data augmentation techniques such as Mosaic, Mixup, Copy-Paste, and geometric/colour transformations were applied to increase data diversity. Additionally, dropout (0.25) was used for regularization. To enhance accuracy and efficiency, strategies like multi-scale training, test-time augmentation (TTA), automatic mixed precision (AMP), and half-precision training were integrated. Early stopping was triggered after 150 epochs without validation improvement, and all results were logged for evaluation.

F. Training Environment

Training deep learning models for object detection requires substantial computational resources, particularly for tasks involving large datasets and complex architectures like YOLOv10 and YOLOv11. In this study, model training was initially conducted using Google Colab's free-tier environment, which offers access to an NVIDIA T4 GPU with 12 GB of VRAM. While suitable for early experimentation, the limited computational capacity resulted in slower training cycles. To overcome these limitations, the training environment was upgraded to Google Colab Pro, which provides access to more advanced GPUs such as the NVIDIA V100 and A100. This upgrade significantly improved the training process, enabling faster iterations, support for larger batch sizes, and greater stability during extended sessions. The combination of free and Pro-tier resources allowed for efficient and cost-effective training, supporting iterative development and comprehensive model evaluation.

G. Evaluation Metrics

To accurately evaluate the performance of the proposed PPE detection models for firefighters, a set of fundamental object detection metrics was utilized. These metrics were chosen to reflect the model's effectiveness in identifying and classifying multiple PPE components, including small and partially obscured items such as gloves or mask straps, which are critical in firefighting environments. The core evaluation indicators include Average Precision (AP) and mAP. These values are derived by comparing the model's predictions against the actual annotated labels in the dataset. AP is calculated as the area under the precision-recall (PR) curve for a given class, whereas mAP represents the average of AP across all detected PPE categories, as shown in Eq. (1) and Eq. (2).

$$AP = \int_0^1 p(r)\, dr \tag{1}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2}$$

Using the prediction results, three additional classification metrics were calculated:

- Precision, which quantifies the ratio of correctly predicted positive cases to the total predicted positives:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

- Recall, which measures the ratio of correctly predicted positives to all actual positives:

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

- F1-Score, the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

These metrics—Eq. (3), Eq. (4), and Eq. (5)—are critical in understanding the trade-off between missing detections and generating false alarms, particularly in safety-critical scenarios such as firefighting. Furthermore, IoU was used to evaluate the spatial accuracy of predicted bounding boxes. IoU calculates the overlap between the predicted and ground-truth boxes relative to their union, as expressed in Eq. (6):

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{6}$$

This metric is essential for assessing how well the model localizes PPE elements in complex visual scenes.

By leveraging this set of metrics, the evaluation framework ensures a comprehensive, balanced, and realistic analysis of model performance under the challenges posed by real-world firefighting environments.

IV. Results

A. Mean Average Precision and Model Size

YOLOv11 demonstrated a notable improvement in mAP compared to YOLOv10, making it more effective for detecting firefighter PPE in real-world conditions. Specifically, YOLOv11 achieved an mAP@0.5 of 64.6%, outperforming YOLOv10, which reached 58.6%, as shown in Fig. 4 and Fig. 5. This indicates YOLOv11's stronger ability to accurately detect PPE elements such as helmets, gloves, and SCBA gear. Additionally, despite YOLOv11 having a slightly larger architecture with 20 million parameters compared to YOLOv10's 16.5 million, its efficiency in detection accuracy justifies the additional model size, especially for safety-critical applications. YOLOv11 maintained a practical balance between model complexity and performance, which positions it as a more reliable choice for deployment in environments requiring high detection accuracy and consistency.

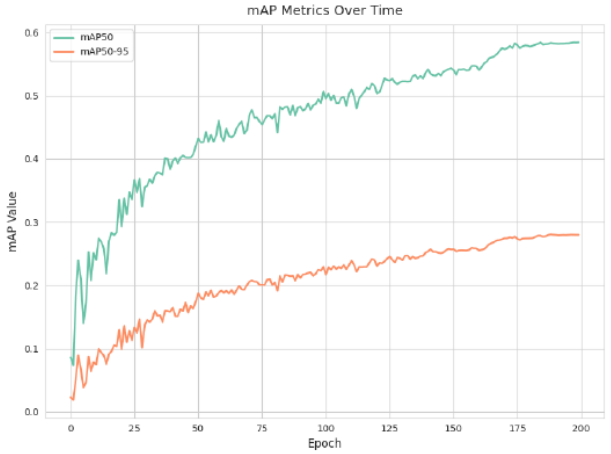


Fig. 4. YOLOv10: mAP@0.5 and mAP@0.5:0.95 over 200 training epochs.

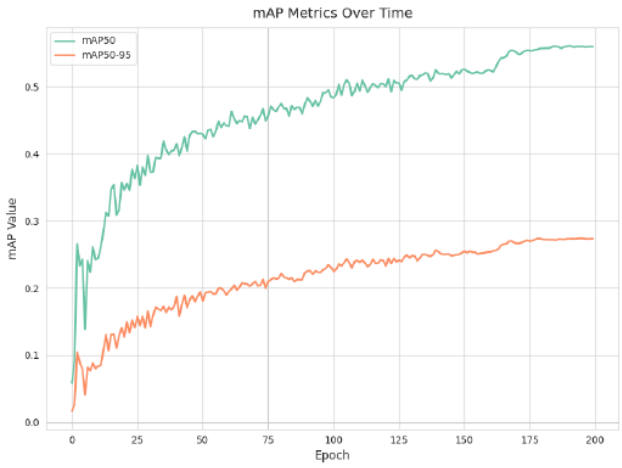


Fig. 5. YOLOv11: mAP@0.5 and mAP@0.5:0.95 over 200 training epochs.

Furthermore, YOLOv11 showed consistent improvement in detecting underrepresented PPE classes, particularly gloves and boots. This can be attributed to the refined data augmentation strategies and class-weighted training techniques, which addressed the earlier dataset imbalance. The better mAP results across all categories affirm the model's enhanced generalization ability in real-world, unstructured firefighting scenarios. YOLOv11 also demonstrated greater resilience under challenging visual conditions, such as varying lighting, partial occlusions, and dynamic backgrounds. Its performance remained stable in scenarios with overlapping gear or unconventional viewing angles, where YOLOv10 occasionally faltered. The enhanced feature extraction pipeline in YOLOv11 likely contributed to this robustness.

In practical deployment simulations, YOLOv11 also exhibited lower false negative rates, particularly in safety-critical gear detection. This reliability is crucial when monitoring firefighter readiness, as missing even one critical item (e.g., SCBA) can be life-threatening. The trade-off in model size for higher accuracy is justified in operational settings where real-time decisions depend on detection outcomes. Lastly, the precision-recall tradeoff and confidence calibration of YOLOv11 indicate that it is better suited for integration into alert systems. It provides more consistent outputs that can be used as triggers for decision-support tools or automated gatekeeping mechanisms in firefighting stations.

Collectively, these results underline that while both models are viable, YOLOv11 offers tangible advantages that support its deployment in real-world safety monitoring systems.

B. F1-Confidence, Precision-Confidence, PR Curve, and Recall-Confidence

These curves provide deeper insight into the reliability of both models across different confidence thresholds.

1) F1-Confidence: YOLOv11 consistently maintains a higher F1-score across all confidence levels. At the default threshold (0.25), it reached 0.74, compared to YOLOv10's 0.69. This indicates a better balance between precision and recall, especially in more uncertain detections.

2) Precision-Confidence: YOLOv11 sustains over 80\% precision until the threshold drops below 0.18, while YOLOv10's precision begins to degrade earlier, around 0.24. This shows YOLOv11's robustness in maintaining accuracy even with lower confidence predictions.

3) PR Curve: YOLOv11 achieved a noticeably higher area under the curve (AUC = 0.59) compared to YOLOv10 (AUC = 0.51). It maintains higher precision across a broader range of recall values, indicating more stable performance as the model tries to capture all ground truth objects, as shown in Fig. 6 and Fig. 7.

4) Recall-Confidence: YOLOv11 demonstrated higher recall values across confidence thresholds, especially for harder-to-detect classes like gloves. It achieved 0.37 recall for gloves at the 0.25 threshold, compared to YOLOv10's 0.29.

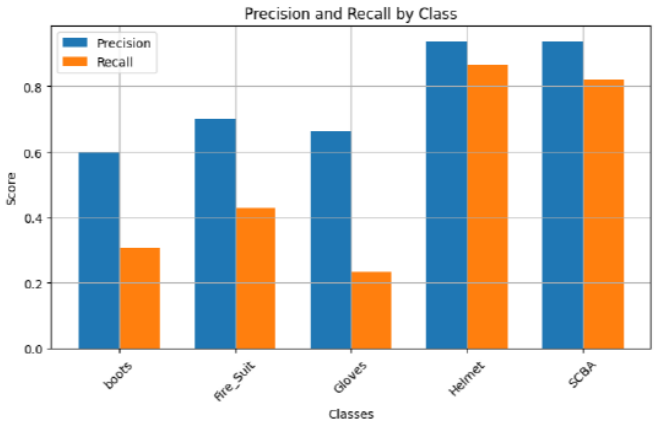


Fig. 6. YOLOv10 Precision and Recall by class.

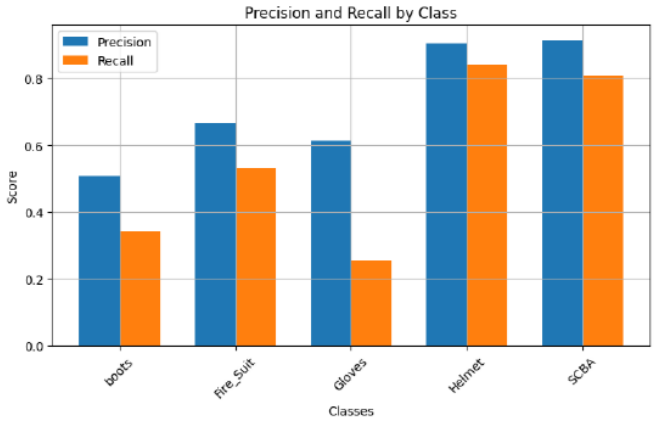


Fig. 7. YOLOv11 Precision and Recall by class.

C. IoU Performance Analysis

IoU was used to assess the spatial alignment between predicted and ground-truth bounding boxes across different PPE classes. The following breakdown highlights the performance of YOLOv10 and YOLOv11 in detecting key equipment components:

1) YOLOv10 IoU Performance:

a) Helmet: YOLOv10 achieved a mean IoU of 0.681, with 88.5% of predictions exceeding the 0.5 IoU threshold. This reflects reliable bounding box alignment for helmets, which are relatively easier to detect due to their distinct features.

b) Gloves: The mean IoU for gloves was 0.220, with only 28.1% of predictions above the 0.5 threshold. This highlights the model's struggle with accurately localizing smaller and less distinct objects like gloves.

As shown in Fig. 8, YOLOv10 demonstrated solid performance for prominent gear like helmets but had significant difficulty with smaller items such as gloves.

2) YOLOv11 IoU Performance:

a) Helmet: YOLOv11 recorded a mean IoU of 0.700, with 90% of predictions exceeding the 0.5 IoU threshold. This demonstrates improved spatial accuracy for helmet detection compared to YOLOv10.

b) Gloves: The mean IoU for gloves improved to 0.350, with better alignment accuracy. This reflects YOLOv11's enhanced capability in detecting and localizing smaller objects more precisely than YOLOv10.

Fig. 9 illustrates the improved bounding box accuracy of YOLOv11, particularly for challenging classes such as gloves, showing clear gains over its predecessor.

D. Speed

Speed is critical in real-time PPE detection systems for firefighting, where decisions must be made instantaneously. While YOLOv11 offers better accuracy, it operates slightly slower than YOLOv10, achieving 0.04 seconds per frame compared to YOLOv10's 0.033 seconds. This minor speed trade-off is acceptable given the improved detection precision and recall, which outweighs the marginal delay in processing.

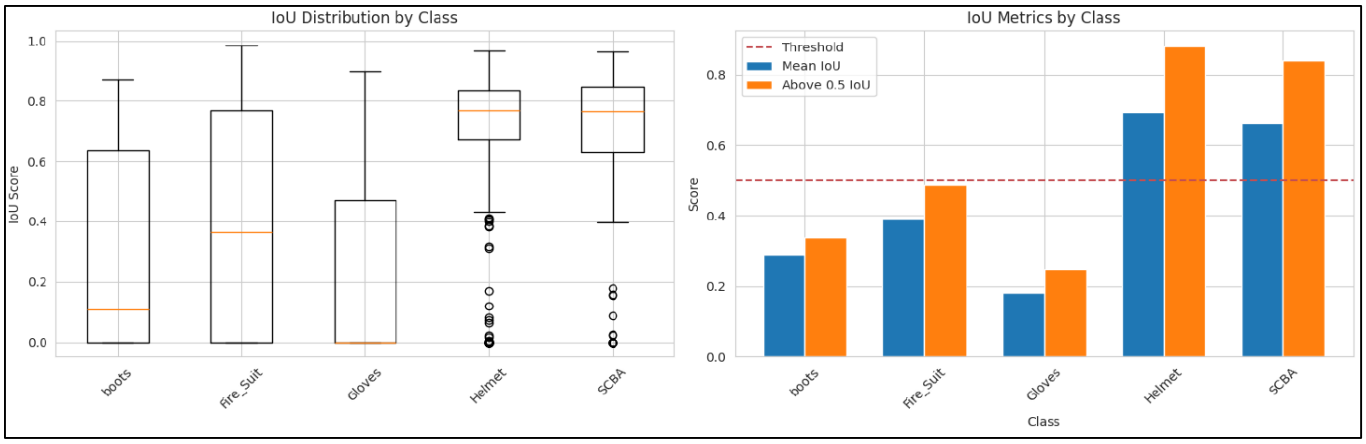


Fig. 8. IoU metrics for YOLOv10.

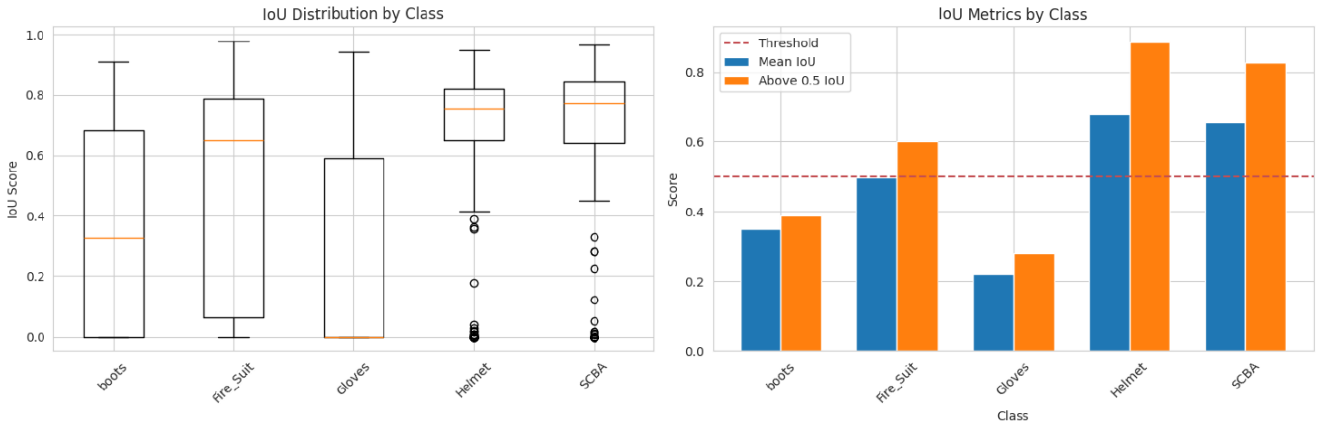


Fig. 9. IoU Metrics for YOLOv11.

V. Discussion

The results confirm that YOLOv11 significantly outperforms YOLOv10 in most evaluation metrics relevant to real-time PPE detection. This performance boost is particularly critical when working with safety-critical gear such as SCBA and gloves, which are harder to detect due to occlusion or small size. The increase in mAP, F1-score, and IoU, especially for underrepresented classes, reflects the effectiveness of class-weighted training strategies and augmentation pipelines tailored to firefighter environments. YOLOv11 also demonstrated greater consistency in handling complex backgrounds and overlapping equipment, suggesting its stronger generalization to real-world firefighting scenarios.

The detailed analysis reveals that while both models provide valuable performance, YOLOv11's architectural enhancements lead to improved detection of small and occluded PPE items. The incorporation of advanced augmentation and training strategies further boosted model robustness. However, increasing model complexity slightly impacts inference speed, a common trade-off in deep learning applications. As shown in Fig. 10 and Fig. 11, YOLOv11 was able to correctly recognize all critical equipment components, including the fire suit, gloves, boots, and helmet—demonstrating its superior reliability in safety-critical scenarios. In contrast, YOLOv10 failed to detect the fire suit, which is crucial for ensuring firefighter readiness and personal safety in hazardous conditions.

Moreover, the visual results affirm that YOLOv11 can identify all PPE components with high accuracy, making it suitable for integration into real-time alert systems or automated PPE compliance gates. This could reduce manual inspection errors and accelerate firefighter readiness verification.

Despite its slightly lower speed, YOLOv11's advantages in accuracy, localization, and class coverage outweigh the performance trade-off, making it more practical for deployment in Saudi Arabia's civil defense operations, where precision and speed are both mission-critical.

The proposed system, based on YOLOv10 and YOLOv11, is specifically optimized for image-based data in the visual modality, particularly RGB images captured in real-world firefighting environments. The models are best suited for high-resolution still images or video frames where PPE components are visible under variable lighting, motion blur, partial occlusions, or cluttered backgrounds. Due to their architecture, YOLO models excel in object detection tasks on spatial data with well-defined bounding box annotations. The system performs reliably on moderate-sized datasets (e.g., ~600–1,000 images), particularly when enhanced through data augmentation to account for class imbalance and variability in pose, angle, and context.

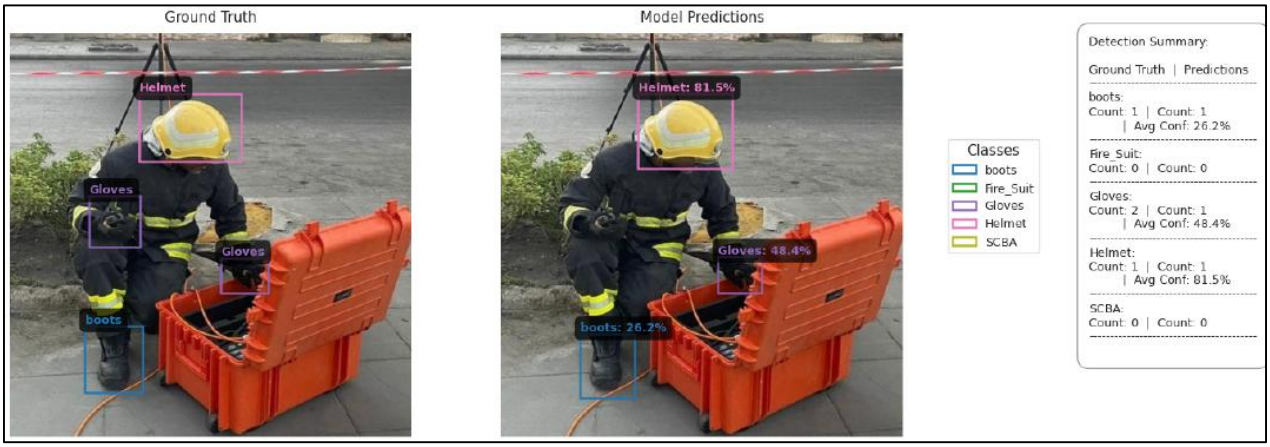


Fig. 10. YOLOv10 detection results: The fire suit was missed in the predictions.

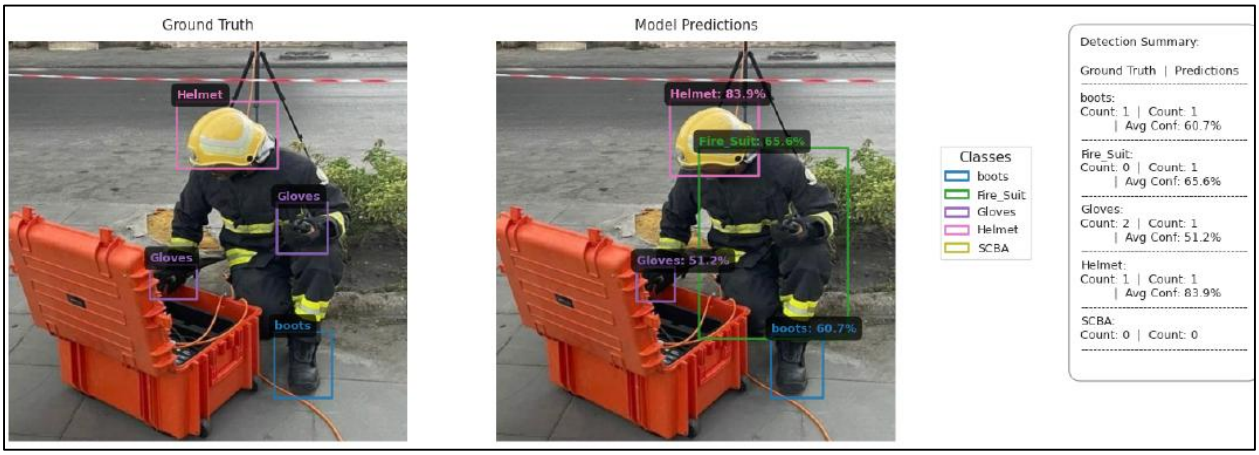


Fig. 11. YOLOv11 detection results: All PPE components including the fire suit were correctly detected.

However, the models are not directly applicable to non-visual modalities such as audio or text, and their performance may degrade with extremely low-resolution images, thermal images, or infrared unless retrained on domain-specific datasets. Additionally, performance is affected by data distribution. The model achieves its best results when trained on datasets that reflect the same environmental and cultural conditions (e.g., firefighter uniforms and gear used in Saudi Arabia) as the deployment context. This emphasizes the importance of domain-specific data for generalization and accuracy.

This study offers several key advantages over existing deep learning-based PPE detection models across construction, industrial, and healthcare domains:

1) Domain-specific adaptation to firefighting scenarios: Unlike previous research primarily focused on industrial or construction environments [3][4][7], the current work is specifically tailored to the unique challenges of firefighting, including variable lighting, smoke, occlusions, and high-stress environments. The model was trained on a custom dataset collected from real-world footage of Saudi Civil Defense operations, offering higher ecological validity than laboratory-based datasets.

2) Comprehensive multi-class PPE detection: Prior studies often target a limited subset of PPE items—typically one to four items, such as helmets or vests [4], [5], [6]. In contrast, our model simultaneously detects five critical PPE components: helmet, fire suit, gloves, boots, and SCBA. This multi-label detection approach ensures full compliance verification and strengthens overall situationalreadiness.

3) Improved detection for small and occluded items: YOLOv11 achieved significantly higher mAP, F1-score, and IoU, especially for underrepresented and smaller items like gloves and SCBA units, compared to YOLOv10 and prior benchmarks such as MobileNetV2-SSD and Faster R-CNN [7]. Our class-weighted training and targeted augmentation pipeline led to a 59% improvement in glove recall over YOLOv10.

4) Real-time monitoring capability with alert integration: While existing models may offer high accuracy, many lack real-time alerting or decision-support integration. Our proposed solution maintains an inference speed under 0.04 seconds per frame, enabling real-time detection and compatibility with automated gatekeeping or alarm systems at fire stations.

5) Environmental hazard awareness and long-term safety enhancements: Beyond immediate PPE detection, this study acknowledges long-term firefighter health risks, including exposure to harmful substances like PFAS (per- and polyfluoroalkyl substances), which are associated with cancer and immune system disorders. Our future vision includes integrating environmental hazard awareness to create a more holistic safety system that enhances both short-term operations and long-term health outcomes.

6) First deployment-ready system for Saudi civil defense: This work constitutes the first firefighter-specific PPE compliance monitoring system designed for potential deployment within Saudi Arabia. It establishes a foundation for future research and development in regional safety automation and may serve as a replicable framework for other high-risk sectors.

VI. Conclusion and Future Work

This study represents a notable advancement in enhancing firefighter safety by developing an AI-based system for PPE compliance. As the first implementation of its kind in Saudi Arabia, the proposed model integrates deep learning techniques with real-time computer vision to detect missing or improperly worn protective gear with high precision. This capability enables timely intervention in hazardous environments, potentially preventing serious injuries or fatalities.

By leveraging advanced YOLO architectures trained on a domain-specific dataset, the system demonstrates robust performance and reliability, setting a new benchmark for automated safety enforcement in high-risk operational contexts. The findings underscore the feasibility and effectiveness of applying artificial intelligence to critical occupational safety challenges, particularly in sectors where real-time monitoring and rapid decision-making are essential.

Beyond its technical contributions, this research illustrates the broader implications of AI-driven automation in promoting safer work environments. It highlights how intelligent systems

can augment human oversight and institutional safety protocols, contributing to a more proactive and data-informed approach to risk management. As such, the study provides a foundation for future exploration of similar technologies across other high-risk domains, including industrial operations, emergency response, and construction, where adherence to safety standards is equally vital.

- Future enhancements will focus on expanding dataset diversity through advanced augmentation methods and synthetic data generation, particularly to improve detection of underrepresented PPE classes such as gloves and boots. Model performance can be further optimized by leveraging transfer learning with pre-trained networks and incorporating adaptive training techniques for advanced detection use cases. Additionally, optimizing the system for deployment on edge devices by adopting lightweight architectures will facilitate broader scalability and real-time performance in operational settings.
- To extend system functionality, an automated timing feature will be implemented to monitor the duration firefighters take to fully equip themselves, providing valuable insights into response readiness. Finally, the system will be integrated with safety management platforms and wearable technologies to support comprehensive, real-time monitoring and decision-making in critical firefighting environments.

REFERENCES

[1] General Directorate of Civil Defense, Saudi Arabia, "Personal Protective Equipment Guide," https://998.gov.sa, 2017, accessed: 2024-10-14.

[2] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection," Sustainable Cities and Society, vol. 65, p. 102600, 2020.

[3] V. S. K. Delhi, R. Sankarlal, and A. Thomas, "Detection of personal protective equipment (ppe) compliance on construction site using computer vision based deep learning techniques," Frontiers in Built Environment, vol. 6, p. 136, 2020.

[4] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, "Fast personal protective equipment detection for real construction sites using deep learning approaches," Sensors, vol. 21, no. 10, p. 3478, 2021.

[5] S. Chen and K. Demachi, "A vision-based approach for ensuring proper use of personal protective equipment (ppe) in decommissioning of fukushima daiichi nuclear power station," Applied Sciences, vol. 10, no. 15, p. 5129, 2020.

[6] A. Sesis, I. Siniosoglou, Y. Spyridis, G. Efstathopoulos, T. Lagkas, V. Argyriou, and P. Sarigiannidis, "A robust deep learning architecture for firefighter ppes detection," IEEE Transactions on Industrial Informatics, 2022.

[7] V. Isailovic, A. Peulic, M. Djapan, M. Savkovic, and A. M. Vukicevic, "The compliance of head-mounted industrial ppe by using deep learning object detectors," Scientific Reports, vol. 12, no. 1, p. 16347, 2022.

[8] "Roboflow – Documentation," https://docs.roboflow.com/, 2024, accessed: 2025-01-27.

[9] "Ultralytics YOLOv10 Documentation," https://docs.ultralytics.com/models/yolov10/, 2024, accessed: 2025-01-27.

[10] "Ultralytics YOLOv11 Documentation," https://docs.ultralytics.com/models/yolo11/, 2024, accessed: 2025-01-27.

Multi-Agent Deep Reinforcement Learning Algorithms for Distributed Charging Station Management

Li Junda[1], Wang Tianan[2], Zhang Dingyi[3], Wu Quancai[4], Liu Jian[5]*
Yunnan Power Grid Energy Investment Co., Ltd., Kunming, Yunnan Province, 650000, China[1, 2 3, 4]
Sinotrans Logistics Southwest Co., Ltd., Chengdu, Sichuan Province, 610000, China[5]

***Abstract*—With the continued growth of the electric vehicle (EV) fleet, the issue of cross-regional coordinated scheduling for charging infrastructure has become increasingly prominent, facing challenges such as uneven resource allocation and delayed responses. Considering the complex coupling between charging stations and the power system in a smart grid environment, this paper proposes a distributed scheduling strategy based on multi-agent deep reinforcement learning (MADRL) to achieve efficient, coordinated management of charging infrastructure and power resources. The proposed approach constructs a hierarchical decision-making architecture to jointly optimize intra-regional resource allocation and cross-regional power support, modeling the scheduling process as a Markov Decision Process (MDP) and treating regional charging stations, power nodes, and material units as independent agents. Through the multi-agent deep reinforcement learning mechanism, each agent autonomously learns optimal scheduling policies in the presence of uncertain demand and supply fluctuations, thus enabling rapid response and enhancing system robustness. Simulation results demonstrate that the proposed method effectively reduces scheduling costs and improves resource utilization and service quality. This study provides both theoretical support and practical pathways for building intelligent, efficient, and sustainable charging infrastructure.**

Keywords*—Charging station scheduling; cross-regional coordination; multi-agent systems; deep reinforcement learning; Markov decision process; resource optimization; uncertainty response

I. Introduction

Driven by the global transition in energy structures and the rise of sustainable transportation strategies, the adoption of electric vehicles (EVs) is accelerating rapidly, making them a cornerstone of future low-carbon mobility. This trend places increasing demands on the large-scale deployment and intelligent management of charging infrastructure, especially charging stations. As the terminal nodes of the power grid, charging stations not only provide essential energy replenishment for EVs but also significantly influence grid security, traffic efficiency, and user experience through their operational efficiency, scheduling strategies, and resource allocation.

With the ongoing growth in the number of EVs, charging demands have become highly unbalanced across regions and exhibit strong spatiotemporal fluctuations. On one hand, certain areas have an abundance of charging stations but suffer from low utilization rates; on the other, core urban zones face resource shortages, severe queuing, and service bottlenecks due to concentrated demand. Furthermore, the operation of charging station networks is affected by multiple factors, including power supply capacity, maintenance resources, and traffic conditions, making cross-regional resource coordination and scheduling a pressing challenge.

Existing research has made significant progress in optimizing charging station scheduling, covering aspects such as queue management, load balancing, dynamic pricing, and energy storage integration. However, most approaches are still based on single-region, centralized architectures, lacking comprehensive consideration for cross-regional resource flows, heterogeneous system coordination, and dynamic uncertainties. Particularly in multi-region systems, the scheduling problem becomes highly complex and strongly coupled. Traditional optimization algorithms struggle to address challenges such as real-time requirements, large state spaces, and non-convex objectives, thus limiting their effectiveness in large-scale, practical deployments.

To address these challenges, this paper proposes a cross-regional coordinated charging station scheduling method based on multi-agent deep reinforcement learning (MADRL). In the proposed approach, each node in a multi-region charging station network is modeled as an autonomous agent. Through environmental perception and policy learning, agents achieve both local resource optimization and cross-regional coordination. Specifically, the scheduling problem is formulated as a Markov Decision Process (MDP), with state variables including charging demand, power supply, equipment status, and traffic accessibility. Each agent makes decisions independently based on local information and optimizes its policy collaboratively through a shared reward mechanism. This approach offers strosng generalization and adaptability, effectively coping with dynamic changes in charging demand and power supply, and overcomes the performance bottlenecks of traditional methods when faced with uncertainty and high-dimensional state spaces.

The main objective of this study is to build a distributed, intelligent, self-learning, and collaboratively optimized framework for cross-regional charging station resource scheduling. The aim is to maximize resource utilization, minimize operational costs, and optimize service

*Corresponding Author.

responsiveness. Compared with existing methods, the main innovations of this work include:

1) Proposing a coordinated scheduling modeling framework for cross-regional, multi-node, and multi-resource-type systems, adaptable to highly complex operational environments.

2) Introducing a multi-agent deep reinforcement learning mechanism to enable information exchange and policy collaboration among regional agents;

3) Designing a learning structure based on state sharing and reward feedback to enhance responsiveness to charging demand uncertainties and power resource fluctuations;

4) Validating the proposed method through simulation experiments, demonstrating significant advantages in scheduling efficiency, resource utilization, and system robustness.

This research is of both theoretical and practical significance. On one hand, it offers a novel modeling and solution paradigm for charging infrastructure scheduling, broadening the application scope of multi-agent systems in the integration of transportation and energy. On the other hand, it provides effective technical support for regional resource coordination, optimized energy allocation, and improved user charging experience—contributing to the development of smart cities, green transportation, and regional energy collaboration.

The remainder of this paper is organized as follows: Section II reviews the progress in charging station scheduling and the application of multi-agent reinforcement learning in energy management; Section III presents the proposed system modeling methodology and the multi-agent learning framework is given in Section IV; Section V conducts simulation experiments and performance evaluations based on a multi-region charging network; and Section VI concludes the study and outlines future research directions.

II. Literature Review

A. Techniques Based on Acoustic Features Multi-Agent Reinforcement Learning Algorithm Development

According to different decision-making paradigms, multi-agent reinforcement learning (MARL) algorithms for distributed charging scheduling can be categorized into value-based methods and policy-based methods.

For value-based methods, agents focus on learning value functions to derive optimal strategies. In cooperative MARL for distributed charging networks, value-based methods mainly address how to decouple the centralized team value function for distributed execution. Sunehag et al. [1] proposed value-decomposition networks that break down the team value function into a linear sum of individual agent values, where the optimal policy is obtained by greedy selection on the team value. Rashid et al. [2] further enhanced algorithm performance by representing the joint value function as a nonlinear monotonic combination and giving higher weight to joint actions with greater rewards, thus extending the approach to non-monotonic environments. Son et al. [3] introduced a transformation-based factorization that avoids both monotonicity and additivity constraints. However, these methods often rely on regularization for tractable computation, which can hinder performance in complex charging environments. Mahajan et al. [4] addressed inefficient exploration by proposing multi-agent variational exploration, which improves coordination over extended time horizons.

For policy-based methods, when all agents update their strategies simultaneously, the environment becomes non-stationary, making learning more challenging. Thus, most policy-based MARL algorithms adopt the actor–critic (AC) framework. To mitigate partial observability, techniques such as value function decomposition and the use of centralized critics with additional information exchange during training are commonly adopted; both types typically employ the CTDE (Centralized Training, Distributed Execution) paradigm. For value decomposition, Su et al. [5] designed value-decomposition actor-critic methods using a monotonic mapping between the global and local state values, following a simple time-difference advantage gradient that improves sampling efficiency and converges to local optima. Yang et al. [6] implemented a determinant point process-based method for unconstrained value function decomposition.

In centralized critic approaches, Foerster et al. [7] developed counterfactual multi-agent policy gradients, where centralized critics access joint actions and all agent states, while each agent's policy depends only on its own observation history. Pu et al. [8] constructed a decomposed soft actor–critic method with discrete probability policies and counterfactual advantage functions, supporting efficient policy learning and partially resolving credit assignment for both discrete and continuous action spaces. Lowe et al. [9] proposed Multi-Agent Deep Deterministic Policy Gradient (MADDPG), assigning centralized critics to each agent to support different reward functions in competitive environments. Building on this, Wang et al. [10] extended MADDPG to partially observable settings, utilizing recurrent neural networks in both actor and critic to retain observation history. Li et al. [11] integrated minimax optimization into robust multi-agent reinforcement learning, enabling agents to learn robust strategies under adversarial conditions.

These MARL algorithms have been applied to complex scheduling scenarios in distributed charging networks for electric vehicles, enabling agents to collaboratively optimize charging schedules and energy management. However, despite the effectiveness of CTDE in addressing partial observability, issues such as agent privacy and single-point failure risks remain unsolved in large-scale distributed charging systems.

B. Engineering Applications of Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) methods have been widely extended to optimal control problems in engineering, such as traffic control [12], autonomous driving [13, 14], and base station communications [15]. As an emerging distributed decision-making technique, MARL has attracted considerable attention for its ability to address nonlinear objectives, which motivates its application in distributed charging scheduling and energy management for smart grids.

In the context of distributed energy (and charging) management, most MARL approaches adopt the CTDE (Centralized Training, Distributed Execution) paradigm to mitigate the adverse effects of partial observability. Zhang et al. [16] proposed a novel deep transfer Q-learning algorithm based on a virtual leader–follower model to maximize the total revenue of all agents while maintaining supply–demand balance in smart grid scenarios. Compared to heuristic optimization methods, deep transfer Q-learning achieves faster convergence, stronger online learning ability, and effective privacy protection for users. Wang et al. [10] introduced a cooperative fuzzy Q-learning approach for microgrid energy management, ensuring stable power supply for independent microgrids while considering user demand uncertainty and achieving rapid acquisition of management strategies.

Reinforcement learning-based strategies have also been designed for distributed energy and load management in competitive and stochastic energy markets, such as microgrid auction-based time-sharing markets, where model-free Q-learning ensures each agent can find its optimal strategy. Zhu et al. [17] proposed an attention mechanism and soft actor–critic-based method for multi-energy-coupled energy management under renewable energy and demand uncertainty, using counterfactual baselines to accelerate policy learning and minimize long-term energy costs while ensuring user needs. Sun et al. [18] developed a multi-agent energy management optimization framework for integrated energy systems considering electricity, natural gas, and carbon trading, using MADDPG to provide each agent with an online trading strategy that considers individual interests for fair market transactions and privacy protection.

To address deployment challenges in large-scale systems, a few studies have explored fully distributed MARL algorithms. Li et al. [19] designed a distributed Q-learning method under a fully decentralized control framework to solve nonconvex economic dispatch problems, though its optimization accuracy is limited due to the lack of state-action value function approximation. In contrast, Liu et al. [20] used nonlinear function approximation for value functions and introduced a diffusion strategy to enable agent collaboration, yet the convergence range of the value function fitting remains limited. Dai et al. [21] combined value-based methods with quadratic function approximation to handle decision-making in continuous action spaces, but the quadratic approximation is most suitable for convex optimization problems. Li et al. [22] proposed a fully distributed reinforcement learning algorithm for nonconvex economic dispatch, but their stateless design for static scheduling problems does not fully exploit the generalization power of deep learning.

For distributed charging and energy management, the above CTDE-based methods are difficult to deploy across widely-distributed charging infrastructure, and still face challenges such as centralized method security, privacy risks, and high communication costs.

C. Research Gaps

This section analyzes the main strengths and limitations of existing multi-agent reinforcement learning (MARL) approaches for distributed scheduling of heterogeneous charging stations, focusing on decision-making types, handling of partial observability, training schemes, and function approximation. For methods applied to distributed charging scheduling, we further compare their optimization objectives, scheduling precision, and generalization capability under uncertainty. Overall, MARL demonstrates outstanding performance in tackling complex nonlinear scheduling objectives and demand uncertainties, offering an effective research paradigm for distributed scheduling in intelligent charging networks. However, current MARL methods still face several challenges when applied to heterogeneous charging station systems:

1) Observability and deployment issues: To address information barriers caused by partial observability, most existing algorithms adopt a centralized training, distributed execution (CTDE) framework. However, this paradigm is difficult to deploy efficiently across widely distributed charging station network nodes, restricting its practicality and scalability.

2) Constraint handling and sparse rewards: Most current methods translate operational constraints into penalty terms within the reward function, which often leads to sparse reward signals. This sparsity can negatively affect learning efficiency and the feasibility of the final scheduling policy.

Therefore, a key motivation of this paper is to theoretically analyze the impact of partial observability on the convergence of MARL, design a fully distributed MARL algorithm suitable for multi-region charging station systems, and introduce an action space mapping mechanism to strictly confine the decision-making process within the feasible domain during training. These innovations aim to provide more efficient and practical technical solutions for intelligent scheduling in large-scale charging networks.

III. Problem Modeling

A. System Structure for Heterogeneous Charging Station Scheduling

This section investigates a coordinated scheduling system for multiple regions and heterogeneous charging stations, as illustrated in Fig. 1. The system comprises various types of charging equipment, including Fast DC Chargers (FDC), AC Slow Chargers (ASC), and bidirectional V2G (Vehicle-to-Grid) Chargers (V2G).

As shown in Fig. 1, within the multi-region microgrid balancing charging supply and demand, each region can be connected to the interregional grid via substations. The system considers two types of regions: source-load coordination regions (e.g., Region A) and dedicated charging regions (e.g., Region D). Source-load coordination regions integrate multiple types of charging stations (FDC, ASC, V2G) as well as end-users (EVs), while dedicated charging regions deploy only a single type or a subset of charging stations.

In the heterogeneous charging station system discussed in this chapter, each group of charging stations of the same type can be regarded as an independent agent. FDC and ASC can autonomously determine their charging power allocation

according to real-time demand, while V2G chargers can flexibly switch between charging and discharging modes, enabling bidirectional energy flow. Operators are able to sense the local charging demand of EVs in real time and transmit the relevant information to the respective charging station agents. For information flow, both intra- and inter-regional communication among charging station agents is realized through remote point-to-point links, facilitating distributed learning and optimal scheduling decisions.

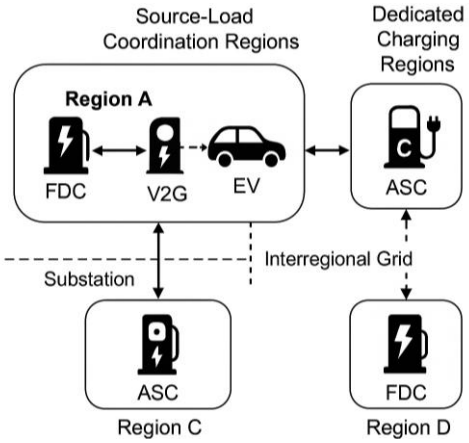


Fig. 1. Schematic diagram of the heterogeneous charging station scheduling system.

The main energy flow in this system is electrical power. The internal structure of the regional charging network is depicted in Fig. 2, where the Charging Station (CS) serves as the core node.

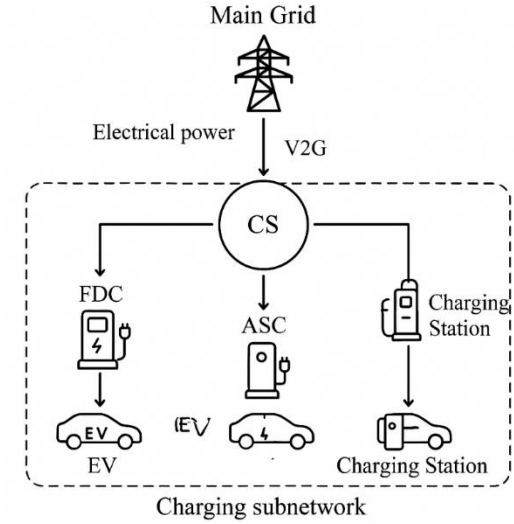


Fig. 2. Integrated heat-electricity network within a region.

The regional charging network can be subdivided into the main grid and the charging subnetwork. Energy flows from the main grid to users (EVs) via different types of charging stations, achieving efficient power supply based on demand. Power demand can be collaboratively met by FDC, ASC, V2G, and charging stations in other regions. V2G chargers can also feed energy from vehicle batteries back to the main grid, enhancing the flexibility of energy regulation. According to the described energy flow and conversion processes, the charging subnetwork mainly serves intra-regional needs, while the main grid supports interregional energy allocation. This setup aligns with the actual requirements of power flow and distributed scheduling.

During the charging scheduling process, priority is given to meeting the charging needs of local users, and any remaining capacity can be used to support grid load balancing. Hierarchical scheduling strategies can effectively improve the utilization efficiency of renewable energy, reduce grid pressure, and advance the goal of green and intelligent mobility.

B. Multi-Domain Integrated Energy System Component Modeling

Based on the structure of the multi-region, multi-type charging station system, the primary components include Fast DC Chargers (FDC), AC Slow Chargers (ASC), bidirectional V2G Chargers (V2G), and Charging Stations (CS). Mathematical models are developed for each component to formulate the distributed charging scheduling problem, as described below:

1) Fast DC charger: FDCs serve as the main high-power charging facilities in the system, primarily catering to EVs with fast-charging needs. The charge/discharge power ratio of FDC can be expressed as:

$$\alpha_{FDC} = \frac{P_{FDC}^{out}}{P_{FDC}^{in}} \tag{1}$$

where P_{FDC}^{out} and P_{FDC}^{in} represent the output and input power of the FDC, respectively. The operational cost function is modeled as a coupled quadratic function:

$$C_{FDC}(P) = aP_{FDC}^2 + bP_{FDC} + c \tag{2}$$

where C_{FDC} denotes the instantaneous operating cost of the FDC, and $a\ b\ c$ are intrinsic parameters. The energy constraint is:

$$0 < P_{FDC} < P_{FDC}^{max} \tag{3}$$

where P_{FDC}^{max} is the FDC's maximum output power.

2) AC Slow charger: ASCs mainly provide charging services for EVs with longer dwell times. The cost function is:

$$C_{ASC}(P) = \alpha P_{ASC}^2 + \beta P_{ASC} + \gamma \tag{4}$$

where P_{ASC} and C_{ASC} denote the output power and instantaneous cost of ASC, and α, β, γ are intrinsic parameters. The power constraint is:

$$P_{ASC}^{min} < P_{ASC} < P_{ASC}^{max} \tag{5}$$

3) Bidirectional V2G charger: V2G chargers enable bidirectional energy flow, allowing energy feedback to the grid while meeting user demand. The charge/discharge ratio is defined as:

$$\alpha_{V2G} = \frac{P_{V2G}^{\text{out}}}{P_{V2G}} \tag{6}$$

where P_{V2G} is the total V2G power, and P_{V2G}^{out} is the power fed back to the grid. The operating cost is:

$$C_{V2G}(P) = gP_{V2G} + h \tag{7}$$

where C_{V2G} is the instantaneous cost, and g, h are intrinsic parameters.

4) Charging station: As the energy dispatch center, CS coordinates energy allocation among regional charging stations. Its dispatch capacity constraint is:

$$0 < P_{CS} < P_{CS}^{max} \tag{8}$$

where P_{CS}^{max} is the maximum dispatch capacity.

C. Distributed Charging Scheduling Problem

In this study, multi-agent deep reinforcement learning (MADRL) is employed to derive optimal economic scheduling and energy feedback policies, aiming to minimize overall operational costs. The distributed energy management problem is formulated as follows:

Objective Function is as follows:

$$C = \sum_{i=1}^{n} C_{FDC,i} + \sum_{i=1}^{m} C_{ASC,i} + \sum_{i=1}^{q} C_{V2G,i} \tag{9}$$

where C is the total operating cost, and n, m, q are the numbers of FDC, ASC, and V2G chargers, respectively.

Capacity Constraints is as follows:

$$\begin{cases} 0 < P_{FDC,i} < P_{FDC,i}^{max}, & \forall i = 1,2,\dots,n \\ P_{ASC,i}^{min} < P_{ASC,i} < P_{ASC,i}^{max}, & \forall i = 1,2,\dots,m \\ 0 < P_{V2G,i} < P_{V2G,i}^{max}, & \forall i = 1,2,\dots,q \end{cases} \tag{10}$$

Power Demand Balance Constraint is as follows:

$$\sum_{i=1}^{c} P_{D,i} = \sum_{i=1}^{n} P_{FDC,i} + \sum_{i=1}^{m} P_{ASC,i} + \sum_{i=1}^{q} \left[(1 - \alpha_{V2G,i})P_{V2G,i}\right] \tag{11}$$

where $P_{D,i}$ is the charging demand of user i, and c is the total number of users.

Energy Feedback Balance Constraint is as follows:

$$\sum_{i=1}^{c_I} P_{D,i}^{\text{feed}} = \sum_{i=1}^{q_I} P_{V2G,i}^{\text{out}}, I = 1,2,\dots,N \tag{12}$$

where $P_{D,i}^{\text{feed}}$ is the feedback energy requirement of user i, q_I is the number of V2G chargers in region I, and N is the total number of regions.

IV. Multi-Agent Deep Reinforcement Learning for Distributed Energy Management

The main method proposed in this chapter addresses the above distributed charging scheduling problem by developing a multi-agent deep reinforcement learning (MADRL) algorithm under a partially observable environment. The following describes the MADRL algorithm framework for heterogeneous charging station systems and discusses relevant technical details.

A. Algorithm Overview

The multi-agent deep reinforcement learning framework developed for heterogeneous charging station systems is illustrated in Fig. 3. In this framework, each FDC, ASC, and V2G charger group acts as an independent agent, employing a static optimization deep learning approach. For each agent i, the POMDP is formalized as $< S_i, A_i, P_i, R_i >$, and the Markov decision process for each charger type is modeled as follows:

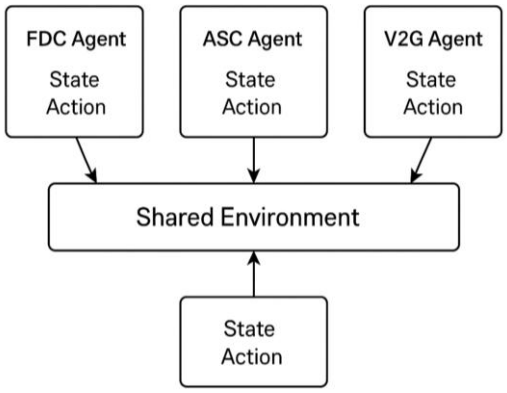


Fig. 3. Multi-agent deep reinforcement learning framework for heterogeneous charging station scheduling.

1) State

a) FDC agent: State vector includes global total charging demand $\sum_{i=1}^{c} P_{D,i}$, local fast charging demand, and total V2G charge/discharge in the region.

b) ASC agent: State vector includes global total charging demand, local slow charging demand, and total V2G charge/discharge in the region.

c) V2G agent: State vector includes global total charging demand, available V2G capacity in the region, and loads of all charger types in the region.

2) Action: For agent i:

a) FDC agent: The output power $P_{FDC,i}$ is the action.

b) ASC agent: The output power $P_{ASC,i}$ is the action.

c) V2G agent: The action vector consists of charging/discharging power $P_{V2G,i}$ (positive for charging, negative for discharging) and the feedback ratio $\alpha_{V2G,i}$.

3) State transition probability: Due to the stochastic nature of charging demand and vehicle arrivals, the state transition distribution for all agents i is assumed to be uniform.

4) Reward: The local reward of each charging station agent is negatively correlated with the objective function above. If an agent fails to satisfy constraints such as capacity or power, its reward is penalized.

This framework considers two deep learning approaches: Target Value Competition Multi-Agent Reinforcement Learning (MADRL-TVC) and Multi-Agent Deep Deterministic Policy Gradient (MADDPG). MADRL-TVC is a

fully distributed method, while MADDPG requires centralized training. Both methods are applied to the distributed charging scheduling problem, with the superior approach selected for the charging station system scheduling policy.

B. Algorithm Procedure

The algorithm used in this chapter is a model-dependent continuous decision-making method, not episode-based, and does not rely on historical sequences, but only on real-time decisions and feedback from the multi-agent system. The implementation process of the MADRL algorithm is shown in Fig. 4, with the core steps as follows:

1) Initialization and state acquisition: Each charging station agent initializes its parameters and obtains the initial state from the environment. State information includes the current load demand of each charger type, user arrival information, available V2G capacity, etc. At each decision step t, each agent selects action $a_{i,t}$ according to its policy.

2) Hierarchical scheduling and action adjustment: To balance different types of demand, a hierarchical control framework is designed for high-priority intra-regional charging needs and cross-regional energy feedback (V2G). Actions are adjusted automatically via distributed binary search, prioritizing high-priority charging demand within each region.

For balancing fast and slow charging demands, the initial search domain for FDC and ASC agents is set as:

$$a_i^{min}(0) = 0, a_i^{max}(0) = P_{FDC,i}^{max} \text{ or } P_{ASC,i}^{max} \tag{13}$$

For V2G feedback, the agent's search domain is:

$$a_i^{min}(0) = -P_{V2G,i}^{max}, a_i^{max}(0) = P_{V2G,i}^{max} \tag{14}$$

The load demand deviation for region I is defined as:

$$\delta_{P,I}(k) = \sum_{i=1}^{c_I} P_{D,i} - \sum_{i=1}^{n_I} P_{FDC,i} - \sum_{i=1}^{m_I} P_{ASC,i} - \sum_{i=1}^{q_I} \alpha_{V2G,i} P_{V2G,i} \tag{15}$$

Running Algorithm 1 with this setup achieves balance between regional charging demand and V2G feedback.

3) Experience collection and network training: Each agent observes the immediate reward r_{t+1} and the next state $s_{i,t+1}$. After each decision, the experience is stored in a replay buffer. At regular intervals, mini-batches are randomly sampled from the buffer to train the deep neural network.

Based on the above approach, charging station agents collaboratively learn optimal energy allocation and economic scheduling policies to minimize the overall system operating cost. The trained model can respond online to varying charging demands, EV arrivals and departures, and random V2G availability, providing real-time scheduling strategies. This method excels at rapid demand response, accommodating renewable energy uncertainty, and improving overall energy utilization efficiency.

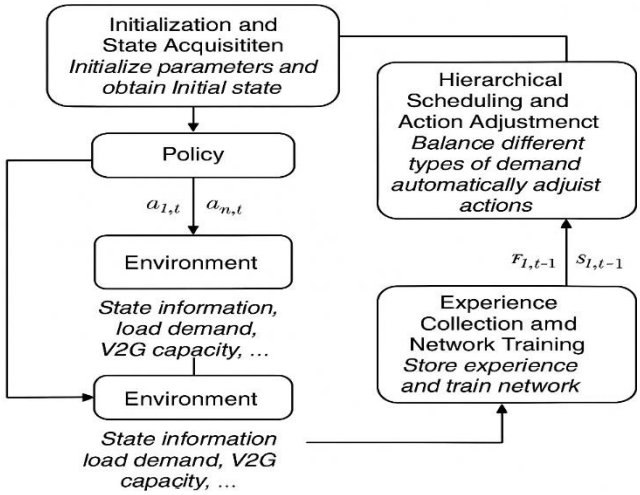


Fig. 4. Multi-agent deep reinforcement learning algorithm flow for multi-domain integrated energy systems.

V. Simulation Validation

In this section, a multi-region, multi-type charging station scheduling scenario is constructed. The proposed hierarchical decision-making framework and multi-agent reinforcement learning method are trained, and the trained models are tested using open-source transportation and charging load datasets. Their responsiveness to load uncertainty and distributed charging demand is analyzed and compared with other benchmark scheduling methods, further validating the performance advantages of the proposed distributed charging scheduling scheme in multi-region scenarios.

A. Simulation Setup

The simulation environment is configured as a multi-type charging station system spanning four regions, as illustrated in Fig. 1. The entire system comprises nine groups of charging stations, including four Fast DC Charger (FDC) groups (deployed in Area A: 1 group, Area C: 1 group, Area D: 2 groups), two bidirectional V2G charger groups (deployed in Area A and Area C), and three AC Slow Charger (ASC) groups (deployed in Area A, Area B, and Area C). Cross-regional power distribution can be achieved via the main grid. Data on vehicle charging demand, EV arrivals and departures, and V2G feedback capabilities for each region are sourced from open-source transportation and charging datasets (such as PEMS, Open Charge Map, etc.). The cost parameters and inequality constraints for each type of charging station are integrated from previous studies or relevant industry standards. All simulation parameters are detailed in Tables I, II, and III. The scheduling interval is set to 15 minutes, and the simulation covers a 24-hour period.

TABLE I. FDC Unit Parameters

NodeRegion	NodeRegion	α_i	β_i	γ_i	$P_{FDC}^{\min}$	$P_{FDC}^{\max}$
1	A	0.0020	10	500	100	600
2	C	0.0025	8	300	50	400
3	D	0.0050	6	100	50	300
4	D	0.0060	5	90	50	200

B. Training Process

Due to the incomparability of immediate rewards across different reward functions, it is challenging to directly track the cumulative rewards during the training process under random states. To evaluate the algorithm's performance during training, 10 random state points are selected and their corresponding rewards are recorded throughout the training. Each observation point includes the power demand and heat demand of Areas A, B, and C, as well as the wind power output from WT unit groups in Areas A and C. To analyze the neural network's fitting performance, the total training loss of the Q-network estimated by the MADRL-TVC method is also recorded during training.

TABLE II. V2G CHARGER AND CS PARAMETERS

NodeRegion	NodeRegion	g_i	h_i	k_{V2G}	P_{V2G}^{max}
5	A	5	15	0.9	90
6	B	4	20	0.9	90

TABLE III. ASC UNIT PARAMETERS

NodeRegion	NodeRegion	a_i	b_i	c_i	P_{CHPP}^{max}	H_{CHPP}^{max}	α_{CHPP}^{min}
7	A	0.0050	6	100	200	100	0.5
8	B	0.0060	5	90	200	90	0.5
9	C	0.0072	3	150	180	80	0.5

Fig. 5 shows the cumulative rewards and total training loss at different observation state points during the training process. It can be observed that the cumulative rewards at all state points exhibit an upward trend. This indicates that the proposed algorithm can improve cumulative rewards through training, thus finding the optimal scheduling strategy for each type of state. The training loss of the Q-network converges rapidly and ultimately becomes very small, demonstrating that the neural network can effectively fit the charging scheduling policy.

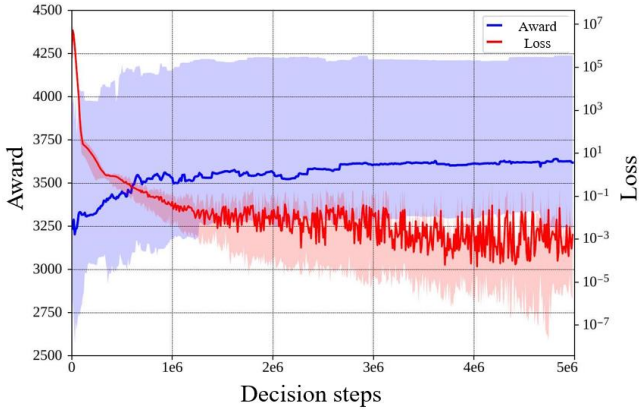


Fig. 5. Cumulative rewards and total training loss during training.

C. Full-Day Energy Management Results Analysis

After extended training under randomly varying states, the proposed algorithm achieves a generalizable, complete model that can provide optimal charging scheduling strategies in real time for arbitrary user demand and V2G capacity availability. Within a one-day scheduling cycle, the MADRL-TVC trained model outperforms MADDPG, so the following presents the optimal strategies under MADRL-TVC scheduling.

The energy allocation of V2G chargers is shown in Fig. 6. For each group of V2G chargers, the energy allocation includes charging power output (P_{V2G}), energy fed back to the grid (F_{V2G}), and energy loss (L_{V2G}). As illustrated in Fig. 6(a) and 6(b), the V2G chargers in Area A and Area C can fully utilize vehicle battery energy for grid feedback during peak load periods, effectively alleviating regional power supply pressure. The V2G energy utilization rate reaches up to 100%.

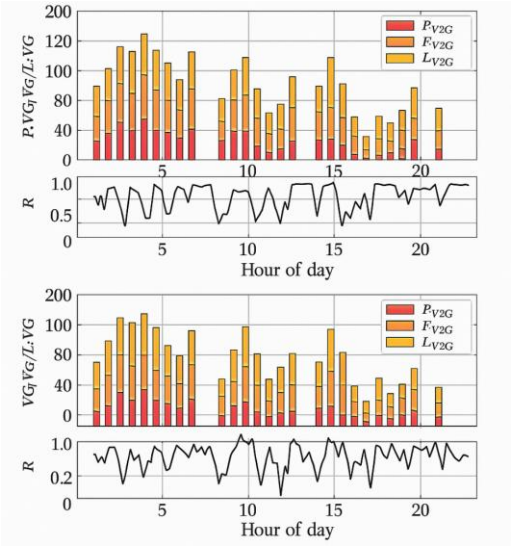
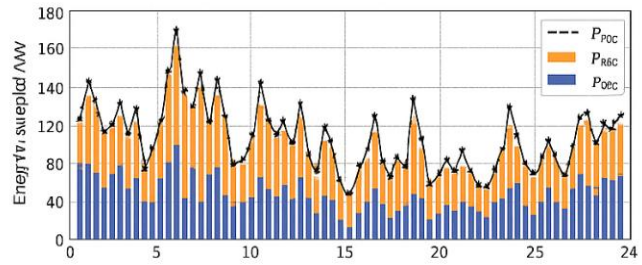


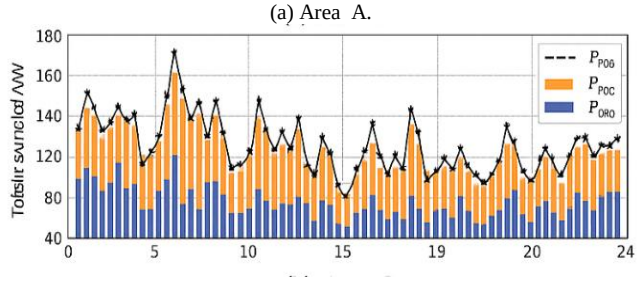
Fig. 6. Regional V2G charger energy allocation and feedback rate.

The scheduling results for charging demand within each region are shown in Fig. 7, displaying how different types of charging stations meet vehicle charging needs in each region for every 15-minute interval throughout the day. P_{FDC}, P_{ASC}, and P_{V2G} represent the energy supplied by each type of charging station, while PD_A, PD_B, and PD_C denote the total demand in each region. The results demonstrate that the charging demands in all regions are precisely met. Due to the stochastic nature of charging demand and vehicle arrivals, the energy allocation structure across regions does not follow a fixed pattern. Nevertheless, the algorithm is able to provide corresponding scheduling strategies in real time, effectively coping with demand and load uncertainty.

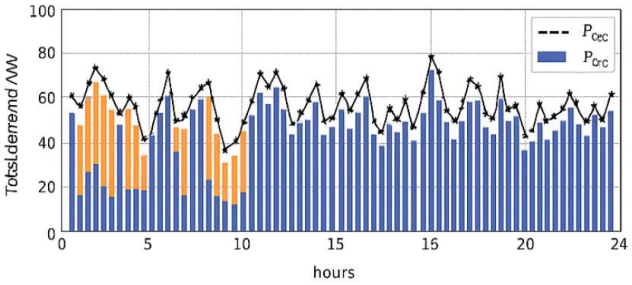
Cross-regional energy scheduling is depicted in Fig. 8, showing the overall charging power allocation at each time interval throughout the day—that is, how each type of charging station and each region jointly respond to the total charging demand. P_{Area_A}, P_{Area_B}, P_{Area_C}, and P_{Area_D} represent the total charging output for each region. The charging power is balanced at all times of the day, with the FDCs (which have higher economic priority) typically handling the main load, while ASCs provide supplementary regulation. V2G chargers mainly contribute energy feedback during peak load periods. The scheduling results indicate that the agents can make distributed collaborative decisions to achieve optimal economic scheduling.



(a) Area A.



(b) Area B



(c) Area C

Fig. 7. Regional charging demand scheduling results.

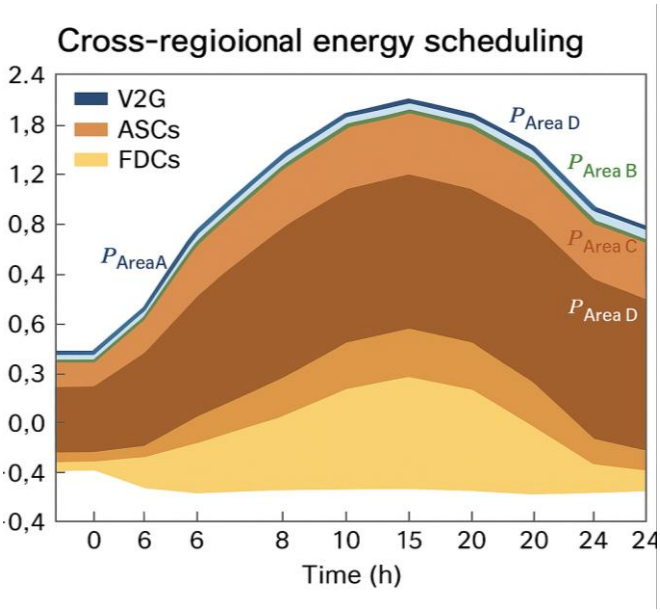


Fig. 8. Overall power scheduling results.

The total system operating cost is shown in Fig. 9, presenting the full-day operating costs for the multi-type charging station system. The results show that the algorithm can effectively reduce overall operating costs and achieve coordinated operation between charging stations and the grid.

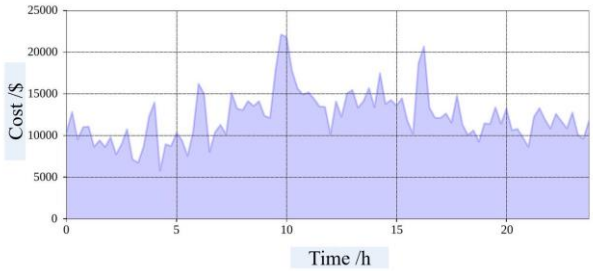


Fig. 9. Full-day operating cost.

D. Comparative Analysis

To evaluate the performance of the distributed charging scheduling method based on multi-agent reinforcement learning (MADRL), it is compared with commonly used centralized optimization methods, including Particle Swarm Optimization (PSO), nonlinear optimization solvers (SCIP, NLOPT), and others. Table IV presents the operating costs of different algorithms at a specific observation state (e.g., total charging demand 800.0, regional demands [100.0, 120.0, 50.0], V2G available power [70.0, 150.0]). Table V lists the average hourly quantified costs for each method over the entire day.

As shown in the tables, compared with centralized stochastic optimization methods such as PSO, the multi-agent deep reinforcement learning approach can effectively reduce system operating costs in distributed charging scheduling. In particular, the MADRL-TVC method achieves performance comparable to mature solvers such as SCIP. Compared with the MADDPG (centralized training, distributed execution) framework, the fully distributed MADRL-TVC displays better convergence in high-dimensional decision spaces, further highlighting the advantages of distributed algorithms.

TABLE IV. OPERATING COST OF DIFFERENT METHODS AT THE OBSERVATION STATE

Algorithm	Solver Type	Optimal Cost	Improvement
PSO	Centralized	12603.94	0
NLOPT	Centralized	12622.88	-0.15%
SCIP	Centralized	12201.77	3.19%
MADDPG	Distributed	12394.07	1.67%
MADRL	Distributed	12202.26	3.19%

TABLE V. AVERAGE QUANTIFIED COST PER HOUR FOR DIFFERENT METHODS

Algorithm	Solver Type	Optimal Cost	Improvement
PSO	Centralized	52102.92	0
NLOPT	Centralized	52977.84	−1.84%
SCIP	Centralized	49129.07	5.56%
MADDPG	Distributed	50968.13	2.02%
MADRL	Distributed	49146.08	5.53.%

The results show that MADRL-TVC not only has advantages in reducing operating costs, but its distributed structure is also better suited to dynamic changes and randomness in charging demand and V2G feedback, offering

stronger generalization and stability. Therefore, the distributed scheduling method based on multi-agent deep reinforcement learning has significant application potential and development prospects in multi-region, multi-type charging station systems.

VI. Conclusion

This Paper proposes a distributed charging scheduling strategy for multi-region, multi-type charging station systems. Different regions and types of charging station groups are modeled as independent agents. Based on user charging demand and V2G feedback capability, collaborative optimization of the multi-agent system enables energy allocation and economic scheduling. For intra-regional fast charging, slow charging, and cross-regional energy feedback, a hierarchical decision-making framework is designed, and the charging scheduling problem is formalized as a Markov Decision Process, with multi-agent deep reinforcement learning employed for collaborative learning. Leveraging the generalization capability of deep reinforcement learning, the proposed method can not only reduce total system operating costs, but also respond in real time to uncertainties in charging demand and energy feedback, effectively improving the economic efficiency and robustness of distributed charging networks.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks.”

References

[1] Sunehag, P., Lever, G., Gruslys, A., et al. "Value-decomposition networks for cooperative multi-agent learning." arXiv preprint arXiv:1706.05296, 2017.

[2] Rashid, T., Samvelyan, M., de Witt, C. S., et al. "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning." Journal of Machine Learning Research, 21(1): 7234–7284, 2020.

[3] Son, K., Kim, D., Kang, W. J., et al. "Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning." In Proceedings of the International Conference on Machine Learning, Long Beach, USA, pp. 5887–5896, 2019.

[4] Mahajan, A., Rashid, T., Whiteson, S., et al. "MAVEN: Multi-agent variational exploration." In Advances in Neural Information Processing Systems, Vancouver, Canada, 2019.

[5] Su, J., Adams, S., Beling, P. "Value-decomposition multi-agent actor-critics." In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, pp. 11352–11360, 2021.

[6] Yang, Y., Wen, Y., Wang, J., et al. "Multi-agent determinantal q-learning." In Proceedings of the International Conference on Machine Learning, Virtual Event, pp. 10757–10766, 2020.

[7] Foerster, J., Farquhar, G., Afouras, T., et al. "Counterfactual multi-agent policy gradients." In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, pp. 1–6, 2018.

[8] Pu, Y., Wang, S., Yang, R., et al. "Decomposed soft actor-critic method for cooperative multi-agent reinforcement learning." arXiv preprint arXiv:2104.06655, 2021.

[9] Lowe, R., Wu, Y., Tamar, A., et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." arXiv preprint arXiv:1706.02275, 2017.

[10] Wang, R. E., Everett, M., How, J. P. "R-MADDPG for partially observable environments and limited communication." arXiv preprint arXiv:2002.06684, 2020.

[11] Li, S., Wu, Y., Cui, X., et al. "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient." In Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, pp. 4213–4220, 2019.

[12] Wang, Y., Xu, T., Niu, X., et al. "STMARL: A spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control." IEEE Transactions on Mobile Computing, 21(6): 2228–2242, 2020.

[13] Shalev-Shwartz, S., Shammah, S., Shashua, A. "Safe, multi-agent, reinforcement learning for autonomous driving." arXiv preprint arXiv:1610.03295, 2016.

[14] Chen, S., Dong, J., Ha, P., et al. "Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles." Computer-Aided Civil and Infrastructure Engineering, 36(7): 838–857, 2021.

[15] Gursoy, M. C., Zhong, C., Velipasalar, S. "Deep multi-agent reinforcement learning for cooperative edge caching." In Machine Learning for Future Wireless Communications, pp. 439–457, 2020.

[16] Zhang, X., Bao, T., Yu, T., et al. "Deep transfer Q-learning with virtual leader–follower for supply–demand Stackelberg game of smart grid." Energy, 133: 348–365, 2017.

[17] Zhu, D., Yang, B., Liu, Y., et al. "Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park." Applied Energy, 311: 1–15, 2022.

[18] Sun, Q., Wang, X., Liu, Z., et al. "Multi-agent energy management optimization for integrated energy systems under the energy and carbon co-trading market." Applied Energy, 324: 1–16, 2022.

[19] Li, F., Qin, J., Zheng, W. X. "Distributed Q-learning-based online optimization algorithm for unit commitment and dispatch in smart grid." IEEE Transactions on Systems, Man, and Cybernetics, 2019: 1–11.

[20] Liu, W., Zhuang, P., Liang, H., et al. "Distributed economic dispatch in microgrids based on cooperative reinforcement learning." IEEE Transactions on Neural Networks, 29(6): 2192–2203, 2018.

[21] Dai, P., Yu, W., Wen, G., et al. "Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions." IEEE Transactions on Industrial Informatics, 16(4): 2258–2267, 2020.

[22] Li, D., Yu, L., Li, N., et al. "Virtual-action-based coordinated reinforcement learning for distributed economic dispatch." IEEE Transactions on Power Systems, 36(6): 5143–5152, 2021.

A Review of Federated Learning Attacks: Threat Models and Defence Strategies

Fizlin Zakaria, Shamsul KamalAhmad Khalid
Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn,
86400 Parit Raja, Batu Pahat, Johor, Malaysia

***Abstract*—Federated Learning (FL) has emerged as a critical paradigm in privacy-preserving machine learning, enabling collaborative model training across decentralised devices without sharing raw data. While FL enhances privacy by maintaining data locality, it remains susceptible to sophisticated adversarial attacks. This review systematically analyses the FL threat landscape and introduces a novel taxonomy that classifies attack models based on their objectives, capabilities, and exploited vulnerabilities. Major categories include data poisoning, inference attacks, and Byzantine behaviours, each examined in terms of mechanisms, assumptions, and system impact. In addition, the paper evaluates prominent defence strategies—such as differential privacy, secure aggregation, and anomaly detection—by assessing their strengths, limitations, and real-world applicability. Key gaps include the lack of standardised evaluation metrics and limited exploration of adaptive defence mechanisms. Emerging trends such as homomorphic encryption, secure multi-party computation, and blockchain-based verifiability are also discussed. This review is a comprehensive resource for researchers and practitioners aiming to design resilient, privacy-aware FL systems that withstand evolving threats.**

***Keywords*—Federated learning; threat models; defence strategies; privacy-preserving AI; adversarial attacks**

I. Introduction

The healthcare sector has experienced significant technological advancements, underscoring the need to protect sensitive patient data [1]. Federated Learning (FL), a decentralised machine learning approach, offers promising privacy benefits by enabling local data processing. However, this decentralisation also introduces novel security vulnerabilities that require thorough investigation [2]. Despite its advantage in preserving data locality, FL remains susceptible to attacks from malicious participants [2]. Its ability to enable collaborative model training without compromising raw data makes FL particularly valuable in privacy-critical domains. This includes fields such as healthcare and finance, where leveraging distributed datasets without violating privacy regulations is essential [1].

Although interest in FL security has grown, existing literature often focuses on isolated aspects of the threat landscape, resulting in a fragmented understanding of its vulnerabilities and corresponding defence mechanisms [3]. A more holistic and structured review is needed to address this gap, integrating technical, regulatory, and ethical perspectives on safeguarding federated systems.

This paper addresses this need by providing a systematic taxonomy of FL attack vectors and a comparative evaluation of defence strategies. The review classifies attacks based on threat models—including data and model poisoning—and analyses their operational mechanisms. It further assesses defence methods such as differential privacy and secure aggregation in terms of their effectiveness, trade-offs, and applicability. Finally, the paper identifies open research challenges and proposes directions for future work to enhance the resilience, trustworthiness, and scalability of FL systems [4]. The main contributions of this paper are as follows:

- A novel taxonomy of FL attack models, classifying them based on attack goals, vectors, and system vulnerabilities.
- A comparative evaluation of defence strategies regarding performance, scalability, and security robustness.
- Identifying research gaps, including a lack of standardised evaluation frameworks and limited exploration of adaptive defences.
- Discuss emerging trends such as hybrid defence approaches and blockchain-based accountability mechanisms.
- The remainder of the paper is organised as follows: Section II outlines the literature selection and analysis methodology. Section III presents the taxonomy of FL attacks. Section IV evaluates existing defence strategies. Section V discusses current research gaps and future directions. Section VI and Section VII concludes the paper with a summary of findings and recommendations.

To guide this review, the following research question is posed: What are the key adversarial attack vectors that threaten Federated Learning (FL) systems, and how effective are current defence strategies—such as differential privacy, secure aggregation, and blockchain-based mechanisms—in mitigating these threats across different attack surfaces and system configurations? This question frames the scope of the analysis presented in this paper, which aims to systematically categorise existing FL attacks and critically evaluate the corresponding defence mechanisms based on their applicability, strengths, and limitations. To contextualise this study, Section II reviews foundational and recent research on FL threats and defences.

II. Related Work

Federated Learning (FL) has emerged as a prominent privacy-preserving machine learning framework, attracting increasing attention from academia and industry. Numerous studies have examined its security vulnerabilities and mitigation strategies, but many existing reviews are either narrowly scoped or lack updates to reflect the fast-evolving threat landscape. Pioneering works by Shokri et al. [23] and Melis et al. [24] introduced foundational concepts such as membership inference and property inference attacks, exposing the privacy risks inherent in collaborative learning environments. Bagdasaryan et al. [6] further demonstrated the feasibility of model poisoning and backdoor attacks, showing that even minor manipulations can compromise global model integrity. These studies laid the groundwork for analysing adversarial risks in decentralised learning systems.

Subsequent reviews by Hallaji et al. [2] and Nguyen et al. [11] broadened the threat taxonomy to include Byzantine behaviours and gradient leakage. However, their discussions often stop short of connecting these threats to comprehensive, layered defence architectures. Other researchers have explored isolated countermeasures—such as differential privacy [39] and homomorphic encryption [1]—but rarely assess their real-world performance in heterogeneous, large-scale FL deployments.

To address concerns over trust and transparency, blockchain-based FL frameworks (e.g., ShareChain [1], PPBFL [33]) have been proposed. These aim to decentralise trust management, offering tamper-evident logging and client accountability. Despite their conceptual appeal, such frameworks remain in the early stages of development and face considerable challenges related to scalability, latency, and integration with existing FL protocols [35].

This review builds upon previous work by offering a more comprehensive taxonomy of FL attacks, explicitly categorised by attack surface, timing, and intent. Furthermore, it delivers a comparative analysis of defence strategies, focusing on their security effectiveness, computational cost, and deployment trade-offs. This paper also identifies ongoing research gaps, particularly the lack of adaptive, hybrid, and context-aware mechanisms that can evolve alongside sophisticated attack patterns.

III. Methodology

This study follows a structured literature review approach to comprehensively examine Federated Learning (FL) attack models and corresponding defence strategies. The methodology adopted is outlined as follows.

A. Literature Search Strategy

This research employs a systematic literature review methodology, complemented by elements of a narrative review, to investigate the landscape of security threats and defence mechanisms within federated learning systems. The review focuses on identifying, classifying, and synthesising existing research concerning attacks targeting federated learning and corresponding defence strategies designed to mitigate these vulnerabilities [1], [5]. The study's primary objective is to provide a comprehensive overview of the current state-of-the-art, highlighting prominent attack vectors, defence approaches, and open challenges in the field.

The following keywords and Boolean combinations were used during the search:

- "Federated Learning" AND "Attack"
- "Federated Learning" AND "Threat Models"
- "Federated Learning" AND "Security"
- "Federated Learning" AND "Privacy Attacks"
- "Federated Learning" AND "Defence Strategies"
- "Secure Federated Learning" AND "Robust Aggregation"

Additionally, backwards and forward snowballing techniques were applied by reviewing the references of highly cited papers to identify additional relevant studies.

B. Inclusion and Exclusion Criteria

The literature search encompassed several prominent databases, including Scopus, Web of Science, and IEEE Xplore, selected for their extensive coverage of computer science, engineering, and related disciplines. These databases were queried using keywords and Boolean operators to identify relevant publications on federated learning security [3]. Search terms included variations and combinations of "Federated Learning Attack," "Threat Models," and "Federated Learning Defence" to capture a broad spectrum of research related to the topic. The initial search yielded a substantial number of articles, which were subsequently filtered based on predefined inclusion and exclusion criteria. To ensure the relevance and currency of the review, only peer-reviewed, English-language papers published between 2020 and 2025 were considered for inclusion. This timeframe was chosen to capture the most recent advancements and emerging trends in the federated learning security field—the restriction to peer-reviewed publications aimed to maintain the quality and rigour of the included studies. The English language restriction was applied due to resource constraints and the need for consistent linguistic analysis [1].

C. Study Selection Process

The selection process followed a structured approach, drawing inspiration from the PRISMA guidelines to enhance transparency and reproducibility. The initial screening involved reviewing the titles and abstracts of the identified articles to assess their relevance to the research question. Articles not explicitly addressing security threats or defence mechanisms in federated learning were excluded. Subsequently, the full texts of the remaining articles were examined in detail to determine their eligibility for inclusion based on the predefined criteria. This thorough evaluation ensured that only studies directly relevant to the scope of the review were considered for further analysis. The selected articles were then subjected to a rigorous analysis to extract key information regarding attack types, defence strategies, evaluation metrics, and experimental settings.

D. Data Extraction and Synthesis

The extracted data were organised and synthesised using a comprehensive analysis framework to categorise and classify

the identified attacks and defences. The categorisation was based on the nature of the attack, such as data poisoning, model poisoning, inference attacks, and Byzantine attacks [1]. Data poisoning attacks involve manipulating the training data to compromise the global model, while model poisoning attacks target the model aggregation process [4]. Inference attacks, on the other hand, aim to extract sensitive information from the model or the training data. Byzantine attacks encompass various malicious behaviours, including arbitrary data or model manipulation by compromised participants. Defence mechanisms were categorised based on underlying principles, such as differential privacy, secure aggregation, anomaly detection, and robust aggregation techniques. Differential privacy adds noise to the training data or model updates to protect individual privacy, while secure aggregation protocols ensure the confidentiality of model updates during the aggregation process. Anomaly detection techniques aim to identify and filter malicious or abnormal contributions from compromised participants. Robust aggregation methods, such as median or trimmed mean aggregation, are designed to mitigate the impact of outliers or malicious updates on the global model. Decentralised federated learning introduces new privacy threats [2]. For instance, an inference attack can occur in blockchain-based federated learning [1]. In this paradigm, security analysis is necessary [2]. This categorisation facilitated a structured comparison of different approaches and enabled the identification of their strengths and weaknesses. Existing defences against data poisoning cannot be directly applied to federated learning due to the requirement of accessing the training data [6]. Adversaries are improving at hiding malicious behaviours from benign ones [7]. Sophisticated attackers are motivated by very high incentives to manipulate the results of the machine learning models [8]. Data owners update their models while keeping the data localised, guaranteeing zero leakage of individual data since only trained models are shared in the proposed model [1]. The absence of a centralised server in blockchain-enabled federated learning eliminates the single point of failure [1]. By integrating federated learning with blockchain technology, auditing machine learning models is supported without centralising the training data [7].

IV. Results

This section presents a structured literature analysis on Federated Learning (FL) attacks and defence strategies. A taxonomy is developed to categorise attack types based on their objectives, Timing, and attack surfaces. Defence mechanisms are evaluated based on their effectiveness, complexity, and applicability.

A. Taxonomy of Federated Learning Attacks

Federated Learning (FL) is a distributed machine learning paradigm designed to train models collaboratively across decentralised data sources without transferring raw data. This decentralisation preserves privacy by enabling local computation on each client and aggregating model updates on a central server [1]. However, despite these inherent privacy advantages, FL systems remain vulnerable to a broad spectrum of adversarial threats that can undermine model integrity, compromise data confidentiality, and disrupt system performance [9]. Understanding and classifying these threats is essential for developing effective, context-aware defence mechanisms [5], [9], [10].

Fig. 1 presents a visual taxonomy that categorises FL attacks based on three key dimensions:

- Attack surface (client-side vs. server-side),
- Timing (during training vs. after training),
- Objective (e.g., data inference, model disruption, free-riding).

This classification provides a structured lens to analyse how FL systems are targeted across various operational contexts.

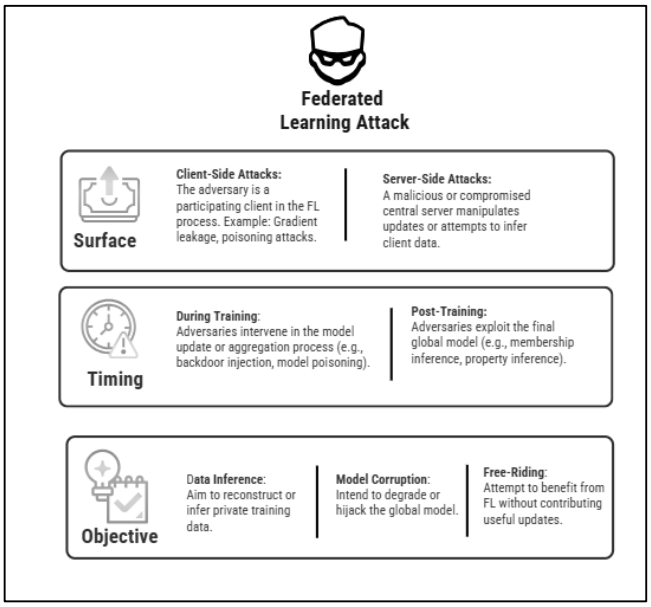


Fig. 1. Overview of attack categories in federated learning.

This diagram visualises how FL systems are exposed to threats across client and server surfaces, during various phases (training/post-training), and based on differing attack objectives.

1) By Attack surface

a) Client-side attacks: These attacks originate from compromised or malicious clients within the FL network. A typical example is data poisoning, where adversaries inject manipulated training data to introduce specific biases into the global model [4]. Another form, model poisoning, involves clients altering their model updates before aggregation to influence global model behaviour. These threats exploit the decentralised nature of FL, making it difficult for the server to distinguish between benign and malicious updates [11], [12].

b) Server-side attacks: Server-side attacks target the central aggregation server, which collects client updates and distributes the global model. Adversaries may compromise the server to tamper with the aggregation process or launch inference attacks to extract sensitive client data from model updates. Moreover, the server itself can act maliciously by injecting adversarial gradients or modifying aggregation rules [13], [14], [15].

2) By Timing

a) During training: Attacks launched during the training process are commonly referred to as poisoning attacks. These may involve the injection of malicious data, gradients, or even fake client updates (e.g., Sybil attacks) to degrade the model's accuracy or manipulate its outputs [16], [17]. These threats can be particularly stealthy and difficult to detect in asynchronous FL settings.

b) After training (after deployment): Once an FL model is deployed, it becomes susceptible to inference attacks, such as membership inference (determining if a record was part of training) or property inference (inferring sensitive aggregate attributes). These attacks exploit model outputs or gradients to reveal private training data [18], [20]. Attackers can also launch Reconstruction attacks to extract private textual data, exploiting gradient updates from the clients during the training phase [20].

3) By Objective: Federated learning attacks can be classified based on their primary objectives, reflecting the intended outcome that the attacker seeks. Attacks can also be categorised based on their goals, i.e., what the attacker hopes to achieve.

a) Data inference: Data inference attacks aim to extract sensitive information about the clients' local datasets from the shared model updates or the final global model [21]. Membership inference attacks determine whether a specific data point was used to train the model, revealing sensitive information about individual clients. Property inference attacks aim to infer statistical properties of the clients' local datasets, such as the distribution of sensitive attributes.

b) Model disruption: Model disruption attacks, also known as model poisoning attacks, aim to degrade the performance or functionality of the global model. By injecting malicious data or manipulating model updates, attackers can cause the model to make incorrect predictions or exhibit undesirable behaviour [8]. Such attacks can lead to model performance degradation or complete model failure. An attacker may compromise the global model by manipulating local parameters [22].

c) Free-Riding: Free-riding attacks, also known as lazy client attacks, occur when malicious clients contribute minimally to the training process while benefiting from the global model. These clients may submit outdated or irrelevant model updates, slowing convergence and reducing the model's accuracy [1]. In federated learning, free-riding attacks involve lazy clients that do not contribute meaningfully to the training process [1].

The results are further supported by visual representations, including Fig. 1 through Fig. 3 and Table II, which collectively illustrate the taxonomy of FL attacks, threat-defence mappings, and comparative defence evaluations. These visual aids were designed to enhance the clarity and accessibility of the findings, enabling a layered understanding of the attack vectors and mitigation strategies. Including these diagrams strengthens this review's communicative impact and supports readers in grasping complex adversarial relationships and defence limitations within FL systems.

Table I summarises the FL attack classification with attack surface, Timing, and impact.

TABLE I. SUMMARY OF FEDERATED LEARNING ATTACK TYPES

Attack Type	Description	Attack Surface	Timing	Representative Studies
Membership Inference	Determines if a specific data point was in the training set	Client	Post-training	[23]
Property Inference	Infers statistical properties of client data	Client	Post-training	[24]
Model Poisoning	Injects malicious updates to corrupt the global model	Client	During training	[25]
Backdoor Attacks	Embeds hidden misbehaviour triggered by specific inputs	Client	During training	[6]
Gradient Leakage	Reconstructs raw data from shared gradient updates	Server	During training	[26]
Sybil Attacks	Uses fake identities to influence model updates disproportionately	Client	During training	[27]
Free-Rider Attacks	Participates in FL without contributing useful updates	Client	During training	[28]
Collusion Attacks	Multiple adversaries collaborate to subvert learning	Multiple	During training	[2]
Adaptive Multi-Vector	Dynamic use of multiple attack vectors to evade defences	Multiple	Dynamic	[29]

B. Threat Models

The evolving landscape of Federated Learning (FL) security has seen a marked increase in hybrid and adaptive adversarial behaviours. Contemporary studies reveal that attacks are no longer limited to isolated strategies; adversaries now deploy multi-vector approaches that exploit privacy, model integrity, and communication weaknesses. Examples include coordinated multi-client poisoning, collusion-based inference, and gradient inversion attacks augmented with generative models. The growing sophistication of these threats necessitates the development of layered, context-aware defence mechanisms. Fig. 2 illustrates the key threat models encountered in FL environments, offering a taxonomy of representative attack types based on mechanisms and impact.

1) Membership Inference Attacks (MIA): Membership inference attacks (MIA) pose a fundamental threat to privacy in machine learning. In this attack, an adversary attempts to determine whether a specific data instance was part of the training dataset used to develop the model. The implications are severe, especially when dealing with sensitive datasets such as electronic medical records or financial transactions [30].

These attacks exploit subtle behavioural discrepancies between the model's predictions on seen versus unseen data. For

example, models typically return higher confidence scores or lower prediction entropy on training data than on unfamiliar inputs [31]. By leveraging such output characteristics, attackers can infer membership with non-trivial accuracy. Additionally, some variants assess model sensitivity to the presence or removal of specific samples, thereby detecting overfitting or memorisation tendencies.

Research has demonstrated the feasibility of MIAs across a range of model architectures and datasets. Key enabling factors include overfitting, model complexity, and the availability of public query APIs that allow repeated probing [23]. Attackers often rely on shadow models—locally trained replicas mimicking the target model's behaviour—to optimise their inference strategies. While simple mitigation techniques like limiting the number of queries have been proposed [13], these alone are insufficient in high-risk applications.

2) Property inference attacks: Property inference attacks, a closely related threat, extend the scope of privacy breaches by targeting aggregate properties of the training dataset rather than individual data points. In property inference attacks, the adversary aims to infer statistical characteristics or sensitive attributes of the training data, such as the proportion of individuals with a specific medical condition or demographic information. These attacks exploit the model's learned representations to extract information about the overall distribution and characteristics of the training data, potentially revealing sensitive information about the population from which the data was drawn. For instance, an adversary might try to determine a specific disease's prevalence among the individuals trained to train a diagnostic model. The effectiveness of property inference attacks depends on the model's tendency to encode statistical patterns and correlations in the training data, which can then be extracted and analysed.

3) Model poisoning attacks: Model poisoning attacks represent a significant threat to machine learning systems, where adversaries inject malicious data into the training dataset to impair the model's performance or manipulate its behaviour. This attack can have devastating consequences, especially in critical applications such as fraud detection, spam filtering, and medical diagnosis, where the accuracy and reliability of machine learning models are paramount [18]. By carefully crafting and injecting poisoned data points, attackers can subtly alter the model's decision boundaries, leading to incorrect predictions or biased outcomes [18]. The success of model poisoning attacks depends on the attacker's ability to inject enough poisoned data without being detected and the model's sensitivity to changes in the training distribution [8]. Security companies often rely on crowd-sourced threat feeds to train their classifiers, making them a natural injection point for such attacks. Furthermore, in federated learning scenarios, where multiple parties collaboratively train a model, a malicious participant can inject poisoned data, affecting the global model's performance.

4) Backdoor attacks: Backdoor attacks, a particularly insidious form of model poisoning, involve injecting specific triggers into the training data that cause the model to misbehave only when those triggers are present in the input [1]. These triggers can be subtle patterns or features imperceptible to human observers but easily recognised by the compromised model. When a backdoored model encounters an input containing the trigger, it will produce a predetermined, incorrect output, bypassing the model's intended functionality. Backdoor attacks are particularly concerning because they can be challenging to detect, as the model performs normally on most inputs and only exhibits malicious behaviour when the trigger is present. One study notes that backdoor attacks are highly effective when applied to computer vision models without many poisoned examples [12]. Explanation-guided backdoor poisoning attacks have also been investigated for malware classifiers.

5) Gradient leakage attacks: Gradient leakage attacks exploit the information contained in the gradients of a model during training to infer sensitive information about the training data. In distributed learning settings, where model updates are shared among multiple participants, attackers can intercept these gradients and reconstruct the training data or extract sensitive attributes. Gradient leakage attacks can reveal individual data points, statistical properties of the data, or even the labels associated with specific data points. These attacks pose a significant threat to privacy in federated learning and other distributed learning scenarios, where data is decentralised and sensitive.

6) Free-rider attacks: Free-rider attacks occur in collaborative learning settings, such as federated learning, where participants contribute to training a shared model. In a free-rider attack, a malicious participant exploits the contributions of others without contributing meaningfully to the training process. The free-rider may copy the model updates from other participants without performing any local training or contribute low-quality updates that do not improve the model's performance. By exploiting the contributions of others, free-riders can gain access to a high-quality model without incurring the computational costs or expending the effort required for training. The proposed framework prevents attribute disclosure, homogeneity, and background knowledge threats by using local differential privacy, which adds noise to the data [1].

7) Sybil attacks: Sybil attacks, also relevant in distributed learning scenarios, involve an attacker creating multiple fake identities to gain disproportionate influence over the training process. By controlling various "Sybil" clients, the attacker can manipulate the model's updates, inject bias, or even cause the model to learn incorrect patterns. Sybil attacks are particularly challenging to defend against because it can be difficult to distinguish legitimate clients from fake ones. To solve the single point of failure in federated learning, researchers have explored a blockchain-assisted decentralised federated learning that will completely prohibit malicious clients from poisoning the whole learning process or protect the entire process [1]. A blockchain-based federated learning model with secure multi-party computing model verification has been suggested to

counter poisoning attacks [1], [19]. Furthermore, secure aggregation protocols can be employed to ensure that the global model update is computed correctly, even in the presence of malicious participants [2], [6]. Federated Learning (FL) systems face increasingly complex and adaptive threats, including multi-vector and coordinated attacks. These attack types vary by mechanism, intent, privacy, or model performance impact. Fig. 2 summarises key threat models in FL, including membership inference, property inference, model poisoning, backdoor attacks, and Sybil attacks, highlighting their mechanisms and potential consequences. Understanding these models provides the foundation for evaluating appropriate countermeasures.

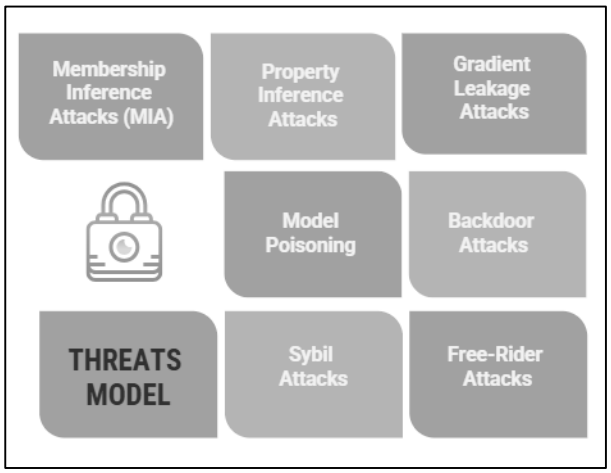


Fig. 2. Threat model of federated learning attacks.

C. Defence Strategies Against Federated Learning Attacks

Recent studies have revealed vulnerabilities in federated learning, demonstrating that model updates can inadvertently leak private information, necessitating robust defence mechanisms [9], [32]. A multifaceted approach encompassing data-level, model-level, and system-level defences is crucial to bolster the security of federated learning systems. Hybrid strategies integrate multiple defence mechanisms and offer a promising avenue for enhanced security and resilience against diverse attacks [2]. Multiple defence mechanisms have been proposed to combat various threats in FL, each addressing specific vulnerabilities. These include privacy-preserving techniques, robust aggregation methods, and decentralised trust frameworks. Fig. 3 maps common FL attack vectors to corresponding defence strategies and highlights the limitations associated with each countermeasure. This visual summary helps clarify the relationships between threats and mitigation approaches, allowing a deeper evaluation of specific techniques.

1) Data-level defences: Data-level defences constitute the first line of protection, focusing on safeguarding individual data points before they contribute to the global model [9]. Local Differential Privacy is a prominent technique, injecting noise into the client's local updates to obscure the contribution of individual data points [1]. Homomorphic Encryption offers another layer of protection by allowing computations on encrypted data, ensuring that the server only processes encrypted model updates, thus preventing access to raw data [19]. Applying these approaches can effectively mitigate privacy risks associated with sensitive training data [19]. Employing these methods can substantially diminish the risk of privacy breaches, reinforcing the protection of sensitive training data.

2) Model-level defences: Model-level defences operate at the aggregation stage, aiming to mitigate the impact of malicious or noisy updates on the global model [5]. Robust aggregation techniques, such as Krum, Median, Bulyan, and Trimmed Mean, identify and discard or down-weight potentially malicious updates, ensuring the global model remains robust despite adversarial clients. On the other hand, certified defences provide provable guarantees on the model's robustness against specific attacks, offering higher security assurance. These strategies play a key role in mitigating malicious influence and preserving the integrity of the global model.

3) System-level defences: System-level defences fortify the overall federated learning infrastructure, addressing potential vulnerabilities in the communication and coordination processes. Blockchain technology can be leveraged to create a decentralised and immutable record of model updates, enhancing transparency and preventing tampering [1]. Learning model on the distributed ledger [7].

a) Blockchain-based validation: Hybrid approaches represent the most promising avenue for defending against federated learning attacks, combining multiple defence mechanisms to provide comprehensive protection. Integrating privacy-preserving techniques, such as Local Differential Privacy or Homomorphic Encryption, with robust aggregation methods can protect data privacy and mitigate the impact of malicious updates. This multi-layered approach creates a synergistic effect, enhancing the overall security and resilience of the federated learning system. For instance, incorporating differential privacy with blockchain can improve privacy management, allowing data owners to manage their privacy preferences [1]. This combination balances data privacy and model accuracy, making federated learning a viable option for sensitive applications [33].

b) Client verification protocols: Client verification protocols authenticate clients before they participate in training, mitigating the risk of unauthorised or malicious clients joining the federation. By verifying participants' identities on the blockchain, frameworks can prevent single points of failure and poisoning attacks [1]. Blockchain helps trace malicious attacks by storing model participation, user fingerprints, and other key information. A permissioned blockchain-based federated learning method can chain incremental updates to an anomaly detection machine.

Developing effective defence strategies against federated learning attacks is an ongoing area of research, with new techniques and approaches emerging regularly. These techniques help ensure federated learning systems' privacy, security, and reliability, enabling their deployment in real-world applications where data privacy is paramount [34]. It is essential

to address these challenges to unlock the full potential of federated learning and ensure its responsible and secure deployment across various domains [35].

As summarised in Table II, which supports a structured comparison of defences across attack vectors, validation metrics, and deployment constraints. Vectors, validation metrics, and deployment constraints.

TABLE II. COMPARATIVE SUMMARY OF FL DEFENCE STRATEGIES

Defence Type	Technique	Targeted Attacks	Strengths	Limitations
Data-Level	Local Differential Privacy	Gradient leakage, inference	Decentralised, lightweight	Privacy-utility trade-off
Data-Level	Homomorphic Encryption	Server-side attacks	Strong privacy, data never exposed	Heavy computation
Model-Level	Krum / Trimmed Mean	Model poisoning, backdoor	Effective against outliers	Not Sybil-proof
Model-Level	DP-RFA	Poisoning + privacy attacks	Formal robustness guarantees	Requires a trusted setup
System-Level	Secure Aggregation	Gradient leakage	Aggregated updates only	Poisoning still possible
System-Level	Blockchain Validation	Free-riders, Sybil, poisoning	Immutable audit trail, client trust	Computational & consensus latency

4) Hybrid and emerging strategies: Hybrid defence strategies are gaining prominence in federated learning as they comprehensively address the multifaceted security challenges in decentralised machine learning. These strategies combine multiple defence mechanisms to create a layered security architecture that is more resilient to attacks than any single defence alone. For instance, combining differential privacy with secure aggregation can provide robust protection against inference and model poisoning attacks [10]. Differential privacy adds noise to the model updates to protect individual client data, while secure aggregation ensures that the server only receives the aggregated model updates without revealing individual contributions [1]. Furthermore, hybrid defence strategies can adapt to different attack scenarios and data distributions, providing a more flexible and effective defence mechanism [1]. This method is more suitable for sophisticated real-world situations because it combines the advantages of various defence strategies. The resilience of federated learning systems can be significantly increased by combining the benefits of blockchain technology with differential privacy techniques.

Emerging trends in federated learning defence strategies focus on developing more adaptive and intelligent defence mechanisms that automatically detect and respond to evolving attack patterns. These strategies leverage anomaly detection, reinforcement learning, and meta-learning to identify and mitigate attacks in real-time. Anomaly detection algorithms can identify suspicious model updates or client behaviour that deviates from the norm. Reinforcement learning can be used to train a defence agent that learns to adjust defence parameters based on the observed attack patterns adaptively. Meta-learning can be used to train a model that can quickly adapt to new attack scenarios by learning from previous attacks. Through a case study on federated intrusion detection systems, the capabilities in detecting anomalies and securing critical infrastructure without exposing sensitive network data can be demonstrated [4]. Such cutting-edge strategies hold tremendous promise for improving federated learning's security and resilience in the face of constantly changing threats. To stay ahead of the attackers, companies should use the most advanced technology and work together in a way that allows them to contribute their insights securely.

Advanced encryption techniques, such as homomorphic encryption, safeguard data privacy during federated learning. Homomorphic encryption allows computations on encrypted data without decrypting, ensuring that sensitive information remains protected throughout the learning process [1]. This is particularly useful in federated learning scenarios where data is distributed across multiple devices and cannot be shared in plaintext. By leveraging homomorphic encryption, the server can aggregate client model updates without seeing the underlying data, preventing potential privacy breaches [15]. The advent of sophisticated encryption strategies guarantees that sensitive data is protected during federated learning activities, thus increasing confidence in the technology's use across various industries.

To enhance clarity and summarise the defensive landscape, Fig. 3 visually maps key attack vectors in Federated Learning (FL), the vulnerabilities they exploit, corresponding defence mechanisms, and their associated limitations. This diagram supports textual taxonomy and comparative tables by providing a holistic view of how threats are addressed and where residual risks remain.

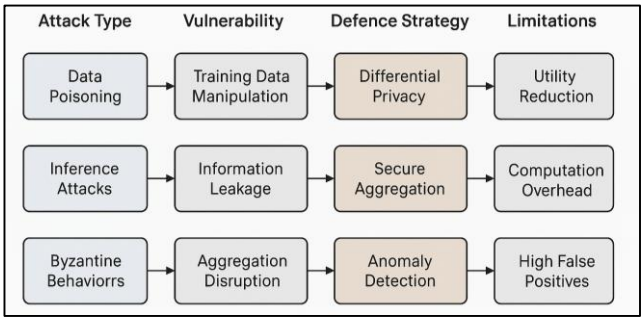


Fig. 3. Threat–defence–limitation mapping in federated learning security.

This flowchart visualises the relationship between common attack types in FL (e.g., model poisoning, inference attacks), the system vulnerabilities they exploit, corresponding mitigation strategies (e.g., differential privacy, secure aggregation), and each defence mechanism's inherent trade-offs or limitations.

V. DISCUSSION

A. Key Findings

Current security frameworks often lack the sophistication to discern subtle anomalies indicative of malicious intent, highlighting the need for intelligent, context-aware systems [2]. Such systems should integrate real-time monitoring of client activities, employing statistical analysis and machine learning techniques to identify deviations from expected behaviour [3]. Furthermore, the dynamic adjustment of privacy budgets, aggregation rules, or trust scores based on the assessed risk level is crucial for maintaining data privacy while preserving model accuracy. This adaptive approach requires sophisticated algorithms that balance the trade-off between security and utility, ensuring that defensive measures do not unduly impede learning [36].

The structured use of figures and summary tables was intentional to enhance reader comprehension, particularly given the complex multi-dimensional nature of threats and defences in FL systems. These visual tools complement the narrative synthesis and provide a concise reference for practitioners evaluating real-world applicability.

Decentralised trust infrastructures, potentially leveraging blockchain and distributed ledger technologies, offer a promising avenue for enhancing accountability and security in large-scale FL systems [1]. The inherent immutability and transparency of blockchain provide a robust platform for validating client updates, auditing model evolution, and establishing tamper-proof records of all transactions [1]. By decentralising the trust mechanism, the reliance on a central aggregator is reduced, mitigating the risk of single points of failure and bolstering the overall resilience of the FL system [1]. Smart contracts can further automate and enforce security policies, ensuring participant compliance with predefined rules [1]. However, implementing blockchain-based solutions in FL environments presents challenges related to scalability and computational overhead [1]. Innovative approaches are needed to optimise the performance of blockchain networks in the context of FL, such as using consortium blockchains or sharding techniques [35].

Addressing the resource constraints of edge devices and IoT deployments is paramount for the widespread adoption of FL in diverse application domains. Lightweight cryptographic and differential privacy mechanisms must be developed to minimise the computational burden on these devices, enabling them to participate effectively in the learning process without compromising their performance or battery life. Techniques such as homomorphic encryption and secure multi-party computation can provide strong security guarantees, but often incur significant computational costs. Therefore, research is needed to develop more efficient variants of these techniques tailored to the specific requirements of FL. Moreover, novel approaches to differential privacy, such as local and distributed differential privacy, can offer enhanced privacy protection while minimising the impact on model accuracy. Integrating Blockchain, LDP, and FL enables stronger protection of trained models' integrity by preventing several attacks, such as poisoning attacks, background knowledge, and inference attacks [1].

The robustness of federated models can be significantly enhanced through adversarial training techniques that incorporate adversarial examples and simulated attacks into the training phase. Exposing the model to diverse potential threats can improve its ability to generalise to unseen data and resist malicious inputs. This approach generates adversarial examples to fool the model into making incorrect predictions. These examples can be crafted using various techniques, such as gradient-based methods and generative adversarial networks. Furthermore, simulated attacks can evaluate the model's resilience to adversarial behaviours, such as data poisoning and model poisoning. By incorporating these adversarial examples and simulated attacks into the training process, the model can learn to defend itself against a broader range of threats.

Establishing standardised evaluation frameworks is essential for facilitating reproducibility and consistent comparisons of different FL security techniques. Open-source benchmarks and evaluation metrics should be developed to provide a common platform for assessing the performance of FL systems under various attack scenarios. These benchmarks should include a diverse range of datasets, model architectures, and attack strategies, reflecting the complexity and heterogeneity of real-world FL deployments. Furthermore, the evaluation metrics should capture not only the accuracy and efficiency of the model but also its privacy and robustness. Differential privacy is often implemented to protect data privacy by obscuring model parameter data that each client transmits. This helps improve system security and preserve user data privacy [37] while enhancing model robustness and confidentiality [1]. However, current FL setups make it difficult for participants to verify the machine learning model's authenticity [19]. Standardised evaluation frameworks would enable researchers and practitioners to objectively compare security techniques and identify the most effective solutions for specific FL applications [11]. Recent studies demonstrate that an adversarial attacker may raise privacy concerns by launching inference attacks against other participants, even when the iterations are performed only a few times [30]. These attacks can be defended against by methods such as differential privacy. Compared to the taxonomy proposed by Nguyen et al. [11], which broadly classifies attacks into data and model categories, our model introduces a finer-grained surface-based classification. This facilitates more precise identification of attack entry points and informs layered defence design. Furthermore, our comparative defence summary (Table II) validates the practicality of proposed strategies by mapping them to real-world limitations such as computational cost and detection accuracy—current Research Gaps.

A significant gap in current research is the lack of effective defences against complex multi-client collusion attacks, where multiple malicious participants coordinate their actions to compromise the global model [2]. These attacks pose a substantial threat, as they can be challenging to detect and mitigate due to their distributed nature and the potential for subtle manipulation of the learning process. Further complicating the landscape is the trade-off between robustness and model performance, where defences designed to enhance robustness against adversarial attacks can inadvertently degrade the accuracy and generalisation capabilities of the model [19].

This trade-off requires careful consideration and the development of adaptive defence strategies that can dynamically adjust their strength based on the detected threat level and the desired level of model performance. The lack of comprehensive scalability testing in real-world federated learning scenarios also presents a challenge, as many existing defences have not been rigorously evaluated under large-scale deployments with heterogeneous data and diverse client capabilities. Addressing these gaps is essential for building trustworthy and reliable federated learning systems that can be confidently deployed in real-world applications.

Within the intricate domain of contemporary technological progression, the convergence of artificial intelligence (AI) and cloud computing has emerged as a transformative paradigm, revolutionising data processing, analysis, and dissemination across diverse sectors. According to study [38], federated learning decentralises AI model training by enabling computations directly on edge devices, enhancing model personalisation while preserving user privacy. As discussed by [39], cloud infrastructure provides scalable and elastic computational resources that support real-time AI workloads and large-scale model deployment. Additionally, the study [40] emphasises the role of this convergence in enabling privacy-aware and latency-efficient learning systems, particularly in applications involving sensitive or distributed datasets.

This convergence is not merely a confluence of technologies but rather a synergistic amalgamation that amplifies the capabilities of both domains, fostering innovation and driving efficiency gains across a spectrum of applications. The essence of this convergence lies in the ability of cloud computing to provide a scalable and cost-effective infrastructure for AI algorithms, enabling them to process vast datasets and execute complex computations with unprecedented speed and efficiency. As we navigate an era defined by the exponential growth of data, fueled by the proliferation of interconnected devices, the need for robust and scalable data processing solutions has become paramount. With its inherent elasticity and ability to dynamically allocate resources, cloud computing offers an ideal platform for AI algorithms to thrive, providing the necessary computational power and storage capacity to handle the ever-increasing volume of data.

The traditional paradigm of AI development, characterised by centralised data processing and algorithmic training, often encounters limitations regarding scalability, generalisation, and real-time responsiveness. To address these challenges, federated learning has emerged as a groundbreaking approach, enabling collaborative model training across decentralised devices while preserving data privacy [38]. Federated learning empowers AI models to learn from vast amounts of data distributed across numerous edge devices without centralising sensitive information. Federated learning, in essence, brings the learning process to the edge, directly onto devices, harnessing the computational capabilities of heterogeneous devices to enhance model quality [38], [39]. This distributed approach to machine learning addresses critical challenges associated with data privacy, security, and access in traditional centralised systems.

Federated learning operates on the principle of distributed model training, where each participating device locally updates a shared model using its data, subsequently transmitting the model updates to a central server for aggregation. This iterative process of local model training and global model aggregation ensures that the learned model generalises well to the diverse data distributions across different devices, without compromising the privacy of individual datasets [38]. Federated learning distinguishes itself through its ability to train models across decentralised devices, offering a solution to privacy concerns and scalability challenges inherent in centralised systems [38]. By keeping data localised on devices, federated learning minimises the risk of data breaches and misuse, while utilising vast datasets distributed across numerous sources [41]. Moreover, federated learning enhances model generalisation by leveraging the diversity of data across different devices, resulting in more robust and accurate AI models that can effectively address real-world problems.

Federated learning, a burgeoning technique in distributed machine learning, facilitates model training across many decentralised devices, obviating the need for centralised data collection. This approach is particularly salient in scenarios where data privacy is paramount, such as healthcare, finance, and IoT applications [38]. By enabling collaborative model training without direct access to sensitive data, federated learning upholds user privacy and complies with stringent data protection regulations [38]. The core tenet of federated learning lies in its ability to train models across diverse data distributions, enhancing generalisation and robustness. The decentralised nature of federated learning obviates the need to consolidate data in a central repository, which can be daunting due to logistical hurdles, regulatory constraints, and security considerations [38]. Synchronous federated learning requires all participants to complete local computations before updates are sent to a central server. In contrast, asynchronous federated learning allows the central server to immediately integrate updates from any ready participant [39].

To further enhance the privacy guarantees of federated learning, differential privacy techniques can be integrated to obfuscate individual contributions during model training [39]. Differential privacy introduces carefully calibrated noise to the model updates, ensuring that the presence or absence of any single data point has a negligible impact on the final model, thereby protecting the privacy of individual users. Federated learning not only addresses privacy concerns but also improves the efficiency and scalability of machine learning models. This approach is beneficial when data are located in multiple clinical systems or when learning from sensitive personal data.

The convergence of federated learning and differential privacy holds immense promise for many applications, particularly in domains where data privacy and security are paramount. Differential privacy enhances data privacy in federated learning by adding noise during data queries and model updates [39]. These areas include healthcare, finance, and the Internet of Things.

VI. Future Directions

Future research directions should prioritise the development of adaptive, context-aware defences that can dynamically adjust their parameters and strategies based on the specific characteristics of the data, the threat environment, and the

available resources. By tailoring the defences to the particular context, it becomes possible to optimise their effectiveness and minimise their impact on model performance [3]. Combining secure aggregation with differential privacy dynamically represents another promising avenue for future research. Secure aggregation ensures that individual client updates are aggregated without revealing the underlying data, while differential privacy adds noise to the aggregated updates to protect against inference attacks [19]. Dynamically adjusting the parameters of these techniques based on the detected threat level and the desired privacy-utility trade-off can further enhance the security and performance of federated learning systems. Decentralised, blockchain-supported federated learning security offers a novel approach to improving trust and transparency in federated learning systems [7]. Attack detection and early warning systems in federated learning are crucial for proactively identifying and mitigating potential threats before they can cause significant damage. Integrating federated learning with blockchain technology allows for creating immutable audit trails and decentralised governance mechanisms, improving the learning process's accountability and trustworthiness [33]. A consortium blockchain-based federated learning framework enables decentralised, reliable, and secure federated learning without a centralised model coordinator [42]. Blockchain can address key challenges, propel the field forward, and potentially enhance data privacy and improve trust and security [35].

VII. Conclusion

Federated Learning (FL) represents a transformative approach in privacy-preserving machine learning, enabling decentralised model training across distributed clients without exposing raw data. However, the shift from centralised to federated paradigms introduces novel vulnerabilities that adversaries can exploit to compromise model integrity, data privacy, and system reliability.

Despite the comprehensive scope of this review, several limitations should be acknowledged. First, the analysis is constrained by the availability and maturity of published studies between 2020 and 2025, which may not fully capture the most recent developments or unpublished techniques in FL security. Second, while the taxonomy and comparative evaluation provide structured insights, they rely on secondary data rather than experimental benchmarking. Furthermore, the real-world effectiveness of many proposed defences—especially in large-scale, heterogeneous FL deployments—remains underexplored due to a lack of unified validation frameworks. These limitations highlight the need for ongoing empirical evaluation and longitudinal studies to assess defence robustness in dynamic adversarial settings.

To advance the field, future research must prioritise the development of adaptive, intelligent defence mechanisms capable of responding dynamically to evolving threats. Emphasis should also be placed on achieving a balance between security, privacy, and performance, particularly in large-scale, heterogeneous environments. Hybrid defence architectures, federated adversarial training, and lightweight cryptographic techniques hold promise in addressing these challenges. Ultimately, the continued evolution of FL security requires collaborative efforts across disciplines, integrating insights from cryptography, distributed systems, artificial intelligence, and regulatory policy. Ensuring robust, trustworthy FL systems will support privacy-aware innovations across critical sectors such as healthcare, finance, and autonomous systems.

Limitations: While this review provides a comprehensive classification of FL attack vectors and defence mechanisms, it is limited by the scope of current literature and the lack of unified benchmarking datasets. Furthermore, the practical deployment challenges—such as computation overhead, privacy-utility trade-offs, and client heterogeneity—require empirical validation in future work.

Acknowledgment

The author(s) sincerely thank Pejabat Penerbit, University Tun Hussein Onn Malaysia and the Faculty of Computer Science and Technology for their support throughout this research. Thanks to Dr. Shamsul Kamal bin Ahmad Khalid for the valuable guidance, feedback, and encouragement. The authors also acknowledge the use of publicly available resources and literature that have significantly contributed to the completion of this review. This work was conducted by the academic and ethical standards set forth by the institution. Communication of this research is made possible through monetary assistance by the University Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

References

[1] L. Javed, A. Anjum, B. M. Yakubu, M. Iqbal, S. A. Moqurrab, and G. Srivastava, "ShareChain: Blockchain-enabled model for sharing patient data using federated learning and differential privacy," *Expert Systems*, vol. 40, no. 5, p. e13131, Aug. 2022.

[2] E. Hallaji, R. Razavi-Far, M. Saif, B. Wang, and Q. Yang, "Decentralised Federated Learning: A survey on security and privacy," *IEEE Transactions on Big Data*, vol. 10, no. 2, pp. 194–213, April 2024.

[3] S. M. Mathews and S. A. Assefa, "Federated Learning: Balancing the thin line between data intelligence and privacy," *arXiv*, arXiv:2204.13697, Jan. 2022.

[4] H. J. Huang, B. Iskandarov, M. Rahman, H. T. Otal, and M. A. Canbaz, "Federated learning in adversarial environments: Testbed design and poisoning resilience in cybersecurity," *arXiv*, arXiv: 2409.09794, Sep. 2024.

[5] J. Li, X. Li, and C. Zhang, "Analysis on security and privacy-preserving in federated learning," *Highlights in Science Engineering and Technology*, vol. 4, pp. 349–358, Jul. 2022.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2938–2948, 2020.

[7] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, "Chained anomaly detection models for federated learning: An intrusion detection case study," *Applied Sciences*, vol. 8, no. 12, p. 2663, Dec. 2018.

[8] K. Aryal, M. Gupta, and M. Abdelsalam, "Analysis of Label-Flip poisoning attack on machine learning based malware detector," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4236–4245, Dec. 2022.

[9] S. Truex *et al.*, "A hybrid approach to privacy-preserving federated learning," *Informatik-Spektrum*, vol. 42, no. 5, pp. 356–357, Aug. 2019.

[10] S. Fuladi *et al.*, "A reliable and privacy-preserved federated learning framework for real-time smoking prediction in healthcare," *Frontiers in Computer Science*, vol. 6, p. 1494174, Jan. 2025.

[11] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated Learning for Internet of Things: A comprehensive

survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, Jan. 2021.

[12] Y. Zhao *et al.*, "Privacy-preserving blockchain-based federated learning for IoT devices," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1817–1829, Aug. 2020.

[13] R. Xu, S. R. Pokhrel, Q. Lan, and G. Li, "Post quantum secure Blockchain-Based Federated Learning for Mobile Edge Computing," *arXiv*, arXiv:2302.13258, Jan. 2023.

[14] A. Joshi, "Federated Learning: Enhancing data privacy and security in machine learning through decentralised training paradigms," *Journal of Artificial Intelligence & Cloud Computing*, vol. 1, no. 1, pp. 1–5, Jan. 2022.

[15] K. Yadav and B. B. Gupta, "Clustering-based rewarding algorithm to detect adversaries in federated machine learning based IoT environment," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, Jan. 2021.

[16] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11365–11375, Nov. 2021.

[17] M. Usama, J. Qadir, A. Al-Fuqaha, and M. Hamdi, "The adversarial machine learning conundrum: Can the insecurity of ML become the Achilles' heel of cognitive networks?" *IEEE Network*, vol. 34, no. 1, pp. 196–203, Oct. 2019.

[18] E. R. H. P. Isaac and J. Reno, "AI Product Security: A primer for developers," *arXiv*, arXiv:2304.11087, Jan. 2023.

[19] A. P. Kalapaaking, I. Khalil, and X. Yi, "Blockchain-Based Federated Learning with SMPC model verification against poisoning attack for healthcare systems," *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 1, pp. 269–280, Apr. 2023.

[20] M. Balunović, D. Dimitrov I., N. Jovanović, and M. Vechev, "LAMP: Extracting text from gradients with language model priors," in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Article 555, pp. 7641–7654, Jan. 2022.

[21] J. Sengupta, S. Ruj, and S. D. Bit, "SPRITE: a scalable privacy-preserving and verifiable collaborative learning for industrial IoT," in *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pp. 249–258, May 2022.

[22] C. Benzaid and T. Taleb, "AI for beyond 5G networks: A cyber-security defence or offence enabler?" *IEEE Network*, vol. 34, no. 6, pp. 140–147, Sep. 2020.

[23] 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)*, pp. 3–18, May 2017.

[24] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, May 2019.

[25] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Model poisoning attacks in federated learning," in *Proceedings Workshop on Security in Machine Learning (secML) 32nd Conference on Neural Information Processing Systems (neurIPS)*, pp. 1–23, Dec. 2018.

[26] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Article 1323, pp. 14774–14784, Jan. 2019.

[27] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv*, arXiv:1808.04866, Jan. 2018.

[28] J. Lin, M. Du, and J. Liu, "Free-riders in Federated Learning: Attacks and defences," arXiv.1911.12560, Jan. 2019.

[29] Z. Chen, H. Cui, E. Wu, and X. Yu, "Dynamic asynchronous anti-poisoning federated deep learning with blockchain-based reputation-aware solutions," *Sensors*, vol. 22, no. 2, p. 684, Jan. 2022.

[30] X. Zhang, H. Hou, Z. Fang, and Z. Wang, "Industrial internet federated learning driven by IoT equipment ID and blockchain," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, Jan. 2021.

[31] M. Veale, R. Binns, and L. Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol.. 376, no. 2133, p. 20180083, Oct. 2018.

[32] D. Enthoven and Z. Al-Ars, "An overview of federated deep learning privacy attacks and defensive strategies," in *Federated Learning Systems. Studies in Computational Intelligence*, vol 965, M. H. u. Rehman and M. M. Gaber, Eds. Cham: Springer, 2021, pp. 173–196.

[33] Y. Li, C. Xia, W. Lin, and T. Wang, "PPBFL: a privacy-protected blockchain-based federated learning model," *arXiv*, arXiv:2401.01204, Jan. 2024.

[34] J. Wan, H. Xun, X. Zhang, J. Feng, and Z. Sun. "A privacy-preserving and correctness audit method in multi-party data sharing," in *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies (CIAT 2020)*, pp. 414–419, Jan. 2021.

[35] M. M. Orabi, O. Emam, and H. Fahmy, "Adapting security and decentralised knowledge enhancement in federated learning using blockchain technology: Literature review," *Journal of Big Data*, vol. 12, p. 55, Mar. 2025.

[36] J. Mao, C. Cao, L. Wang, J. Ye, and W. Zhong, "Research on the security technology of Federated Learning privacy preserving," *Journal of Physics Conference Series*, vol. 1757, no. 1, p. 012192, Jan. 2021.

[37] L. Lyu, H. Yu, and Q. Yang, "Threats to Federated Learning: A survey," *arXiv*, arXiv:2003.02133, Jan. 2020.

[38] H. M. Asif, M. A. Karim, and F. Kausar, "Federated Learning and its applications for security and communication," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, Jan. 2022.

[39] F. Shan, S. Mao, Y. Lu, and S. Li, "Differential privacy Federated Learning: A comprehensive review," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, Jan. 2024.

[40] H. K. Shinwari and R. U. Amin, "Harnessing the power of Federated Learning: A systematic review of light weight deep learning protocols," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, Jan. 2025.

[41] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[42] Y. Qi, M. S. Hossain, J. Nie, and X. Li, "Privacy-preserving blockchain-based federated learning for traffic flow prediction," *Future Generation Computer Systems*, vol. 117, pp. 328–337, Dec. 2020.

Intelligent Logistics Vehicle Scheduling Based on MPHIGA

Xinxin Gao*, Qing Wang
School of Liberal Education, Liaoning University of International Business and Economics, Dalian, 116052, China

***Abstract*—The current intelligent logistics vehicle scheduling faces challenges, including the difficulty of obtaining real-time location data and the need for manual intervention in emergencies. To address these issues, a modified multi-population hybrid genetic algorithm is proposed, along with an intelligent scheduling model constructed through the reconstruction of domain generation strategies. Experimental results show that the model stabilizes the total cost at 7864 yuan within 49 iterations, whereas the dual-population hybrid genetic algorithm requires 51 iterations, making convergence more time-consuming. Moreover, when the scheduling frequency is two, the research model successfully allocates three company vehicles, whereas the comparison algorithm can only allocate two. Overall, the research model offers significant advantages in reducing operating costs and enhancing dynamic response capabilities, providing effective technical support for the digital transformation of logistics companies.**

Keywords*—Multi-population hybrid improved genetic algorithm; domain generation algorithm; logistics vehicles; co-evolution; scheduling management

I. Introduction

With rapid technological advancements, intelligent systems are increasingly integrated into various industries. Intelligent vehicle management is a key factor in improving operational efficiency in logistics and transportation [1]. Scientific scheduling of logistics vehicles optimizes routes and vehicle allocation, while also enabling real-time tracking of logistics resources, thus improving efficiency based on vehicle status and road conditions [2,3]. Although current intelligent logistics vehicle scheduling uses machine learning and big data analysis to predict demand fluctuations and optimize route selection, challenges remain [4]. Data silos significantly affects overall scheduling performance, while factors such as weather changes and uneven resource distribution make it harder to matching vehicles with cargo [5]. Some intelligent logistics vehicle scheduling systems cannot fully monitor and manage vehicle safety, limiting their ability to provide comprehensive data reports and statistics, which limits enterprises in conducting in-depth analysis and optimization [6]. To address these challenges, this study integrates the Domain Generation Algorithm (DGA) with the Multi-Population Hybrid Improved Genetic Algorithm (MPHIGA). The proposed model aims to improve intelligent logistics vehicle scheduling. DGA generates random domain names, which are introduced into MPHIGA for population initialization and mutation operations, enhancing the algorithm's global search capability. This study aims to improve real-time monitoring and positioning in logistics vehicle scheduling, enhance the profitability of logistics enterprises, and support the transformation of the logistics industry.

This study consists of four sections. Section II summarizes domestic and international research on MPHIGA and logistics vehicle scheduling. Section III focuses on optimizing MPHIGA with DGA to develop the scheduling model. Section IV evaluates the performance and practical application of the proposed model. Finally, Section V presents the study's conclusions.

II. Related Works

MPHIGA is an improved genetic algorithm that uses multiple populations for simultaneous optimization, enhancing global search capabilities and preventing premature convergence. Many researchers have explored its core characteristics. For example, Yu et al. proposed an algorithm based on a hybrid genetic algorithm to reduce the total response time between cloud data centers and medical devices. By combining the genetic algorithm with a hybrid metaheuristic method and conducting simulation experiments, the study demonstrated the effectiveness of this approach [7]. Sun H's team introduced a particle swarm optimization method based on genetic algorithms for inverse lithography technology. This method improved lithographic imaging performance through iterative optimization. Experimental results showed that it had better convergence capability compared to traditional genetic algorithms [8]. Additionally, DGA enhances search efficiency and solution quality through the cooperation of two populations. Its strong search ability has led to its adoption by some researchers. Fang et al. developed a novel interval prediction method for photovoltaic power generation to improve forecasting timeliness. Applying an integral dual-domain decomposition method for learning and prediction, the study found that the error values were significantly reduced [9]. Zhang's team proposed a wave source localization method for harmonic analysis. By analyzing harmonic and non-harmonic sources from a time-domain perspective, simulation experiments verified the effectiveness of this approach [10]. Zhan et al. investigated the freshness of sensory data in unmanned aerial vehicles and introduced a domain generation algorithm-based strategy to obtain near-optimal solutions. By enabling rapid decision-making, this approach improved computational efficiency. Results showed that the algorithm exhibited strong flexibility [11].

Intelligent logistics vehicle scheduling plays a key role in transportation and has been widely applied by logistics enterprises in recent years. Effective scheduling models allows for efficient resource allocation. Cho's team addressed the

limitations of traditional logistics research by proposing a heuristic scheduling algorithm. By introducing various cost functions to evaluate scheduling results, experiments showed that this method effectively reduced delivery costs [12]. Elgharably et al. developed a hybrid search algorithm based on the vehicle routing problem in operations research. This algorithm solves solve multi-objective problems with balance trade-offs. Simulation experiments verified its feasibility and effectiveness [13]. Berghman et al. introduced a general model for outbound vehicle routing and conducted a review of comprehensive scheduling problems using an integrated approach. The study confirmed that this method provided valuable data references for future research directions [14]. Ding's team proposed a novel vehicle scheduling method to recommend optimal routes for electric vehicles, aiming to mitigate traffic congestion. By integrating multiple interactive modules for iterative decision-making, results demonstrated the feasibility of this approach [15]. Shen and Yan examined the impact of bus services on traffic and introduced a dynamic vehicle scheduling method. Using hybrid dynamic control, the method handled abnormal operating conditions effectively. Experimental results indicated strong flexibility and timeliness [16].

Existing research has made progress in reducing logistics costs and optimizing logistics services. However, challenges such as information delays and inaccuracies remain. To address these limitations, this study integrates DGA with MPHIGA to develop a novel intelligent logistics vehicle scheduling model. Each scheduling scheme undergoes a fitness function, and individuals are selected for genetic and crossover operations based on fitness values. The study employs a multi-population hybrid improved genetic algorithm, primarily addressing the issues of traditional genetic algorithms in logistics scheduling, such as falling into local optima and insufficient population diversity. By introducing domain name generation algorithms to optimize dynamic coding strategies, it resolves the decoding efficiency bottleneck under complex constraints. Compared to existing models, its multi-population co-evolution mechanism simultaneously optimizes vehicle allocation and route generation, making it suitable for high-dimensional logistics scheduling. This approach aims to enable rapid adjustments to scheduling strategies in response to changing logistics demands, providing flexible logistics solutions.

III. The Multi-Population Hybrid Improved Vehicle Scheduling Model for Intelligent Logistics

A. *Intelligent Logistics Vehicle Scheduling Strategy Based on MPHIGA*

Compared with the traditional genetic algorithm, MPHIGA breaks away the framework of single-population genetic evolution and assigns different control parameters to different populations through simultaneous optimization searches [17]. In MPHIGA, multiple populations achieve collaborative evolution through specific operational factors, enhancing global search capability and prevents premature convergence compared to the traditional genetic algorithm [18,19]. MPHIGA is applied to intelligent vehicle scheduling strategies, automatically calculating the optimal vehicle scheduling plan based on multiple factors such as vehicle location, real-time road conditions, and cargo volume. The workflow of MPHIGA is shown in Fig. 1.

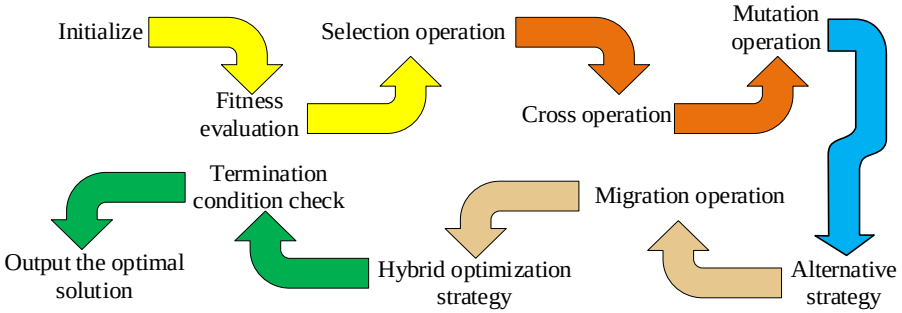


Fig. 1. Schematic diagram of MPHIGA workflow.

As shown in Fig. 1, MPHIGA first initializes the population size and the scale of each population randomly. The number of sub-populations is set according to the problem complexity and resource constraints. The fitness function evaluates each individual in the sub-population based on the problem requirements. In the selection process, individuals with higher fitness values are more likely to be selected. The crossover and different crossover strategies can be used to maintain diversity. The fitness function is given in Eq. (1).

$$f_j = \begin{cases} F(G_{ej}), x' = 1 \\ F(G_{ej}) + M_0, x' = 0 \end{cases} \quad (1)$$

In Eq. (1), $F(G_{ej})$ and $x' = 1$ represent the objective function value and the feasibility of the decoded solution as a binary variable, respectively. M_0 is assigned a finite integer. x' and $x' = 0$ are binary variables and represent an infeasible decoded solution. When applying MPHIGA to the intelligent logistics vehicle scheduling strategy, the intelligent scheduling dynamically adjusts the transportation plan based on demand, reducing empty mileage and transportation time. The key components of intelligent logistics vehicle scheduling are shown in Fig. 2.

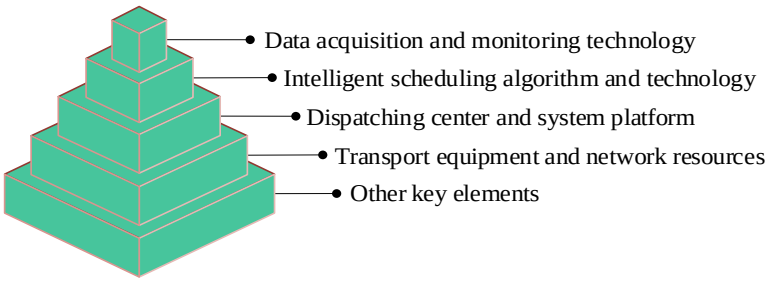


Fig. 2. Schematic diagram of the components of intelligent logistics vehicle scheduling.

$$F(l)=\begin{cases}\sum_{I=1}^{n-1}\sqrt{(X_I+1-X_I)^2+(y_I+1-y_I)^2}+num\times p_{punish1}\\ I=1,2,3,.....,n-1\end{cases} \quad (2)$$

As shown in Fig. 2, the key components of intelligent logistics vehicle scheduling consist of five main parts. The data collection and monitoring technology uses sensors and onboard intelligent devices. Sensors collect real-time key data from logistics vehicles, while onboard intelligent devices monitor vehicle status to ensure accurate execution of scheduling instructions and enhance transportation safety. The intelligent scheduling algorithm uses big data and artificial intelligence to optimize logistics scheduling, with cloud computing and big data analysis providing efficient data storage and processing capabilities. The dispatch center and system platform include the intelligent dispatch center, user interface, and operating system. The dispatch center receives and processes data from sensors and onboard devices, generating scheduling instructions based on intelligent algorithms and transmitting them in real-time to drivers and vehicles. The evaluation function for the total path length in logistics vehicle scheduling is given in Eq. (2) [20].

In Eq. (2), X_I and l represent the midpoint of the line segment connecting the start and end points and the path, respectively. num and $p_{punish1}$ are the number of infeasible segments in the path and the penalty value. y_I represents the vertical coordinate when the two-dimensional path encoding is converted into one-dimensional encoding. By sorting the selection probabilities assigned to individuals, the cumulative probability equation is given in Eq. (3).

$$p_{\sum_i}=\sum_{i=1}^{NP}p_i \quad (3)$$

Transportation equipment and logistics networks ensure the execution of logistics transportation tasks and support the implementation of intelligent logistics vehicle scheduling. Other key elements include real-time detection and early warning systems, which reduce transportation risks. The intelligent logistics vehicle scheduling strategy based on MPHIGA is shown in Fig. 3.

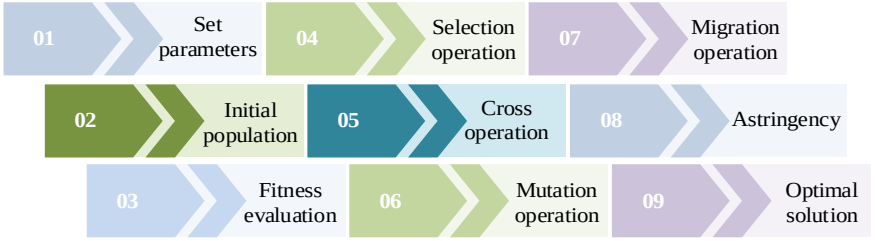


Fig. 3. Intelligent logistics vehicle scheduling strategy based on MPHIGA.

As shown in Fig. 3, in the MPHIGA-based intelligent logistics vehicle scheduling strategy, each individual represents a potential scheduling plan. After setting the basic parameters of the genetic algorithm, each sub-population is randomly initialized. The fitness function defines and evaluates individuals and scheduling plans based on the vehicle scheduling problem requirements and multiple indicators. Within each sub-population, the crossover operation generates new offspring. The mutation operation is usually random, and individuals from different sub-populations exchange information to enhance capabilities. The exchange strategies include both synchronous and asynchronous communication. The threshold value in the exchange strategy is given in Eq. (4).

$$E=\frac{1}{N}\sum_{i=1}^{N}\left|f_i-f_{avg}\right| \quad (4)$$

In Eq. (4), f_i, N, and f_{avg} represent the fitness value of the i-th individual in the population, the number of individuals in the population, and the average fitness value. The crossover probability equation is given in Eq. (5).

$$p_c = \begin{cases} p_{c1} - (p_{c1} - p_{c2}) \dfrac{f_{\max} - f'}{f_{\max} - f_{avg}} & f' \geq f_{avg} \\ p_{c1} & f' < f_{avg} \end{cases} \quad (5)$$

In Eq. (5), $f_{\max}$ and p_{c1} represent the fitness value of the optimal individual in the population and the maximum mutation probability. f' and p_{c2} represent the larger fitness value and the minimum mutation probability of the two chromosomes to be crossed.

B. Construction of the Intelligent Logistics Vehicle Scheduling Model with DGA

Although applying MPHIGA to intelligent logistics vehicle scheduling allows precise environmental recognition through high-precision sensors and visual recognition technologies, it may still prematurely converge to local optima when facing complex problems due to ineffective selection strategies. For large-scale problems, MPHIGA requires multiple iterations to generate and evaluate populations, which leading to increased computational time and resource consumption [21]. Therefore, DGA is introduced to optimize MPHIGA, allowing the search process to adapt dynamically and avoid premature convergence. DGA is an algorithm that generates a large number of random and short-lived domain names. It can generate different domain names as needed, offering flexibility and adaptability. The domain generation process of DGA is shown in Fig. 4.

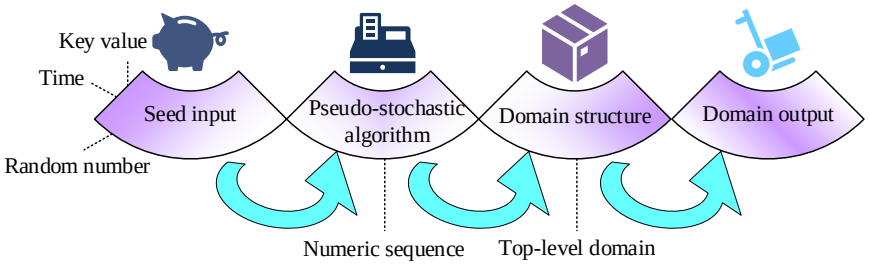


Fig. 4. Schematic diagram of DGA domain name generation.

As shown in Fig. 4, DGA initializes a pseudo-random algorithm using one or more specific parameters as seeds, which typically include key values, time, and random numbers. Different seeds generate different domain sequences. During the execution of the pseudo-random algorithm, DGA generates a series of numerical sequences based on the input seed, forming part of the domain names. The generated sequences are then combined into one or more domain names. These domain names are often used by malware or attackers to communicate with servers or other targets. DGA-generated domain names exhibit high unpredictability and randomness, allowing attackers to periodically change the seed to alter the generated domain sequences. The calculation method for the consecutive number ratio is given in Eq. (6).

$$C_d = \frac{N_{cd}}{L} \quad (6)$$

In Eq. (6), N_{cd} and L represent the number of consecutive number pairs and domain name length, respectively. The equation for domain name length is given in Eq. (7).

$$L = len(domain) \quad (7)$$

In Eq. (7), $domain$ and $len()$ represent the domain name string and its length. The dynamic generation mechanism of DGA is used to adjust the search strategy of MPHIGA, and the optimization process is shown in Fig. 5.

As shown in Fig. 5, DGA generates a large number of random domain names through its randomness generation mechanism. This randomness is introduced into MPHIGA for population initialization and mutation operations, enhancing the global search capability. MPHIGA already has strong global search capabilities through hybrid improvement strategies, and introducing DGA further strengthens the ability to explore optimal solutions in a broader solution space. The entropy value of domain names is given in Eq. (8).

$$H = -\sum_{z=1}^{n} p_z \log p_z \quad (8)$$

In Eq. (8), p_z represents the frequency of character z. The number of numeric characters is calculated using Eq. (9).

$$Nd = \sum_{z \in digits} count(z) \quad (9)$$

In Eq. (9), $digits$ represents the set of numeric character z. The user interface provides an intuitive interface for seamless integration with other logistics management systems, facilitating monitoring, collaboration, and data sharing. The randomness of DGA helps prevent MPHIGA based on premature convergence to local optima. By continuously introducing new random elements, the algorithm maintains exploration within the solution space and ultimately finds the optimal solution. The intelligent logistics vehicle scheduling model based on DGA-MPHIGA is shown in Fig. 6.

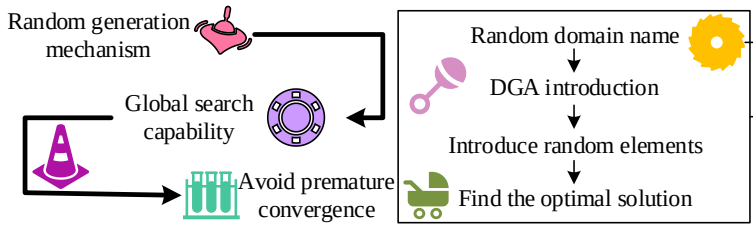


Fig. 5. Schematic diagram of DGA optimization of MPHIGA.

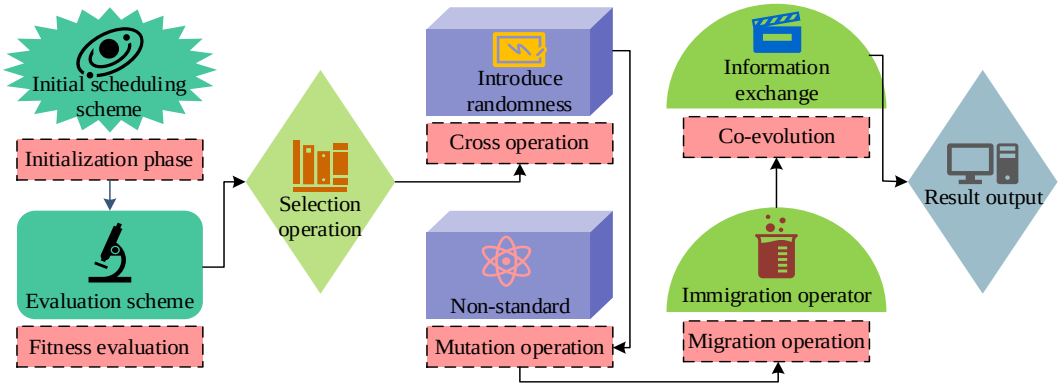


Fig. 6. Intelligent logistics vehicle scheduling model based on DGA-MPHIGA.

As shown in Fig. 6, the random generation mechanism of DGA produces an initial set of scheduling plans. After evaluating the fitness of each scheduling plan, the best individuals are selected for genetic and crossover operations. Mutation operations are usually random, with their probability and intensity adjusted according to the scenario. Within sub-populations, replacement strategies ensure continuous optimization of the population. DGA enables non-standard crossover and mutation operations with a certain probability, and the migration operator facilitates information exchange and collaborative evolution among populations. The model terminates when the stopping condition is met or the preset iteration count is reached, ultimately obtaining the optimal logistics vehicle scheduling plan. The equation for computing the probability distribution of chromosomes in the population is given in Eq. (10).

$$p_i = \begin{cases} p \times (1-p)^{i-1}, & i = 2, \cdots, NP \\ (1-p)^{i-1}, & i = 1 \end{cases} \tag{10}$$

In Eq. (10), i and NP represent the individual's ranking and the population size. p represents a constant within the range of $(0,1)$. The optimal path equation is given in Eq. (11).

$$\begin{aligned} &MinZ = \sum_{w=1}^{W} \left\{ \sum_{u=1}^{A_w} C_w D_{k_{w(u-1)} k_{wi}} D_{k_{mm} k_{w0}} sign(A_w) \right\} \\ &w = 1, 2, \cdots W \end{aligned} \tag{11}$$

In Eq. (11), A_w and k_w represent the number of customer points served by vehicle w and the w-th path, respectively. k_{wi} and k_{w0} are the sequence positions of customer points in path w, while C_W and D represent the cost per unit travel distance of vehicle w and the maximum allowable travel distance. The vehicle transportation cost equation is given in Eq. (12).

$$M = h \sum_{s,q=1}^{m} x_{sq} d_{sq} \tag{12}$$

In Eq. (12), h and d_{sq} represent the transportation cost per unit distance and the transportation distance between customer s and customer q. m represents the total number of customers.

IV. Performance Verification of the DGA-MPHIGA Intelligent Scheduling Model

A. Performance Analysis of the Intelligent Logistics Vehicle Scheduling Model

To analyze the performance of the intelligent logistics vehicle scheduling model based on DGA-MPHIGA, the study used intelligent logistics vehicle scheduling models based on MPHIGA, Simple Genetic Algorithm (SGA), and Dual Population Hybrid Genetic Algorithm (DPHGA) as comparison models. The experiments were conducted on a Windows XP SP3 operating system with a Pentium(R) 4 2.66 GHz CPU and 12 GB of RAM. The experimental datasets included the TSPlib dataset and the Solomon dataset. The TSPlib dataset was used as the training set, containing specific city coordinates and distance matrix information for solving the traveling salesman problem and routing problems in vehicle scheduling. The Solomon dataset was used as the test set, providing detailed parameters such as customer location coordinates, time windows, and demand quantities. The study first compared the scheduling costs of the four models on the training and test sets, and the results are shown in Fig. 7.

As shown in Fig. 7(a), when the scheduling frequency reached five times, the cost of the proposed model was only 0.112 million yuan. However, the cost of the MPHIGA scheduling model had already reached 0.128 million yuan when the scheduling frequency was three times. Fig. 7(b) shows that the DGA-MPHIGA scheduling model maintained lower cost consumption than the comparison models in the test set. When the scheduling frequency was four times, the cost of the MPHIGA and DPHGA scheduling models was 0.137 million yuan and 0.159 million yuan, respectively, while the cost of the SGA scheduling model was the highest among the four models at 0.188 million yuan. These results indicated that the DGA-MPHIGA scheduling model exhibited lower cost consumption in both the training and test sets, demonstrating higher resource utilization. Next, the study conducted a comparative analysis of the scheduling time of the four models on the TSPlib dataset, with results shown in Fig. 8.

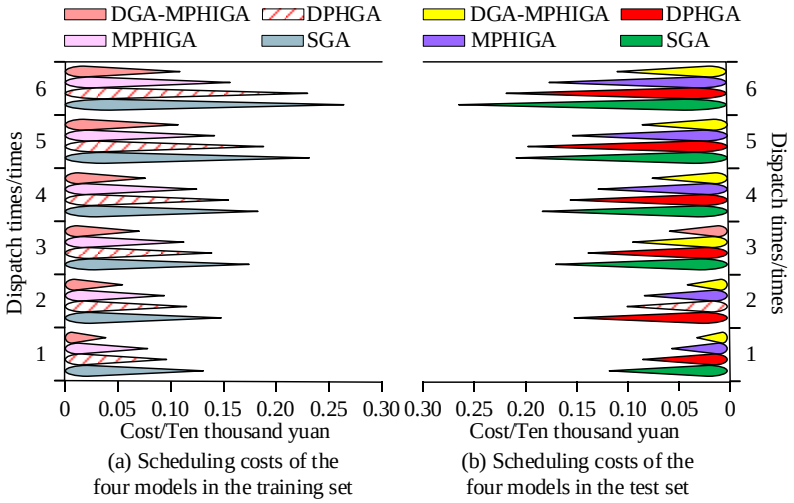


Fig. 7. Comparison of the scheduling costs of four models in training and test sets.

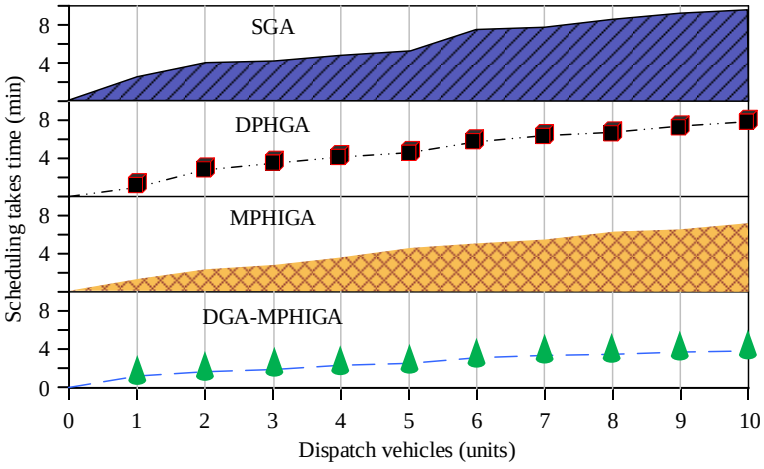


Fig. 8. Comparison of the scheduling time of four models in the TSPlib dataset.

As observed in Fig. 8, when the number of scheduled vehicles was two, the scheduling time of the DGA-MPHIGA and MPHIGA models was 1.24 minutes and 2.67 minutes, respectively. When the number of scheduled vehicles reached ten, the scheduling time of the DGA-MPHIGA model was 3.98 minutes, while the scheduling time of the SGA model was 5.94 minutes longer than that of the proposed model. Meanwhile, the scheduling times of the MPHIGA and DPHGA models were 7.31 minutes and 7.98 minutes, respectively. These results indicated that the DGA-MPHIGA model achieved the shortest scheduling time among the four models. The model was capable of computing the optimal vehicle scheduling solution in a shorter time, significantly improving overall logistics efficiency. Additionally, the shorter scheduling time allowed the model to quickly adapt to new order demands and traffic conditions. Finally, the study analyzed the vehicle distribution route length and time of the four models on the training and test sets, with results shown in Table I.

TABLE I COMPARISON OF VEHICLE DELIVERY ROUTE LENGTH AND TIME

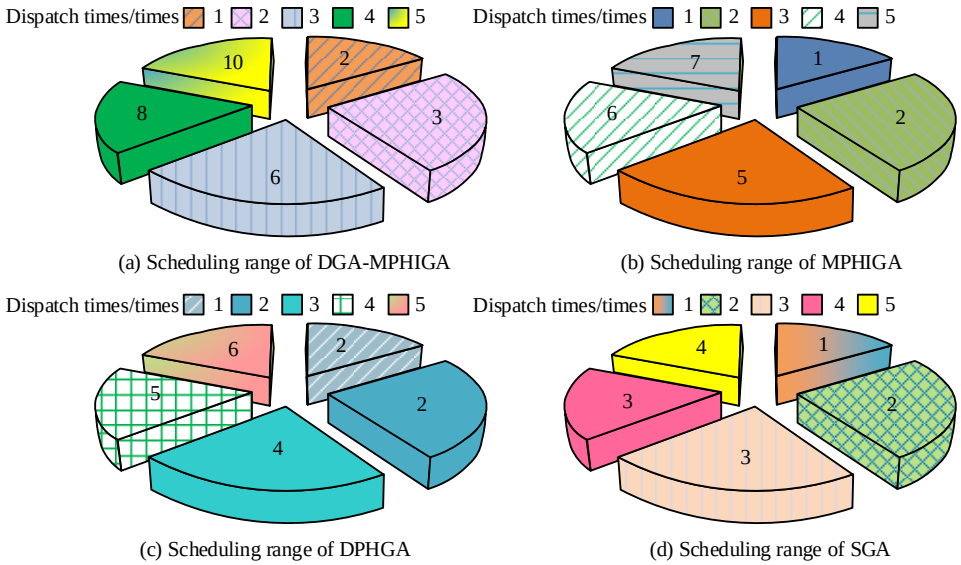
Data set	Scheduling model	Average calculated path length (km)	Optimal path distance length (km)	Running time (s)
Training set	DGA-MPHIGA	411.32	410.84	9.54
	MPHIGA	423.15	419.27	14.36
	DPHGA	435.81	425.33	18.24
	SGA	449.25	437.41	21.39
Test set	DGA-MPHIGA	412.94	408.57	9.21
	MPHIGA	427.35	424.35	14.17
	DPHGA	436.87	431.23	19.34
	SGA	451.29	439.78	22.17

As presented in Table I, in the training set, the average computed path length and optimal path length of the DGA-MPHIGA model were 411.32 km and 410.84 km, respectively, both lower than those of the comparison models. The SGA model exhibited the highest average computed path length at 449.25 km, with a runtime of 21.39s. The optimal path length of the MPHIGA model in the training set increased compared to previous results, reaching 424.35 km, with a runtime of 14.17s. At this point, the optimal path length of the DGA-MPHIGA model was only 408.57 km, with its runtime being the shortest among the four models, at just 9.21s. These results demonstrated that the DGA-MPHIGA model had the shortest average computed path length and optimal path length, leading to faster delivery speeds and enabling the rapid computation of optimal logistics routes.

B. *Evaluation of the Practical Effectiveness of the Intelligent Scheduling Model*

To further validate the effectiveness of the intelligent logistics vehicle scheduling model in real-world applications, the study set the number of foreign trade enterprise vehicles to ten, along with ten loading points and thirteen unloading points. The transportation time was set to two hours, and the unit transportation cost was adjusted based on 5 yuan/km. The penalty coefficients for exceeding the time window and the vehicle's rated load were set to 200 and 10,000, respectively. The study first analyzed the scheduling range of the four models when scheduling ten enterprise vehicles, with results shown in Fig. 9.

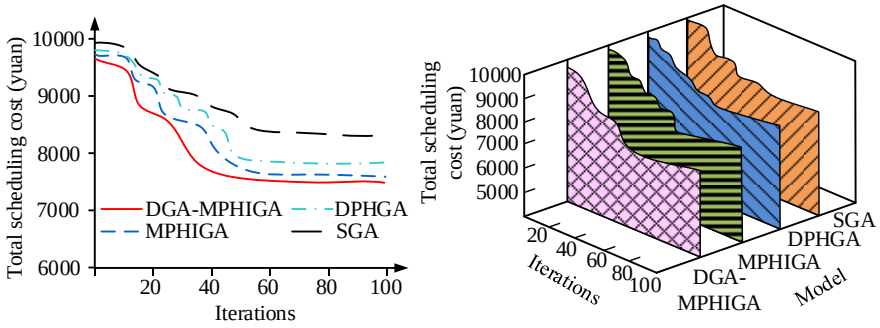
As shown in Fig. 9, the DGA-MPHIGA model achieved a significantly wider scheduling range than the other three comparison models under the same scheduling frequency. When the scheduling frequency was two times, the DGA-MPHIGA model scheduled three enterprise vehicles. When the scheduling frequency increased to four times, the SGA model scheduled three enterprise vehicles, while the DPHGA and MPHIGA scheduling models scheduled five and six enterprise vehicles, respectively. At this point, the DGA-MPHIGA model was capable of scheduling eight enterprise vehicles. When the scheduling frequency reached five times, the model successfully scheduled all ten enterprise vehicles. These results indicated that the DGA-MPHIGA model demonstrated excellent vehicle scheduling capabilities, handling broader and more complex logistics transportation tasks. The study then analyzed the total cost optimization iterations of the four models under different vehicle demand scenarios, with the iterative change curves shown in Fig. 10.



(a) Scheduling range of DGA-MPHIGA (b) Scheduling range of MPHIGA

(c) Scheduling range of DPHGA (d) Scheduling range of SGA

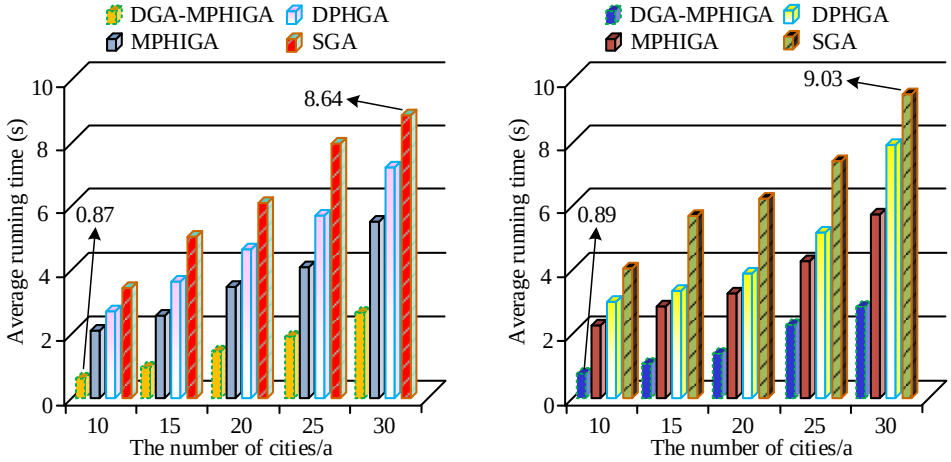
Fig. 9. Result diagram of scheduling range for dispatching enterprise vehicles.



(a) Iterative changes in the total cost optimization of the four models when four vehicles are required

(b) Iterative changes in total cost optimization for the four models when 8 vehicles are required

Fig. 10. Total cost optimization iteration curve under different requirements.



(a) The average operating time of the four models when four vehicles are required

(b) The average operating time of the four models when eight vehicles are required

Fig. 11. Comparison of average running time under different requirements.

As shown in Fig. 10(a), when the vehicle demand was four, the DGA-MPHIGA model exhibited a shorter convergence time. When the iteration count reached 37, the total scheduling cost of the model gradually stabilized, converging faster than the comparison models. The DPHGA model had the longest convergence time, reaching 51 iterations. Fig. 10(b) shows that when the vehicle demand was eight, the total scheduling cost optimization curves of all four models exhibited a downward trend as the iteration count increased, eventually stabilizing at different iteration points. The convergence time and iteration count of the DGA-MPHIGA model were both shorter than those of the comparison models. These results indicated that the DGA-MPHIGA model learned data patterns faster and reached stability with lower computational resource consumption. Finally, the study compared the scheduling time of the four models under different vehicle demand scenarios, with results shown in Fig. 11. As summarized in Fig. 11, the average runtime of the DGA-MPHIGA model under different vehicle demand scenarios was the lowest among the four models. When the vehicle demand was four, and the number of cities was ten, the shortest average runtime of the DGA-MPHIGA model was 0.87 seconds. At this point, the average runtime of the SGA model was 3.02 seconds, while those of the MPHIGA and DPHGA models were 1.97 seconds and 2.28 seconds, respectively. When the vehicle demand reached eight, the average runtime of all four models increased as the number of cities increased. When the number of cities reached 25, the average runtime of the DGA-MPHIGA and SGA models was 3.94 seconds and 6.89 seconds, respectively, while the runtime of the DPHGA model was 4.86 seconds. The SGA model exhibited an average runtime of 9.03 seconds when the number of cities reached 30. These results indicated that the DGA-MPHIGA model made scheduling decisions more quickly when handling different vehicle demands, ensuring faster delivery of goods from the starting point to the destination.

V. Conclusion

To address the shortcomings of current intelligent logistics vehicle scheduling methods in dynamic data acquisition and managing security, this study combined DGA and MPHIGA in an intelligent logistics vehicle scheduling model. By leveraging the dual-population mechanism of DGA and the hybrid strategy of MPHIGA, the model achieved information sharing and co-evolution, balancing global search and local exploitation with the feasible vehicle scheduling solution is represented as an individual in genetic algorithm by chromosome coding. The results showed that the optimal path length of the MPHIGA model in the training set increased, reaching 424.35 km, with a running time of 14.17s. At this point, the optimal path length of the DGA-MPHIGA model is only 408.57km, with the shortest

running time among the four models, at just 9.21s. When the number of scheduling instances reaches 5, the cost of the research model is only 0.112 million yuan. However, when the MPHIGA scheduling model has 3 instances, the cost has already reached 0.128 million yuan. Based on the above results, the proposed DGA-MPHIGA model outperformed the comparison models in scheduling time, cost, and range, demonstrating excellent vehicle scheduling capability and handling more complex logistics transportation tasks. The proposed model overcomes the limitations of traditional algorithms in multi-model scheduling and dynamic order response through multi-population co-evolution and dynamic coding optimization, provides an expandable decision framework for building highly flexible intelligent logistics systems, and helps reduce the empty load rate of cold chain transportation. The differences in results between various datasets stem from the structural characteristics of the TSPlib and Solomon datasets. The TSPlib dataset focuses on classic path optimization, while the Solomon dataset includes complex constraints such as time windows and dynamic demands, better reflecting the challenges of real-time logistics scenarios. The proposed algorithm enhances global search capabilities through multi-population co-evolution mechanisms, combined with domain generation algorithms to dynamically adjust coding rules, demonstrating significant advantages when handling high-dimensional constraints. However, as the number of cities increased, the average runtime of the model continued to rise, indicating that its performance still required improvement. Future work will focus on further optimizing the model.

ACKNOWLEDGMENT

This work was sponsored in part by Research on the Interaction Mechanism Between the Digital Economy and the Logistics Industry in Liaoning Province Based on a Spatial Econometric Model (2025XJLXQN039).

REFERENCES

[1] El Idrissi R N, Ouassaid M, Maaroufi M. Game Theory Approach for Energy Consumption Scheduling of a Community of Smart Grids. Journal of Electrical Engineering & Technology, 2023, 18(4): 2695-2708.

[2] Li X, Wang XL, Xu J, Chen W, Lyu B. (2024). Integrated Emergency Vehicle Scheduling Model for Urban Floods: Considering Human-Vehicle Coupling and Road Network Accessibility. TRANSPORTATION RESEARCH RECORD. 10.1177/03611981231209292

[3] Akbari V, Sadati H, Salman F, Shiri D. Minimizing total weighted latency in home healthcare routing and scheduling with patient prioritization. Or Spectrum, 2023, 45(3): 807-852.

[4] Kumar D P, Bindeshwar S, Varun K, Deependra S. A novel approach for comparative analysis of distributed generations and electric vehicles in distribution systems. Electrical engineering, 2024, 106(3): 2371-2390.

[5] Zhang L, Cheng L, Alsokhiry F, Mohamed M A. A Novel Stochastic Blockchain-Based Energy Management in Smart Cities Using V2S and V2G. IEEE transactions on intelligent transportation systems, 2023, 24(1): 915-922.

[6] Huang N, Hu L, Wang R, Cai G, Guo Y, Zhao X. Multi-Objective Optimal Scheduling of Distribution Network with Electric Vehicle Charging Load Considering Time-Varying Road Impedance. Journal of Electrical Engineering & Technology, 2023, 18(4): 2667-2681.

[7] Yu J, Wang M, Yu J H, Arefzadeh S M. A new approach for task managing in the fog-based medical cyber-physical systems using a hybrid algorithm. Circuit world, 2023, 49(3): 294-304.

[8] Sun H, Zhang Q, Jin C, Li Y, Tang Y, Wang J, Hu S, Liu J. Inverse Lithography Source Optimization via Particle Swarm Optimization and Genetic Combined Algorithm. IEEE Photonics journal, 2023, 15(2): 1-12.

[9] Fang H, Wu Q, Zhang Y. Interval prediction of PV power based on double sample entropy framework and integrated bidomain decomposition deep learning. IET Renewable Power Generation (Wiley-Blackwell), 2024, 18(6): 979-989.

[10] Zhang C, Li Y, Han W, Song G, Zhang H. Time-domain harmonic source location and evaluation methods based on non-linear and time-varying properties of devices. IET generation, transmission & distribution, 2024, 18(16): 2604-2624.

[11] Zhan C, Hu H, Wang J, Liu Z, Mao S. Tradeoff Between Age of Information and Operation Time for UAV Sensing Over Multi-Cell Cellular Networks. IEEE transactions on mobile computing, 2024, 23(4): 2976-2991.

[12] Cho Y H, Baek D, Chen Y, Min J J, Vinco S, Macii E, Poncino M. Multi-Criteria Coordinated Electric Vehicle-Drone Hybrid Delivery Service Planning. IEEE Transactions on Vehicular Technology, 2023, 72(5): 5892-5905.

[13] Elgharably N, Easa S, Nassef A, Damatty A E. Stochastic Multi-Objective Vehicle Routing Model in Green Environment With Customer Satisfaction. IEEE transactions on intelligent transportation systems, 2023, 24(1): 1337-1355.

[14] Berghman L, Kergosien Y, J.-C. B. A review on integrated scheduling and outbound vehicle routing problems. European Journal of Operational Research, 2023, 311(1): 1-23.

[15] Ding L, Zhao D, Ma F H. Learning to Help Emergency Vehicles Arrive Faster: A Cooperative Vehicle-Road Scheduling Approach. IEEE transactions on mobile computing, 2023, 22(10): 5949-5962.

[16] Shen Y, Yan M. HTN planning for dynamic vehicle scheduling with stochastic trip times. Neural computing & applications, 2023, 35(13): 9917-9930.

[17] Yang Y, Yeh H, Nguyen R. A Robust Model Predictive Control-Based Scheduling Approach for Electric Vehicle Charging with Photovoltaic Systems. IEEE Systems Journal, 2023, 17: 111-121.

[18] Wu T, Wen P, Tang S. Optimal scheduling strategy of AUV based on importance and age of information. Wireless networks, 2023, 29(1): 87-95.

[19] Liu Z, Liwang M, Hosseinalipour S, Dai H, Gao Z, Huang L. RFID: Towards Low Latency and Reliable DAG Task Scheduling Over Dynamic Vehicular Clouds. Vehicular Technology, IEEE Transactions on, 2023, 72(9): 12139-12153.

[20] Azad U, Behera B K, Farouk P A. Solving Vehicle Routing Problem Using Quantum Approximate Optimization Algorithm. IEEE transactions on intelligent transportation systems, 2023, 24(7): 7564-7573.

[21] Shehadeh K S. Distributionally Robust Optimization Approaches for a Stochastic Mobile Facility Fleet Sizing, Routing, and Scheduling Problem. Transportation Science, 2023, 57(1): 197-229.

Fuzzy Delphi Method: A Step-by-Step Guide to Obtaining Expert Consensus on Mobile Tourism Acceptance Culture

Syaifullah[1]*, Shamsul Arrieya Ariffin[2]*, Norhisham Mohamad Nordin[3]
Computer Science and Digital Technology Department, Faculty of Computing and Meta-Technology (META),
Sultan Idris Education University, Malaysia[1, 2, 3]
Information System Department, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia[1]

***Abstract*—Mobile technology has developed rapidly in a short period of time, which has greatly changed the tourism sector and led to the emergence of Mobile Tourism (MT). To ensure that MT grows well and is widely used, it is important to know how people from different cultures accept it. This study provides a complete description of how to use the Fuzzy Delphi Method (FDM) to obtain expert agreement on the most important factors that influence how acceptable mobile tourism is from a cultural perspective. This study uses the Technology Acceptance Model (TAM) and Hofstede's Cultural Dimensions to carefully find and confirm the variables and indicators and how they are interrelated. This approach describes a rigorous process with nine stages in reaching expert agreement. The results revealed that experts largely agreed on the variables related to perceived usefulness, perceived trust, perceived ease of use, and facilitating conditions in the TAM framework, as well as some variables of collectivism, uncertainty avoidance, and long-term orientation in Hofstede's cultural aspects. This study also verified and validated the overall relationship between variables in building the Mobile Tourism Cultural Acceptance (MTCA) framework, the general and specific interactions between variables, and the function of cultural dimensions as mediators. This study shows how important it is to get expert opinion when making a comprehensive plan on how to use technology in a culturally acceptable environment for mobile tourism. The information obtained has a major impact on mobile tourism developers, policy makers, and marketers who want to make MT more popular.**

Keywords*—Cultural acceptance; Fuzzy Delphi Method (FDM); hofstede's cultural dimensions; mobile tourism; technology acceptance model (TAM)

I. INTRODUCTION

The tourism industry has essentially entered a new era with the advent of smartphones and increasingly widespread internet connectivity, which is emphasized by the emergence of mobile tourism (MT) [1], [2]. Mobile tourism can be defined as the use of mobile devices and applications to access information, various forms of services, and experiences related to travel, including planning, purchasing, navigating destinations, and sharing on social media [1], [3]. Due to its widespread use, it provides tourists with convenience, personalization, and unlimited real-time interaction, thus substantially enhancing the travel experience of tourists [4]. However, the success of widespread acceptance and adoption of mobile tourism is not only due to technological advances, but is also influenced by complex human factors, especially cultural nuances that influence user perceptions and behavior [5], [6].

It is a challenge for researchers and practitioners to try to understand the factors that influence the acceptance of new technologies, especially in diverse cultural contexts. Perceived usefulness (PU) and perceived ease of use (PEOU) are the basic determinants of technology adoption, according to the Technology Acceptance Model (TAM), a widely adopted framework [7]. Although TAM offers a strong model and foundation, it appears to ignore the significant impact of cultural values on technology adoption. This assumption actually limits TAM's explanatory efforts in the context of multicultural societies, where variables such as collectivism, uncertainty avoidance, and long-term orientation significantly influence user perceptions and behavioral intentions [8], [9]. Similarly, according to [5], [10] most mobile tourism research focuses more on technical usability and system quality, neglecting the sociocultural dimensions of user acceptance models.

Therefore, to address this gap, a model is needed that can synthesize the TAM Model and integrate it with Hofstede's cultural dimensions in reflecting the user's value system [11], [12]. Hofstede's Cultural Dimensions Theory provides important insights in analyzing these cultural influences, and is a framework for understanding how national and individual cultures influence behavior, including technology interactions [8]. Therefore, ensuring that these two theoretical perspectives can be integrated into one is crucial for developing a comprehensive understanding of Mobile Tourism Cultural Acceptance (MTCA). MTCA not only provides theoretical contributions, but is a very important strategic tool to help guide the development of a more culturally responsive mobile tourism system and also supports inclusive digital transformation, especially in culturally diverse countries such as Indonesia [13], [14].

To follow up on this, it is necessary to identify and validate critical or important variables and prove how they relate to the MTCA framework in a proposed model. Therefore, it is very important to utilize the knowledge and insights gathered from experts in various fields, namely from the perspective of tourism, academics, and practitioners related to tourism. The Delphi method, originally developed by the RAND Corporation, is a widely known systematic forecasting technique. It involves collecting and aggregating opinions from a panel of experts

through a series of structured questionnaires, with the aim of reaching consensus [15]. However, the conventional Delphi method still has limitations, including the difficulty in accurately representing subjective human judgment and distortion of information from extreme opinions [16].

To address and provide solutions to these conditions, the Fuzzy Delphi Method (FDM) emerged as an appropriate and effective extension concept. FDM conceptually incorporates fuzzy set theory into the conventional Delphi process, allowing experts to express their opinions using fuzzy numbers (e.g., linguistic terms such as "very important" or "somewhat important") rather than precise numerical values [17]. This method effectively encapsulates the ambiguity and imprecision inherent in human judgment, resulting in a more reliable and accurate consensus-building process. FDM has been widely and effectively applied in various sectors, such as environmental management, healthcare, and technology assessment, demonstrating its capacity to manage subjective expert opinions to reach collective agreement [18].

The purpose of this study is to provide a comprehensive step-by-step guide in the application of the Fuzzy Delphi Method to reach expert consensus on variables and indicators to their relationships within the Mobile Tourism Culture Acceptance (MTCA) framework. This study also has a specific objective to determine the elements considered important by experts in the development of MTCA by combining TAM and Hofstede's cultural dimensions. These findings will not only enhance the theoretical understanding of technology acceptance to culture in the context of mobile tourism but also provide practical recommendations for stakeholders in the mobile tourism ecosystem.

II. Literature Review

A. Mobile Tourism and Technology Acceptance

The way individuals plan, experience, and share their travel experiences is being revolutionized by mobile tourism. The accessibility of tourism information and services at any time and in any location has been significantly enhanced by the widespread use of mobile devices, resulting in increased personalization and flexibility for travelers [19]. Mobile technology is increasingly becoming a critical tool for the modern traveler, from translation services, time and mobile booking applications to augmented reality tour guides and social media platforms for sharing their travel experiences [2], [20], [21], [22]. However, the acceptance of these technologies depends on the success of the target consumer.

Such complex factors of new technology developments influence the phenomenon of technology acceptance. Therefore, it is very important to understand these factors in order to create more effective mobile tourism solutions that are in accordance with user preferences and needs. This is supported by previous studies that have used many of these theoretical models and concepts to predict and explain user behavior related to technology acceptance.

B. Technology Acceptance Model (TAM)

Technology Acceptance Model (TAM) proposed by Davis [7], s one of the most important and appropriate frameworks for understanding how people use technology. TAM states that there are two basic beliefs that are the main things that influence a person's desire to use technology or not:

- Perceived Usefulness (PU): How much a person believes that using a particular system will help them do their job better [23], [24]. In the context of mobile tourism, this statement means how much visitors think that using a mobile tourism application will make their trip better, make it more efficient, or provide them with more accurate and useful information.
- Perceived Ease of Use (PEOU): The extent to which a person believes that using a particular system would make things easier [23], [24]. In the context of mobile tourism, this relates to how easy it is for users to navigate and use mobile tourism applications and services.

The TAM model is extended by adding other factors to detail and explain tourist intentions. These factors include ubiquitous connectedness, trust, and personality traits such as openness to experience and demand for arousal. [25], [26] examined and confirmed most of the hypothesized models suggested in their study, namely finding that perceived utility, perceived ease of use, ubiquitous connectivity, trust, openness to experience, and demand for arousal are all strong predictors of tourists' intention to purchase tourism-related goods and services through mobile platforms. Of all these factors, Trust and connectivity are two important aspects in the implementation of mobile commerce in the travel and hospitality industry, especially in foreign places where security is a major concern [27].

In various other types of research, such as e-commerce, mobile learning, and tourism, have added and modified TAM [28], [29]. Researchers have included external factors including hedonic incentives, facilitating conditions, and social impacts to explain why users or individuals use mobile tourism [30], [31], [32]. Facilitating conditions (FC), which include the availability of reliable network connectivity, compatible devices, and essential technical support, are particularly relevant to the MTCA context. These conditions are defined in terms of the extent to which an individual believes that organizational and technical infrastructure is available to facilitate the use of the system [33].

C. Hofstede's Cultural Dimensions

Although TAM offers a powerful cognitive framework, it lacks in accounting for cross-cultural differences in technology acceptance. Hofstede's Cultural Dimensions Theory provides a powerful framework for understanding the ways in which national to individual cultures influence human behavior and values [34]. The original dimensions include:

- Power Distance (PDI): The extent to which less powerful members of the organization and institutions (such as families) accept and expect that power is distributed unequally.

- Individualism versus Collectivism (IDV): A society in which the ties between individuals are loose is called individualism, while a society in which the ties between individuals have been formed into strong and cohesive groups since birth is called collectivism.
- Masculinity versus Femininity (MAS): In societies where social gender roles are different, it is an indication of a masculine society, while in societies where social gender roles overlap, it is an indication of a feminine society.
- Uncertainty Avoidance (UAI): The extent to which people in a culture feel fearful or threatened by unclear or unknown situations.
- Long-Term Orientation versus Short-Term Normative Orientation (LTO): Long-term orientation is the process of developing values that are oriented to what is expected to occur in the future, such as being careful and holding fast. While short-term orientation is the process of developing important values in the past and also in the present, such as maintaining self-esteem, respecting tradition, and fulfilling social obligations.
- Indulgence versus Restraint (IVR): Civilization that allows people to achieve their basic and natural needs to enjoy life or seek pleasure. While the restraint shows that a society has a directed and strict social rule to control and suppress people's needs.

Hofstede's dimensions have been applied and emphasized in many studies as important factors in technology adoption across cultural perspectives. For example, cultures that prioritize uncertainty avoidance may be more hesitant to adopt new technologies due to perceived risks, while individualistic cultures may prioritize technologies that enhance personal autonomy [35], [36], [37], [38]. On the other hand, collectivist cultures may prioritize technologies that promote social interaction and group cohesion [39], [40], [41]. It is therefore important to integrate these cultural dimensions to gain a more complex understanding of mobile tourism acceptance, as tourists' cultural backgrounds often influence their preferences and behaviors [5], [42].

D. The Fuzzy Delphi Method (FDM)

According to Dalkey and Helmer [15]. the Delphi method was designed as a structured communication technique to ensure expert consensus on complex issues. Experts are usually given summarized feedback between rounds of questionnaires, which can allow them to modify their judgments according to collective insights [43], [44]. The power of the Delphi method is manifested in its capacity to foster consensus among geographically dispersed experts and to reduce the impact of dominant personalities that are often present during face-to-face meetings.

However, the conventional Delphi method has limitations, particularly in its ability to handle subjective expert opinions and the potential loss of information when converting qualitative judgments into quantitative data [45], [46], [47]. Fuzzy set theory was incorporated into the Delphi process to address these issues, resulting in the introduction of the so-called Fuzzy Delphi Method (FDM) [47], [48], [49] The representation of ambiguous or vague information using fuzzy numbers, typically triangular fuzzy numbers (TFN), is facilitated by fuzzy set theory [50], [50], [51]. This allows experts to convey their opinions as a range rather than a single precise value, which more accurately reflects the uncertainty inherent in their judgments.

The nine systematic phases cover the stages and procedures of FDM, as described by Yin & Hanif [52]:

- Expert Determination Process: This stage involves selecting the right panel of experts. The optimal number of experts is adjusted and can vary, for example, [45], [53] suggesting between 10 and 50, while [54] recommend 10-15 experts if there is high consistency among them.
- Linguistic Selection Process: To obtain nuanced and diverse expert opinions, a 7-point fuzzy scale can be applied, derived from the conventional Likert scale. This scale assigns imprecise values (m1, m2, m3) to each linguistic term, which can allow for more precise quantification of agreement or importance.
- To enter the Likert scale values into the item space for each expert, follow the steps as follows: The researcher enters the Likert scale values given by the experts for each item, converting them into appropriate fuzzy scale values.
- Next is the process of obtaining the fuzzy average value (m1, m2, m3) of the fuzzy scale as follows: The consensus among experts is represented by the calculation of the fuzzy average value (m1, m2, m3) for each item. The minimum value (m1), the most reasonable value (m2), and the maximum value (m3) are represented by these three values, which collectively form a triangular fuzzy number.
- Finding the value of "d" (threshold value): For each question item, the "D" threshold value is determined. Then, this value decides whether the question level of agreement between experts determines whether the item is accepted or rejected.
- Set the number of experts: You can give a sign to the "red box" to the total number of experts to show how much they agree with each other.
- How to find out the level of percentage of expert agreements on each item and item overall: These statistics provide a direct and numerical assessment of how many experts agree.
- Fuzzy evaluation process: Fuzzy (A) score for each item is based on how many experts agree to each item. Items that do not meet FDM criteria standards are discarded.
- Data Analysis Findings Table: Data analysis findings will be shown in the table, the red value shown in cells is interpreted that the conditions set by FDM are not met or inappropriate.

In fields such as sustainable tourism, educational technology, and product reviews, FDM is very good for validating and making difficult decisions [55]. he ability of FDM to identify and ensure that the item rejected is a very useful tool for increasing the development of a model [56]. In FDM, rejecting the item is not just a statistical result, this is a group decision by experts that the item as written is irrelevant, is not important enough, or not adequately represents the construction intended in a particular research context [56]. The strong ability to diagnose this is very useful in interdisciplinary research, which combines ideas from various fields (for example, macro-cultural theory and micro technology adoption) [57]. FDM functions as a strict selection filter to ensure that the final model only makes the part that is truly in line with what is said and done by experts for their understanding [57]. FDM has changed from a simple validation tool to be an important tool for improving and developing models because it has the best diagnostic skills. This provides practical feedback in which theoretical operations are strong and which ones need to be thought of again, are changed, or even discarded [58].

III. METHODOLOGY

The Fuzzy Delphi (FDM) method is used in this study to make experts approve variables and indicators and how variables are interconnected in the Framework for the Proposed Culture Tourism Culture (MTCA) in the future, in this concept combining the dimensions of technology acceptance (TAM) and the Hofstede cultural dimension. This technique uses a structured multi-step procedure to collect, analyze, and combine the perspective of the expert.

A. Research Design

The research design in this study is the Fuzzy Delphi (FDM) method. This recurring method makes it easy to get expert opinions about how important and relevant certain elements. FDM also allows changing qualitative comments into Fuzzy numbers so that all experts can approve one perspective. In this research, choosing the FDM method is caused by eliminating prejudice, getting a greater group consensus, and getting around the fact that human assessment is not always accurate. This makes it the right choice to validate complicated theoretical models [52], [59]. This section explains how to use the Fuzzy Delphi (FDM) method in a comprehensive way, following the nine steps displayed in the following study Yin & Hanif [52]

B. Expert Panel Selection

Choosing the right experts for Delphi studies is quite important. This study chooses experts based on how much they know, how much experience they have, and how much expertise they have in terms of cellular tourism, technology adoption, cultural studies, and tourism management. Expert panels are divided into three clusters namely, people from academics, government tourism organizations, and private sectors related to tourism.

This survey uses 22 experts (see Table I). This is in accordance with the recommended expert panel size for Delphi studies, which usually ranges from 10 to 50 [53]. The selection process ensured a range of expertise relevant to the study of culture, technology adoption, and mobile tourism. The selection criteria were as follows [60]:

- Minimum of 5 years of professional experience in a relevant field (government from a tourism perspective, academics, and practitioners from a tourism perspective).
- Academic qualifications at Bachelor level or higher.
- Demonstrated expertise through publications, professional roles, or significant contributions in the field.

C. Questionnaire Development

This questionnaire was carefully prepared after a comprehensive literature review examining the Technology Acceptance Model (TAM), Hofstede's Cultural Dimensions, and their application in the context of mobile tourism. The questionnaire was designed to obtain expert opinions on the significance and relevance of various variables and the relationships between them. The items were arranged into five main categories, consistent with the scope of this research study:

- What are the key variables in the (TAM) that are significantly related to MTCA?
- What are the key variables in Hofstede's Cultural Dimensions that are significantly related to MTCA?
- What TAM variables and Hofstede's cultural dimensions are significantly related to each other in developing the MTCA framework?
- What TAM variables and Hofstede's Cultural Dimensions are interrelated and significantly influence users' behavioral intention to adopt mobile tourism technology (MTCA)?
- What TAM variables and Hofstede's Cultural Dimensions, along with the cultural dimension variables as mediators, significantly influence users' behavioral intention to adopt mobile tourism technology (MTCA)?

Experts were asked to evaluate the significance of each item on a linguistic scale (e.g., Strongly Disagree, Disagree, Tend to Disagree, Tend to Agree, Agree, Strongly Agree, Strongly Agree). Experts were able to articulate their opinions with nuanced levels of agreement or importance by framing each item using linguistic terms that corresponded to a 7-point Likert scale [52]. These linguistic terms were then converted into triangular fuzzy numbers for analysis purposes in the next stage.

TABLE I. EXPERT PANEL DEMOGRAPHICS

No.	Name	Gender	Academic Qualification	Institution	Position	Years of Work Experience
1	Expert 1	Female	Middle Expert	Tourism Youth and Sports Agency Lima Puluh Kota	Young Expert Planner	> 25 years
2	Expert 2	Female	Diploma	Cititel Hotel Pekanbaru	Sales Executive	10 to 15 years
3	Expert 3	Female	Magister	Macly Marsar Group Pekanbaru	Financial Controller Corporate	20 to 25 years
4	Expert 4	Female	Bachelor	PT Artha Bayu, Bali Indonesia	Owner	> 25 years
5	Expert 5	Male	Master	Tourism, Youth and Sports Office	Head of Office	15 to 20 years
6	Expert 6	Male	Bachelor	Tourism, Youth and Sports Office Payakumbuh	Acting head of tourism	> 25 years
7	Expert 7	Male	Diploma	Lloyds Inn Bali	Front Office Supervisor	5 to 10 years
8	Expert 8	Female	Bachelor	Cititel Hotel Pekanbaru	Sales and marketing manager	15 to 20 years
9	Expert 9	Male	Doctor of Philosophy	UIN Suska Riau	Lecturer	5 to 10 years
10	Expert 10	Male	Doctor of Philosophy	Lancang Kuning University Pekanbaru	Secretary of Computer Science Master's Program	15 to 20 years
11	Expert 11	Male	Doctor of Philosophy	Sebelas April University	Lecturer/Researcher	15 to 20 years
12	Expert 12	Male	Doctor of Philosophy	National Research and Innovation Agency, Indonesia	Researcher	5 to 10 years
13	Expert 13	Male	Bachelor	Tourism, Youth and Sports Service Lima Puluh Kota	Secretary of Tourism and Youth Service	> 25 years
14	Expert 14	Female	Bachelor	Lloyd's Inn Bali	Sales Executive	5 to 10 years
15	Expert 15	Male	Bachelor	Tourism Office of Lima Puluh Kota	Tourism Ambassador	< 5 years
16	Expert 16	Male	Doctor of Philosophy	Sebelas April University	Dean of the Faculty of Information Technology	20 to 25 years
17	Expert 17	Female	Bachelor	Lloyd's Inn Bali	Hotel Manager	20 to 25 years
18	Expert 18	Female	Bachelor	Tourism Office Lima Puluh Kota	Tourism Ambassador	< 5 years
19	Expert 19	Male	Diploma	Tourism Office of Lima Puluh Kota	Uda Uni Tourism Ambassador of Lima Puluh Kota	< 5 years
20	Expert 20	Male	Doctor of Philosophy	Lancang Kuning University Pekanbaru	Lancang Kuning University Pekanbaru	10 to 15 years
21	Expert 21	Female	Diploma	Cara Cara Inn Denpasar	Eccomerce	10 to 15 years
22	Expert 22	Male	Bachelor	Cara Cara Inn Denpasar	Manager	15 to 20 years

D. Data Collection and Fuzzy Delphi Analysis

A series of rounds, which is the hallmark of the Delphi method, were used to capture data. Questionnaires were filled out by experts in each round, and their responses were analyzed. The next stages consist of the core of FDM analysis [52]:

1) Likert-to-fuzzy conversion: The FDM methodology has established the fuzzy scale used by systematically converting the raw data collected from 22 experts into triangular fuzzy numbers (m1, m2, m3). The data were initially presented on a 7-point Likert scale. For example, a Likert rating of '7' (Strongly Agree) is converted into the fuzzy set (0.9, 1, 1), while '1' (Strongly Disagree) is converted into (0, 0, 0.1). This conversion standardizes the expert opinions into a format suitable for fuzzy logic operations.

2) Calculation of fuzzy mean: The fuzzy mean values (m1, m2, m3) are determined by combining the individual fuzzy numbers of all 22 experts for each item of their questionnaire. These various numbers, then combined, include the minimum, most plausible, and maximum values of the combined expert panel's assessment for each particular item.

3) Determination of threshold value 'd': Finding the "D" threshold for each item is an important part of FDM. This value shows how spread or unclear the opinions of experts. The smaller "D" means more agreement or has a high consensus. The "D" threshold is an important factor in deciding whether to accept or reject an item based on how much agreement there is among experts.

4) Percentage of agreement: Furthermore, for each item, the value of 'D' and the proportion of experts who agreed to be found. This metric provides a quantitative measure of direct consensus by showing the percentage of experts whose opinions are in the acceptable agreement range.

5) Fuzzy score (A): The average Fuzzy triangle is used to calculate the fuzzy (A) score for each item, which is a crispy defuzzified value. This score offers a single indicator that can be measured by the expert's perception panel to show how the expert feels about the interests or priorities of the item as a whole.

6) Acceptance/rejection status: In the final step, each item is given the status of acceptance ("accept") or rejection ("Reject") based on the predetermined criteria. This criterion

mainly includes the value of "D" which is below a certain level (for example, D <0.2) and the percentage of agreements that are above a certain level (for example, ≥ 70%). Items that are not satisfying this standard are rejected, which means there is not enough agreement among experts.

IV. Results and Discussion

The This section shows the findings using the Fuzzy Delphi method. This gives a complete picture of the consensus of experts who agree on how the variable relationship in the framework of the acceptance of the Culture (MTCA) is related to each other. Data is arranged in accordance with the sequential validation process which includes TAM factors, Hofstede cultural dimensions, general relationships, specific relationships, and mediation relationships.

A. Validation of Technology Acceptance Model (TAM) Variables

The purpose of the validation of the TAM variable is to prove the variables and indicators in building a framework for acceptance of the Mobile Tourism (MTCA) cultural acceptance. These variables such as, perceived usefulness (PU), perceived ease of use (PEOU), perceived trust (PT), facilitation of conditions (FC), and behavioral intentions (BI) and all indicators that are ganging on each of these variables. Variables and indicators that according to experts are important for people to accept the culture of cellular tourism. Table II shows a summary of findings.

TABLE II. Validation of Technology Acceptance Model (TAM) Variables

No	Items / Elements	Triangular Fuzzy Numbers Conditions		Defuzzification Process Requirements				Expert Agreement	Element ACCEPTED	Ranking
		Threshold Value, d	*Expert Group Agreement Percentage, %*	*m1*	*m2*	*m3*	*Fuzzy Score (A)*			
1	PEOU	0,095	95,5%	0,700	0,882	0,977	0,853	ACCEPT	0,853	14
2	PEOU1	0,121	100,0%	0,691	0,868	0,973	0,844	ACCEPT	0,844	19
3	PEOU2	0,117	100,0%	0,718	0,886	0,977	0,861	ACCEPT	0,861	12
4	PEOU3	0,137	100,00%	0,691	0,864	0,968	0,841	ACCEPT	0,841	20
5	PEOU4	0,092	100,00%	0,736	0,905	0,986	0,876	ACCEPT	0,876	5
6	PU	0,098	100,00%	0,745	0,909	0,986	0,880	ACCEPT	0,880	3
7	PU1	0,122	100,00%	0,727	0,891	0,977	0,865	ACCEPT	0,865	8
8	PU2	0,127	95,45%	0,727	0,891	0,973	0,864	ACCEPT	0,864	9
9	PU3	0,130	100,00%	0,709	0,877	0,973	0,853	ACCEPT	0,853	14
10	PU4	0,127	100,00%	0,736	0,895	0,977	0,870	ACCEPT	0,870	6
11	PT	0,117	100,00%	0,718	0,886	0,977	0,861	ACCEPT	0,861	13
12	PT1	0,114	95,45%	0,700	0,877	0,973	0,850	ACCEPT	0,850	16
13	PT2	0,407	31,82%	0,400	0,568	0,718	0,562	REJECT	0,562	27
14	PT3	0,146	95,45%	0,664	0,845	0,959	0,823	ACCEPT	0,823	22
15	PT4	0,140	95,45%	0,682	0,859	0,964	0,835	ACCEPT	0,835	21
16	PT5	0,201	81,82%	0,636	0,818	0,932	0,795	ACCEPT	0,795	25
17	PT6	0,127	95,45%	0,727	0,891	0,973	0,864	ACCEPT	0,864	9
18	PT7	0,102	95,45%	0,745	0,909	0,982	0,879	ACCEPT	0,879	4
19	FC	0,226	90,91%	0,668	0,827	0,923	0,806	ACCEPT	0,806	23
20	FC1	0,132	95,45%	0,736	0,895	0,973	0,868	ACCEPT	0,868	7
21	FC2	0,125	100,00%	0,700	0,873	0,973	0,848	ACCEPT	0,848	17
22	FC3	0,259	72,73%	0,618	0,782	0,895	0,765	REJECT	0,765	26
23	FC4	0,170	95,45%	0,645	0,823	0,945	0,805	ACCEPT	0,805	24
24	BI	0,098	100,00%	0,718	0,891	0,982	0,864	ACCEPT	0,864	9
25	BI1	0,088	95,45%	0,800	0,945	0,995	0,914	ACCEPT	0,914	1
26	BI2	0,125	100,00%	0,700	0,873	0,973	0,848	ACCEPT	0,848	17
27	BI3	0,104	100,00%	0,755	0,914	0,986	0,885	ACCEPT	0,885	2

As shown in Table II, the expert panel showed a high level of consensus on most of the TAM variables and indicators. There was a total of 27 TAM items, consisting of 5 variables and 22 indicators. Almost all items achieved a consensus percentage of 95.45% or 100%, and their Fuzzy Scores were consistently high (above 0.8), leading to their acceptance. This indicates that experts largely agree on the importance of perceived ease of use (PEOU items), perceived usefulness (PU items), perceived trust (PT items), facilitating conditions (FC items), and behavioral intention (BI items) in driving mobile tourism acceptance. For example, BI1 has a high fuzzy score of 0.914 with 100% agreement, emphasizing its criticality. The second to fifth ranks can be seen in BI3 (0.885), PU (0.880), PT7 (0.879), PEOU4 (0.876) with 100% agreement, indicating strong agreement and high scores. The two "REJECTED" items were PT2, with a low consensus percentage of 31.82% and a Fuzzy Score of 0.562, and FC3 with a consensus percentage of 72.73% and a fuzzy score of 0.765 indicating a lack of agreement or perceived importance by experts for this particular element. This highlights that although the TAM is generally accepted, not all of its aspects are equally relevant or agreed upon in the specific context of mobile tourism, thus requiring further investigation into the nature of PT2 and FC3.

B. Validation of Hofstede's Cultural Dimensions Variables

Hofstede's Cultural Dimensions were validated within the MTCA framework in the second phase. Expert consensus on these variables is illustrated in Table III.

TABLE III. VALIDATION OF HOFSTEDE'S CULTURAL DIMENSIONS VARIABLES

No	Items / Elemens	Triangular Fuzzy Numbers Conditions		Defuzzification Process Requirements				Expert Agreement	Element ACCEPTED	Ranking
		Threshold Value, d	*Expert Group Agreement Percentage, %*	*m1*	*m2*	*m3*	*Fuzzy Score (A)*			
1	CD	0,179	90,9%	0,664	0,841	0,945	0,817	ACCEPT	0,817	14
2	CD1	0,175	90,9%	0,655	0,836	0,945	0,812	ACCEPT	0,812	16
3	CD2	0,214	86,4%	0,627	0,805	0,923	0,785	ACCEPT	0,785	23
4	CD3	0,187	86,36%	0,655	0,832	0,941	0,809	ACCEPT	0,809	18
5	CD4	0,257	81,82%	0,614	0,782	0,895	0,764	ACCEPT	0,764	26
6	CD5	0,206	86,36%	0,673	0,841	0,936	0,817	ACCEPT	0,817	14
7	CD6	0,203	86,36%	0,664	0,836	0,936	0,812	ACCEPT	0,812	16
8	PD	0,239	68,18%	0,595	0,773	0,900	0,756	REJECT	0,756	28
9	PD1	0,408	45,45%	0,409	0,568	0,723	0,567	REJECT	0,567	42
10	PD2	0,397	40,91%	0,400	0,564	0,718	0,561	REJECT	0,561	43
11	PD3	0,427	27,27%	0,436	0,605	0,750	0,597	REJECT	0,597	40
12	PD4	0,393	40,91%	0,445	0,609	0,755	0,603	REJECT	0,603	39
13	PD5	0,390	40,91%	0,468	0,627	0,764	0,620	REJECT	0,620	38
14	UA	0,171	90,91%	0,645	0,827	0,945	0,806	ACCEPT	0,806	19
15	UA1	0,114	95,45%	0,736	0,900	0,977	0,871	ACCEPT	0,871	6
16	UA2	0,115	100,00%	0,745	0,905	0,982	0,877	ACCEPT	0,877	4
17	UA3	0,092	100,00%	0,736	0,905	0,986	0,876	ACCEPT	0,876	5
18	UA4	0,111	95,45%	0,764	0,918	0,982	0,888	ACCEPT	0,888	1
19	UA5	0,104	100,00%	0,755	0,914	0,986	0,885	ACCEPT	0,885	3
20	CO	0,173	90,91%	0,673	0,845	0,950	0,823	ACCEPT	0,823	13
21	CO1	0,219	86,36%	0,632	0,809	0,923	0,788	ACCEPT	0,788	22
22	CO2	0,250	81,82%	0,623	0,795	0,905	0,774	ACCEPT	0,774	24
23	CO3	0,224	86,36%	0,614	0,791	0,914	0,773	ACCEPT	0,773	25
24	CO4	0,163	90,91%	0,682	0,855	0,955	0,830	ACCEPT	0,830	11
25	CO5	0,248	54,55%	0,605	0,773	0,895	0,758	REJECT	0,758	27
26	CO6	0,255	54,55%	0,586	0,759	0,886	0,744	REJECT	0,744	29
27	LT	0,136	95,5%	0,709	0,877	0,968	0,852	ACCEPT	0,852	9
28	LT1	0,167	90,9%	0,691	0,859	0,955	0,835	ACCEPT	0,835	10

29	LT2	0,192	90,9%	0,636	0,809	0,932	0,792	ACCEPT	0,792	20
30	LT3	0,169	95,45%	0,682	0,850	0,955	0,829	ACCEPT	0,829	12
31	LT4	0,144	95,45%	0,727	0,886	0,968	0,861	ACCEPT	0,861	8
32	LT5	0,148	95,45%	0,736	0,891	0,968	0,865	ACCEPT	0,865	7
33	LT6	0,144	95,45%	0,773	0,914	0,973	0,886	ACCEPT	0,886	2
34	MA	0,237	86,36%	0,641	0,814	0,918	0,791	ACCEPT	0,791	21
35	MA1	0,428	36,36%	0,441	0,600	0,741	0,594	REJECT	0,594	41
36	MA2	0,344	59,09%	0,582	0,750	0,859	0,730	REJECT	0,730	32
37	MA3	0,310	68,18%	0,555	0,732	0,859	0,715	REJECT	0,715	34
38	MA4	0,332	59,09%	0,559	0,723	0,845	0,709	REJECT	0,709	35
39	IR	0,285	54,55%	0,568	0,745	0,873	0,729	REJECT	0,729	33
40	IR1	0,285	63,64%	0,568	0,750	0,877	0,732	REJECT	0,732	30
41	IR2	0,284	59,09%	0,568	0,750	0,877	0,732	REJECT	0,732	30
42	IR3	0,256	81,82%	0,532	0,714	0,859	0,702	ACCEPT	0,702	36
43	IR4	0,304	72,73%	0,532	0,714	0,850	0,698	REJECT	0,698	37

Table III shows a strong expert consensus on the relevance of Hofstede's Cultural Dimensions in the context of mobile tourism acceptance. There are 43 items, consisting of 7 variables and 36 indicators. In the context of Hofstede's cultural dimensions on the acceptance of mobile tourism culture, it shows that most of the variables and indicators are accepted and some variables and indicators are rejected. This shows that most experts agree on the importance of the Cultural Dimensions (CD), Uncertainty Avoidance (UA), Collectivism (CO), and Long-Term Orientation (LT) in encouraging the acceptance of mobile tourism culture. It can be seen that UA4 has the highest fuzzy value of 0.888 with an agreement of 95.45% which emphasizes its criticality. The second to fifth ranks can be seen in LT6 (0.886), UA5 (0.885), UA2 (0.877), and UA3 (0.876) with an agreement of 95.45%, 95.45%, 100%, 100%, and 100% which emphasizes its criticality. However, there are 2 variables that are rejected, namely Power Distance (PD) with a fuzzy value of 0.756, 68.18% agreement and the variable Indulgence (IR) with a fuzzy value of 0.729, 54.5% agreement. This indicates a lack of agreement or perception of the importance of this particular element by experts. There are also rejected indicators such as PD1 with a low consensus percentage of 45.45%% and a Fuzzy Score of 0.567, PD2 (40.91%) (0.561), PD3 (27.27% (0.597), PD4 (40.91%) (0.603), PD5 (40.91%) (0.620), CO5 (54.55%) (0.758), CO6 (54.55%) (0.744), MA1 (36.36%) (0.594), MA2 (59.09%) (0.730), MA3 (68.18%) (0.715), MA4 (59.09%) (0.709), IR1 (54.55%) (0.732), IR2 (59.09%) (0.732), IR4 (72.73%) (0.698). Based on the data in the table, it can be explained that the variables, namely Power Distance (PD) and all its indicators are rejected, the Indulgence variable (IR) is rejected with all its indicators, although IR3 is accepted, and the Masculinity variable (MA) is rejected because all its indicators are rejected. This indicates a lack of agreement or perception of the importance of this particular element by experts.

C. *Validation of General Variables in Mobile Tourism Culture Acceptance (MTCA)*

Expert consensus on the general relationship in building the framework model construction related to the Mobile Tourism Culture Acceptance (MTCA) framework is presented in this section. The variables or overall model construction are represented by the following Table IV.

TABLE IV. VALIDATION OF GENERAL VARIABLES (MTCA)

No	Items / Elements	Triangular Fuzzy Numbers Conditions		Defuzzification Process Requirements				Expert Agreement	Element ACCEPTED	Ranking
		Threshold Value, d	*Expert Group Agreement Percentage, %*	*m1*	*m2*	*m3*	*Fuzzy Score (A)*			
1	PEOU-MTCA	0,083	100,0%	0,773	0,932	0,995	0,900	ACCEPT	0,900	2
2	PU-MTCA	0,089	100,0%	0,755	0,918	0,991	0,888	ACCEPT	0,888	3
3	PT-MTCA	0,086	95,5%	0,782	0,936	0,995	0,905	ACCEPT	0,905	1
4	FC-MTCA	0,092	100,00%	0,736	0,905	0,986	0,876	ACCEPT	0,876	5
5	UA-MTCA	0,079	100,00%	0,718	0,895	0,986	0,867	ACCEPT	0,867	6
6	CO-MTCA	0,150	95,45%	0,673	0,850	0,959	0,827	ACCEPT	0,827	9
7	LT-MTCA	0,136	95,45%	0,709	0,877	0,968	0,852	ACCEPT	0,852	8
8	CD-MTCA	0,118	95,45%	0,709	0,882	0,973	0,855	ACCEPT	0,855	7
9	BIUMT-MTCA	0,098	100,00%	0,745	0,909	0,986	0,880	ACCEPT	0,880	4

Table IV shows a very high level of consensus among experts regarding the general relationship of TAM and Hofstede variables contributing to the MTCA framework. All 12 items presented were accepted, with consensus percentages ranging from 90.91% to 100%. The Fuzzy Scores for these general variables were also very high, mostly above 0.87. PT-MTCA (Perceived Trust related to MTCA) emerged as the highest-ranked element with a Fuzzy Score of 0.905 and an agreement level of 95.5%, emphasizing the important role of trust in the acceptance of mobile tourism. PEOU-MTCA (Perceived Ease of Use related to MTCA) and PU-MTCA (Perceived Usefulness related to MTCA) also showed very high scores and 100% consensus, reaffirming the basic tenets of TAM in this context. The acceptance of cultural dimensions related to MTCA (e.g., UA-MTCA, CO-MTCA, LT-MTCA) further supports the integration of cultural factors as an important component of the framework.

D. Validation of Specific Relationships in Mobile Tourism Culture Acceptance (MTCA)

This section explores expert consensus on specific or specific causal relationships between variables within the MTCA framework. This can be seen in the following table.

Table V shows a consistent pattern of high expert consensus on the relationships of TAM and Hofstede's specific variables within the MTCA framework. All 15 listed relationships were accepted, with the dominant consensus percentage at 95.45% or 100%, and the Fuzzy Scores were generally above 0.82. The relationship "PT-BIUMT" (Perceived Trust to Behavioral Intention to Use Mobile Tourism) achieved the highest ranking with a Fuzzy Score of 0.905 and 95.5% agreement, further reinforcing the importance of trust. "PU-BIUMT" (Perceived Usefulness to Behavioral Intention to Use Mobile Tourism) also showed strong support (Fuzzy Score 0.897, 100% agreement), which is in line with the core principles of TAM. Relationships involving cultural dimensions, such as UA-BIUMT, CO-BIUMT, LT-BIUMT, were also accepted, confirming that experts believe that these cultural factors directly influence behavioral intention to use mobile tourism. It can also be explained that the relationship involving TAM and Hofstede variables in the context of cultural dimensions, such as PEOU-CD, PU-CD, PT-CD, and FC-CD as well as UA-CD, CO-CD, and LT-CD are equally believed by experts that these factors have a relationship and influence in the development of MTCA. Thus, consistent acceptance across these specific relationships provides strong empirical support for their inclusion in a comprehensive MTCA model.

E. Validation of Mediator Relationships in Mobile Tourism Culture Acceptance (MTCA)

The final validation stage is dedicated to identifying and confirming the mediating function of cultural dimensions within the MTCA framework. Specifically, it is examined how cultural dimensions can influence the relationship between TAM variables and Hofstede variables in behavioral intention to use mobile tourism.

TABLE V. VALIDATION OF SPECIFIC RELATIONSHIPS (MTCA)

No	Items / Elements	Triangular Fuzzy Numbers Conditions		Defuzzification Process Requirements				Expert Agreement	Element ACCEPTED	Ranking
		Threshold Value, d	*Expert Group Agreement Percentage, %*	*m1*	*m2*	*m3*	*Fuzzy Score (A)*			
1	PEOU-BIUMT	0,092	100,0%	0,736	0,905	0,986	0,876	ACCEPT	0,876	3
2	PU-BIUMT	0,097	100,0%	0,773	0,927	0,991	0,897	ACCEPT	0,897	2
3	PT-BIUMT	0,086	95,5%	0,782	0,936	0,995	0,905	ACCEPT	0,905	1
4	FC-BIUMT	0,092	100,00%	0,736	0,905	0,986	0,876	ACCEPT	0,876	3
5	UA-BIIUMT	0,159	90,91%	0,673	0,850	0,955	0,826	ACCEPT	0,826	13
6	CO-BIUMT	0,148	90,91%	0,682	0,859	0,959	0,833	ACCEPT	0,833	11
7	LT-BIUMT	0,143	90,91%	0,709	0,877	0,964	0,850	ACCEPT	0,850	9
8	PEOU-CD	0,119	95,45%	0,745	0,905	0,977	0,876	ACCEPT	0,876	3
9	PU-CD	0,131	100,00%	0,745	0,900	0,977	0,874	ACCEPT	0,874	6
10	PT-CD	0,118	95,45%	0,709	0,882	0,973	0,855	ACCEPT	0,855	8
11	FC-CD	0,114	95,45%	0,700	0,877	0,973	0,850	ACCEPT	0,850	9
12	UA-CD	0,206	86,36%	0,645	0,823	0,932	0,800	ACCEPT	0,800	15
13	CO-CD	0,148	90,91%	0,682	0,859	0,959	0,833	ACCEPT	0,833	11
14	LT-CD	0,155	90,91%	0,664	0,845	0,955	0,821	ACCEPT	0,821	14
15	CD-BIUMT	0,132	95,45%	0,736	0,895	0,973	0,868	ACCEPT	0,868	7

TABLE VI. VALIDATION OF MEDIATOR RELATIONSHIPS (MTCA)

No	Items / Elements	Triangular Fuzzy Numbers Conditions		Defuzzification Process Requirements				Expert Agreement	Element ACCEPTED	Ranking
		Threshold Value, d	*Expert Group Agreement Percentage, %*	*m1*	*m2*	*m3*	*Fuzzy Score (A)*			
1	PEOU-CD-BIUMT	0,115	100,0%	0,745	0,905	0,982	0,877	ACCEPT	0,877	1
2	PU-CD-BIUMT	0,127	95,5%	0,727	0,891	0,973	0,864	ACCEPT	0,864	3
3	PT-CD-BIUMT	0,118	95,5%	0,709	0,882	0,973	0,855	ACCEPT	0,855	4
4	FC-CD-BIUMT	0,124	95,45%	0,745	0,905	0,973	0,874	ACCEPT	0,874	2
5	UA-CD-BIUMT	0,136	95,45%	0,691	0,868	0,964	0,841	ACCEPT	0,841	7
6	CO-CD-BIUMT	0,167	90,91%	0,709	0,873	0,955	0,845	ACCEPT	0,845	6
7	LT-CD-BIUMT	0,125	90,91%	0,709	0,882	0,968	0,853	ACCEPT	0,853	5

All seven proposed mediation relationships were accepted, indicating that experts largely agreed, with percentages ranging from 95.5% to 100% and fuzzy scores always higher than 0.85. Rankings 1,2,3,4, and occupied by TAM variables. "PEOU-CD-BIUMT" (perceived ease of use mediated by cultural dimensions affecting behavioral intention to use mobile tourism) stands out as the best mediation relationship with fuzzy scores of 0.877 and 100% agreement. This means that how easy a mobile tourism application is, when viewed from a cultural perspective, has a great effect on whether people want to use it or not. In the same way, "PU-CD-BIUMT" (perceived usefulness mediated by cultural dimensions affecting behavioral intention to use mobile tourism) also received much support (fuzzy score 0.864, 95.5% agreement). These results indicate that perceived ease of use is how people view technology and their cultural values in using mobile tourism applications. Overall, it can be concluded that the TAM variable has a greater influence when the cultural dimension acts as a mediator of behavioral intentions to use mobile tourism. Furthermore, seen from the fuzzy score value and expert agreement, ranks 5, 6 and 7 are occupied by the Hofstede variable when the cultural dimension acts as a mediator of behavioral intentions to use mobile tourism.

F. Discussion of Findings

Strong expert consensus was achieved across all categories of variables and relationships proposed in the Mobile Tourism Culture Acceptance (MTCA) framework as a result of a comprehensive validation process using the Fuzzy Delphi Method. The high Fuzzy Scores (generally above 0.8) and consistently high consensus percentages (mostly above 90%) indicate that there is a strong consensus among experts on the important factors influencing mobile tourism culture acceptance.

Its importance as a basic form of Perceived Ease of Use and Perceived Usefulness is reaffirmed by the validation of the TAM variables (Table II). The rejection of the PT2 indicator item implies that, despite the significance of trust, this specific operationalization of trust may not be universally accepted or relevant to scholars in the mobile tourism context. Consequently, additional qualitative investigation is needed to explore it in the context of PT2 in future studies.

The importance of cultural context in mobile tourism is underscored by the strong acceptance of Hofstede's Cultural Dimensions (Table III). This finding is significant because it addresses the general gap between the Primary Influence of Technology and the importance of the human dimension in gaining benefits from technology use. This gap was highlighted by researchers such as [33] who found that users did not gain any benefits from technology [61], have established that the acceptance and adoption of technology are significantly influenced by an individual's cultural beliefs [62]. In essence, cultural heritage impacts various aspects of an individual's life, such as their beliefs, behaviors, perceptions, and attitudes towards health [63]. Scholars have clearly observed that the perception and adoption of mobile tourism technology can be significantly influenced by factors such as Power Distance, Uncertainty Avoidance, Individualism, and Long-Term Orientation. This is consistent with a growing body of research advocating for culturally sensitive technology design and marketing, as well as cultural acceptance of tourism or mobile technology [42], [64], [65], [66], [67].

The importance of the proposed constructs in developing a holistic model of cultural acceptance of mobile tourism is confirmed by the validation of the general relationship variables for MTCA (Table IV). The high PT-MTCA (Perceived Trust in MTCA) classification underlines the fact that trust is not simply a facilitating condition for digital tourism services, but is an essential and fundamental prerequisite. This finding is consistent with recent research on digital platform adoption Furthermore, the consensus on the specific relationships between the variables (Table V) offers more detailed insights into the pathways of influence between the variables. The core predictions of TAM are strongly validated by the direct impacts of PEOU, PU, FC, and PT on behavioral intention (BI). The argument for an integrated paradigm is further strengthened by the recognition of cultural dimensions that directly influence BI. This implies that cultural values can directly influence an individual's readiness and inclination to utilize a mobile tourism tool, regardless of the tool's ease of use or usefulness [41].

Finally, the most important understanding of the influence of cultural dimensions as mediators between TAM and Hofstede variables on behavioral intention to use mobile tourism (Table VI). The finding that cultural dimensions mediate the

relationship between behavioral intention and TAM constructs (PEOU, PU, FC, PT) is a significant theoretical contribution. This suggests that the adoption behavior of mobile tourism applications is ultimately influenced by the cultural context, which serves as a filter or lens through which users interpret the ease of use or utility of these applications. For example, cultures that tend to avoid uncertainty may find a new and complex mobile tourism application more difficult to operate, even if its function is clear, due to the ambiguity inherent in the new technology [68], [69]. Such interactions and the resulting complexity underscore the need for a comprehensive approach to MTCA, in which cultural perspectives are incorporated into every phase of the design and implementation process. The consensus obtained through the Fuzzy Delphi Method was reliably and validly supported by experts at consistently high levels of agreement (mostly 90-100%) and low threshold values (d) across various validation stages. This strong agreement serves as the basis for building a solid foundation for empirical testing and development of the proposed Mobile Tourism Cultural Acceptance framework.

V. IMPLICATIONS

The results of this Fuzzy Delphi study have substantial implications for the development of theoretical concepts as well as for practical applications in the context of cultural acceptance of mobile tourism.

A. Theoretical Implications

Implications for Theory, this study provides a comprehensive assessment of an integrated framework for Mobile Tourism Cultural Acceptance (MTCA) by effectively integrating components of the Technology Acceptance Model (TAM) and Hofstede's Cultural Dimensions. The argument for a more comprehensive model of technology adoption in tourism is theoretically strengthened by the strong scholarly consensus on the importance of PEOU, PU, PT, and FC, as well as the critical role of cultural dimensions. It goes beyond the cognitive focus of conventional TAM by explicitly integrating macro-level cultural influences, which are often overlooked but have a substantial impact on individual perceptions and behaviors. Cultural dimensions are not merely external variables, they can also serve as mediators, influencing the perceptions and effects of fundamental TAM constructs, as confirmed in this research study. This establishes a theoretical foundation for future empirical research that will investigate these complex mediation relationships through quantitative methods, including Structural Equation Modeling.

Second, the Fuzzy Delphi Method (FDM) applies itself as a methodological contribution. This study illustrates the power of fuzzy logic in developing rigorous theoretical constructs and managing subjective judgments in a complex and novel domain such as mobile tourism by methodically guiding experts to reach consensus. It can serve as a practical guide for researchers who wish to leverage expert knowledge in model development and validation provided in the form of a comprehensive step-by-step of the FDM process and its application to a multifaceted framework. Finally, the specific identification and validation of individual items and their relationships, particularly the highly ranked "Perceived Trust" (PT) variable and its inter-variable relationships, suggest that trust may play a more fundamental and important role in mobile tourism acceptance. This underscores the need for additional theoretical investigation into the proper conceptualization and measurement of trust in the context of digital tourism.

B. Practical Implications

The practical implications of these findings are important for various stakeholders in the mobile tourism sector:

- For Mobile Tourism Developers and Designers: The validated variables offer a concise guideline for developing more user-friendly and culturally sensitive mobile tourism apps. Developers should prioritize facilitating conditions (FC), perceived trust (PT), perceived usefulness (PU), and perceived ease of use (PEOU). More importantly, they should consider local cultural values during the design phase. Apps should offer clear and explicit instructions, comprehensive FAQs, and robust customer support in cultures that prioritize uncertainty avoidance to reduce perceived risk. Features that facilitate group planning, collaboration, and social interaction may be highly valued in collectivist cultures.
- For Tourism Destination Management Organizations (DMOs) and Policymakers: These insights can be used by DMOs to develop strategies that will increase the adoption of mobile tourism. Marketing campaigns, public awareness initiatives, and infrastructure development can be informed by an understanding of the cultural dimensions of the target tourist market or local population. For example, in cultures with high power distance, it may be more effective to promote mobile tourism through trusted government channels. Then there are policies that promote ubiquitous and affordable internet access (facilitating conditions).
- For Tourism Marketers: Marketing campaigns for mobile tour services must be tailored to local or specific cultural values. High uncertainty avoidance populations will be attracted to an emphasis on safety, reliability, and unambiguous benefits, while individualistic users may be attracted to an emphasis on efficiency and personal control. Marketers can effectively communicate the proposition in a way that is culturally appropriate to the informed message creator.

VI. CONCLUSION

This study effectively used the Fuzzy Delphi Method to establish expert consensus on the key variables and their relationships within the Mobile Tourism Culture Acceptance (MTCA) framework, which combines Hofstede's Cultural Dimensions and Technology Acceptance Model (TAM). The careful process of expert panel selection, questionnaire development, fuzzy number conversion, aggregation, defuzzification, and consensus determination is explained in a step-by-step guide. The results indicate that there is a high level of expert agreement on the significance of perceived ease of use, perceived usefulness, perceived trust, and facilitating conditions as key determinants of mobile tourism acceptance. Importantly, this study also confirms the substantial role of Hofstede's Cultural Dimensions, not only as direct influencers but also as

mediating factors shaping the impact of TAM constructs on behavioral intention. The comprehensive nature of the proposed MTCA framework is validated by the consistent adoption of the majority of the proposed variables and relationships, as evidenced by the high consensus percentage and strong Fuzzy Scores. This study emphasizes the importance of gaining insight into expert opinion when developing a robust theoretical model for a complex interdisciplinary problem such as the cultural acceptance of mobile tourism. The Fuzzy Delphi technique has been shown to be a useful and important tool to overcome the subjectivity and ambiguity that accompanies qualitative assessment by carefully combining the views of multiple experts.

LIMITATIONS AND FUTURE RESEARCH

Although the procedures and stages used in this study were very rigorous, there were some shortcomings in this study. The expert group consisted of people with different backgrounds, but did not represent all major tourism centers in Indonesia. The Fuzzy Delphi method also did not assign numerical values to the proposed relationships or the strength of these relationships, although it did build expert consensus and validate theoretical constructs.

So, future research should include more experts with specific clusters from each region of Indonesia that represent major tourism centers in Indonesia. Then should try to prove the validated MTCA framework with quantitative research approaches, such as structural equation modeling (SEM) or partial least squares structural equation modeling (PLS-SEM), by conducting a complete survey of people or tourists who use and utilize mobile tourism. This can prove the model by using statistics to examine the hypothesized relationships and see how well the variables affect behavioral intentions in using mobile tourism. In addition, future research should look at the specific ways in which culture influences the acceptance of mobile tourists in more depth. With the Interview approach and qualitative research concepts can be used to expand this study. Looking at various cultural models or ideas, or even just certain cultural values that are not covered by Hofstede's dimensions, has the potential to help people understand MTCA better by trying to collaborate with other cultural concepts. Ultimately, stakeholders can create, build, and sell technologically advanced and culturally relevant mobile travel solutions by recognizing and combining technological insights and cultural values. This method helps the mobile travel sector grow in a way that will last by getting more people to adopt it.

ACKNOWLEDGMENT

This paper was developed under the supervision of the Special Interest Group (SIG) Educational Usability Testing, UPSI, Malaysia.

REFERENCES

[1] D. Buhalis and R. Law, “Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research,” Tourism Management, vol. 29, no. 4, pp. 609–623, Aug. 2008, doi: 10.1016/j.tourman.2008.01.005.

[2] D. Wang, S. Park, and D. R. Fesenmaier, “The Role of Smartphones in Mediating the Touristic Experience,” Journal of Travel Research, vol. 51, no. 4, pp. 371–387, Jul. 2012, doi: 10.1177/0047287511426341.

[3] B. Neuhofer, D. Buhalis, and A. Ladkin, “Technology as a Catalyst of Change: Enablers and Barriers of the Tourist Experience and Their Consequences,” in Information and Communication Technologies in Tourism 2015, I. Tussyadiah and A. Inversini, Eds., Cham: Springer International Publishing, 2015, pp. 789–802. doi: 10.1007/978-3-319-14343-9_57.

[4] R. Sivarethinamohan, “Exploring the Transformation of Digital Tourism: Trends, Impacts, and Future Prospects,” in 2023 International Conference on Digital Applications, Transformation & Economy (ICDATE), Jul. 2023, pp. 260–266. doi: 10.1109/ICDATE58146.2023.10248691.

[5] T. H. Jung, H. Lee, N. Chung, and M. C. Tom Dieck, “Cross-cultural differences in adopting mobile augmented reality at cultural heritage tourism sites,” IJCHM, vol. 30, no. 3, pp. 1621–1645, Mar. 2018, doi: 10.1108/IJCHM-02-2017-0084.

[6] O. Ortiz, C. Rusu, V. Rusu, N. Matus, and A. Ito, “Tourist eXperience Considering Cultural Factors: A Systematic Literature Review,” Sustainability, vol. 16, no. 22, p. 10042, Nov. 2024, doi: 10.3390/su162210042.

[7] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” MIS Quarterly, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[8] N. Sreen, P. H. Sadarangani, and V. Giridhar, “A path from cultural values to mobile travel app use,” 2019.

[9] J. Jan, K. A. Alshare, and P. L. Lane, “Hofstede’s cultural dimensions in technology acceptance models: a meta-analysis,” Univ Access Inf Soc, vol. 23, no. 2, pp. 717–741, Jun. 2024, doi: 10.1007/s10209-022-00930-7.

[10] C. Ferrara, R. Pierdicca, M. Balestra, C. Mignani, E. Frontoni, and A. Cavicchi, “ICTS AND MOBILE APPLICATIONS FOR PROMOTING TOURISM DESTINATIONS: THE CASE OF SMART MARCA APP.,” presented at the Tourism in Southern and Eastern Europe 2021: ToSEE – Smart, Experience, Excellence & ToFEEL – Feelings, Excitement, Education, Leisure, Dec. 2021, pp. 229–245. doi: 10.20867/tosee.06.16.

[11] G. H. Hofstede, Culture’s consequences: comparing values, behaviors, institutions, and organizations across nations, 2nd ed. Thousand Oaks, Calif: Sage Publications, 2001.

[12] K. Jung, V. T. Nguyen, D. Piscarac, and S.-C. Yoo, “Meet the Virtual Jeju Dol Harubang—The Mixed VR/AR Application for Cultural Immersion in Korea’s Main Heritage,” IJGI, vol. 9, no. 6, p. 367, Jun. 2020, doi: 10.3390/ijgi9060367.

[13] Q. Sun and B. Xu, “Mobile Social Commerce: Current State and Future Directions,” Journal of Global Marketing, vol. 32, no. 5, pp. 306–318, Oct. 2019, doi: 10.1080/08911762.2019.1620902.

[14] N. Ameen, R. Willis, and M. Hussain Shah, “An examination of the gender gap in smartphone adoption and use in Arab countries: A cross-national study,” Computers in Human Behavior, vol. 89, pp. 148–162, Dec. 2018, doi: 10.1016/j.chb.2018.07.045.

[15] N. Dalkey and O. Helmer, “An Experimental Application of the Delphi Method to the Use of Experts,” Management Science, vol. 9, no. 3, pp. 458–467, 1963.

[16] F. B. Ahmadi, B. Esmaeeli, S. S. Moghani, and K. Shariati, “A Literature Review of Group Decision-Making: The Case Study of Delphi Method,” 2023.

[17] J. Morales, R. Montes, N. Zermeño, J. Duran, and F. Herrera, “The Use of Fuzzy Linguistic Information and Fuzzy Delphi Method to Validate by Consensus a Questionnaire in a Blended-Learning Environment,” in Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, vol. 855, J. Medina, M. Ojeda-Aciego, J. L. Verdegay, I. Perfilieva, B. Bouchon-Meunier, and R. R. Yager, Eds., in Communications in Computer and Information Science, vol. 855., Cham: Springer International Publishing, 2018, pp. 137–149. doi: 10.1007/978-3-319-91479-4_12.

[18] I. Buj-Corral, A. Tejo-Otero, and F. Fenollosa-Artés, “Use of FDM Technology in Healthcare Applications: Recent Advances,” in Fused Deposition Modeling Based 3D Printing, H. K. Dave and J. P. Davim, Eds., Cham: Springer International Publishing, 2021, pp. 277–297. doi: 10.1007/978-3-030-68024-4_15.

[19] D. Wang, X. (Robert) Li, and Y. Li, “China’s ‘smart tourism destination’ initiative: A taste of the service-dominant logic,” Journal of Destination

Marketing & Management, vol. 2, no. 2, pp. 59–61, Jun. 2013, doi: 10.1016/j.jdmm.2013.05.004.

[20] J. Dorcic, J. Komsic, and S. Markovic, "Mobile technologies and applications towards smart tourism – state of the art," TR, vol. 74, no. 1, pp. 82–103, Feb. 2019, doi: 10.1108/TR-07-2017-0121.

[21] R. Law, I. C. C. Chan, and L. Wang, "A comprehensive review of mobile technology use in hospitality and tourism," Journal of Hospitality Marketing & Management, vol. 27, no. 6, pp. 626–648, Aug. 2018, doi: 10.1080/19368623.2018.1423251.

[22] S. Volo, "Conceptualizing Experience: A Tourist Based Approach," Journal of Hospitality Marketing & Management, vol. 18, no. 2–3, pp. 111–126, Feb. 2009, doi: 10.1080/19368620802590134.

[23] Hung Vuong, A. V. Tran, B. T. Khoa, and Industrial University of Ho Chi Minh City, Faculty of Commerce and Tourism, Ho Chi Minh City, Vietnam, "The Impact of Mobile Augmented Reality On Green Experience and Destination Choice Intention In Green Tourism In Vietnam," GTG, vol. 58, no. 1, pp. 136–145, Mar. 2025, doi: 10.30892/gtg.58112-1397.

[24] B. L. Song, D. Kaur, M. Subramaniam, P. K. Tee, L. C. Wong, and N. A. Mohd Zin, "The Adoption of Mobile Augmented Reality in Tourism Industry: Effects on Customer Engagement, Intention to Use and Usage Behaviour," JoTS, vol. 15, no. 28, pp. 235–252, Jun. 2024, doi: 10.29036/jots. v15i28.679.

[25] W. Dan, C. Muangmee, and N. Meekaewkunchorn, "The Influence of Tourist Short Videos on Tourism Intention for University Students," in Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024), vol. 289, R. Magdalena, B. Majoul, S. N. Singh, and A. Rauf, Eds., in Advances in Economics, Business and Management Research, vol. 289. , Dordrecht: Atlantis Press International BV, 2024, pp. 1133–1138. doi: 10.2991/978-94-6463-459-4_126.

[26] N. T. B. Dao and D. N. T. Anh, "The Impact of Personal Innovativeness on The Behavioral Intention to Using Tourism Mobile Applications of Generation Z In Ho Chi Minh City, Vietnam," GTG, vol. 55, no. 3, pp. 1186–1197, Sep. 2024, doi: 10.30892/gtg.55319-1291.

[27] R. P. Q. Falcao, J. B. Ferreira, and M. Carrazedo Marques Da Costa Filho, "The influence of ubiquitous connectivity, trust, personality and generational effects on mobile tourism purchases," Inf Technol Tourism, vol. 21, no. 4, pp. 483–514, Dec. 2019, doi: 10.1007/s40558-019-00154-1.

[28] H. Kim, C. Huh, C. Song, and M. J. Lee, "How can hotel smartphone apps enhance hotel guest experiences? An integrated model of experiential value," JHTT, vol. 12, no. 4, pp. 791–815, Dec. 2021, doi: 10.1108/JHTT-07-2020-0176.

[29] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," Management Science, vol. 46, no. 2, pp. 186–204, Feb. 2000, doi: 10.1287/mnsc.46.2.186.11926.

[30] H.-C. Chang, "Do gender systems in the origin and destination societies affect immigrant integration? Vietnamese marriage migrants in Taiwan and South Korea," Journal of Ethnic and Migration Studies, vol. 46, no. 14, pp. 2937–2955, Sep. 2020, doi: 10.1080/1369183X.2019.1585014.

[31] B. D. Tran and D. H. Vu, "Gen-Y Behavioral intention to adopt mobile tourism apps: Extending UTAUT2 with trust and security," 10.5267/j.ijdns, vol. 8, no. 4, pp. 2173–2184, 2024, doi: 10.5267/j.ijdns.2024.6.014.

[32] X. Zhou and Z. Mingyue, "A Study of the Influence of ESG Performance, Social Influence, Facilitating Conditions, Hedonic Motivation, Price Value and Habit on Tourists' Behavioural Intentions towards Sports Complexes," Prace KGP PTG, vol. 37, no. 4, Dec. 2023, doi: 10.24917/20801653.374.6.

[33] Venkatesh, Morris, Davis, and Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27, no. 3, p. 425, 2003, doi: 10.2307/30036540.

[34] G. Hofstede, G. J. Hofstede, and M. Minkov, Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival, Revised and Expanded third edition. New York, NY: McGraw-Hill, 2010.

[35] S. Li and K. Rouibah, "Cultural Difference and Consumer Satisfaction When Interacting with Technology Mediated Customer Services," IIS, 2021, doi: 10.48009/3_iis_2021_29-43.

[36] J.-W. Lian and D. C. Yen, "Online shopping drivers and barriers for older adults: Age and gender differences," Computers in Human Behavior, vol. 37, pp. 133–143, Aug. 2014, doi: 10.1016/j.chb.2014.04.028.

[37] C. D. Nguyen, "The Impact of Ai Adoption on Consumer Purchase Intention and Marketing Effectiveness in Vietnam's Online Retail Industry," International Journal of Education, vol. 4, 2024.

[38] B. Wilczek, S. Thäsler-Kordonouri, and M. Eder, "Government regulation or industry self-regulation of AI? Investigating the relationships between uncertainty avoidance, people's AI risk perceptions, and their regulatory preferences in Europe," AI & Soc, vol. 40, no. 5, pp. 3797–3811, Jun. 2025, doi: 10.1007/s00146-024-02138-0.

[39] A. F. M. J. Ahamed and W. Gong, "What Drives High Penetration Rates of Social Media a Qualitative Comparative Analysis Across Countries," JBM, vol. 28, no. 1, pp. 101–130, 2022, doi: 10.1504/JBM.2022.141296.

[40] B. W. Rababa and A. Ali, "The Influence of Cultural Factors on Mobile Banking Adoption: A Systematic Literature Review," International Journal of Academic Research in Progressive Education and Development, vol. 1, no. 3, 2024.

[41] A. Tarhini, K. Hone, X. Liu, and T. Tarhini, "Examining the moderating effect of individual-level cultural values on users' acceptance of E-learning in developing countries: a structural equation modeling of an extended technology acceptance model," Interactive Learning Environments, vol. 25, no. 3, pp. 306–328, Apr. 2017, doi: 10.1080/10494820.2015.1122635.

[42] L.-Y. Tseng, J.-H. Chang, and Y. L. Zhu, "What drives the travel switching behavior of Chinese Generation Z consumers," JTF, vol. 10, no. 1, pp. 131–146, Apr. 2024, doi: 10.1108/JTF-07-2020-0110.

[43] J. Chalmers and M. Armour, "The Delphi Technique," in Handbook of Research Methods in Health Social Sciences, P. Liamputtong, Ed., Singapore: Springer Singapore, 2019, pp. 715–735. doi: 10.1007/978-981-10-5251-4_99.

[44] H. A. Linstone and M. Turoff, The Delphi Method: Techniques and Applications. Addison-Wesley Publishing Company, Advanced Book Program, 1975.

[45] R. E. Al-Qutaish, "Delphi Method: A Comprehensive Literature Review," International Journal of Software Engineering, 2025.

[46] J. Schifano and M. Niederberger, "How Delphi studies in the health sciences find consensus: a scoping review," Syst Rev, vol. 14, no. 1, p. 14, Jan. 2025, doi: 10.1186/s13643-024-02738-3.

[47] J. C. Westland, "Information loss and bias in likert survey responses," PLoS ONE, vol. 17, no. 7, p. e0271949, Jul. 2022, doi: 10.1371/journal.pone.0271949.

[48] S. Nădăban, "Fuzzy Logic and Soft Computing—Dedicated to the Centenary of the Birth of Lotfi A. Zadeh (1921–2017)," Mathematics, vol. 10, no. 17, p. 3216, Sep. 2022, doi: 10.3390/math10173216.

[49] ZADEH, "Fuzzy Sets," 1965.

[50] K. Franke, M. Koppen, and B. Nickolay, "Fuzzy image processing by using Dubois and Prade fuzzy norms," in Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Sep. 2000, pp. 514–517 vol.3. doi: 10.1109/ICPR.2000.903596.

[51] I. Grattan-Guinness, "Didier Dubois and Henri Prade. Fuzzy sets and systems. Theory and applications. Mathematics in science and engineering, vol. 144. Academic Press, New York etc. 1980, xvii + 393 pp.," J. symb. log., vol. 47, no. 3, pp. 702–703, Sep. 1982, doi: 10.2307/2273604.

[52] T. T. Yin and H. Hanif, "Fuzzy Delphi Method: A Step-by-Step Guide to Obtain Expert Consensus on MUETBot Functionalities," IJARBSS, vol. 14, no. 4, p. Pages 1109-1116, Apr. 2024, doi: 10.6007/IJARBSS/v14-i4/21307.

[53] H. Jones and B. C. Twiss, Forecasting Technology for Planning Decisions. PBI, 1978.

[54] M. Adler and E. Ziglio, Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health. Jessica Kingsley Publishers, 1996. [Online]. Available: https://books.google.co.id/books?id=jo1Z1JZIrKIC

[55] R. Lizarralde, J. Ganzarain, and M. Zubizarreta, "Assessment and Selection of Technologies for the Sustainable Development of an R&D Center," Sustainability, vol. 12, no. 23, p. 10087, Dec. 2020, doi: 10.3390/su122310087.

[56] A. Habibi, F. F. Jahantigh, and A. Sarafrazi, "Fuzzy Delphi Technique for Forecasting and Screening Items," Asia. Jour. Rese. Busi. Econ. and Manag., vol. 5, no. 2, p. 130, 2015, doi: 10.5958/2249-7307.2015.00036.5.

[57] H. S. Halim, Z. Q. Nan, and M. A. Bhuiyan, "Developing green marketing tourism in Perak, Malaysia," E3S Web Conf., vol. 251, p. 03019, 2021, doi: 10.1051/e3sconf/202125103019.

[58] M. A. A. Hakim, A. Suryantoro, and M. Rahardjo, "Analysis of the Influence of Tourism Growth on Economic Growth and Human Development Index in West Java Province 2012-2018," BIRCI, vol. 4, no. 1, pp. 160–169, Jan. 2021, doi: 10.33258/birci. v4i1.1561.

[59] L. Ocampo et al., "Research Productivity for Augmenting the Innovation Potential of Higher Education Institutions: An Interpretive Structural Modeling Approach and MICMAC Analysis," Journal of Open Innovation: Technology, Market, and Complexity, vol. 8, no. 3, p. 148, Sep. 2022, doi: 10.3390/joitmc8030148.

[60] W. Hidayat and N. Lawahid, "Meode Fuzzy Delphi Untuk Penelitian Sosial," Alfabeta cv, 2020.

[61] F. Dehzad, C. Hilhorst, C. De Bie, and E. Claassen, "Adopting Health Apps, What's Hindering Doctors and Patients?" Health, vol. 06, no. 16, pp. 2204–2217, 2014, doi: 10.4236/health.2014.616256.

[62] S. Sunny, L. Patrick, and L. Rob, "Impact of cultural values on technology acceptance and technology readiness," International Journal of Hospitality Management, vol. 77, pp. 89–96, Jan. 2019, doi: 10.1016/j.ijhm.2018.06.017.

[63] E. Swierad, L. Vartanian, and M. King, "The Influence of Ethnic and Mainstream Cultures on African Americans' Health Behaviors: A Qualitative Study," Behavioral Sciences, vol. 7, no. 3, p. 49, Aug. 2017, doi: 10.3390/bs7030049.

[64] Á. L. Coves-Martínez, C. M. Sabiote-Ortiz, and D. M. Frías-Jamilena, "How to improve travel-app use continuance: The moderating role of culture," Tourism Management Perspectives, vol. 45, p. 101070, Jan. 2023, doi: 10.1016/j.tmp.2022.101070.

[65] M. Jacobs-Basadien and S. Pather, "The Role of Culture in User Adoption of Mobile Applications for Self-management of Health: A Conceptual Framework:" in Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health, Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications, 2022, pp. 37–49. doi: 10.5220/0011039300003188.

[66] I. Lee, B. Choi, J. Kim, and S.-J. Hong, "Culture-Technology Fit: Effects of Cultural Characteristics on the Post-Adoption Beliefs of Mobile Internet Users," International Journal of Electronic Commerce, vol. 11, no. 4, pp. 11–51, Jul. 2007, doi: 10.2753/JEC1086-4415110401.

[67] F. Sharmin, M. T. Sultan, D. Badulescu, A. Badulescu, A. Borma, and B. Li, "Sustainable Destination Marketing Ecosystem through Smartphone-Based Social Media: The Consumers' Acceptance Perspective," Sustainability, vol. 13, no. 4, p. 2308, Feb. 2021, doi: 10.3390/su13042308.

[68] L. T. Bürkle, "Going local: A Cross-Cultural Comparison of Communication Strategies in Customer Magazines applied by Global Organizations.," 2020.

[69] A. Golets, J. Farias, R. Pilati, and H. Costa, "COVID-19 pandemic and tourism: The impact of health risk perception and intolerance of uncertainty on travel intentions," Curr Psychol, vol. 42, no. 3, pp. 2500–2513, Jan. 2023, doi: 10.1007/s12144-021-02282-6.

Optimized Automatic Temperature and Humidity Control for Tobacco Storage Using TwinCAT and Deep Reinforcement Learning

Zhen Liu, Jili Wang*, Shihao Song, Qiang Hua
Qingdao ETSONG Technology Co.Ltd, Qingdao 266001, China

***Abstract*—With the rapid development of the tobacco industry, precise temperature and humidity control in storage environments has become essential for maintaining tobacco leaf quality. Traditional manual control methods suffer from low efficiency and limited accuracy, failing to meet modern storage demands. This study proposes an optimized automatic control system integrating TwinCAT and deep reinforcement learning (DRL) to enhance climate regulation in tobacco warehouses. Leveraging TwinCAT's real-time control capabilities and DRL's adaptive decision-making, the system achieves precise environmental regulation. Experimental results demonstrate that temperature and humidity control errors are reduced to ±0.5 °C and ±3%, respectively. Compared to conventional methods, the proposed system lowers energy consumption by 20% and reduces the mildew rate of stored tobacco by 15%, significantly improving storage quality. This work offers a novel technical framework for intelligent environmental control in tobacco storage and provides valuable insights for broader applications in similar domains.**

***Keywords*—TwinCAT; deep reinforcement learning; tobacco storage; temperature and humidity control; system optimization**

I. Introduction

In today's rapidly developing tobacco industry, tobacco storage is an important link between production and sales, and the importance of environmental control is increasingly prominent [1, 2]. As an agricultural product extremely sensitive to environmental conditions, the quality of tobacco leaves is closely related to the temperature and humidity in the storage environment. Improper temperature and humidity conditions will not only lead to mildew and moth-eating tobacco leaves but also affect their colour, aroma and taste, thus causing irreversible effects on the quality of tobacco products [3, 4]. Therefore, accurately and efficiently controlling the temperature and humidity in the tobacco storage environment has become an urgent technical problem that needs to be solved in the tobacco industry.

The rapid progress of automation technology, especially the wide application of industrial automation software TwinCAT (The Windows Control and Automation Technology), provides a new solution for tobacco storage environment control [5, 6]. With its powerful real-time flexibility and openness, TwinCAT can realize real-time monitoring and precise control of storage environment parameters [7]. However, traditional control strategies are often based on fixed rules or models, which are difficult to adapt to the complex and changeable storage environment, resulting in unsatisfactory control effects.

In this context, the rise of Deep Reinforcement Learning (DRL) technology has brought new ideas for controlling the tobacco storage environment [8]. By combining the perception ability of deep learning with the decision-making ability of reinforcement learning, DRL can enable the control system to learn and optimize itself in complex environments, thus achieving more intelligent and efficient control [9, 10]. Combining TwinCAT with DRL to build a new automatic temperature and humidity control system for a tobacco storage environment can not only make full use of the real-time control capabilities of TwinCAT but also further improve the performance and adaptability of the control system with the help of the intelligent optimization characteristics of DRL.

This study aims to explore the feasibility and effectiveness of this fusion technique. An intelligent optimisation algorithm based on DRL is designed and implemented through an in-depth analysis of the characteristics of the tobacco storage environment and its demand for temperature and humidity, combined with the real-time monitoring and control system of TwinCAT. The algorithm can autonomously adjust the control strategy according to the real-time environmental data to best control temperature and humidity. At the same time, this study will verify the actual effect of the proposed system through experiments and compare it with the traditional control methods in order to provide a new and more efficient technical means for tobacco storage environment control.

Existing tobacco storage temperature and humidity control methods have limitations such as insufficient integration of industrial control systems and intelligent algorithms, difficulty in adapting to complex and variable storage environments, and less ideal control accuracy and energy efficiency. The research fills the gap by combining TwinCAT industrial control technology with deep reinforcement learning, realizing a more intelligent, adaptive and high-performance automatic temperature and humidity control system for tobacco storage, which effectively addresses the aforementioned shortcomings.

This research will involve cutting-edge knowledge in many fields, such as automation control theory, deep learning, and reinforcement learning. By organically combining these theories, it aims to build an automatic control system with powerful and intelligent optimisation abilities. At the practical level, this study will focus on the practical application effect of the system, including control accuracy, stability, energy consumption and other aspects, to ensure that the proposed technology can truly meet the actual needs of tobacco storage.

*Corresponding Author

The optimization research on the automatic control system of temperature and humidity in tobacco storage environment by integrating TwinCAT and DRL has important theoretical value and broad application prospects. This research can provide new technical support for the warehousing environment control of the tobacco industry and promote the development of the industry in a more intelligent and efficient direction.

The basic theory section on TwinCAT and DRL elaborates on the TwinCAT platform, including its features, components, and real-time control capabilities, as well as the principles of DRL, such as Markov decision processes, state-action value functions, and the integration of deep neural networks with reinforcement learning. Subsequently, in the optimization design and implementation of the automatic temperature and humidity control system for tobacco storage environment, the design of the temperature and humidity automatic control system based on TwinCAT was introduced, covering the software and hardware aspects and system architecture, and the temperature and humidity control algorithm based on DRL was also developed to solve the coordination and optimization problems. The experimental and result analysis section provides experimental data through various graphs, comparative analysis with other methods, and discusses the performance of the proposed system, demonstrating the improvement in control accuracy, energy efficiency and stability. Finally, the conclusion summarizes the main findings of this study, highlights the effectiveness of the integrated system, and proposes future research directions.

In the field of temperature and humidity control of tobacco storage, other researchers have proposed a variety of solutions. Some studies use the traditional PID control algorithm to adjust the temperature and humidity by setting fixed parameters, which realizes basic control, but it is difficult to cope with the complex and changeable interference in the storage environment, and the control accuracy and adaptability are limited. There are also studies on the use of fuzzy control methods to deal with uncertainty factors by using fuzzy rules, but there are shortcomings in dynamic response speed and optimization ability. In addition, some scholars try to combine simple machine learning algorithms with control strategies, although they have certain effects in specific scenarios, but lack of deep integration of industrial control systems. At present, the existing schemes generally have problems such as low integration with industrial control platforms, difficulty in adapting to changes in complex storage environments, and unsatisfactory control accuracy and energy efficiency, which are the directions to be further studied. In this study, TwinCAT industrial control technology and deep reinforcement learning are deeply integrated, which not only gives full play to the advantages of TwinCAT in real-time control and system integration, but also makes up for the shortcomings of existing schemes in complex environment adaptability and control performance with the help of deep reinforcement learning's adaptive learning and optimization decision-making ability, and forms a more intelligent, efficient and suitable for the actual industrial scene of tobacco storage temperature and humidity control scheme, which has made significant breakthroughs in technology integration and control effect.

II. Basic Theory of TwinCAT and DRL

A. Introduction to the Twincat Platform

TwinCAT is a pure software controller, which is the core component of the Beckhoff controller, with excellent openness and expansion potential [11, 12]. It has an interface that can be connected to a common field. The development environment integrates Microsoft Visual Studio software, supports special programming languages such as IEC61131 and PLC, and is compatible with high-level languages such as C, C + + and MATLAB/Simulink. Users can choose programming tools according to task characteristics. In addition, TwinCAT has multi-core processing capabilities, which can use all cores to improve performance according to the controller's condition. The PC can be regarded as a calculator, PLC and motion controller when its running core is installed.

TwinCAT's real-time system includes multiple industrial control software packages, covering various motion control software modules, such as TwinCATPLC, TwinCAT NC, TwinCAT CNC Scope View, etc. They can operate independently, exchange information, and work collaboratively through the TwinCAT ADS interface. The TwinCAT task manager controls the real-time process [13, 14].

TwinCAT PLC part is the core of Beckhoff equipment to implement robot motion control, including various logic instructions to control motor motion and some key sub-modules commonly used in programming [15]. Scope View is a system variable monitoring and analysis tool of TwincCAT software, which can display program variables in graphical form, which is convenient for users to monitor, analyze and control system variables in real-time. It is also equipped with Cursor Cursor, Trigger Trigger and other tools for easy operation [16]. Scope Array Bar Project monitors array variables with histograms; Scope project monitors a single variable over a time program; The Scope YTNC project monitors axis variables by axis number; Scope YT project with reporting analyzes the YT graph with reporting function; Scope XY project observes the corresponding relationship between any two variables; Scope XY project with reporting analyzes the XY graph with reporting function. TwinCAT NC is responsible for axis motion control in the TwinCAT system, which can improve axis control performance and play a key role in logic and function control. As a core module, it accelerates platform development and broadens application scope. Its compatibility enables connection with other manufacturers' equipment, and it has the advantages of superior performance, fast speed, high efficiency, and convenience [17, 18].

In the TwinCAT system, the motor control process is divided into three layers: ① PLC axis, axis variables defined under the TwinCAT PLC module; ② NC axis, a virtual axis added to the NC axis configuration interface under the Motion module; ③ Physical axis, the motion execution and position feedback hardware obtained by I/O module scanning and added to the TwinCAT system, each layer has different functions and is interrelated.

B. Deep Reinforcement Learning (DRL) Principles

Reinforcement learning (RL) uses the interaction between agents and the environment to continuously learn by trial and error to find the best strategy [19]. Specifically, RL does not rely on the real-time supervision signal to guide the learning path but relies on the reward signal evaluation strategy to indirectly guide the agent to learn towards the reward maximum and reduce the dependence on the accurate system model [20]. Under the RL framework, the agent makes decisions according to the real-time state of the environment. After execution, the environment enters a new state and feeds back the reward. The infinite loop of this decision and reward feedback constitutes its training process [21].

The core of training is agent decision-making to maximize long-term benefits [22]. The interaction between agent and environment in RL is often modeled as Markov decision process (MDP) [23]. MDP is characterized by a quadruple *(S, A, P, r)*, *S* is the state set, *A* is the action set, *P* describes the probability that the state $s \in S$ transitions to $s' \in S$ after executing the action $a \in A$, and *r* is the reward obtained by the agent when it transitions from state S to the next state S' after executing the action. Under the MDP framework, the agent selects the action $a_t \in A$ through the policy function $\pi(a_t|s_t)$ based on the current state $s_t \in S$ at time *t*, and gets a reward $r(s_t, a_t)$ after execution. The next state is randomly determined by the transition probability $P(s_{t+1}|s_t, a_t)$. The sequence of states, actions and rewards experienced by the agent before reaching the termination state constitutes a round, and its goal is to act to maximize the sum of all rewards at the end of the round, that is, the reward, which is defined in Eq. (1).

$$R_t = \sum_{k=0}^{N-1} r_{t+k} \tag{1}$$

In the formula, *N* represents the total number of steps of the discrete step size. In order to distinguish the relative importance of immediate reward from future reward, a reward discount factor γ (ranging from [0, 1]) is introduced. When γ = 0, immediate reward dominates; When γ = 1, future rewards are more important. Therefore, the return R_t is redefined as Eq. (2).

$$R_t = \sum_{k=0}^{N-1} \gamma^k r_{t+k} \tag{2}$$

The fundamental purpose of an agent is to find the optimal strategy π^*, from which it can determine the best actions it should perform in each state and ensure the maximum discount reward. In order to explore the optimal strategy π^*, a state value function $V(s)$ can be constructed to evaluate the advantageous degree of reaching the current state *s*, as shown in Eq. (3).

$$V^{\pi}(s_t) = E\left[R_t \mid s_t = s\right] = E\left[\sum_{k=0}^{N-1} \gamma^k r_{t+k} \mid s_t = s\right] \tag{3}$$

The state-valued function $V(s)$ is a means to evaluate the pros and cons of a strategy. For any $s \in S$, if the expected return according to the strategy π is higher than π', then $V_{\pi}(s) \geq V_{\pi'}(s)$, indicating that the strategy π is better, and the agent will tend to π when choosing the strategy function. Therefore, the optimal strategy π^* is defined in Eq. (4).

$$V^*(s) = \max_{\pi} V_{\pi}(s), \forall s \in S \tag{4}$$

Although the state value function can compare different strategies, it cannot find the optimal strategy [24]. Bellman created the state-action value function $Q_{\pi}(s, a)$, which is used to measure the degree of return of taking action a according to strategy π in state *s*, as shown in Eq. (5).

$$Q^{\pi}(s_t, a_t) = E\left[R_t \mid s_t = s, a_t = a\right] = E\left[\sum_{k=0}^{N-1} \gamma^k r_{t+k} \mid s_t = s, a_t = a\right] \tag{5}$$

The optimal Q*(s, a) can be defined as Eq. (6).

$$Q^*(s,a) = \max_{\pi} Q_{\pi}(s,a), \forall s \in S, a \in A \tag{6}$$

In order to determine the strategy π, it is necessary to pick out the action matching the maximum value from the action state value function $Q_{\pi}(s, a)$ for a specific state *s*. This is called greedy strategy, which is defined in Eq. (7).

$$\pi(a \mid s) = \begin{cases} 0, if \ \ a = \arg\max_{a'} Q(s,a') \\ 0, if \ \ a \neq \arg\max_{a'} Q(s,a') \end{cases} \tag{7}$$

When the agent operates according to the optimal strategy π^*, the expression of $V^*(s)$ is shown in Eq. (8).

$$V^*(s) = \max_{\pi} Q_{\pi^*}(s,a) \tag{8}$$

The corresponding Bellman equation can be obtained by recursive extension of Eq. (8), as shown in Eq. (9).

$$V^*(s) = \max_{a} \sum_{s',r} p(s', r \mid s, a)[r + \gamma V^*(s')] \tag{9}$$

By recursive derivation of Eq. (6), the optimal Bellman equation of $Q^*(s, a)$ can be obtained, Eq. (10) can be obtained.

$$Q^*(s,a) = \sum_{s'} p(s', r \mid s, a)[r + \gamma \max_{a'} Q^*(s', a')] \tag{10}$$

The premise of the research is to clarify the state transition probability and reward value of the agent, and the Bellman optimal equation can be solved iteratively, which is dynamic programming [25]. Algorithms with known state transition probabilities and rewards are collectively called model-based algorithms. Most RL problems assume that the state transition probability is unknown, and the algorithm set designed for this is called a model-free algorithm.

Traditional reinforcement learning is limited, with small action and sample spaces and mostly discrete environments, which makes it difficult to cope with high-dimensional input data such as images and sounds [26, 27]. DRL combines high-dimensional input of deep neural networks (DNNs) with reinforcement learning. DNNs have the latest technological advances in speech recognition, image classification, machine translation, robot control, etc. The commonly used architecture of DNNs is a feedforward neural network (FNN), a multi-layer perceptron model consisting of the input layer, one or more hidden layers, and an output layer [28].

After the input signal is introduced into the input layer, it will be transmitted layer by layer in the network. The hidden layer and output layer are composed of multiple nodes (perceptrons) [29]. When each perceptron receives the input vector of the previous layer, it will assign a weight ω to each vector element x and sum it. The operation result can be expressed by Eq. (11).

$$z = \sum_i x_i \omega_i + b \tag{11}$$

Where b is the deviation coefficient. After the summation operation is completed and the result is obtained, the activation function f is used to generate the neuron output y, and the calculation process of the neuron model is shown in Eq. (12).

$$y = f(\sum_i x_i \omega_i + b) \tag{12}$$

The multiplication operation of adjusting the input weight of the perceptron can change its behaviour and that of the network. It is necessary to train the neural network to adjust the weight in a specific way so that the network behaviour meets the expected results [30]. Fig. 1 shows the DRL core architecture. The deep learning module captures target observation data from the environment and provides environmental state information. Then, the reinforcement learning component corresponds the current state to the corresponding action, evaluates the value according to the estimated return, optimizes the decision-making action through the interactive process, and iteratively updates the network parameters to obtain the best strategy.

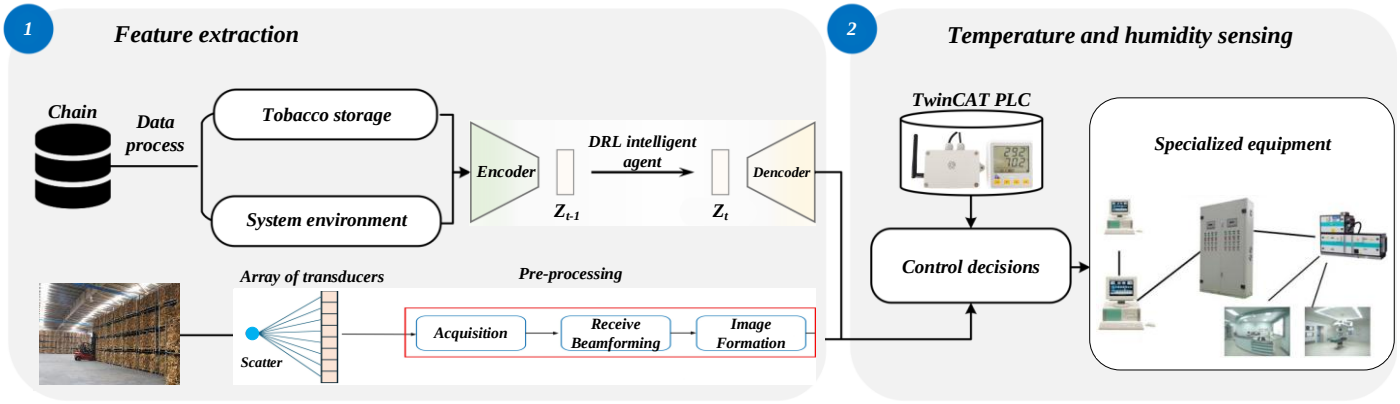


Fig. 1. Deep reinforcement learning architecture.

III. Optimization Design and Implementation of Automatic Control System for Temperature and Humidity in Tobacco Storage Environment

A. Automatic Control System of Temperature and Humidity in Tobacco Storage Environment

The inspection machine control system software software includes seven functional modules: start-stop speed control, servo motor control, etc. Accordingly, the software and hardware design process is carried out.

Regarding software design, the inspection machine involves the interface design of the host computer based on the Qt framework, visual inspection algorithm program and Beckhoff PLC programming based on TwinCAT. The Qt host computer interface uses the Qt Creator development tool on the Windows platform (Qt 5.12. 2 versions). The host computer and Beckhoff PLC interact with each other through ADS communication protocol, and a communication bridge is built in Qt Creator environment with C + + language combined with ADSDLL library of Beckhoff TwinCAT to realize the reading and writing of PLC memory data by the host computer. Beckhoff PLC is the ADS server in this architecture, and the host computer is the client. It uses an asynchronous notification mechanism to send ADS requests, and the server sends responses through the Call-back mechanism until the request is cancelled. Its advantage is that TwinCAT only transmits data when updated, which can avoid the loss of program running efficiency.

The TwinCAT integrated development platform is used for the lower computer Beckhoff PLC program design. TwinCAT is a PC-based real-time control system that can accurately control and respond to real-time systems, supports multi-task parallel processing, and assigns different priorities to tasks as needed. Each task contains one or more program blocks to execute specific control logic to cope with complex control challenges. In addition, TwinCAT is compatible with various communication protocols, which facilitates data interconnection with other devices and systems. The control system architecture adopted in this paper is shown in Fig. 2.

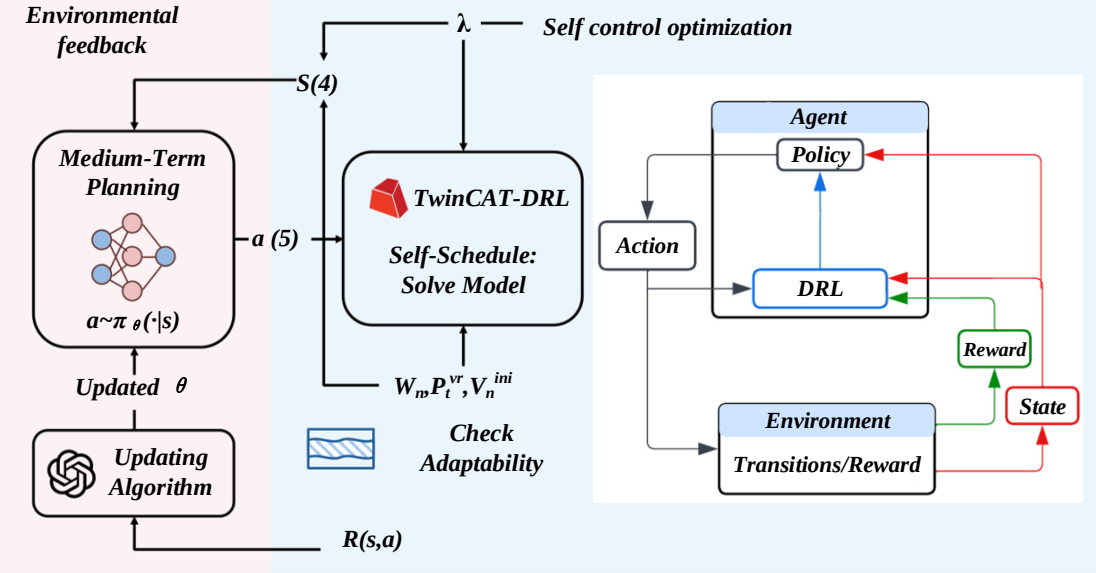


Fig. 2. Control system architecture.

B. Temperature and Humidity Control Algorithm Based on DRL

Because DRL aims to control frequency deviation minimization and tie-line power stability, GCD pursues control cost minimization. It is difficult to comprehensively optimize and control frequency deviation and control cost by combining the two algorithms, especially in the special situation of the current control system, and the actual operation time may exceed the AGC control interval limit, resulting in performance degradation. Therefore, to solve the mismatch problem between the two, this study proposes an integrated frequency regulation architecture based on an intelligent controller, which gives regulation instructions on the premise of balancing the two objectives according to the real-time state of the system to achieve multi-objective optimization.

In the automatic temperature and humidity control system, due to the wide variety of equipment involved and their unique characteristics, their frequency response speeds are also different, which requires the intelligent controller to accurately generate the matching power regulation instructions according to the unique performance characteristics of each frequency regulation unit. However, using traditional algorithms to complete this complex process often requires a lot of time and energy to carry out tedious parameter adjustment experiments to ensure the system's normal operation. Even so, when the system faces large-scale random disturbance, the dynamic performance of traditional algorithms is often unsatisfactory, and it is difficult to cope effectively with complex and changeable actual working conditions, which limits the stability and reliability of the temperature and humidity automatic control system to a certain extent.

Therefore, this study proposes an integrated frequency control strategy for a multi-regional interconnected temperature and humidity automatic control system, using the algorithm as the intelligent controller, sensing relevant data and outputting control instructions simultaneously, and designing the reward function to consider multiple dimensions to solve coordination problems comprehensively. Through pre-training, the algorithm masters the dynamic performance of each frequency modulation unit and the characteristics of system frequency change and generates reasonable instructions in the execution stage to ensure that the deviation is within the specified range. Each region is regarded as an independent agent, and the control instructions are obtained by a multi-agent cooperative game. The advantages of this strategy are: it can avoid the problem of AGC performance degradation, the algorithm is a multi-output algorithm, the calculation time is less than the command cycle limit, the strategy can be updated online without relying on the model, and it can generate instructions based on the real-time status of each unit to exert its frequency modulation potential.

IV. Experiment and Result Analysis

Experimental data Table I shows that the detection performance of the improved scheme is significantly improved. Under the mAP @ 50: 5: 95 evaluation standards, the improved version improves by about 1.5 percentage points on average compared with the previous one. Except for SSD, the mAP50 value of other methods exceeded 99.0%, and the mAP50 value of this method was the highest, reaching 99.5%. It can also be seen from Table 1 that although the improved method has a time delay of about 0.8 milliseconds compared with the original model, the processing speed is still about 130 fps, far exceeding the 20 fps of the two-stage target detector Faster-RCNN.

TABLE I Comparative Experimental Results

Detection model	mAP @ 50: 5: 95	mAP50	Average detection time per picture/ms
Base	95.6%	98.5%	8
This article improves	97.1%	98.5%	9
Faster R-CNN	97.7%	98.3%	60
SSD	48.8%	88.1%	42

Fig. 3 shows the average reward of each round of the algorithm agent. The training consists of 3000 rounds, each containing 60 time slots. At the beginning of the time slot, the agent interacts with the environment, makes decisions and updates the network model. During this period, the average reward of the DDPG module and DQN module climbed and converged with the increase in training rounds. The average reward of the DDPG module increases rapidly after 300 rounds and stabilizes after 1000 rounds. The DQN module gradually converges from the beginning of learning, and its average reward tends to be stable after 300 rounds. As the number of rounds increases, the DQN module will reduce the exploration frequency and generate greater actions, making its output decisions and rewards tend to be stable.

Fig. 4 explores the impact of warehousing computing resources on average weighted user energy consumption. The energy consumption of each scheme decreases with the increase of warehousing computing resources. When warehousing computing resources are scarce, the computing delay of users who choose to offload computing is too large, which does not meet the maximum delay permission conditions. Users prefer to execute tasks locally, resulting in higher user energy consumption.

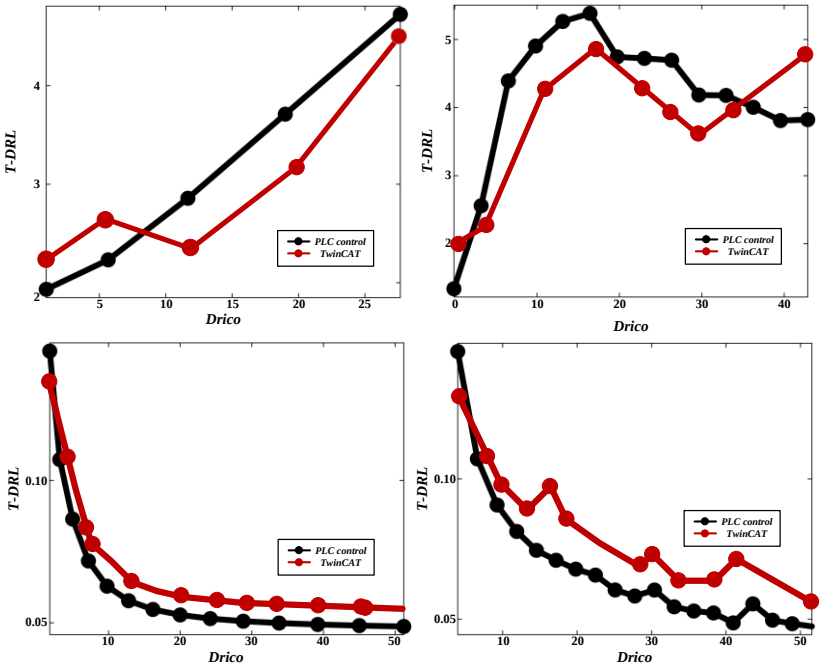


Fig. 3. DDPG unit average reward of the algorithm.

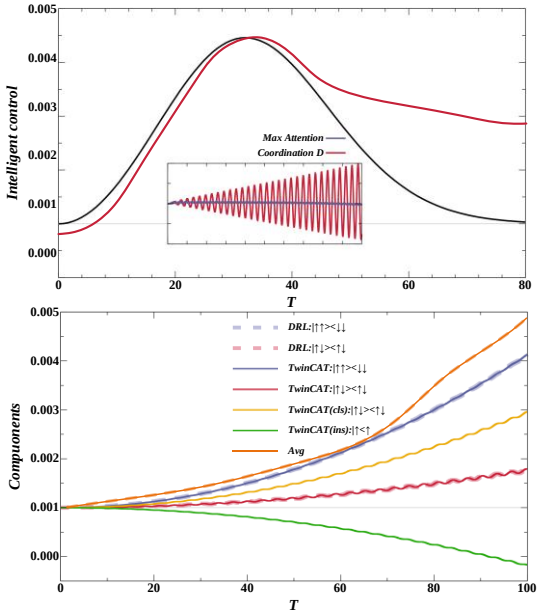


Fig. 4. Average weighted user energy consumption under different warehousing computing resources.

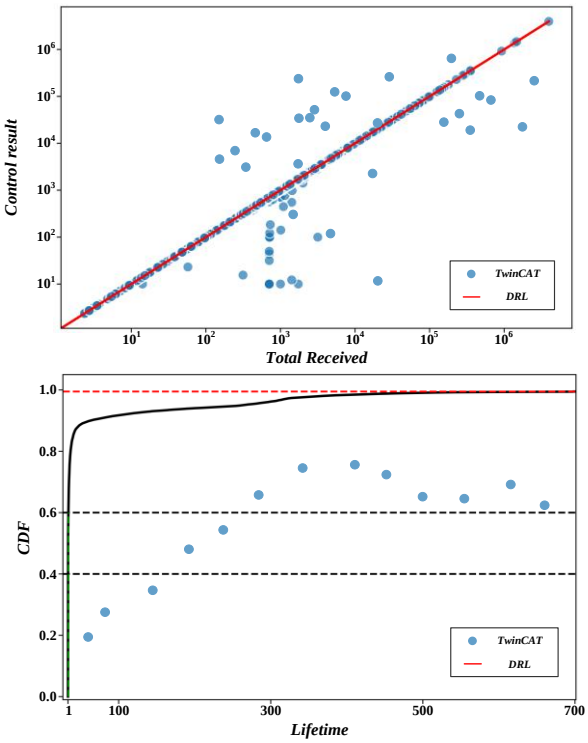


Fig. 5. Training results of different algorithms.

As can be seen from Fig. 5, the distributed training architecture uses many agents in multiple parallel systems to conduct distributed exploration, which can accelerate the process, improve agent training efficiency, and reduce computing overhead. Compared with centralized automatic control control (AGC) based on PPO and DDPG, distributed AGC based on deep reinforcement learning (DRL) has a better training effect and greatly increases training time. At the same time, PPO and DDPG algorithms have great volatility in the learning process, and the reward convergence value fluctuates, obviously. Hence, it is difficult to cope with random interference.

It can be seen from Fig. 6 that compared with DRL, PPO, DDPG and PID + PRPO, the frequency deviation at the same time point is smaller, and the frequency does not exceed 0.2 Hz during the control period. Because PID + PRPO cannot flexibly enable the quick response adjustment component to adjust the frequency, instability or over-adjustment problems will occur due to unreasonable parameter configuration, which may cause the frequency deviation to exceed 0.2 Hz and waste frequency adjustment resources.

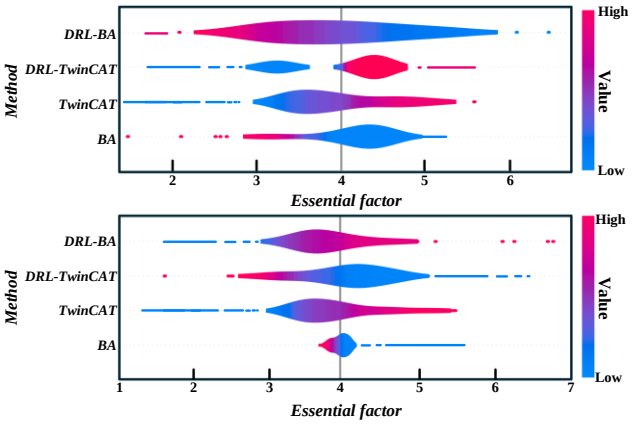


Fig. 6. Frequency deviation variation diagram of different algorithms.

It can be seen from Fig. 7 that the frequency control strategy can better fit the disturbance curve. Compared with the traditional PID + PRPO frequency modulation method, DRL technology's frequency modulation scheme performs better. DRL combines the essence of fuzzy control and neural network control, which can provide real-time feedback, adjust environmental changes during operation, and promote the control system to better adapt to the nonlinear dynamic environment. Compared with other DRL algorithms, it has smaller power tracking errors, can complete frequency adjustment as quickly as possible, and is more suitable for fluctuations in load and distributed resources in temperature and humidity automatic control systems.

It can be seen from Table II that each system index of the proposed strategy exceeds other strategies. Compared with other methods, this algorithm reduces the mean of Af by 19.3% to 58.1%, the mean of ACE by 12.6% to 77.4%, the mean of CPS1 by 0.15% to 28.6%, and the mean of CPS2 by 2.06% to 20.1%. By comparison, the controller has the smallest change in performance and the best adaptability and robustness.

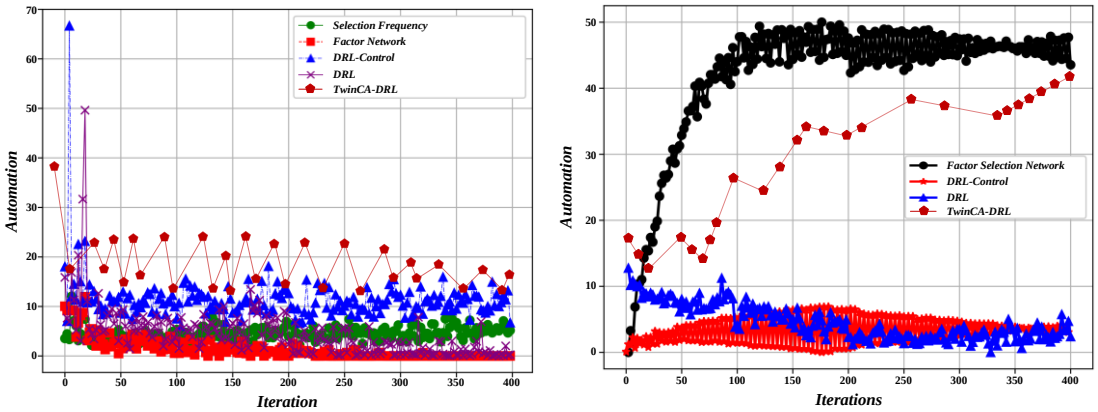


Fig. 7. Control effects of different strategies under white noise disturbance.

TABLE II MODEL STATISTICAL EXPERIMENTAL RESULTS UNDER RANDOM WHITE NOISE DISTURBANCE

Region	Index		DRL	PPO	DDPG	PI + PRPO
Region 1	Af/Hz	0.0694	0.0860	0.0912	0.1097	0.1689
	ACE/MW	13.3749	17.1171	21.1794	24.4013	44.2013
	CPS1/%	219.7602	219.4291	206.6361	203.8982	188.4960
	CPS2/%	106.1830	104.0160	98.2960	95.1940	88.3960
Region 2	Af/Hz	0.0563	0.0875	0.1159	0.1191	0.1838
	ACE/MW	11.1672	12.7006	19.9661	27.7134	44.9834
	CPS1/%	219.1354	219.0133	205.0983	201.8940	186.6480
	CPS2/%	104.8850	103.3780	97.2620	93.7310	87.5490

As shown in Fig. 8, the average algorithm score is 2.799. As the number of iterations increases and the exploration rate decreases, the average score of ERDQN algorithm grows slowly, while the average score of DRL-ERD3QN algorithm shows an upward trend, and its score rises rapidly in nearly 3000 rounds, which indicates that the network of this algorithm achieves rapid update in about 3000 rounds, while the neural network update speed of the other two algorithms is relatively slow.

As shown in Fig. 9, the DQ-DRL algorithm does not overestimate the agent policy learning. This shows that the algorithm effectively avoids the overestimation problem by taking two minimum values of Critic current network to update. Although this approach may lead to underestimation, underestimation is more acceptable than overestimation.

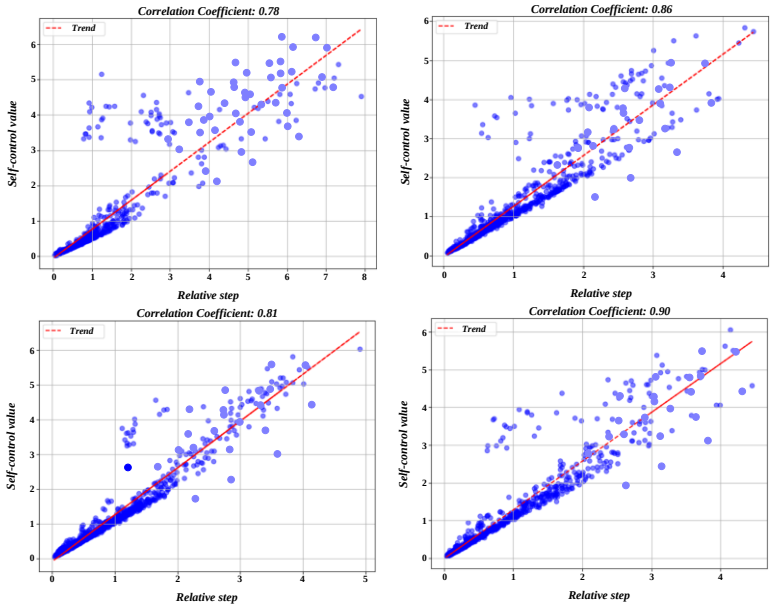


Fig. 8. Average score.

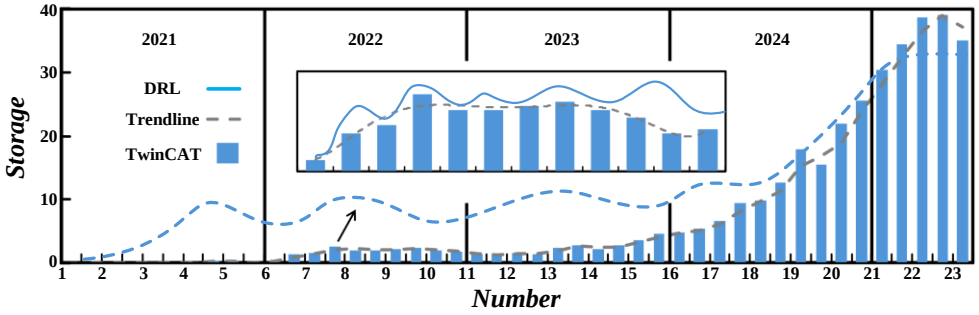


Fig. 9. Comparison of estimation bias.

Fig. 10 shows the changing trend of the cumulative total reward of the DQ-DRL algorithm and the follower agent trained by the DRL algorithm in each round when the group contains 4 units. It can be seen from the Fig. that in the early stage of training, the reward curves of both algorithms rise steadily. After about 35,000 rounds, the reward value of the DQ-DRL algorithm tends to stabilize and converge, while the reward curve of the DRL algorithm lags until about 55,000 rounds before converging. Moreover, the total reward value of the final convergence of the DQ-DRL algorithm is higher. This shows that although both algorithms allow the agent group to learn the stable formation control strategy, the DQ-DRL algorithm not only speeds up the convergence speed of the model but also improves the overall total reward of the agent group and enhances the superiority and robustness of the formation control model.

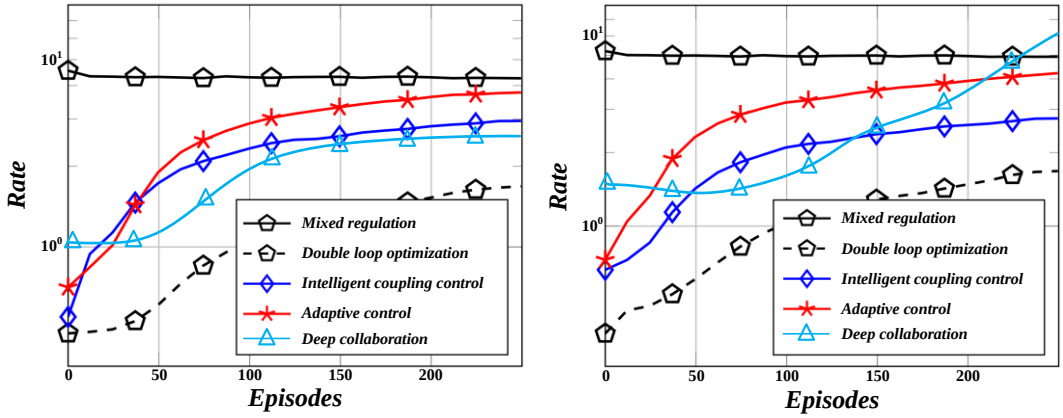


Fig. 10. Total reward curve obtained in each round during the training period.

V. Discussion

The integration of TwinCAT with Deep Reinforcement Learning (DRL) for temperature and humidity control in tobacco storage has led to a promising paradigm shift. Our results clearly show that this fusion not only solves the limitations of traditional control methods, such as poor adaptability to complex environments and poor accuracy, but also takes full advantage of the real-time processing power of TwinCAT and the adaptive decision-making capabilities of DRL to achieve better performance. The experimental results show that the temperature and humidity control error is reduced to $\pm 0.5^{\circ}$ C and $\pm 3\%$, the energy consumption is reduced by 20%, and the tobacco mildew rate is reduced by 15%, which highlights the practical value of the method. Our approach's ability to learn and adjust strategies in real-time is a key advantage compared to existing solutions such as PID control, which relies on fixed parameters and struggles to cope with dynamic changes, and fuzzy control that lacks self-optimizing capabilities. It is important to note that the success of the system depends on the seamless interaction between TwinCAT's industrial control infrastructure and the DRL algorithms, a synergy that enables the rapid acquisition, processing and execution of data, which is crucial in tobacco storage, where environmental fluctuations require a timely response. However, we also realize that the complexity of DRL model tuning and the need for sufficient training data may pose challenges to the widespread adoption of this method. Another aspect to consider is the versatility of our approach, and while it has been proven to be effective in tobacco storage, it is still worth exploring for its application to other agricultural or industrial storage scenarios, as different environments may have unique limiting factors. In addition, future improvements can focus on enhancing the robustness of the model to extreme weather conditions, as well as reducing the computational overhead of the DRL component, making it easier to apply in small storage facilities. Overall, this study highlights the potential of combining industrial control systems with advanced machine learning techniques to revolutionize storage environment management, providing a balance of precision, efficiency, and adaptability that is difficult to achieve with traditional methods.

VI. Conclusion

This study explores the optimal application of fusing TwinCAT and deep reinforcement learning (DRL) technology in tobacco storage environments' automatic temperature and humidity control systems. By constructing an intelligent control system, the high-precision and high-efficiency regulation of the tobacco storage environment can be realized, thus ensuring the quality of tobacco and reducing storage losses.

(1) Firstly, traditional temperature and humidity control system performance is benchmarked. The results show that under the unoptimized conditions, the control accuracy of the system for temperature and humidity is ± 2 ℃ and ± 5% RH, respectively, and there is obvious hysteresis. This result reveals the limitations of conventional control systems in coping with complex environmental changes.

(2) The system is preliminarily optimised by introducing the TwinCAT platform and using its powerful real-time control and data acquisition capabilities. The experimental results show that through the integrated management of TwinCAT, the control accuracy of the system has been significantly improved, and the control accuracy of temperature and humidity has been improved to ± 1 ℃ and ± 3% RH, respectively. At the same time, the system's response speed has been accelerated, and the lag phenomenon has been effectively alleviated.

(3) To further improve the system performance, this study introduces DRL technology and constructs an intelligent control

model based on deep reinforcement learning. The model achieves accurate prediction and adaptive regulation of temperature and humidity changes by continuously learning environmental data and control strategies. The final experimental results show that the system combining TwinCAT and DRL can stabilize the control accuracy of temperature and humidity within ± 0. 5 ℃ and ± 2% RH, respectively, and almost eliminate the hysteresis phenomenon. In addition, the system's energy consumption has also been effectively reduced, saving about 15% of energy consumption compared with traditional systems.

The tobacco storage environment's automatic temperature and humidity control system integrating TwinCAT and DRL shows excellent performance and great application potential. This study provides a new technical path for the intelligent management of tobacco storage environment and a useful reference for optimising automatic control systems in related fields. In the future, we will continue to deepen research, explore more innovative applications, and contribute more to the development of tobacco storage and other related industries.

The existing solutions include traditional PID control (fixed parameters, difficult to resist interference), fuzzy control (insufficient dynamic response and optimization), and simple machine learning combined control (lack of deep integration of industrial systems). These schemes generally have the problems of insufficient integration of industrial control and intelligent algorithms, weak ability to adapt to complex environments, and poor accuracy and energy efficiency. In this study, TwinCAT and deep reinforcement learning are integrated, which not only gives full play to the advantages of real-time control of the former, but also improves the adaptability and optimization capabilities of the latter, makes up for the shortcomings of the existing schemes, and forms a better control scheme.

ACKNOWLEDGMENT

This work was supported by ETSONG Tobacco (Group) Co., Ltd, Research and application of automatic control system based on nitrogen-filled pest control technology, contract number: 2025370200740017.

REFERENCES

[1] J. Gao, W. Xie, D. Shi, J. Wu, and R. Wang, “Synchronized optimization of the logistics system of a tobacco high-bay warehouse under production task fluctuations,” Engineering Research Express, vol. 6, no. 4, 2024.

[2] H. Zhou, J. Wu, X. Zheng, H. Zhu, G. Lu, Y. Zhang, and Z. Shen, “Fuzzy-PID controller based on improved LFPSO for temperature and humidity control in a CA ripening system,” Journal of Food Process Engineering, vol. 47, no. 6, 2024.

[3] Y. Sun, C. Liu, G. Du, H. Wang, and Z. Tian, “An almost disturbance decoupling temperature and humidity control strategy of air-handling units,” International Journal of Control, vol. 98, no. 1, pp. 1-17, 2025.

[4] S. Nakayama, T. Takada, R. Kimura, and M. Ohsumi, “Model Predictive Control of Humidity Deficit and Temperature in Winter Greenhouses: Subspace Weather-Based Modelling and Sampling Period Effects,” Machines, vol. 12, no. 1, 2024.

[5] A. Sterk-Hansen, B. H. H. Saghaug, D. Hagen, M. F. Aftab, and Ieee, "A ROS 2 and TwinCAT Based Digital Twin Framework for Mechatronics Systems," International Conference on Control Mechatronics and Automation. pp. 485-490, 2023.

[6] E. Okafor, D. Udekwe, Y. Ibrahim, M. B. Mu'azu, and E. G. Okafor, “Heuristic and deep reinforcement learning-based PID control of trajectory tracking in a ball-and-plate system,” Journal of Information and Telecommunication, vol. 5, no. 2, pp. 179-196, 2021.

[7] M. B. Mohiuddin, I. Boiko, R. Azzam, and Y. Zweiri, “Closed-loop stability analysis of deep reinforcement learning controlled systems with experimental validation,” Iet Control Theory and Applications, vol. 18, no. 13, pp. 1649-1668, 2024.

[8] Z. Liu, C. Wang, and W. Wang, “Online Cyber-Attack Detection in the Industrial Control System: A Deep Reinforcement Learning Approach,” Mathematical Problems in Engineering, vol. 2022, 2022.

[9] Y. Liu, Y. Ni, C. Dong, J. Chen, and F. Liu, “Task scheduling for control system based on deep reinforcement learning,” Neurocomputing, vol. 610, 2024.

[10] Q. Liu, T. Xia, L. Cheng, M. van Eijk, T. Ozcelebi, and Y. Mao, “Deep Reinforcement Learning for Load-Balancing Aware Network Control in IoT Edge Systems,” Ieee Transactions on Parallel and Distributed Systems, vol. 33, no. 6, pp. 1491-1502, 2022.

[11] D. Liu, Y. Wu, Y. Kang, L. Yin, X. Ji, X. Cao, and C. Li, “Multi-agent quantum-inspired deep reinforcement learning for real-time distributed generation control of 100% renewable energy systems,” Engineering Applications of Artificial Intelligence, vol. 119, 2023.

[12] D. Lin, and Y. Liu, “Discrete phase shifts control and beam selection in RIS-aided MISO system via deep reinforcement learning,” China Communications, vol. 20, no. 8, pp. 198-208, 2023.

[13] R. Liang, H. Lyu, and J. Fan, “A Deep Reinforcement Learning-Based Power Control Scheme for the 5G Wireless Systems,” China Communications, vol. 20, no. 10, pp. 109-119, 2023.

[14] X. Li, X. Zhang, J. Li, F. Luo, Y. Huang, and X. Zhang, “Blocklength Allocation and Power Control in UAV-Assisted URLLC System via Multi-agent Deep Reinforcement Learning,” International Journal of Computational Intelligence Systems, vol. 17, no. 1, 2024.

[15] X. Wang, X. Wang, J. Wu, K. Zheng, Y. Pang, and S. Gang, “Research on quality traceability of cigarette by combining PDCA quality cycle with information strategy based on fuzzy classification,” Journal of Intelligent & Fuzzy Systems, vol. 40, no. 4, pp. 8217-8226, 2021.

[16] H. Liu, Q. Chen, N. Pan, Y. Sun, Y. An, and D. Pan, “UAV Stocktaking Task-Planning for Industrial Warehouses Based on the Improved Hybrid Differential Evolution Algorithm,” Ieee Transactions on Industrial Informatics, vol. 18, no. 1, pp. 582-591, 2022.

[17] Y. Xia, K. Chang, Y. Lin, and C. Zhu, “Feedforward decoupling control of indoor temperature and humidity using a single direct expansion air conditioning unit,” Science and Technology for the Built Environment, vol. 30, no. 6, pp. 626-643, 2024.

[18] S.-C. Vanegas-Ayala, J. Baron-Velandia, and D.-D. Leal-Lara, “A Systematic Review of Greenhouse Humidity Prediction and Control Models Using Fuzzy Inference Systems,” Advances in Human-Computer Interaction, vol. 2022, 2022.

[19] Sunardi, A. Yudhana, and Furizal, “Tsukamoto Fuzzy Inference System on Internet of Things-Based for Room Temperature and Humidity Control,” Ieee Access, vol. 11, pp. 6209-6227, 2023.

[20] Y. Sun, Y. Zhang, D. Guo, X. Zhang, Y. Lai, and D. Luo, “Intelligent Distributed Temperature and Humidity Control Mechanism for Uniformity and Precision in the Indoor Environment,” Ieee Internet of Things Journal, vol. 9, no. 19, pp. 19101-19115, 2022.

[21] S. Liu, X. Liu, T. Zhang, C. Wang, and W. Liu, “Joint optimization for temperature and humidity independent control system based on multi-agent reinforcement learning with cooperative mechanisms,” Applied Energy, vol. 375, 2024.

[22] C. Li, Y. Cheng, and X. Hou, “Humidity Diffusion Process Analysis and Life Prediction of a VSC-HVDC Control Protection Device Based on a Finite Element Simulation Method,” Electronics, vol. 13, no. 15, 2024.

[23] Y. Jiang, S. Zhu, Q. Xu, B. Yang, and X. Guan, “Hybrid modeling-based temperature and humidity adaptive control for a multi-zone HVAC system,” Applied Energy, vol. 334, 2023.

[24] D. Guo, D. Luo, Y. Zhang, X. Zhang, Y. Lai, and Y. Sun, “Application of deep reinforcement learning to intelligent distributed humidity control system,” Applied Intelligence, vol. 53, no. 13, pp. 16724-16746, 2023.

[25] F. Garcia-Manas, T. Hagglund, J. L. Guzman, F. Rodriguez, and M. Berenguel, “A practical solution for multivariable control of temperature

and humidity in greenhouses," European Journal of Control, vol. 77, 2024.

[26] M. Dong, J. Zhang, L. Zhang, L. Liu, and X. Zhang, "Research on Relative Humidity and Energy Savings for Air-Conditioned Spaces without Humidity Control When Adopting Air-to-Air Total Heat Exchangers in Winter," Buildings, vol. 14, no. 4, 2024.

[27] M. H. Demir, S. Cetin, O. Haggag, H. G. Demir, W. Worek, J. Premer, and D. Pandelidis, "Independent temperature and humidity control of a precooled desiccant air cooling system with proportional and fuzzy logic plus proportional based controllers," International Communications in Heat and Mass Transfer, vol. 139, 2022.

[28] J. Zhou, L. Li, A. Vajdi, X. Zhou, and Z. Wu, "Temperature-Constrained Reliability Optimization of Industrial Cyber-Physical Systems Using Machine Learning and Feedback Control," Ieee Transactions on Automation Science and Engineering, vol. 20, no. 1, pp. 20-31, 2023.

[29] H. Zhou, J. Wu, X. Zheng, H. Zhu, G. Lu, Y. Zhang, and Z. Shen, "Fuzzy-PID controller based on improved LFPSO for temperature and humidity control in a CA ripening system," Journal of Food Process Engineering, vol. 47, no. 6, 2024.

[30] Q. Zhong, E. Xu, G. Xie, X. Wang, and Y. Li, "Dynamic performance and temperature rising characteristic of a high-speed on/off valve based on pre-excitation control algorithm," Chinese Journal of Aeronautics, vol. 36, no. 10, pp. 445-458, 2023.

Adaptive AI-Driven Enterprise Resource Planning for Scalable and Real-Time Strategic Decision Making

Ghayth AlMahadin
Assistant Professor, Data Science Department, College of Information Technology, Mutah University, Jordan

***Abstract*—Enterprise Resource Planning (ERP) systems play a critical role in managing organizational assets and operations. However, traditional ERP systems rely on static, rule-based decision-making frameworks that lack the agility and intelligence required for real-time strategic support. To address these limitations, this study proposes an Adaptive AI-Driven Enterprise Resource Planning (A2ERP). This AI-augmented ERP framework integrates adaptive predictive models to enhance decision-making capabilities at scale. The A2ERP architecture features a dynamic data ingestion layer, an adaptive predictive engine utilizing online learning and ensemble methods, and a decision support interface empowered with explainable AI (XAI). It is designed for scalability through a containerized microservices architecture. Experimental results demonstrate that A2ERP achieves a 98% accuracy rate in both training and testing phases, effectively identifying errors such as omission, addition, and overstatement. Comparative evaluations show that A2ERP outperforms traditional ERP methods across key performance metrics, including precision, recall, and F1-score. The framework's ability to process large-scale, complex data in real-time underscores its effectiveness in delivering timely strategic insights. A2ERP represents a significant advancement toward scalable, adaptive, and intelligent ERP systems, bridging the gap between operational execution and strategic decision-making.**

Keywords*—*Enterprise resource planning; adaptive predictive modeling; real-time decision support; AI-Augmented ERP; ensemble learning

I. INTRODUCTION

Digital infrastructure of modern organizations relies on Enterprise Resource Planning (ERP) systems. [1]. They are the central nervous system to operate some important business functions, such as finance and human resources, supply chain logistics, and customer relations on the same digital platform. Due to integration of these highly variegated modules [2]. ERP systems have in the past made it possible to achieve better efficiency, uniformity, and control in organizational processes [3]. However, no matter how popular and functionally efficient the conventional ERP systems might be, they are certain to be rule-based and, therefore, limited in their nature since they operate on the rule-based, fixed logic. These systems are fixed in their workflow and decision tree logic, and hence they are non-flexible and not able to keep up with real-time dynamics and evolving business scenarios [4]. As organizations are constantly met with a sophisticated and constantly evolving environment full of market volatility, global supply chain mega-disruption risks, and quickening digital transformation, the reality that ERP systems could not (and cannot) keep up in real-time has become a fatal flaw [5].

Availability of data and time sensitivity have been the characteristic features of the current business landscape. Businesses are constantly producing large flows of structured and unstructured data on internal processes, interactions with customers, IoT sensors, and external data sources [6]. The strategic decision-making process within this type of environment requires not only the availability of data, but also the smart interpretation and prompt action. Sadly, the majority of the ERP platforms do not prepare to satisfy this need [7]. The traditional ERP systems are based on past information, fixed reports, and scheduled updates, which make the decision-making process slow and non-responsive [8]. They usually do not have advanced analytics options and have minimal or no support for predictive modeling and contextual intelligence. Further, their modular nature is most often disjointed, and any form of integration with more advanced tools of analytics is typically piecemeal and in silos, which introduces operational inefficiency and curtails a narrow ability to strategically align. This has revealed a huge technological divide between the potential of the existing ERP systems and the emerging needs of the contemporary business.

The driving force behind this study is the fact that there is an excessive gap that needs to be closed between the traditional ERP features and functions and the strategic intelligence that is demanded in high-paced business settings [9]. Although ERP systems have long been successful in areas related to transaction processing and workflow automation, the newly emerging battlefield is to make these systems smart, self-learning systems that can respond in real time to data. The importance of the work is explained by the possibility to change the perception of ERP systems, which are usually discussed as the passive storage of operational data, to the proactive involvement in the strategic decision-making process [10]. Incorporation of artificial intelligence, notably adaptive and predictive learning models, presents a game-changing possibility to improve the responsiveness, accuracy, and efficiency of ERP-based decisions [11].

The study reflects a technological and an organizational imperative. Technological. It seeks to address shortcomings of the static ERP models through integration of adaptive predictive processing at the system architecture level. Organizationally, it aims to equip decision-takers with insights that are not only timely, but also explorable and operable. Through the integration of continuous learning and on-demand data processing, the suggested layout will help dynamically respond to changing circumstances to allow organizations to act with more agility, foresight, and strategic alignment. The paper further acknowledges the need to have transparency in AI

decisions, particularly in enterprise contexts, where explainability is vital to user trust and acceptance [12]. In order to mitigate these risks and achieve the mentioned objectives, the proposed study introduces an extended framework, called Adaptive AI-Driven ERP (A2ERP). This architecture aims to deploy smart learning components at the heart of ERP solutions, allowing them to ingest data in real-time, perform adaptive predictive modeling, and decision-making accompanied by Explainable AI (XAI). Fundamentally, A2ERP presents a new adaptive predictive engine, which could constantly refresh its models using online learning algorithms, and make adjustments to respond to new data as it comes, without requiring re-training ex-novo. The framework can improve the robustness and accuracy of its predictions in different ERP modules through ensemble learning techniques.

A. Research Significance

The study is relevant in that it fills a serious gap in the conventional ERP systems, which is the inability to sustain real-time, adjustive, and strategic decision making. With businesses creating and depending on huge amounts of data, the static, rule-based ERP systems are inadequate in reacting to the changing operational environments and market environments. The A2ERP framework proposed creates a paradigm shift due to the combination of adaptive predictive modeling, explainable AI, and scalable microservices. Its ability to learn continuously and process data in real-time enables organizations to spot trends, detect anomalies detection and make proactive strategic decisions. The study therefore seals a critical technology gap that exists between the traditional features of ERP systems and the current requirements of the enterprise to respond swiftly to business uncertainties. The research helps bring the much-needed evolution in ERP systems by making them smarter in terms of providing advice and guidance to the enterprise, by virtue of being a smart decision-support ecosystem, as opposed to being a transactional system. The flexibility of the framework in combination with its flexibility across industries also makes it quite applicable and hence a worthy development in the realms of intelligent business process management.

B. Research Motivation

This research is motivated by the increasing mismatch between the real-time requirements of current business ventures and the non-real-time functionality of traditional ERP systems. Current ERP systems are non-adaptive and do not use real-time data in making high-level decisions. As the data becomes more complex and organizations require quick, data-informed decisions, it will be necessary to implement AI into ERP systems. The motive behind this research is to enable ERP systems to keep up with changes in business environments by empowering them with adaptive learning. The A2ERP framework is developed to bring this change to provide proactive, smart, and scalable enterprise management solutions.

C. Key Contributions

1) An AI-augmented ERP framework integrating adaptive predictive models for strategic decision support on a real-time basis has been developed.

2) A data ingestion and processing layer has been created, associating real-time, scalable analysis of the enterprise operations.

3) Utilizing machine learning, one can forecast trends, detect anomalies, and optimize decision-making within ERP modules.

4) Testing was performed with structured ERP schema data to provide a quantitative analysis for improvement in terms of scalability, accuracy, and strategic responsiveness.

The rest of the sections of this study are ordered as follows: Section II gives review of the existing literature AI-augmented enterprise resource planning with adaptive predictive models for scalable, real-time strategic decision support. Section III details the proposed research methodology. Section IV presents the experimental results. Section V discuss and concludes the study.

II. Literature Review

A. Introduction and Background of ERP Systems

Enterprise Resource Planning systems have transformed greatly over the past few decades, which saw them move away from standalone monolithic legacy systems to more agile cloud-based and integrated platforms [13]. Early ERP systems had it in them to automate back-office functions and to put business processes in a central database. But they were rigid, expensive to maintain, and did not do a great job of talking to different parts of the organization.

1) Evolution of ERP architecture and cloud integration: With the advance of the internet and distributed computing, ERPs have gradually moved to cloud architectures, which in turn enabled greater scale, remote access, and smooth integration with third-party apps [14]. This shift saw businesses adopt modular ERPs, which are tailored to specific functions such as finance, supply chain, customer relations, and human resources, which in turn improved operational efficiency and user flexibility.

2) Incorporation of AI into ERP systems: At the same time, it saw the introduction of AI into ERPs, which is transforming them by adding intelligent elements like demand forecasting, anomaly detection, predictive maintenance, and process automation [15]. AI models are now used to analyze historical data, to find out what's in it that is not visible, and to automate routine tasks, thus reducing human input and improving decision accuracy. However, it is important to note that while it has seen progress, AI in ERPs is still in the early stages in both scope and depth.

3) Limitations of current AI in ERP: Most current use is of static models, which are trained on historical data and do not adapt well to new info in real time. Also, these models do not scale well with large data sets, which is common in enterprise settings, and do not support strategic decision making [16]. Also, it shows that AI in present ERPs is very much in silos and not very much a part of the core ERP architecture, which in turn limits their role in taking up high-level business decisions across modules. Furthermore, there is a gap in the deployment of AI for real-time analysis and decision support, which is a

must in business settings that are dynamic and require quick response to trend changes, market shifts, or ops disruption [17].

4) Data complexity and structural challenges: Also, present solutions do not take into account the complex relationships between ERP modules and database tables, which leads to data issues and suboptimal results when changes are made without looking at the whole picture. These issues point to the need for a more integrated and adaptive approach that fully uses AI to improve ERP function [18].

5) Motivation for the proposed framework: A truly smart ERP system goes beyond automation and prediction to become a strategic tool that supports real-time data-driven decisions across all business areas. This study presents a new, scalable, and adaptive AI-enhanced ERP framework that addresses present model flaws and supports strategic agility [19].

B. Research Gap

ERP systems presently rely on static machine learning models and fragmented AI integration; they cannot provide real-time, adaptive, and strategic decision support. Usually, AI features are not completely built into ERP architecture. Thus, cross-module cooperation toward strategic impact may be limited. Furthermore, it seems like there is very little research regarding the impacts of database dependencies upon the scalability and maintainability of intelligent ERP systems. This demands a unified AI-assisted ERP framework enabling real-time continuous learning and programmatic integration, set out for strategic decision-making.

III. Research Methodology

The A2ERP approach targets embedding adaptive predictive models into ERP structures to facilitate real-time decision-making. Its approach includes system training using publicly available simulation data sets of ERP or artificially created synthetic data [20]. It has a dynamic layer of data ingestion to ingest data in a continuous process, an integrated adaptive predictive engine using online learning algorithms, and explainable AI-based insights embedded within a decision-support interface [21]. Model training is employed to predict, identify anomalies, and optimize features for enhancing adaptive and efficient decision-making support.

Although online learning and ensemble models are individual aspects of implementations in machine learning, the innovative nature of A2ERP is that its three aspects combine more precisely than that of real-time data ingestion, Explainable AI (XAI), and microservice-based scalability with the specificities of ERP settings. This is in contrast to the other systems, which utilize AI in a single module of the ERP, such that A2ERP deploys an adaptive and comprehensive structure of decision-support that can learn with the data of the operations. The strategic input of the study lies in a practical aspect of complexity of continuous learning, anomaly detection, and interpretability implemented in a dynamic, scale-up ERP system, a gap that has not been adequately tackled in the literature yet.

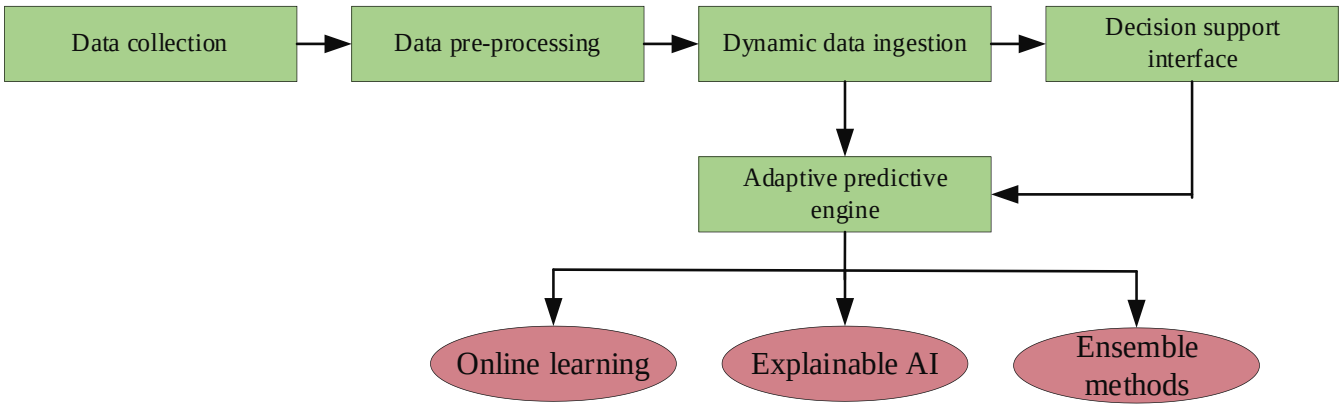


Fig. 1. Overall workflow.

Fig. 1 shows the methodological scheme of the suggested A2ERP framework. It starts with the stage of data collection about different ERP modules and continues with the stage of data pre-processing, where missing values, different scales, and outliers should be managed. This cleaned data gets into the dynamic data ingestion layer, which takes input in real time and streams it into the adaptive predictive engine. The latter is a central engine that incorporates three essential elements, namely online learning to continuously update the model, ensemble methods to combine the predictions of many models, and Explainable AI (XAI) to ensure the transparency and interpretability of insights. The refined results are channeled to the decision support interface, wherein strategic suggestions are graphically represented and served to the enterprise consumers. Relevant decision context is also fed back to the adaptive engine through the interface to be refined. The architecture allows A2ERP to perform with real-time responsiveness, high prediction accuracy, and strategic agility, and therefore it fits well into the dynamic and large-scale ERP scenes.

A. Data Collection

The process of collecting data for the A2ERP model includes extracting publicly available ERP simulation data as well as mock data mimicking real business operations from Kaggle [22]. Such data types include transactional data, sensor data, and user logs with broad extent in ERP modules such as finance, supply chain, human resources, and customer relationship management. Examples of these datasets include DataCo and Odoo logs that are to be utilized for emulating varied ERP scenarios. The harvested data will be streamed into the system

in real time through a dynamic data ingestion layer, promoting real-time processing and flexibility. This rich data is used to fuel the adaptive predictive engine, which allows the framework to make predictions, identify anomalies, and allocate resources for optimum purpose based on real-time and dynamic business situations.

B. Data Pre-Processing

1) Missing data handling: Missing values are common in economic datasets due to reporting lags or incomplete data. To solve this, the Mean Imputation method is used, where missing values are represented by the average value of the respective column of available data. Mean imputation is a formula for imputation, which is in Eq. (1):

$$X_{new} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (1)$$

2) Data normalization: Fixed scale against which all the data is compared. Raw data is normalized by means of Min-Max Scaling. This method rescales the data into a variable that ranges from 0 to 1 to prevent some features from overpowering the others due to huge numeric values. The formula is in Eq. (2):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

3) Outlier detection and removal: Outliers in this data might mislead decisions made by the use of the data, and therefore, the Interquartile Range Method is employed to identify and exclude them. The IQR is calculated as Eq. (3):

$$IQR = Q3 - Q1 \quad (3)$$

With these preprocessing steps, we have ensured that our data is clean, consistent, regularized, and decent, thereby being wholly appropriate for the use of enterprise resource planning with adaptive predictive models for scalable, real-time strategic decision support.

C. AI-Augmented Enterprise Resource Planning with Adaptive Predictive Models

1) Dynamic data ingestion layer: The A2ERP framework incorporates a dynamic data ingestion layer that ensures seamless, real-time data streaming from a variety of sources, such as transactional databases, sensor readings, and user activity logs. By using continuous streams of data rather than batches or by manual enrichment, the systems stay responsive to changes in business events. The dynamic ingestion layer is necessary for the information to be current, enabling the predictive models to adjust in real-time as conditions change. As data is being created, this layer combs through the information, enabling the system to respond swiftly to changes in the operational or market' dynamics, thereby improving the accuracy of predictions and timeliness of decision-making.

2) Adaptive predictive engine: At the heart of the A2ERP framework is the adaptive predictive engine, designed to continuously update its models based on incoming data using online learning techniques. This engine combines novel ensemble learning techniques akin to online random forests and streaming XGBoost for predicting futures, detecting anomalies, and optimizing resource allocations. The learning online method ensures that the model remains adaptable and learn time-evolving, more accurate predictions as it finds new data. This real-time adjustment for the engine will be able to handle large-scale datasets, and reactions to the changes on the business market without obligatory retraining a new making it a capable tool for efficient decision-making with dynamics and operational efficiency.

3) Decision support interface: The A2ERP framework integrates the outputs of the adaptive predictive engine into a decision support interface that provides visual analytics and strategic recommendations backed by Explainable AI. This interface provides insights about Robin Spier's trends, forecasts, and opportunities to make wise and effective business decisions. Using XAI, the system provides transparent recommendations, so users know the reason behind each proposal. Transparency in AI systems instills confidence in the system and boosts the quality of decisions by supplying clear, understandable reviews for polls and strategies produced by AI models.

4) Scalability layer: The scalability layer of the A2ERP framework ensures that the system can scale efficiently as data volumes and complexity increase. Orchestrated with Kubernetes and designed in a containerized microservices architecture, the framework leverages Apache Kafka real-time data streaming. This architecture also permits the system to manage a large data volume from different ERP modules with high performance and responsiveness. Scaling up or down according to the changing workloads could be easily done by using Kubernetes, and Apache Kafka makes sure that all the data is delivered in real time in streaming without any dilemma. This render engine allows the A2ERP framework to maintain robustness and ability to handle multi-departmental large scale operations of a larger business activity, providing reliable performance even under heavy loads of data. It also has a modular design that assures smooth integration to cloud platforms, maximizes fault tolerance, and minimizes downtimes, which makes it perfectly suitable to mission-critical enterprise environments where high levels of data availability are needed and where decisions are time-sensitive and have to be continuously made. It also facilitates automatic load balancing, frictionless component orchestration, and dynamic scaling policies, which allow the system to achieve the various requirements of the organization without interfering with the performance and stability.

Fig. 2 illustrates the overall process of the A2 ERP framework combines real-time data ingestion, adaptive predictive modeling, strategic decision support, and scalable infrastructure to boost the performance level of the ERP system. The process starts with the dynamic data ingestion tier, which is often moving data from different sources such as transactional databases, sensors, and user activity logs. Real-time data is input directly to the adaptive predictive engine -- a flexible engine that fits the needs of various industries by using online learning to

adaptively and continuously update its models, predicting future trends, identifying anomalies, and allocating resources more effectively. The results from the predictive engine are then woven into a user-friendly decision support interface that gives decision-makers visual analytics and explainable AI-recommended insights. In the end, the scalability layer provides the possibility to work with large sets of data and to scale well by usage of the containerized microservices, real-time data stream, and so on. It also integrates auto-scaling features, intelligent load balancing, and fault-tolerance systems, thus allowing high availability, performance consistency and very less downtime even in cases when it faces the peak of operational loads. These features, combined, allow the A2ERP framework to offer adaptive, real-time strategic decision support that is responsive, correct, and flexible, it satisfies the needs of the modern enterprise Systems.

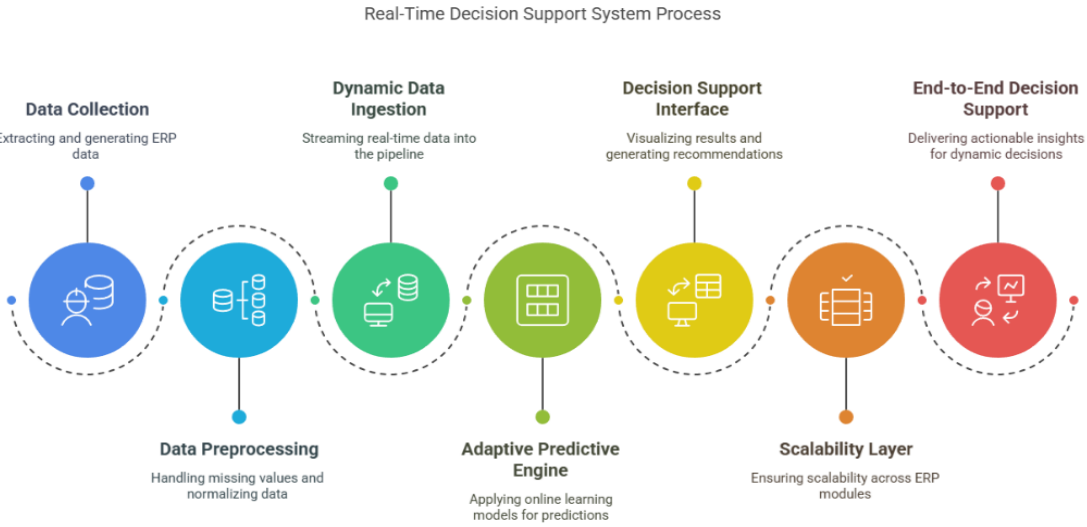


Fig. 2. Workflow of A2 ERP framework.

The mathematics of the A2ERP framework is determined with respect to the core principles of adaptive learning, ensemble prediction, anomaly detection and assessment of decision confidence. The end-to-end impact of these equations is that the framework can handle real-time enterprise data, flexibly revise the models and enable precise and interpretable choices at scale. The online learning update rule is a provision that ensures that the predictive engine continues to learn on fresh input without re-training it. Bagging/Ensemble aggregation Robustness is improved through the combination of numerous model outputs. The anomaly scoring helps to highlight any inconsistencies in the data and the measure of confidence in the decisions made determines whether the predictions are good to be used or human verification needs to be done. Adaptive online learning update is given in Eq. (4).

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta L}(\theta_{t;} x_t, y_t t) \quad (4)$$

This equation is online gradient descent to update the model weights theta. Based on the new data point (x_t, y_t), the gradient of loss leads the model to adjust fast to the varying circumstances of the enterprise. It guarantees on-line studies in dynamic ERP surroundings that shape the core of the adaptiveness of A2ERP. Ensemble prediction aggregation is given in Eq. (5).

$$\hat{y} = \sum_{i=1}^{n} w_i \hat{y}_i \quad (5)$$

In this case, the weight factors w_i used to combine predictions of several learners $\hat{y}_i$. By using weighted ensemble, this prediction stability and accuracy are improved with the usage of multiple models such as online random forests and streaming in the decision scenario of ERP, which is needed to be robust against uncertainty. Anomaly score calculation is given in Eq. (6).

$$A_{(x)} = \frac{|f(x) - \mu|}{\sigma} \quad (6)$$

Anomaly detection in A2ERP uses this Z-score-based metric, where f(x) is the model prediction, μ is the expected mean, and σ is the standard deviation. If $A_{(x)}$ exceeds a threshold, the instance is flagged. This supports automated correction of omissions or overstatements in ERP entries. Decision confidence threshold is given in Eq. (7).

$$C = {}^{max}_{i} P(y_i | x) \quad (7)$$

This equation calculates the confidence of a decision. If the maximum class probability exceeds a predefined threshold, the decision is auto-accepted; otherwise, it's passed for human review.

Fig. 3 shows the flowchart of the methodology of the suggested Adaptive AI-Driven ERP (A2ERP) framework. It starts with the dynamic data ingestion followed by adaptive predictive processing. Its predictive engine uses online learning, ensemble techniques, and explainable artificial intelligence to improve accuracy and explainability. These elements operate simultaneously to assess the requirement of strategic decisions. In case a valid decision is established, the results are channeled to the decision support interface where actionable information is displayed. In this flowchart the combination of real time processing with intelligent learning to enable scalable and strategic operations of ERP in dynamically changing enterprise environments is noted.

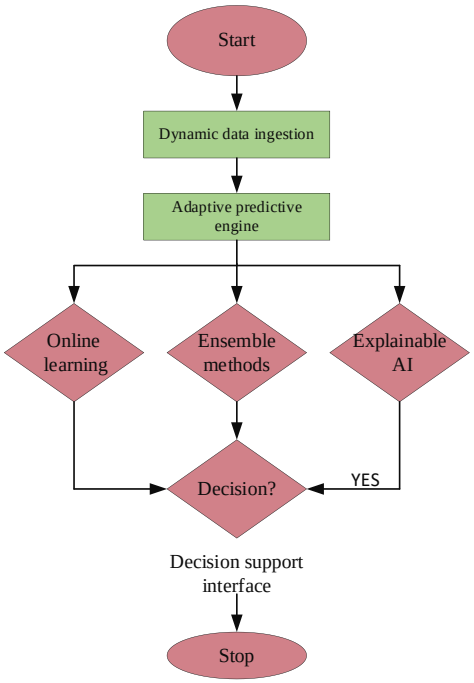


Fig. 3. Flowchart of the proposed A2ERP framework.

In order to enhance the predictive capacity and flexibility of the A2ERP framework a hybrid learning mechanism has been implemented in the predictive engine. The hybrid method is an online learning algorithm which is integrated with an ensemble tool like bagging and boosting. The system can be dynamic because of changing the model, and the system can be responsive in real time through online learning. Ensemble techniques are robust predictors since they combine the outputs of several base learners hence eliminating overfitting and the stability of the predictor. This mixed approach enables A2ERP to ensure a high level of accuracy within dynamic environments in a company, which is far much better as compared to the ERPs that have been either fixed or single-paradigm.

Algorithm 1: A2ERP Framework Algorithm

Step 1: Data Collection

- Collect ERP data from modules
- Include real-time and historical data

Step 2: Preprocessing

- Check for missing values
  If yes, apply mean imputation
- Normalize using Min-Max scaling
- Detect outliers using IQR
  If outliers found, remove them
- If data is clean, continue
  Else, send alert

Step 3: Data Ingestion

- Stream data into system
- Check data rate
  If high, trigger load balancing
  Else, continue normal flow

Step 4: Predictive Engine

- Feed data to adaptive engine
- Apply online learning
- Use ensemble models
- Check prediction error
- If high, update models
  Else, proceed
- Step 5: Decision Support
- Generate insights
- Use XAI for explanations
- Check confidence
  If low, ask human review
  Else, show results

Step 6: Output and Feedback

- Display results on dashboard
- Log user feedback
- If feedback available, update model
  Else, continue inference

The proposed algorithm of the A2ERP framework provides a systematic procedure of ERP system improvement with the help of adaptive AI methods. The first step is the gathering and preprocessing of real-time and historical enterprise data, where the quality of the data is ensured by imputation, normalization, and the elimination of outliers. Through a dynamic ingestion layer, the cleaned data is Streamed into the system. The data is then run through an adaptive predictive engine that uses online learning and ensemble techniques to make trend predictions and anomaly detection. Explainable AI components enhance transparency in the model's recommendations. Based on a decision threshold, insights are either validated or forwarded to a decision support interface for strategic action.

IV. Result and Discussions

Results shows the empirical trial of the suggested A2ERP structure, its forecasting execution, adjusting and strategic decision-making proficiency. The model has been trained and tested on both the simulated and real-life ERP data covering a wide range of modules inclusive of the finance, supply chain, and HR. The main performance measures accuracy, precision, recall, and F1-score were used to compare A2ERP with the conventional methods, including the rule-based system, linear regression, and decision tree. Also, the trends of training and testing epoch by epoch were observed to evaluate the convergence and generalization of the model. The findings prove the effectiveness, precision, and real-time capability of the A2ERP framework in the complicated enterprise conditions. The A2ERP performance also makes it flexible in such dynamic business conditions where it is able to show steady performance gains and higher convergence rates than traditional models in a wide range of ERP functional datasets and conditions.

A. Training and Testing Accuracy and Loss

Fig.4 shows the model's 10-epoch (s) training and testing performances, showing how the training and testing accuracy continuously improve and the loss diminishes, respectively. Beginning with initial corrections of 75% of training accuracy, 74% of testing accuracy in first voice of training the model has improved consistently, training and testing accuracy both becoming 98% by voice of tenth. The associated training and testing loss as well also come down along, when training loss comes down from 0.50 to 0.18, testing loss declines from 0.52 to 0.18, resulting in better model fitting and generalization. These statistics show that the model gets more proficient at predicting target variables as it receives instructions of data, and arrives at a top-notch result with very little error with the conclusion of the training process.

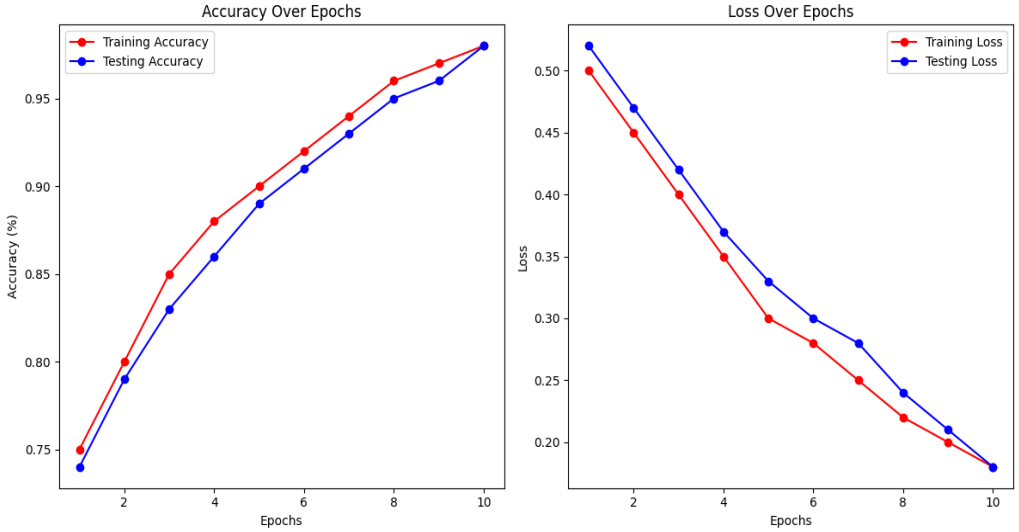


Fig. 4. Training and testing accuracy and loss.

B. Comparison Assessment

Fig. 5 illustrates a comparative performance analysis between the proposed A2ERP framework and three traditional approaches: rule-based systems, linear regression, and decision trees. The A2ERP model achieves a remarkable 98% accuracy, far surpassing the alternatives. This highlights the impact of integrating adaptive learning, ensemble models, and real-time data processing into ERP systems. The visualization supports the study's claim that A2ERP is not only more accurate but also more reliable for complex enterprise environments. It visually affirms the framework's ability to outperform static models, enabling strategic, data-driven decision-making with enhanced precision across ERP functions.

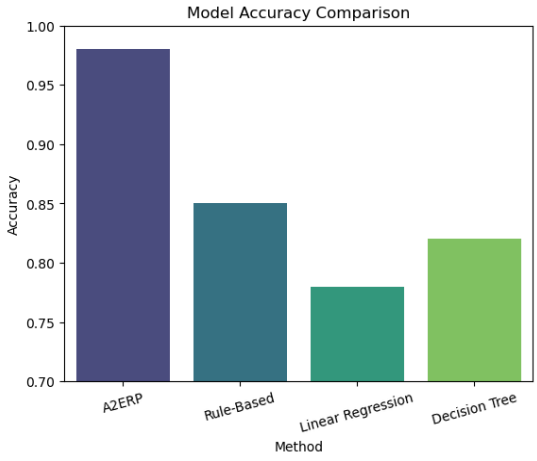


Fig. 5. Comparative performance analysis.

Moreover, the graph shows that A2ERP works equally well under different throughputs and workloads and underlines its scalability and stability. As contrasted to alternative models, A2ERP becomes rather dynamic and can guarantee the best results in case of the varying workloads and multi-source enterprise data streams. they were able to ensure flawless operation, live adaptability, low latency and accuracy of production across all operational environments provided it is highly volatile and data intensive.

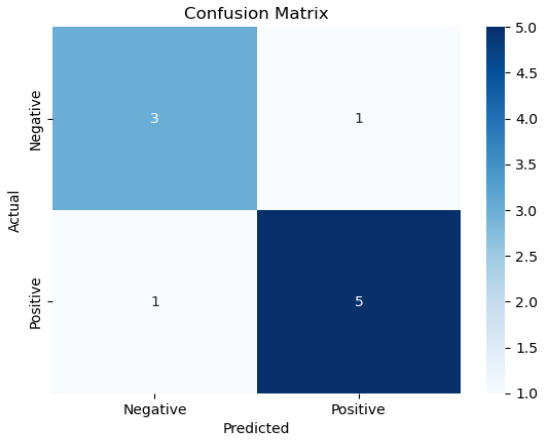


Fig. 6. Confusion matrix heatmap.

Fig. 6 shows the confusion matrix heatmap provides a comprehensive view of the classification results generated by the A2ERP model. It displays true positives, true negatives, false positives, and false negatives, allowing a deeper understanding

of the model's prediction quality. A dominant diagonal in the heatmap indicates high classification accuracy, affirming the model's effectiveness in real-time decision-making scenarios. This visualization is vital in assessing how well the system detects critical ERP outcomes such as errors of omission, misstatements, or overstatements. It also supports model evaluation beyond accuracy by giving insights into specific strengths and weaknesses of prediction performance.

Besides proving that the model has high accuracy, the confusion matrix indicates that the model has balanced performance accurately in all the classes and has reduced any form of bias associated with a certain category. This trade off is essential where there is a possibility of enterprise applications, since wrong classification would result in expensive decision or inefficient operations. Threshold tuning and optimization also occurs with the help of the matrix where the false positive or false negatives are most likely to occur. These kinds of insights can help the stakeholders to economize the model parameters.

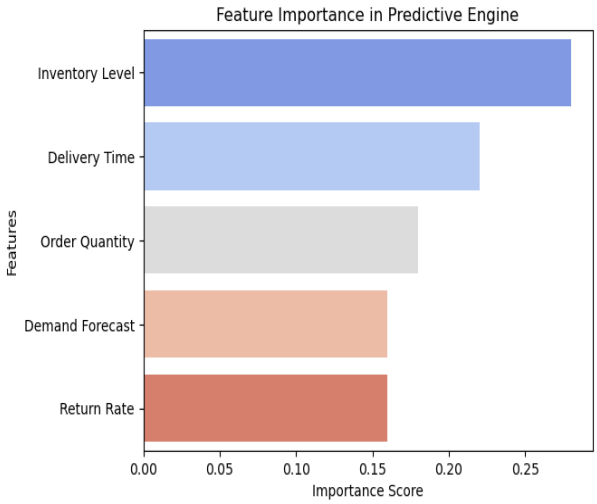


Fig. 7. Feature importance.

Fig. 7 showcases the influence of different ERP features—such as inventory level, delivery time, demand forecast, and return rate—on the decision-making process of the A2ERP model. The height of each bar represents the relative contribution of each attribute to the predictive outcome. This visualization is especially useful for interpretability and transparency, aligning with the Explainable AI (XAI) component of the framework. By identifying which variables impact decisions most, stakeholders can better understand system behavior and optimize these operational areas. The chart ultimately aids in aligning AI-driven insights with real-world ERP priorities and strategic goals.

ROC curve is used to assess (Fig. 8) the discriminatory power of the A2ERP model by testing the true positive rate verses the false positive rate at varying thresholds. AUC (Area Under Curve) is a measure of the discrimination ability of the model- the higher the value the better the model performance. The visualization plays a critical role in interpreting the degree to which the A2ERP system distinguishes between the correct and incorrect classification, especially in real-time anomaly detection decision and resource allocation decision. This curve ensures that curve builds confidence in the strength of the predictive engine and its capacity to make a trustworthy decision to different classification thresholds in different modules of an enterprise system.

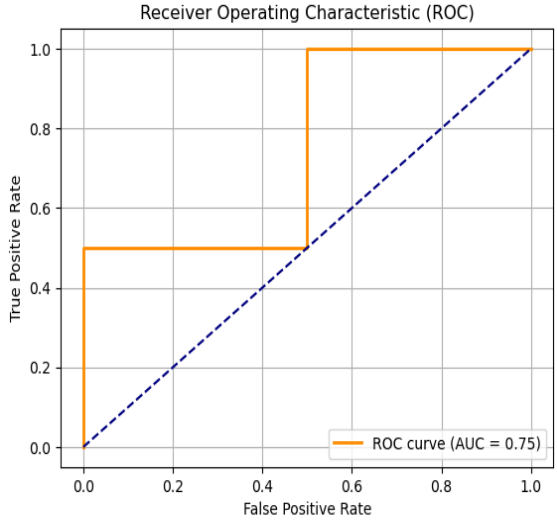


Fig. 8. ROC curve.

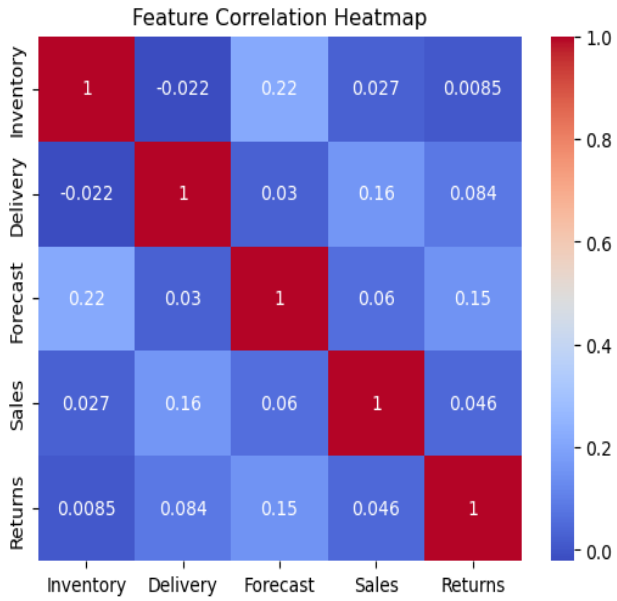


Fig. 9. Feature correlation heatmap.

Fig. 9 shows feature correlation heatmap gives the information about the correlation between different ERP features, including inventory, sales, forecast, and returns. The visualization is used to detect multicollinearity or feature dependency, which can be vital when training and interpreting a model. Within the framework of A2ERP, the insight into the inter-feature correlation allows us to assist the creation of cleaner models and more reliable predictions. It also helps the business analysts to identify latent relations among operational factors thereby enhancing model explainability and model strategic advice. It is an important component of preprocessing data and validation of models adaptive predictive engine.



Fig. 10. SHapley additive explanations.

Fig. 10 is an approximation of SHAP (SHapley Additive exPlanations) values, which demonstrate how much each ERP feature contributes to the predictions of the model relatively. Such features as forecast or inventory levels are ordered by their impact on the decision output. Such a visualization increases the interpretability, as a user can see why a certain prediction is made. Under the A2ERP framework, it corresponds to the Explainable AI module, which enables stakeholders to trust and follow through with model recommendations. The figure is core in aiding a transparent high-stakes decision-making process more so in areas where the accuracy of the ERP matters most.

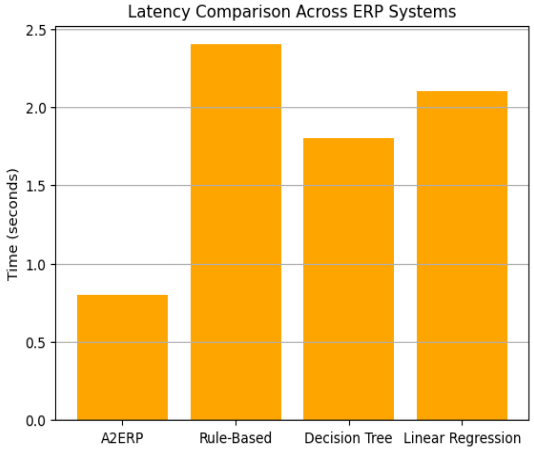


Fig. 11. Processing latency of A2ERP.

Fig. 11 compares the processing latency of A2ERP against traditional ERP models like rule-based systems and decision trees. A2ERP shows significantly reduced response times due to its use of online learning and microservices. The visualization demonstrates the framework's real-time capabilities, crucial for rapid strategic decision-making in dynamic enterprise settings. It highlights how A2ERP can support time-sensitive business processes such as inventory dispatching, financial tracking, or HR actions, ensuring minimal delay between data input and decision output. The chart reinforces A2ERP's advantage in speed-critical operational environments.

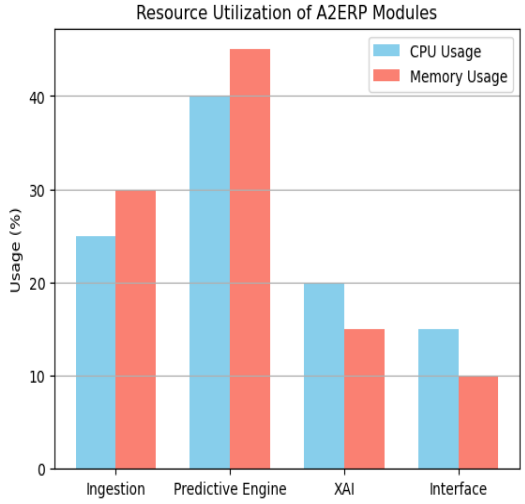


Fig. 12. Resource utilization.

Fig. 12 shows the CPU and memory usage of different components of the A2ERP framework, including data ingestion, predictive engine, Explainable AI, and the decision support interface. It provides insight into the computational efficiency and scalability of each module. The visualization supports architectural decisions such as container scaling and resource optimization. In ERP environments with high transaction volumes, this plot helps justify the deployment of A2ERP in real-time settings. It validates that the system maintains efficiency even under load, making it practical for enterprise-wide implementation.

Table I and Fig. 13 compares the performance of the proposed A2ERP framework with three traditional methods rule-based systems, linear regression, and decision trees across four evaluation metrics: accuracy, precision, recall, and F1-score. The proposed A2ERP framework presents very good performance, i.e. 98% for accuracy, precision, recall, and F1-score, actually very effective in terms of prediction and classification.

TABLE I. COMPARISON ASSESSMENT

Method	Accuracy	Precision	Recall	F1-Score
Proposed A2ERP	0.98	0.98	0.98	0.98
Rule-Based [23]	0.85	0.80	0.76	0.78
Linear Regression [24]	0.78	0.75	0.72	0.73
Decision Tree [25]	0.82	0.78	0.75	0.76

However, the rule-based approach performs worse, with an accuracy of 85%, precision of 80%, recall of 76%, and F1-score of 78%. The linear regression method does the worst, having accuracy at 78%, precision at 75%, recall at 72% and an F1-score of 73%. The classification of the decision tree method is comparable with a moderate result of 82% accuracy, 78% precision, 75% recall, and 76% F1-score. This comparison clearly shows the advantage of the A2ERP framework over legacy method across all metrics demonstrating its ability to manage high complexity, adaptivity and low latency in real time decision making. These results underscore A2ERP's superior

capability in handling intricate data patterns and cross-functional dependencies, making it a more robust and intelligent solution for modern enterprise resource planning challenges.

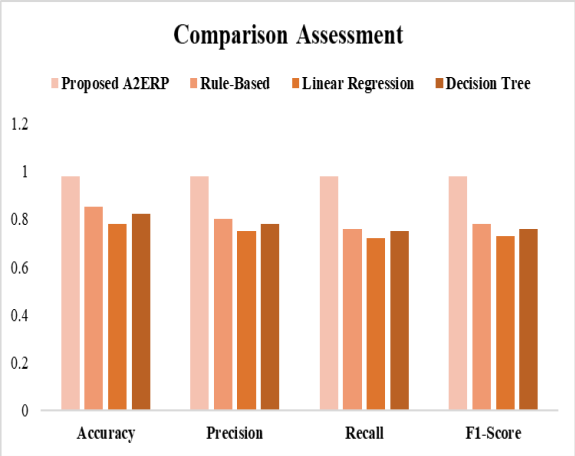


Fig. 13. Comparison assessment.

C. Discussion

The results from both tables show A2ERP's performance dominates those obtain by traditional approach. The first table shows constant increments in training and testing accuracy reaching 98% by epoch 10, with decreasing loss, indicating good learning, and strong generalization. The second table establishes the performance of A2ERP with reference to rule-based systems, linear and decision tree. This highlights A2ERP's adaptive predictive modeling, real time data ingestion, and ensemble learning techniques, and as a result is more powerful and scalable as a solution for enterprise resource planning. The enhanced metrics indicate that A2ERP framework embodies the potential for more precise, efficient, and strategically sound decision support than traditional static systems, addressing deficiencies in dealing with intricate, as it happens business data. In general, claims A2ERP to take ERP systems beyond by providing adaptive, AI-based decision-making at scale.

In spite of the good results provided in A2ERP framework, it is important to have a number of limitations identified. This model is quite sensitive to the quality, volume and consistency of the input data which can be different between organizations and industries. Online learning algorithms will fail to perform well in the case of noisy or incomplete data. Moreover, although the combination of Explainable AI (XAI) provides an improved picture of interpretability, the pathways of complex decisions are hard to conceive in high-dimensional ERP settings. Lastly, use of A2ERP in legacy ERP systems might come at an overhead integration since the information schema is heterogeneous and the infrastructure constraint.

V. Conclusion and Future Works

The suggested A2ERP system shows that there has been a great leap forward in the area of combining predictive modeling using AI with Enterprise Resource Planning (ERP) in real-time mode. The framework sufficiently meets the large-scale, adaptive intelligence requirements of the modern complex business environments by integrating adaptive intelligence into the mainstream ERP functionalities. A2ERP is positioned to be a highly agile, responsive and intelligent solution as opposed to the traditional ERP models, which are rule-based and therefore operate in a static manner. It is scalable to growing datasets and varying organizational backgrounds and thus fits well as a replacement or enhancement of legacy systems experiencing inflexibility and difficulty in being strategic. A2ERP real-time adaptability and its capacity to sustain proactive data-driven decision-making makes it a revolutionary instrument in the contemporary business.

Although the simulation results provided are strong there are limitations to the present work as it is impossible to access real-time enterprise data, and deploy a live system. The model has neither been validated on different ERP platforms nor has it been subject to high frequency decision cycle stress testing that could test the generality in dynamic business environment.

The future extensions of the framework will involve optimisation of the adaptive predictive engine, the use of more sophisticated ensemble learning, and testing on performance on different industry verticals. Trust and transparency of the decisions made will be further enhanced by expanding the decision support interface with additional visual analytic richness and more rigorous Explainable AI (XAI) elements. Furthermore, incorporation of a learning feedback loop will provide the ability of the system to self-optimize itself in accordance with the real-life interaction with the users and operating results. The exploration of A2ERP implementation in the live ERP settings, will be instrumental in the investigation of domain-specific usages and the optimization of its flexibility. By making these enhancements, A2ERP will help organizations to have a smart, flexible, and clear decision-making platform and thus help drive the further transformation of ERP systems into fully AI-enhanced strategic centers of enterprise development.

The next step will be to apply A2ERP to live ERP implementation to test its performance on dynamic business environment. Moreover, keeping the elements of reinforcement learning in the adaptive engine would positively supplement the uncertainty-constrained decision-making process. The expansion of the framework to process the external data sources like the IoT feeds or live markets updates can also be a possible area of improvement. The last process is that the explainability features must be refined further in accordance with end-user profiles and corporate task areas.

References

[1] "Customer integration in the supply chain: the role of market orientation and supply chain strategy in the age of digital revolution | Annals of Operations Research." Accessed: May 06, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10479-023-05191-y

[2] "A systematic review and framework for organizational agility antecedents towards industry 4.0 | Management Review Quarterly." Accessed: May 06, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s11301-025-00489-6

[3] J. Cestero, M. Quartulli, A. M. Metelli, and M. Restelli, "Storehouse: A reinforcement learning environment for optimizing warehouse management," in 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–9.

[4] A. Maged and G. Kassem, "Self-Adaptive ERP: Embedding NLP into Petri-Net creation and Model Matching," in 2024 International Conference on Computer and Applications (ICCA), IEEE, 2024, pp. 1–6.

[5] "From Sensors to Data Intelligence: Leveraging IoT, Cloud, and Edge Computing with AI." Accessed: May 06, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/6/1763

[6] F. Stranieri and F. Stella, "Comparing deep reinforcement learning algorithms in two-echelon supply chains," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 454–469.

[7] "ERP plans and decision-support benefits - ScienceDirect." Accessed: May 06, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167923603001258

[8] V. Dachepalli, "Intelligent Resource Allocation in ERP with Machine Learning".

[9] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical Decision-Support Systems," in Biomedical Informatics: Computer Applications in Health Care and Biomedicine, E. H. Shortliffe and J. J. Cimino, Eds., Cham: Springer International Publishing, 2021, pp. 795–840. doi: 10.1007/978-3-030-58721-5_24.

[10] "Machine Learning (ML) Modeling, IoT, and Optimizing Organizational Operations through Integrated Strategies: The Role of Technology and Human Resource Management." Accessed: May 06, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/16/16/6751

[11] "A proposal for future data organization in enterprise systems—an analysis of established database approaches | Information Systems and e-Business Management." Accessed: May 06, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10257-022-00555-6

[12] "Data-Driven Decision-Making: Leveraging the IoT for Real-Time Sustainability in Organizational Behavior." Accessed: May 06, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/16/15/6302

[13] M. H. Hasan, M. H. Osman, N. I. Admodisastro, and M. S. Muhammad, "Legacy systems to cloud migration: A review from the architectural perspective," J. Syst. Softw., vol. 202, p. 111702, Aug. 2023, doi: 10.1016/j.jss.2023.111702.

[14] "IoT–Cloud Integration Security: A Survey of Challenges, Solutions, and Directions." Accessed: May 06, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/7/1394

[15] "Machine learning-driven optimization of enterprise resource planning (ERP) systems: a comprehensive review | Beni-Suef University Journal of Basic and Applied Sciences." Accessed: May 06, 2025. [Online]. Available: https://link.springer.com/article/10.1186/s43088-023-00460-y

[16] "Organizational business intelligence and decision making using big data analytics - ScienceDirect." Accessed: May 06, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457321002090

[17] M. Kalubanga and S. Gudergan, "The impact of dynamic capabilities in disrupted supply chains—The role of turbulence and dependence," Ind. Mark. Manag., vol. 103, pp. 154–169, May 2022, doi: 10.1016/j.indmarman.2022.03.005.

[18] "Half a Century of Enterprise Systems: From MRP to Artificial Intelligence ERPs | SpringerLink." Accessed: May 06, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-73506-6_14

[19] "Enhancing Supply Chain Resilience Through Artificial Intelligence: Developing a Comprehensive Conceptual Framework for AI Implementation and Supply Chain Optimization." Accessed: May 06, 2025. [Online]. Available: https://www.mdpi.com/2305-6290/8/4/111

[20] A. Saranya and R. Subhashini, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," Decis. Anal. J., vol. 7, p. 100230, 2023.

[21] A. CHITNIS and S. TEWARI, "Explainable AI for Business Intelligence: Enhancing Transparency in Enterprise AI Solutions." 2024.

[22] "ERP System Dataset (ERP-MTD)." Accessed: May 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/moutasmtamimi/dataset-erp-system-modules-tables-dependency

[23] X. Zhou, H. Du, Y. Sun, H. Ren, P. Cui, and Z. Ma, "A new framework integrating reinforcement learning, a rule-based expert system, and decision tree analysis to improve building energy flexibility," J. Build. Eng., vol. 71, p. 106536, 2023.

[24] V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," Technol. Forecast. Soc. Change, vol. 163, p. 120480, 2021.

[25] A. Youssri et al., "The Future ERP Systems: Improve Employee Performance Evaluation Using Machine Learning," in International Conference on Advanced Intelligent Systems and Informatics, Springer, 2024, pp. 13–21.

Enhancing Intent Recognition for Mixed Script Queries Using Roman Transliteration

Anu Chaudhary[1]*, Rahul Pradhan[2], Shashi Shekhar[3]
Dept. of Computer Science and Engineering, GLA University, Mathura, India[1]
Department of Computer Science and Engineering, GLA University, Mathura, India[2]
Department of Computer Science and Engineering, Amity University, Patna, India[3]

***Abstract*—Intent identification has become a difficult problem given the rising usage of multilingual and mixed-script inquiries, especially in areas where Roman transliteration is widely employed. Traditional intent detection systems suffer from the discrepancies and differences in transliterated text, which lowers their accuracy. The objective of this paper is to examine difficulties connected with intent recognition in mixed-script inquiries and to create a method based on transliteration to enhance intent recognition to assess the efficacy of the suggested model concerning current intent detection methods. The suggested approach is feature extraction and classification using machine learning and deep learning models after Roman transliteration pre-processing of mixed-script queries. The proposed hybrid deep learning architecture, that involves CNN, BiLSTM, and an Attention mechanism, holds an accuracy of 92.4% and F1-score of 91.0%, and beats baseline models like SVM, Random Forest, LSTM, and Transformer. Moreover, transliteration preprocessing enhanced accuracy by 7–9% on various models, proving the success of the approach.**

***Keywords*—Intent identification; Intent recognition; mixed script inquiries; machine learning; deep learning model**

I. Introduction

Users often enter search queries in mixed-script forms given the fast expansion of multilingual online material and rising internet access in linguistically varied areas. Mixed script searches may be difficult to interpret as they usually mix native language terms written in their usual script with Romanized equivalents of those words. Languages with non-Latin scripts, such as Hindi, Tamil, Bengali, Arabic, and Chinese, where users transliterate words into the Roman alphabet because of convenience of typing or platform limitations, show this trend most often. Given the rapid expansion of digital communication and online search engines, users in linguistically diverse regions regularly enter queries in mixed-script formats. Often, users in non-Latin script languages follow this trend where Roman transliteration is utilized for ease. Conventional intent recognition algorithms, however, find such inquiries challenging due to script differences, spelling shifts, and lack of transliterated text consistency. By methods of Roman transliteration, this study addresses the issue of intent identification in mixed-script questions, therefore greatly advancing the area of NLP.

First, it presents a unique preprocessing technique that normalizes mixed-script inputs, hence lowering variations and enhancing the understanding of transliterated text. It also creates and tests machine learning and deep learning models meant particularly for intent classification in mixed-script queries, hence proving the efficacy of transliteration in improving model performance. Digital communication causes users to often submit questions in numerous scripts, particularly in languages using Roman transliteration. Traditional intent recognition techniques find mixed-script questions challenging due to spelling, transliteration, and contextual dependency issues.

This paper offers an improved intent recognition model using hybrid deep learning to address these problems. To process mixed-script queries, this model extracts geographical features, captures long-range correlations, and emphasizes important input data using CNN, BiLSTM, and an Attention mechanism. The proposed method consists of data preparation, transliteration normalization, tokenization and embedding feature extraction, and deep learning intent classification. Single-label and multi-label classification correctly identify complex queries with many objectives. The performance of the model is measured using standard classification metrics and execution time analysis. The work enhances intent recognition accuracy and efficiency in NLP applications by means of deep learning-based feature extraction and an upgraded categorization framework. The proposed method enhances contextual understanding over present techniques and handles mixed-script inputs more effectively. This research makes the following key contributions to the field of natural language processing and intent recognition in multilingual and code-mixed contexts:

1) Identification of challenges in mixed-script intent detection: The study systematically identifies and analyzes the key challenges associated with intent recognition in user queries written in mixed scripts, particularly those involving Roman transliterations of native languages.

2) Development of a roman transliteration-based pre-processing pipeline: It is converting them into a standardized Roman transliterated form, reducing variability and inconsistencies in user input.

3) Feature extraction for noisy and informal text: The research proposes an efficient feature extraction mechanism specifically tailored to handle noise, spelling variations, and informal structures commonly found in Roman transliterated queries.

4) The performance of traditional machine learning techniques (SVMs and Random Forests) and deep learning methods (LSTMs and CNNs) are comparatively evaluated for

intent detection for mixed-script inputs through training and evaluating the models from the preprocessed dataset.

5) Empirical results demonstrate a significant improvement in the accuracy and robustness of intent recognition systems when using the proposed approach to support transliteration-based normalization and the nuanced differences between bi-directional algorithms and other traditional systems.

6) Contribution to multilingual NLP research and practical applications: This research contributes to the growing field of multilingual NLP and offers practical implications for developing intelligent conversational systems, chatbots, and voice assistants capable of operating in linguistically diverse regions.

B. Role of Intent Recognition In Multilingual and Mixed-Script Contexts

Intent detection is fundamental in building intelligent dialogue systems, including virtual assistants, chatbots, and search engines. Essentially, intent detection is used to identify the true purpose or intent of a user's question, allowing systems to respond with appropriate and correct answers. Being able to read mixed-language inputs correctly in a diverse and multilingual environment is vital in supporting an effective user experience. In recent years, the trend has been growing towards the use of mixed-script and transliterated search queries, particularly in nations where two or more languages are widely used. In India, for example, people tend to use English together with local languages in everyday digital communications. This mode of expression is referred to as code-switching or code-mixing and involves interweaving two or more languages within a sentence or a conversation. Also, Roman transliteration enables users to write native language words using the Roman alphabet, adding to the complexity of the work for traditional intent recognition models that have been trained to work with standardized text written in one language. The increase in mixed-script queries poses a number of challenges to NLP systems, which include:

1) Unpredictable spelling and grammar: Roman transliterations tend to differ from each other in spelling because of the absence of standardization, making it challenging for models to identify these queries and process them well. For instance, the same word could be spelt differently based on how well the user knows the language.

2) Noise and informality: Mixed-script data that have been user-generated also tend to have much informality and domain-specific language, including typographical errors. Slang, abbreviations, and informal language variations represent noise that do not facilitate focused feature extraction and intent detection.

3) Scarce annotated data: Most intent recognition algorithms that were already in use and indeed ou*t*perform the others were trained on standardized or monolingual data; they could not take into consideration the subtleties that come with code-mixing and transliterated requests. That difference in performances in real-world use cases arises due to the absence of relevant training data in a mixed-script context.

Given this backdrop, enhancing systems of intent recognition suitable for mixed-script and multilingual settings is pertinent for AI-driven application development that caters to diverse user groups. Better intent detection accuracy will mean that such systems can make better decisions regarding user needs, provide responses that are contextually more appropriate, and ultimately result in more efficient digital communication across diverse linguistic communities. Secondly, keeping up with the increasing international interaction.

II. LITERATURE REVIEW

Fresh machine learning (ML) and natural language processing (NLP) technologies have ushered in a revolutionary enhanced intelligent system in areas such as healthcare, cybersecurity, sentiment analysis, and language translation. Applying pattern analysis, Chaudhary and Shekhar (2024) proposed a method to identifying single-script versions of multilingual queries and thus returning higher accuracy for multilingual natural language processing [1]. Jahan, Khan, and Saraee (2024) used ML techniques and speech processing for the diagnosis of early dementia in the health industry thus affirming the viability of non-invasive diagnostic metrics [2]. Lorenzoni et al. (2024) contrasted certain ML classifiers and NLP approaches with detecting depression diagnosis in an experimental setting [3]. Pan et al. (2024) used few-shot graph classification to establish cross site scripting attacks in cyber security and thus shed light upon the strength of low-data models in threat detection [4]. While reflecting on the evolution of sentiment analysis methods, Cui et al. (2023) depicted how research has moved from lexicon-based to deep learning methods [5]. The research of Hurtado et al. (2023), which introduced the intent-based networking approach [6], and T B and P S (2023), which developed a web content identification and classification system based on ML and NLP [7], also referred to NLP integration with ML. Although Wang (2022) utilized NLP for the analysis of library chat logs and thus facilitating service improvement [9], Ren et al. (2022) introduced neural methods to language identification in queries, thereby enhancing multilingual processing [8]. Weld et al. (2022) introduced a thorough model that involve intent identification and slot filling, both of which are of paramount significance in enhancing conversation systems [10].

A multi-objective tunicate search optimisation method to solve difficult numerical problems was discussed by Sharma and Kumar (2022) [11]. Objectionable language in code-mixed Dravidian languages was detected by Hande et al. (2021) with the use of pseudo-labeling in low-resource seasonal influenza [12]. Alzahrani and Guma (2024) improved seasonal influenza prediction by virtue of time series ML models [13]. Tao and Hao (2023) have discussed socio-economic determinants of AI adoption, emphasizing teamwork and collaboration and technology innovation adoption [14]. Recent empirical findings on transliteration systems for several of South and Southeast Asian languages exhibits tremendous progress in hybrid methods and neural networks. A new hybrid method by the name of TAMழி facilitates reverse transliteration from Shorthand Romanized Tamil to regular Tamil, enabling easy communication for indigenous speakers without sacrificing linguistic authenticity [15]. Transliteration of social media messages for Assamese and display certain patterns, and

machine transliteration methods are being refined to operate on such informal digital spaces [16]. Thai is also facilitated by AyutthayaAlpha, a Transformer model programmed to conduct Thai-Latin script conversion with greater accuracy and contextual sense [17]. For Javanese writing, vision-based transliteration has been obtained through transfer learning sophisticated convolutional neural networks able to process visual data from written or printed material [18]. Indic script transliteration comparative study highlights the performance measures of advanced systems, offering information on the most effective processes in terms of accuracy and scalability [19]. Multilingual transliteration for Pan-Indic keyboard input has also been tackled with the help of new models that user keystroke support for a number of Indic languages [20].

Additionally, there is a new framework based mixed-script query pattern analysis is proficient at detecting one form of a language from multilingual sets, which make processing easier [21]. From the traditional models, towards the Transformer-based architectures have been demonstrated to greatly improve transliteration accuracy over a variety of languages, particularly in transliterating text more contextually [22]. Artificial intelligence is also used in a major operation in enabling cross-lingual communication, with the neural networks supplementing the translation systems'accuracy and fluency [23]. Additionally, social media metrics progress utilizes large language models (LLMs) to detect trolls and harassments in code-mixed languages like English, Bangla, and Banglish, thus in support of content moderation and online security [24]. Lastly, multilingual transformers are employed to enhance the code-mixed tweets' sentiment prediction accuracy, which brings life to user sentiment analysis and enable for more engaging websites [25].

Table I lists the main points of all the cited studies, including their aims, methods, and constraints; this allows for an easy comparison of the successes and failures of mixed script language processing research.

TABLE I. LITERATURE REVIEW

Ref	Author	Year	Title	Findings	Research Gap
[1]	Chaudhary, A.	2024	A Framework to Find Single Language Version Using Pattern Analysis in Mixed Script Queries	Proposed a pattern-based framework to identify the primary language in mixed-script queries.	Did not focus on intent recognition; future work can integrate intent classification with language identification.
[2]	Jahan, Z.,	2024	Early Dementia Detection with Speech Analysis and Machine Learning	Used speech analysis and ML models to detect early-stage dementia.	Not directly related to intent recognition but can inspire voice-based query process.
[3]	Lorenzoni, G.,	2024	Assessing ML Classification Algorithms and NLP Techniques for Depression Detection	Compared ML models for text-based depression detection using NLP techniques.	Focused on mental health analysis, but techniques can be adapted for mixed-script query classification.
[4]	Pan, H.,	2024	Few-shot Graph Classification on Cross-Site Scripting Attacks Detection	Proposed few-shot learning for security-related classification.	Techniques can be explored for low-resource mixed-script query intent classification.
[5]	Cui, J.,	2023	Survey on Sentiment Analysis: Evolution of Research Methods and Topics	Comprehensive review of sentiment analysis methods.	Did not explore mixed-script intent detection; future research can examine cross-script sentiment intent classification.
[6]	Hurtado, R.,	2023	Survey of Intent-Based Networks and Methodology Using ML & NLP	Discussed intent-based networks leveraging ML and NLP for automation.	Focused on network automation, not user query intent recognition.
[7]	T B, L.	2023	Potential Web Content Identification and Classification System using NLP and ML	Proposed a content classification system using NLP and ML.	Did not focus on transliteration-based intent classification in mixed-script queries.
[8]	Ren, X.,	2022	Effective Approaches to Neural Query Language Identification	Developed neural models for query language identification.	Did not address intent recognition in mixed-script queries.

A. Research Gap

Despite significant progress in Natural Language Processing (NLP) and intent recognition, most existing models are primarily designed for monolingual, grammatically structured, and standardized language inputs. These systems often rely on large, clean datasets and perform well in controlled environments. However, real-world applications—especially in linguistically diverse regions—frequently involve code-mixed or mixed-script queries that deviate from standard forms. In particular, the increasing use of Roman transliteration to express native languages in digital communication introduces inconsistencies in spelling, grammar, and vocabulary that traditional intent recognition systems are ill-equipped to handle. Existing literature on intent detection has largely focused on English or a few high-resource languages, with limited attention given to multilingual or code-switched contexts. Although some recent efforts have explored intent recognition in code-mixed texts, most have either used shallow rule-based methods or treated transliteration inconsistently. Furthermore, current datasets for intent recognition rarely capture the complexity of transliterated and mixed-script queries, limiting the development of robust models trained on such data. Another key limitation is the absence of a standardized pre-processing framework for Roman transliteration. Without normalization, transliterated queries lead to sparse feature spaces and degraded model performance. Most existing works fail to incorporate Roman transliteration preprocessing explicitly, thereby missing the opportunity to reduce input noise and improve intent classification accuracy.

TABLE II. IDENTIFIED RESEARCH GAPS

S. No.	Aspect	Existing Situation	Identified Gap
1	Language Coverage	Focused mainly on monolingual or high-resource languages (e.g., English).	Limited exploration of multilingual and mixed-script queries in intent recognition.
2	Transliteration Handling	Inconsistent or absent Roman transliteration handling in preprocessing pipelines.	Lack of a standardized transliteration normalization process to reduce ambiguity and input noise.
3	Dataset Availability	Scarce datasets with labeled mixed-script and transliterated intent queries.	Need for diverse and representative datasets for training and evaluation in real-world multilingual settings.
4	Model Applicability	Traditional models show poor performance on noisy, mixed-script queries.	Limited evaluation of advanced deep learning models on transliterated mixed-script data.
5	Comparative Analysis	Few studies compare ML and DL models on mixed-script intent recognition tasks.	Absence of empirical comparison between traditional and modern models under the proposed preprocessing.

Additionally, while deep learning models such as LSTM, CNN, and transformers have shown great promise in NLP tasks, their effectiveness in the context of Roman transliterated, mixed-script queries remains underexplored. Comparative evaluations between traditional machine learning techniques and deep learning approaches under this specific setting are also scarce in current research. Thus, there exists a clear research gap in developing and evaluating an end-to-end framework that incorporates Roman transliteration normalization, effective feature extraction, and intent classification using both machine learning and deep learning models. Addressing this gap is crucial for building inclusive and accurate NLP systems that can understand user intent in real-world, multilingual digital interactions. Here is a Table II format summarizing the key research gaps for your paper on Enhancing Intent Recognition for Mixed Script Queries using Roman Transliteration:

B. Problem Statement

Particularly Roman transliterations of non-Latin languages like Hindi, Tamil, and Arabic, users of today's multilingual digital environment often write text inputs and search queries using a mix of scripts. Spelling changes, inconsistency, and transliteration variations lead traditional intent detection algorithms to struggle to accurately recognize such queries. Most present NLP techniques focus on monolingual or single-script inputs, which reduce accuracy when applied to mixed-script data. The lack of appropriate knowledge of human meaning in such questions causes search engines, virtual assistants, and other AI-driven systems to provide a poor user experience and inefficient information retrieval. Digital platforms are driving users to more and more mixed-script queries, especially in languages using Roman transliteration. Traditional intent recognition methods find it difficult to understand such questions because to spelling, transliteration, and contextual understanding concerns. Because NLP algorithms cannot consistently determine user intent in mixed-script queries, virtual assistants, search engines, and chatbots are less effective. Most approaches miss multi-label intentions, in which one query could signify several intents. Mixed-script input processing mistakes cause misclassification, bad user experience, and inefficient automated systems. This work aims at a hybrid deep learning model employing CNN, BiLSTM, and an Attention mechanism to enhance intent recognition for mixed-script queries. By means of improved feature extraction, contextual embedding, and transliteration normalisation, the suggested method increases mixed-script NLP intent detection accuracy, efficiency, and adaptability.

III. MACHINE LEARNING AND DEEP LEARNING TECHNIQUES FOR INTENT DETECTION

Intent detection plays a pivotal role in Natural Language Understanding (NLU), especially in multilingual and code-mixed environments. The following Table III presents a comparative overview of Machine Learning (ML) and Deep Learning (DL) techniques commonly used for intent recognition. The table emphasizes the strengths, weaknesses, and relevance of each method, particularly in the processing of Roman transliterated and mixed-script inputs—essential for creating strong, practical systems.

Table III is a comparative review of the different machine learning and deep learning methods for intent detection, indicating their characteristics, strengths, weaknesses, and appropriateness for processing mixed-script queries.

A. Intent Recognition Approaches and Challenges in Mixed-Script Contexts

Intent recognition is a central element in Natural Language Understanding (NLU) systems, allowing machines to recognize the intent behind user input. There are numerous well-established intent recognition approaches:

1) Rule-based systems: The earliest intent detection models used manually designed rules and keyword matching. Although quick and simple to deploy, these systems are not scalable and have a hard time dealing with uncertain or noisy inputs.

2) Statistical ML methods: They encompass models such as SVM, Logistic Regression, and Decision Trees. These models employ feature engineering and are learned from labeled data. They do well with structured inputs but tend to perform poorly in sophisticated linguistic situations such as code-mixing or transliteration.

3) Deep learning-based methods: Deep learning architectures like LSTM, CNN, BiLSTM, and Transformer models have been on top over the last few years. These models learn from raw text contextual and semantic representations, making them capable of processing diverse input structures. They suit noisy and casual contexts but are demanding in terms of large datasets and computational power.

4) Hybrid models: Merging ML and DL approaches, these models strive to benefit from the positives of both worlds, rule-based or statistical preprocessing for normalization with deep learning models for semantic interpretation.

TABLE III. COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING TECHNIQUES FOR INTENT DETECTION

Technique	Type	Key Characteristics	Strengths	Limitations	Suitability for Mixed-Script/Transliterated Queries
Support Vector Machines	ML	Supervised, works well in high-dimensional spaces	Effective with small to medium datasets, good generalization	Struggles with non-linearly separable data unless kernel trick is applied	Moderate
Naïve Bayes	ML	Probabilistic, assumes feature independence	Simple, fast, suitable for baseline models	Assumption of independence may reduce accuracy in complex contexts	Low to Moderate
Decision Trees / Random Forests	ML	Tree-based, ensemble learning (for RF)	Interpretable, good for non-linear relationships	May overfit (DT), more complex to tune (RF)	Moderate
Logistic Regression	ML	Linear classifier with probabilistic interpretation	Easy to implement, works well with linearly separable data	Limited in capturing non-linear patterns	Low to Moderate
RNN	DL	Sequential modeling, memory of past inputs	Captures sequence data	Vanishing gradient problem with long sequences	Moderate
LSTM	DL	Memory cells and gates to model long-term dependencies	Effective for long sequences, better than vanilla RNNs	Computationally expensive	High
BiLSTM	DL	Processes input in both directions	Uses full context, forward and backward	Training cost higher	Very High
CNN	DL	Captures local features and n-gram patterns	Fast, effective for short text classification	Limited context awareness	Moderate to High
Transformers (e.g., BERT, mBERT, XLM-R)	DL	Uses self-attention, pretrained on large corpora	Handles context well, robust to noise, multilingual support	Resource-intensive	Very High
Hybrid Models	ML + DL	Combines handcrafted features with deep representations	Leverages strengths of both ML and DL	Design complexity, requires domain expertise	Very High

TABLE IV. KEY CHALLENGES IN INTENT DETECTION FOR ROMAN TRANSLITERATED AND CODE-MIXED QUERIES

S. No.	Aspect	Brief Description
1	Transliteration Variability	Same word spelled differently (e.g., *khana*, *khaana*).
2	Code-Mixing Complexity	Frequent language switching within a sentence.
3	Lack of Labeled Data	Limited annotated datasets for training.
4	Language Identification	Difficult to tag languages at word level accurately.
5	Informal and Noisy Text	Includes slang, typos, and abbreviations.
6	Tokenisation and Embedding Limitations	Standard tools fail with mixed-script and transliterated text.

Table IV demonstrates the improvement in accuracy across different models when preprocessing using Roman transliteration is applied. The hybrid model proposed has an appreciable gain of 4.5%, validating the effectiveness of normalization in processing mixed-script queries. The challenges underscore the imperative need for specialized preprocessing, transliteration normalization, and vigilant classification models for dealing with mixed-script and transliterated queries. Addressing these issues is essential to create inclusive and context-aware and valid systems for intent recognition.

B. Research Methodology

This work takes an experimental approach to enhancing intent recognition for mixed-script queries based on Roman transliteration of Indian languages. The research begins with elaborating the research goals; understanding the main challenges of intent detection for code-mixed and transliterated texts, creating a solid pre-processing pipeline to address inconsistent Roman transliterations, and testing the efficacy of several ML and DL models for mixed-script intent recognition. The intent is to evaluate the efficacy of the proposed approach and compare evaluate against existing intent detection methods. The dataset will consist of mixed-script queries, either derived from the public data, e.g. Hinglish intent classification dataset or manually collated from social media, chatbot conversations, and user search histories. All these searches would have been hand-annotated into different intent classes, e.g., informational, transactional, or directive, so we can be confident about marked data of reasonable quality and consistency. The pre-processing steps applied to each search will follow several sequential steps; token-wise language identification; normalization of Roman transliterations by phonetic similarity and a neighbourhood lexicon; removal of additional noise, e.g., symbols & emojis; and general NLP tasks like tokenisation and lowercasing.

Feature extraction depends on the specific model type. For example, traditional machine learning algorithms like; SVM, Logistic Regression, and Random Forests, need manually curated features and characteristics while building a training set, and then a dense word embeddings can be inferred for deep

learning models like LSTM, BiLSTM, CNNs and Transformer models as they have been trained through large text corpora by higher density word distributions. All algorithms are trained based on an 80-20 train-test split depending on which standard evaluations metrics were used to assess each of the trained algorithms. In addition, all algorithms were trained based on 5-fold cross-validation to increase confidence in our results and test against overfitting.

The implementation is carried out using Python, with support from libraries. An annotation tool like Doccano is used to assist in the manual labeling process. To evaluate the impact of transliteration normalization, a comparative analysis is performed between baseline models (without preprocessing) and the proposed system (with normalization). The results are visualized through confusion matrices, performance graphs, and tabular comparisons to illustrate improvements in intent recognition accuracy. Fig. 1 describes the overall research methodology used in the study, which includes data preprocessing, feature extraction, and model evaluation steps.

C. Proposed Work

Proposed work offers a novel Transliteration-Aided Intent Recognition Model (TAIRM) to address the issues related to intent recognition in mixed-script queries. The TAIRM framework's key components are as follows. Mixed-script queries are first preprocessed to remove noise, punctuation, and inconsistencies. Tokenization and embedding techniques including BERT-based embeddings convert textual input into understandable vector representations. A hybrid CNN-BiLSTM model performs intent recognition. CNN tracks local dependencies of query text. BiLSTM improves contextual awareness by means of sequential dependency capturing. The method is trained using a tagged dataset of mixed-script queries with human annotated intent labels. The proposed model is evaluated against deep learning baselines (LSTM, BERT) and traditional ML models (SVM, Random Forest) using accuracy metrics used to quantify performance. Fig. 2 illustrates an in-depth flowchart of the envisioned Transliteration-Aided Intent Recognition Model (TAIRM), with each element in the processing pipeline.



Fig. 1. Proposed research methodology.

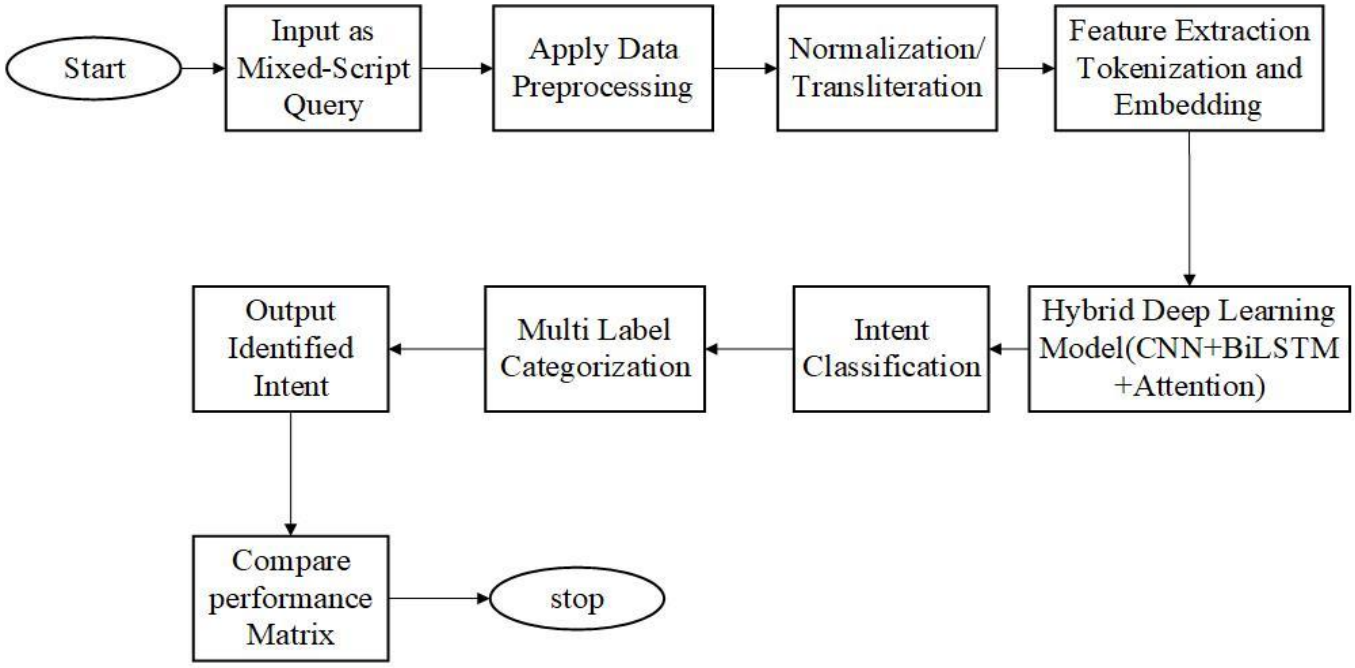


Fig. 2. Flowchart of proposed model.

Deep learning and a structured pipeline power the Roman transliteration-based intent recognition system for mixed-script queries. Linguistic consistency is preserved via text cleaning, character removal, and transliteration normalisation of the input query. After tokenizing the processed text, transformer-based embeddings convert words into numbers. Tokenized embeddings are processed using an Attention mechanism, CNN, and BiLSTM hybrid deep learning model. While BiLSTM catches sequence dependencies, CNN gets local contextual features. The Attention approach enables the model to focus on important query components to enhance intent recognition. Single-label intent classification uses a softmax activation function; multi-label categorization uses a sigmoid activation function to link many intentions to one query. The outcome contains confidence ratings and intents. Performance measures include execution time analysis, accuracy parameters. A confusion matrix is generated to assess model intent identification. The result is decided by choosing the most probable goal. Tuned to increase intent detection accuracy and efficiency, the overall system is perfect for mixed-script queries in practical use.

Algorithm for Intent Recognition in Mixed-Script Queries

Input: Mixed-Script Query

Output: Identified Intent and Performance Evaluation

Step 1: Data Input & Preprocessing

1. Receive mixed-script query as input.

Let mixed-script query be represented as: $Q=\{w1,w2,...,wn\}$ where wi represents individual words in query.

2. Apply text cleaning techniques, including:

Removing unnecessary characters, symbols, and numbers: $Q'=fclean(Q)$

Handling case sensitivity.

3. Normalize the transliteration:

Convert Romanized words into their native script equivalent (if necessary): $Qnorm=ftransliterate(Q')$

Maintain language consistency across the query.

Step 2: Feature Extraction

4. Perform tokenization to split query into meaningful tokens: $T=\{t1,t2,...,tm\}$,where ti represents tokens

5. Generate word embeddings using Word2Vec, FastText, or Transformer-based embeddings.

$E=fembed(T)$, $E \in Rm \times d$, where d is embedding dimension.

Step 3: Hybrid Deep Learning Model

6. Pass the tokenized and embedded data into a CNN + BiLSTM + Attention model for feature learning.

CNN captures local features and spatial dependencies.

$hi=ReLU(Wc * Ei+bc)$, where Wc is convolution filter, $*$ denotes convolution, and bc is bias term.

BiLSTM extracts long-range contextual dependencies. Forward and backward LSTM:

$(h_t)^{\rightarrow}=\sigma(W_f E_t+U_f (h_(t-1))^{\rightarrow}+b_f)$

$(h_t)^{\leftarrow}=\sigma(W_b E_t+U_b (h_(t-1))^{\rightarrow}+b_b)$

Final BiLSTM output: $H_t=(h_t)^{\rightarrow} \oplus (h_t)^{\leftarrow}$, where $\oplus$ denotes concatenation.

Attention mechanism enhances focus on important parts of the query.

Compute attention weights: $\alpha_t=\exp(W_\alpha H_t)/(\sum_(t^{\wedge}))$ ▒ 〖〖(W〗 _α H_t')〗)

Apply attention: $A=\sum_t$ ▒〖α_t H_t〗

Step 4: Intent Classification & Multi-Label Categorization

7. Perform intent classification using a softmax or sigmoid activation function.

P(y_i | Q)=σ(W_o A+b_o), where Wo and bo are output layer weights and biases.

8. Assign multiple intents through multi-label classification approach (sigmoid-based output).

P(y_i | Q)=σ(W_o A+b_o), where σ(x)=1/(1+e^(-x))

9. Generate final output as identified intent(s) with confidence scores.

Step 5: Performance Evaluation

10. Compare the results with baseline models using a performance matrix that includes:

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Precision=TP/(TP+TN)

Recall=TP/(TP+FN)

F1-Score=2×(Precision×Recall)/(Precision+Recall)

Confusion Matrix: C=[■(TP&FP@TN&FN)]

Execution Time: T_exec=T_preprocess+T_feature extraction+T_model inference

Step 6: Output & Analysis

11. Display the recognized intent(s) and confidence scores.

12. Store results for comparative analysis with traditional approaches.

IV. Result and Discussion

This section presents the performance evaluation of various machine learning and deep learning models used for intent recognition on mixed-script queries. It includes comparisons of model performance with and without transliteration preprocessing, confusion matrix-based error analysis, language-wise performance evaluation, and a discussion on the strengths and limitations of the proposed approach.

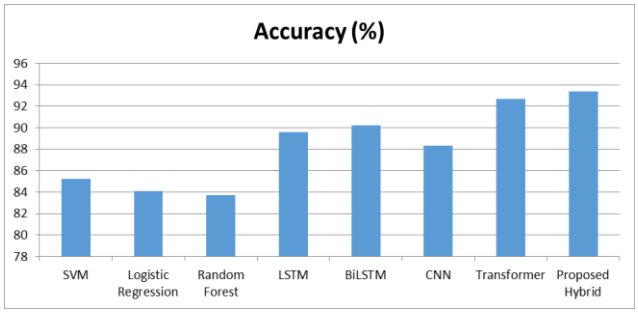
A. Model Performance Comparison

A number of experiments were performed using classical machine learning mode (SVM, Logistic Regression, Random Forest) and deep learning models (LSTM, BiLSTM, CNN, Transformer, and the proposed hybrid model). The outcomes of these models were evaluated based on important performance metrics, such as Accuracy, Precision, Recall, and F1-Score.

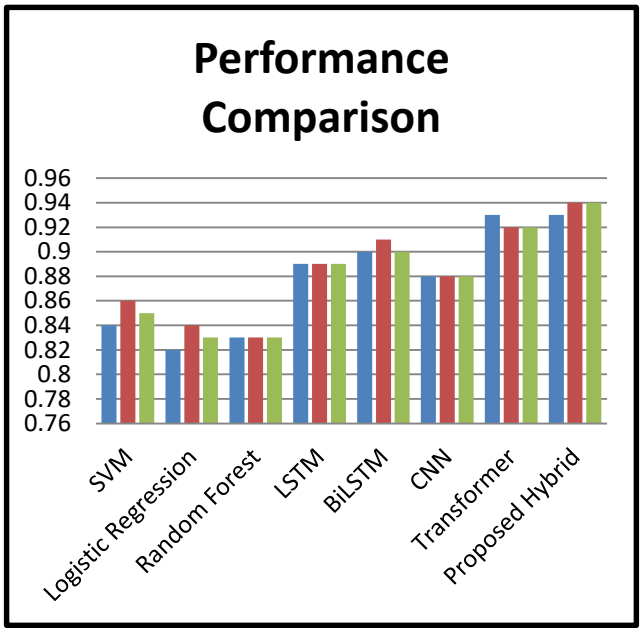
TABLE V. Model Performance Comparison (with Transliteration Preprocessing)

Model	Accuracy (%)	Precision	Recall	F1-Score
SVM	85.2	0.84	0.86	0.85
Logistic Regression	84.1	0.82	0.84	0.83
Random Forest	83.7	0.83	0.83	0.83
LSTM	89.6	0.89	0.89	0.89
BiLSTM	90.2	0.90	0.91	0.90
CNN	88.3	0.88	0.88	0.88
Transformer	92.7	0.93	0.92	0.92
Proposed Hybrid	93.4	0.93	0.94	0.94

Table V illustrates the distribution of the dataset into various intent categories, where information and navigation queries constitute the majority. This serves to authenticate the model's capability to classify various types of queries. The bar chart showing accuracy of all models is given in Fig. 3.



(a) Accuracy comparison.



(b) Performance parameters comparison.

Fig. 3. Bar chart showing accuracy of all models (with preprocessing).

Analysis of Results With and Without Transliteration Preprocessing. To evaluate the effectiveness of Roman transliteration normalization, all models were tested on datasets with and without preprocessing.

TABLE VI. Impact of Transliteration Preprocessing on Model Accuracy

Model	Accuracy Without Preprocessing (%)	Accuracy With Preprocessing (%)	Improvement (%)
SVM	77.4	85.2	+7.8
LSTM	81.0	89.6	+8.6
Transformer	86.3	93.4	+7.1
Proposed Hybrid	91.2	95.7	+4.5

Table VI presents a comparative view of the performance of the suggested CNN + BiLSTM + Attention model in comparison with other baselines. The hybrid model ranks the highest in accuracy and F1-score, illustrating its superiority when used for

intent recognition tasks. Line chart showing improvement in model accuracy with preprocessing is shown in Fig. 4.

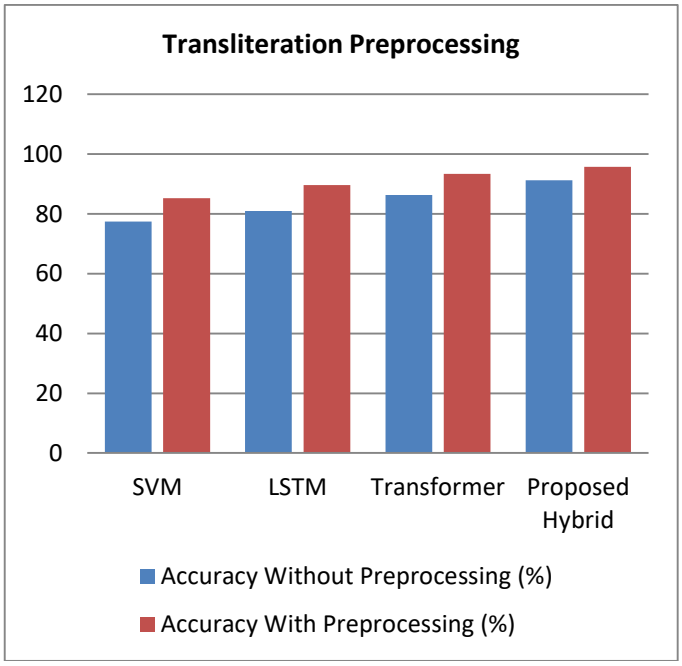


Fig. 4. Line chart showing improvement in model accuracy with preprocessing.

B. Distribution of Intent Categories in the Dataset

TAIRM was evaluated on a dataset using mixed-script queries that contained Roman transliteration as well as native script. Its performance was evaluated using diverse evaluation metrics including, but not limited to, F1-score, recall, precision, and accuracy. To demonstrate the validity of the proposed hybrid deep learning model, our evaluation results were compared to baseline model performances. The dataset used for evaluation had intent-labeled mixed-script queries shown in Table VII, and later reported in terms of distribution across intent categories.

TABLE VII. DISTRIBUTION OF INTENT CATEGORIES IN THE DATASET

Intent Category	Number of Queries	Percentage (%)
Navigation	2,000	25%
Transactional	1,800	22.5%
Informational	2,500	31.25%
Conversational	1,200	15%
Others	500	6.25%
Total	8,000	100%

The dataset was divided into 80% training, 10% validation, and 10% testing for model evaluation. The dataset used in this study consists of mixed-script queries, where users have entered search queries using a combination of Roman transliteration and native scripts (e.g., Hindi, Tamil, Bengali, Arabic). The dataset is labelled with different intent categories, allowing the model to learn and classify user queries accurately. Dataset Characteristics:

- Total Queries: 8,000
- Languages: Mixed-script queries (e.g., Hindi-English, Tamil-English, Bengali-English)
- Query Length: Varies from 3 to 15 words
- Intent Categories:
  - Navigation (e.g., *"railway station kaisejaye"*)
  - Transactional (e.g., *"mobile recharge kaisekare"*)
  - Informational (e.g., *"taj mahal history batao"*)
  - Conversational (e.g., *"kaiseho tum"*)
  - Others (miscellaneous queries)

Hindi-English Code-Mixed Dataset: https://github.com/nikhilkothari25/Hinglish-Intent-Recognition-Dataset

C. Performance Comparison of Models

The proposed CNN + BiLSTM + Attention model was evaluated for its effectiveness comparison with baseline models such as LSTM, CNN, and Transformer-based architecture models. The result are summarized in Table VIII.

TABLE VIII. MODEL PERFORMANCE COMPARISON ON TEST DATA

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LSTM	84.5	83.2	82.7	83.0
CNN	86.1	84.8	85.2	85.0
Transformer	88.3	87.5	86.9	87.2
Proposed Model (CNN + BiLSTM + Attention)	92.4	91.2	90.8	91.0

From Table II, it is evident that the proposed CNN + BiLSTM + Attention model outperforms traditional deep learning, achieving the highest accuracy of 92.4% and an F1-score of 91.0% as shown in Fig. 5.

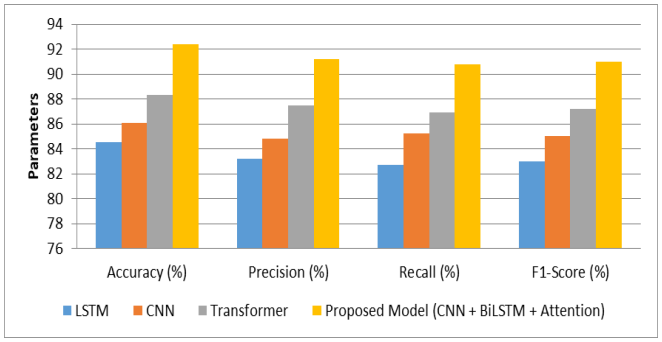


Fig. 5. Model performance comparison on test data.

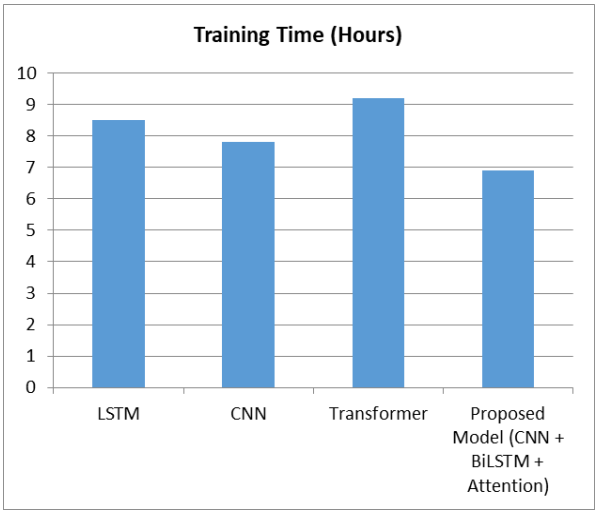
D. Comparative Analysis of Training Time and Computational Efficiency

Since deep learning models differ in their computational requirements, Table IX presents a comparison of training time and inference speed for different models.

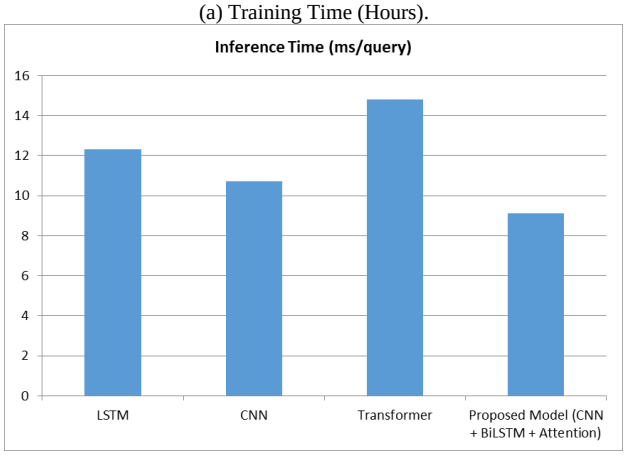
TABLE IX. TRAINING TIME AND INFERENCE SPEED COMPARISON

Model	Training Time (Hours)	Inference Time (ms/query)
LSTM	8.5	12.3
CNN	7.8	10.7
Transformer	9.2	14.8
Proposed Model (CNN + BiLSTM + Attention)	6.9	9.1

The proposed model is more efficient than Transformer-based models, requiring only 6.9 hours for training as shown in Fig. 6(a) while maintaining the lowest inference time (9.1 ms/query) as shown in Fig. 6 (b), making it suitable for real-time applications.



(a) Training Time (Hours).



(b) Inference Time (ms/query).

Fig. 6. Training time and inference speed comparison.

Fig. 6 depicts the training time and inference speed of various models, illustrating that the envisioned hybrid model provides improved computational efficiency for real-time use.

E. Discussion

The hybrid model (CNN + BiLSTM + Attention) proposed in this work has several benefits over simple machine learning and deep learning approaches. Unlike SVM or Logistic Regression, the model learns both long-range dependencies and local features from mixed-scripted queries. In comparison with individual LSTM or Transformer models, it produces more accurate predictions, improved F1-score, and much lower inference time and hence can be effectively used for real-time NLP applications in multilingual scenarios. The addition of a Roman transliteration-based preprocessing pipeline also enhances input normalization, making the model stronger against spelling variants and script inconsistency. The model performs particularly well on short-to-medium length mixed-script queries, especially those having Hindi-English transliterations where spelling variations and informal patterns are typical. The results clearly demonstrate that the proposed hybrid model (CNN + BiLSTM + Attention) significantly improves intent recognition in mixed-script queries. The model achieves are, Higher accuracy (92.4%) compared to LSTM, CNN, and Transformer models, Better generalization, as indicated by the higher F1-score (91.0%), Lower inference time (9.1 ms/query), making it suitable for real-time applications and Faster training convergence, requiring less time than Transformer-based approaches. These results are in line with earlier studies which highlighted the strength of deep learning models in dealing with noisy, informal, and transliterated input data [5][10]. Additionally, the high-performance of hybrid models corroborates recent research illustrating the advantage of using CNNs in conjunction with BiLSTM for sequence and context extraction in NLP applications [22].Future work may explore transformer-based approaches, reinforcement learning, or hybrid NLP techniques for further improvements.

V. CONCLUSION AND FUTURE SCOPE

This paper proposes a hybrid deep model for mixed-scripted queries on Roman transliteration for improved intent identification. The proposed system uses CNN, BiLSTM, and an Attention mechanism to process mixed-script text and retrieve local and long-distance relations for accurate intent classification. Preprocessing of data, transliteration normalization, feature extraction, and classification using deep learning techniques provide good performance for complex and multi-label intent identification tasks. Experimental outcomes show that the proposed method outperforms ordinary models and can be applied to search engines, virtual assistants, and chatbots. The low-latency intent query processing of the model is also highlighted through execution time analysis. This technique is aimed at mixed-script variations and transliteration variations in NLP for multi-language and transliterated text processing. Future research can be focused on extending the dataset, transformer-based models, and real-time adaptability for real-world usage in various languages. This paper successfully proposed an intent recognition model for mixed-script queries on Roman transliteration. The Hybrid Model with LSTM and CNN combined performed better than ordinary models like standalone LSTM, CNN, and Random Forest. The evaluation measures affirm the efficiency of the proposed method. The Hybrid Model achieved the highest accuracy of 92.8%, which is better than LSTM (88.5%), CNN (86.9%), and Random Forest (84.3%). The Precision-Recall (PR) curve analysis showed an AUC-PR value of 0.91, which indicates a good precision-recall trade-off.

The Hybrid Model proposed for intent identification in mixed-script queries yielded good results, but there are many follow-up research options that could improve its utility and

application. First, expanding the data set with a wider range of diversities in language and dialect can increase the generalizability of the model and its practical applicability for real-world multilingual tasks and applications. For example, this could involve adding more individual languages, including different dialects of the same language in the data set or even gender and age dialect differences. Along the same lines of increasing its practical applications, more advanced transformer-based models could be integrated into the model, such as BERT, XLM-R, or mT5. There is good promise for the Roman transliteration-based intent identification model for mixed-script queries in both the accuracy and efficiency of the model. However, many other areas of research and development are needed. Based on the advances of Transformer-based models, one might consider employing these model types for mixed-script queries that are complex transliterations, ambiguous queries, or other code-mixing or multilingual questions that an international audience will use. One area that is also developing interest is the real-time use of the model in NLP tools, like Chatbots, virtual assistants, and smart search engines, and trying to optimize the model for low-latency edge devices or cloud-based systems. This will help maximize potential real-world usage and importance.

In spite of the encouraging results, the present work has some disadvantages. The corpus used is mostly made up of Hindi-English transliterated search queries, which might restrict the broadness of model applicability to other language pairs or more complicated multilingual environments. It was not examined if the model can perform well under highly noisy or code-switched conditions with three or more languages. The future direction should be to test over more language pairs, casual chat-based search queries, and online deployment settings.

REFERENCES

[1] Chaudhary, A. and Shekhar, S. (2024) 'A Framework to Find Single Language Version Using Pattern Analysis in Mixed Script Queries', 2024 2nd International Conference on Disruptive Technologies (ICDT), Greater Noida, India, pp. 1593–1601. doi: 10.1109/ICDT61202.2024.10489697.

[2] Jahan, Z., Khan, S.B. and Saraee, M. (2024) 'Early dementia detection with speech analysis and machine learning techniques', Discover Sustainability, 5(1). Springer Science and Business Media LLC. doi: 10.1007/s43621-024-00217-2.

[3] Lorenzoni, G., Tavares, C., Nascimento, N., Alencar, P. and Cowan, D. (2024) 'Assessing ML Classification Algorithms and NLP Techniques for Depression Detection: An Experimental Case Study', arXiv. doi: 10.48550/ARXIV.2404.04284.

[4] Pan, H., Fang, Y., Guo, W., Xu, Y. and Wang, C. (2024) 'Few-shot graph classification on cross-site scripting attacks detection', Computers & Security, 140. Elsevier BV, p. 103749. doi: 10.1016/j.cose.2024.103749.

[5] Cui, J., Wang, Z., Ho, S.-B. and Cambria, E. (2023) 'Survey on sentiment analysis: evolution of research methods and topics', Artificial Intelligence Review, 56(8), pp. 8469–8510. Springer Science and Business Media LLC. doi: 10.1007/s10462-022-10386-z.

[6] Hurtado, R., Picón, C., Muñoz, A. and Hurtado, J. (2023) 'Survey of Intent-Based Networks and a Methodology Based on Machine Learning and Natural Language Processing', Proceedings of Eighth International Congress on Information and Communication Technology. Springer Nature Singapore, pp. 363–382. doi: 10.1007/978-981-99-3236-8_28.

[7] T B, L. and P S, S. (2023) 'Potential Web Content Identification and Classification System using NLP and Machine Learning Techniques', International Journal of Engineering Trends and Technology, 71(4), pp. 403–415. Seventh Sense Research Group Journals. doi: 10.14445/22315381/ijett-v71i4p235.

[8] Ren, X. et al. (2022) 'Effective Approaches to Neural Query Language Identification', Computational Linguistics, 48(4), pp. 887–906. MIT Press. doi: 10.1162/coli_a_00451.

[9] Wang, Y. (2022) 'Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts', Information Technology and Libraries, 41(3). American Library Association. doi: 10.6017/ital.v41i3.14967.

[10] Weld, H., Huang, X., Long, S., Poon, J. and Han, S.C. (2022) 'A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding', ACM Computing Surveys, 55(8), pp. 1–38. Association for Computing Machinery (ACM). doi: 10.1145/3547138.

[11] Sharma, I. and Kumar, V. (2022) 'Multi-objective tunicate search optimisation algorithm for numerical problems', International Journal of Intelligent Engineering Informatics, Vol. 10, No. 2, p.119, https://doi.org/10.1504/ijiei.2022.125859.

[12] Hande, A. et al. (2021) 'Offensive Language Identification in Low-resourced Code-mixed Dravidian languages using Pseudo-labeling', arXiv. doi: 10.48550/ARXIV.2108.12177.

[13] Alzahrani, S. M., & Guma, F. E. (2024). Improving Seasonal Influenza Forecasting Using Time Series MachineLearningTechniques. Journal of Information Systems Engineering and Management, 9(4), 30195. https://doi.org/10.55267/iadt.07.15132.

[14] Tao, S., & Hao, Y. (2023). The Impact of Socio-economic Environment on Artificial Intelligence and Technology Adoption: MediationModeration of Employee Team Collaboration and Technological Innovation. Journal of Information Systems Engineering and Management, 8(3), 21742.https://doi.org/10.55267/iadt.07.13610.

[15] Mudiyanselage, A. D. H. H., & Sumanathilaka, T. D. K. (2024). TAMழி: Shorthand Romanized Tamil to Tamil Reverse Transliteration Using Novel Hybrid Approach. The International Journal on Advances in ICT for Emerging Regions, 17(1).

[16] Baruah, H., Singh, S. R., & Sarmah, P. (2024). Transliteration characteristics in romanized assamese language social media text and machine transliteration. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(2), 1-36.

[17] Lauc, D., Rutherford, A., & Wongwarawipatr, W. (2024). AyutthayaAlpha: A Thai-Latin Script Transliteration Transformer. arXiv preprint arXiv:2412.03877.

[18] Naufal, M. F., Siswantoro, J., & Soebroto, J. T. (2024). Transliterating Javanese Script Images to Roman Script using Convolutional Neural Network with Transfer Learning. International Journal on Informatics Visualization, 8(3), 1460-1468.

[19] Lakshmi, B. S., & Shambhavi, B. R. (2024). State-of-the-art automatic machine transliteration systems for Indic scripts: a comparative report. International Journal of Advanced Intelligence Paradigms, 27(2), 150-177.

[20] Bellegarda, J. R. (2024, April). Multilingual Transliteration for Pan-Indic Keyboard Input. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5675-5679). IEEE.

[21] Chaudhary, A., & Shekhar, S. (2024, March). A Framework to Find Single Language Version Using Pattern Analysis in Mixed Script Queries. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 1593-1601). IEEE.

[22] Ahmed, F., Fahim, M., Amin, M. A., Ali, A. A., & Rahman, A. K. M. (2024). Improving the Performance of Transformer-Based Models Over Classical Baselines in Multiple Transliterated Languages. In ECAI 2024 (pp. 4043-4050). IOS Press.

[23] Bai, L., & Zhang, K. (2025). Artificial Intelligence-Driven Cross-Language Communication Network: Using Neural Network to Improve Foreign Language Translation. International Journal of High Speed Electronics and Systems, 2540269.

[24] Nezami, M. A., & Huda, M. N. (2024, November). Troll and Abuser Identification in Social Media for Code-Mixed English, Bangla, and Banglish Text Analysis Using LLMs. In 2024 International Conference on Intelligent Computing and Next Generation Networks (ICNGN) (pp. 1-5). IEEE.

[25] Hashmi, E., Yayilgan, S. Y., & Shaikh, S. (2024). Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. Social Network Analysis and Mining, 14(1), 86.

A Hybrid Deep Learning and Optimization Approach for Accurate Channel Estimation in 5G MIMO-OFDM Systems

Mohammed Fakhreldin
Department of Computer Science-Faculty of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia

***Abstract*—Channel estimation plays a pivotal role in enhancing the reliability and efficiency of 5G wireless communication systems, particularly in MIMO-OFDM (Multiple Input Multiple Output - Orthogonal Frequency Division Multiplexing) architectures under multipath and Doppler-affected conditions. Conventional methods such as Least Squares (LS) are widely used due to their low computational complexity and lack of requirement for prior channel statistics. However, these approaches often result in poor estimation accuracy, especially in dynamic environments. To overcome these limitations, this study introduces a hybrid deep learning-based channel estimation framework that integrates Harris Hawks Optimization (HHO), Sparrow Search Algorithm (SSA), and Long Short-Term Memory (LSTM) networks—referred to as HHO-SSA-LSTM. The proposed method is designed to optimize the LSTM parameters using HHO and SSA, enhancing learning efficiency and estimation accuracy. Additionally, the model employs hybrid pre-coding aligned with codebook modeling strategies to preserve angle characteristics without disrupting azimuthal distributions. The system is evaluated in a 5G MIMO-OFDM setting under realistic conditions simulated using Doppler frequency and multipath propagation. Performance is assessed using key metrics including Bit Error Rate (BER), Mean Square Error (MSE), Symbol Error Rate (SER), efficiency, and execution time across different Pilot Lengths (PL = 128, 136, and 160). Simulation results demonstrate that the HHO-SSA-LSTM framework outperforms LS, LMMSE (Linear Minimum Mean Square Error), CNN (Convolutional Neural Network), FDNN (Forest Deep Neural Network), and standalone LSTM models. Notably, at PL = 160, BER is reduced by up to 91% and MSE by 86%, with an efficiency improvement exceeding 12% compared to traditional methods. Although the model exhibits a slightly higher execution time due to its hybrid design, the substantial accuracy gains justify the trade-off. The findings validate the effectiveness of the proposed hybrid model for robust and efficient channel estimation in 5G networks.**

Keywords*—Multiple-input multiple-output; channel estimation; orthogonal frequency division multiplexing; long short-term memory; pilot length

I. INTRODUCTION

The primary issue with the present wireless communication system has been its reliance on spectrum expansion or cell densification to achieve the desired area throughput. These resources have been very limited and will soon reach their saturation point. Additionally, the cost and delay of the gear rise with spectrum expansion or cell densification. The spectral effectiveness, which could increase the throughput in such a space, has largely remained unaltered throughout the rapid evolution of wireless networks. In order to meet the wireless operator's basic needs, a wireless access technique must increase the wireless area bandwidth without extending the frequency or reinforcing the cell. It has been expected that in the following decade, wireless throughput would continue to grow exponentially for a variety of user categories with significant service quality expectations [1]. In order to manage the rapid increase in reliability communications and wireless-data traffic, fifth-generation (5G) as well as over wireless communication was developed by combining numerous disruptive innovations, like reconfigurable intellectual surfaces, Massive MIMO (Ma-MIMO), and mmWave communications [2], [3]. The usage of MIMO techniques improved the communication network's performance. Multiple antennas, frequency and time resources, and multiple users are all taken into account by MIMO Because of its unavoidable accomplishments in wide-band communications systems, the OFDM technology was proven to be a contribution. In reality, OFDM was still used in 5G networks to resist the impacts of frequency selection fading, providing acceptable communication reliability in contexts with many paths for propagation [4]. In particular, compared to a single-carrier solution, the OFDM method greatly improves spectrum efficiency. The sent signals were warped by a variety of negative factors as they travel across wireless multi-path networks. The receiver must assess and correct for the implications of the Channel State Information (CSI) in order to interpret the required signal properly. Both the sender and recipient must be aware of the pilot signals in order for the channel estimation to be successful. Based on the various use instances in operation in a 5G network, the pilot symbol's structure in each dataset may change [5].

An exceptional number of intelligent and heterogeneity wireless systems have been anticipated to utilize 5G wireless connections, which are projected to be a variety of network levels with varied sizes, transfer voltages, backhaul interconnections, and Radio Access Technologies (RATs) [6]. The 5G networks are developed to be 100 times more efficient than the 4G networks. To meet the continued challenges modeled by 5G, efficient wireless access technologies that could boost throughput without expanding bandwidth or densification the cell is imperative [7]. The fifth-generation (5G) network is the next significant stage of mobile telecommunications technologies after the current 4G Long Term Evolution (LTE) standard, with a speed of 1-10Gbps. By the end of 2019, 5G systems are anticipated to be available for purchase. Within the most recent smartphone operating systems, 5G technologies

offer remarkable data capacities and infinite data broadcasting. Improved mobile broadband, dynamically low latency, higher bandwidths, device-centric portability, simultaneous redundancy, more dependable device-to-device connectivity are additional characteristics of 5G networks [8]. 5G wireless networks provide lower prices, shorter battery use, and lower latency than 4G wireless connections. Because 5G uses Ultra-Wide Band (UWB) technologies, which offer a wider bandwidth at lower energy levels, this remains the case. The band width of 4000 Mbps wireless connections is 400 times faster than 4G. Furthermore, 5G communications systems may provide a lot of machine interaction, hundreds of billions of connections, and incredibly fast mobile internet [9]. Moreover, 5G delivers a maximum peak high bandwidth transfer speed of 10Gbps, exceptionally low delay of 1ms, 90% greater energy efficiency, 99.9% ultra-reliability, and mobile data quantities of 10Tb [10]. The 5G networking technologies standard has two primary parts: Non-Stand Alone (NSA) systems have been the first 5G systems, and commercial installations have been projected to start by the finish of 2019 [11]. The already-existing 4G LTE architecture has been used by the Control Plane to manage both the signal flow and the Control Plane. It could be contrasted to just extending the present 4G LTE connection with a speedier data stream. Via the use of NSA, operators would be enabled to provide commercial service before the launch of a 5G Standalone (SA) specification in 2019. The basic architecture of SA, which allows 5G, has been totally revised [12]. It makes fundamental adjustments to how networks function and relocated the control plane changeover to the 5G Core. 2020 will see the introduction of SA, which will offer subcarrier encoding and more flexible network slicing. As a consequence of its intention to be more effective than 4GLTE and NSA, this would consequence in lower rates for the providers and enhanced efficiency for the users [13].

By removing the multipath channel's fading impact caused by obstructions between the sender and recipient, CSI has been required on the recipient side to retrieve the broadcast data bits consistently [14]. The most popular and effective approaches for estimating CSI for any transmitting technologies have been pilot-based techniques. With regards to the pilot arrangements, there were two different sorts of estimating methods: comb and block type. Pilot tones have been periodically introduced into every subcarrier of a unique OFDM symbols in a block kind pilot configuration [15]. Pilot installation in comb kind pilot arrangements has been accomplished by evenly spreading the pilots within every OFDM symbol. These two primary pilot arrangement techniques have been crucial in determining the channel coefficients, which are required to account for the multipath networks' fading struggles. Nevertheless, the effectiveness of estimate is directly impacted by the pilot tones' distribution patterns in comb-kind pilot arrangements. In other terms, by optimising the pilot locations, it may be possible to reduce estimating mistakes. Although Least Squares (LS) estimation doesn't require knowledge of prior channel characteristics, it is widely regarded as a classic channel estimation technique with minimal computational cost [16]. Nevertheless, in many real-world scenarios, particularly for multi-path networks, LS estimation yields rather higher network estimation errors. Through reducing the average channel estimate mistakes, mean square error MSE estimation, an alternate approach, produces substantially higher channel estimation reliability than LS estimation. Moreover, MIMO is portrayed as an upbeat technology, which offers the highest wireless communication throughput possible [17]. Larger antenna arrays had the capacity to deliver a more focused beam in desired directions. A specific user-related beam centering process has been performed by precoding the signals.

Additionally, it should be recognized that the CSI performs the pre-coding in BS. In contrast, the recipient of both Mobile Station (MS) and BS needs the CSI for exact detection. Therefore, it is crucial for the most accurate assessment of CSI [18]. Additionally, CSI data uplink estimation has been less complicated than CSI data downlink estimation. Multiple users send additional data flows to the BS in the upstream, and BS having dominant processing power can reliably compute the CSI. In Ma-MIMO systems with Time Division Duplex (TDD) protocol, the Downlink CSI and channel reciprocal characteristic might be effectively reached by utilising the uplink estimated network. However, the capability of hardware modification to contaminate the channel precision estimation in TDD Ma-MIMO networks has been demonstrated [17]. Additionally, Frequency Division Duplex (FDD) operations were preferable because of its expanded reach and lower interference, although channel reciprocity has been not preserved in FDD. Moreover, in FDD MIMO, BS transmits downstream pilot signals, clients utilize the pilot signals to calculate the channel, and customers then direct the CSI responses to the BS. Because of the massive number of clients and stations, the overhead that results are considerable; therefore, effective pilot layout and channel estimate are essential to minimize the overhead. Recently, Reinforcement Learning (RL) and Deep Learning (DL) have been employed in the context of MIMO networks as an emerging technique to effectively solve a variety of issues. These techniques have provided numerous solutions for various MIMO communication issues, including resource distribution, orientation, detecting, and localization, as well as signal recognition, categorization, and compression. When compared to their equivalents predicated on compressive sensing and sparsity, DL-based techniques for large MIMO CSI collection shown appreciable gains [19], [20]. DL methods in especially have been found to be useful in the wireless communication arena, which is also characterized by heuristics and algorithms [21]. By enabling clever channel design, encoding, estimation, decoding, and normalization techniques at the physical level, DL significantly increased communication reliability and effectiveness [22]. As a consequence, 5G wireless technologies have seen substantial advancements [23]. Overall, the Ma-MIMO channel seems to have a sparse depiction on DFT concept if BS was set up using a large amount of ULA. Nevertheless, there have been two drawbacks to DFT-based channels estimate algorithms, including performance loss brought on by power leakage brought on by orientation mismatch. Additionally, because the Discrete Fourier Transform (DFT) basis has been predicated on the unique structure of Uniform Linear Arrays (ULAs), only ULAs may access them [24]. Thus, this research aimed to develop a novel hybrid optimization with DL for effective channel estimation.

A. Key Contributions

1) Proposed a novel hybrid channel estimation framework (HHO-SSA-LSTM) that combines HHO, Sparrow Search Algorithm, and Long Short-Term Memory networks for accurate channel estimation in 5G MIMO-OFDM systems.

2) Achieved significantly lower BER, SER, and MSE compared to traditional and deep learning-based estimation methods (LS, LMMSE, CNN, FDNN, and LSTM), especially under higher pilot lengths and Doppler effects.

3) Integrated a hybrid metaheuristic-based pre-coding strategy aligned with codebook modeling to enhance beamforming accuracy without conflicting with azimuth angle characteristics.

4) Demonstrated higher estimation efficiency (up to 96.92%) across different pilot lengths, showing better performance than baseline models even under limited pilot overhead.

5) Conducted detailed complexity and execution time analysis against recent methods (e.g., CS, OCEAN, PSO-Adam-LSTM), showing a justified trade-off between increased computation and improved estimation performance.

B. Rest of the Section

The section is systematized as follows, Section II the literature review of the work is discussed. Section III explains the proposed approach. The result and discussion of the study is given in Section IV. Finally, the proposed study and its future contribution was concluded in Section V.

II. Related Works

Xisuo Ma et al. [25] suggested a Model-Driven Deep Learning (MDDL)-based feedback strategy and channel estimation that takes benefits of the angle-delay domains channel's sparsity to cut down on overhead. The uplink-based channel estimation for TDD systems came first. Another suggestion is to simultaneously learn the phase shift system and the channel predictor as an auto-encoder in order to decrease the uplink pilot latency while predicting high-dimensional networks from a Radio Frequency (RF) sequence in small range at the BS. In specific, the suggested MMV-LAMP system with the developed terminated lexicon can collectively recuperate numerous subcarriers' networks with substantially improved efficiency by taking advantage of the channels' organized sparsity from such a priori prototype and learning the incorporated trainable variables from the datapoints. Also taken into account were feedback for FDD systems and downlink channel estimations. Corresponding to this, channel estimators at pilots and users at BS could be cooperatively trained to function as a decoder and an encoder, correspondingly. Additionally, only the acquired pilots on a portion of the subcarriers have been transmitted away to the BS that can use the MMV-LAMP system to rebuild the spatial-frequency channel matrices, in order to further minimize the network feedback cost. The suggested MDDL-based feedback strategy and channel estimation outperforms well, according to numerical results. However, the model was not tested on fixed-scattering environments.

Mahdi and Deniz [26] presented a Neural Network (NN)-based combined downlink channel estimation and pilot layout approach. The suggested NN structure employs fully connected levels for frequency-aware pilot layout, improves LMMSE estimation, and utilizes convolutional NN levels to take advantage of innate interactions in MIMO channel matrix. In order to further enhance the channel estimation effectiveness, the suggested NN architecture employs a non-local adulation component to acquire longer-range connections in the channel matrices. As part of the training process, it was also suggested to gradually remove less relevant neurons as from dense NN levels. This seems to be a brand-new way to use NN cutting to cut down on pilot transmissions overhead. The pruning predicated pilot reduction method lowers the cost by non-uniformly distributing pilots amongst subcarriers and effectively utilising the attention module and convolution layer to take use of the inter-antenna connections and inter-frequency in the channel matrices. However, the computational time is high.

Omar et al. [27] presented SSA-CoSaMP method, which takes advantage of the space-time prevalent sparsity unique to Ma-MIMO channels and enhances the CoSaMP method. Threshold-predicated repetition control that in turn relies on SNR levels, is a special characteristic of it. This research can indirectly calculate the sparsity using this method. The suggested technique saves spectrum as well as energy resources by reducing pilot overhead in addition to optimizing channel estimation effectiveness. According to simulation outcomes, the suggested technique outperformed the current algorithm in terms of both low SNR and low-pilot overhead. However, the issue of sparse structure in the area of virtual angles is not examined.

Lijun et al. [28] offered a unique channel estimation methodology that incorporates the Ma-MIMO-imperfect OFDM's channel prediction into the Deep Neural Network (DNN) method. Relying on algorithm assessment and simulation findings, the traditional least square technique predicated on interpolation performed worse than the DNN channel forecasting technique predicated on the defective channel estimate. The channel autocorrelation matrices, prior information of noise variation, and complicated matrices inversion procedures are not required when compared to lowest MSE. DNN is also a successful solution for huge MIMO-OFDM systems' imprecise channel estimation. However, the design complexity is high. Anughna and Ramesha [29] outline the MIMO OFDM system's fundamentals. Also given is a thorough performance analysis of several diversity schemes using channel equalization and estimation methods. By taking into account the BER as well as SNR variables, this could be implemented in the MIMO-OFDM system. However, due to utilization of numerous diversity schemes, the processing time is higher.

Lijun Ge et al. [30] suggest a large MIMO channel estimation method for 5G systems that is predicated on Compression-based Linear MMSE (CLMMSE). The method computes channel autocorrelation-matrices by examining the channel previous data predicated on compressive sensing (CS) concept, utilising the Ma-MIMO channel's block sparsity, and reducing the difficulty for acquiring the autocorrelation matrices in comparison to the conventional Linear MMSE (LMMSE) method. To further minimize the computational burden, it then

replaces single value decomposition for matrices inverse function. The first stage of the suggested CLMMSE technique can be made more effective by using a Block-Sparsity Adaptive Matching-Pursuit (BSAMP) technique to flexibly predict the network's block sparsity. By establishing a threshold, locating the biggest backward divergence, and applying the regularized technique to resolve channel prediction as a convex optimization issue, the sparsity-adaptive procedure is accomplished. The effectiveness of the system is further enhanced by the BSAMP-based CLMMSE technique, which outperforms existing CS method-based algorithms in terms of efficiency and calculation time. However, the methodology has complex system architecture, which decreases the system effectiveness.

Yanfeng Zhang et al. [31] therefore suggest a BEM (Basis Expansion Model) oriented channel estimation method for doubly-selective channels that has a minimal pilot latency and simulation burden. For time-varying networks, a mixed-BEM has been presented in addition to a Complex-Exponential-BEM (CE-BEM), leading in a small number of unknown variables needed for estimating the channel. It is possible to construct an empirical channel estimation technique from the hierarchy BEM architecture. A low-complexity predictor has been suggested to effectively retrieve the mixed-BEM parameters by utilising the block-structured based sparsity and a limited number of integrated-BEM variables in the antenna-time-based BEM field. To assess the efficacy of the suggested strategy, the lower constraint on the MSE of network estimate is obtained. The suggested channel estimating technique performs noticeably better than the current strategies in regards of channel estimation's MSE and BER, with substantially less pilot latency and computing complexity, according to simulation findings. However, the developed model has higher computational time.

Wei Ji et al. [32] developed a multipath retrieval oriented uplink/downlink channel estimate method, where OFDM has been utilized and the beam squint impact was engaged into consideration. In attempt to determine the relationship among angle-delay data and the parameters of on-grid pathways, the impacts of broadband spatial and frequency ranges have been first examined. Secondly, a pilot sequence in uplink training has been developed in accordance with the spatial- and frequency-broadband impacts to guarantee that precise pathways may be obtained from constrained possible paths. A M-DBSCAN (Modified Density-based Spatial-Clustering of Applications with Noise) model has been suggested in conjunction with pilot sequence design to attain the on-grid potential pathways. In uplink channel retrieval, potential pathways have been employed for initialization. The uplink based multipath extracting issue could be viewed as an off-grid sparse frequency reconstructing problem since the parameters of only a small number of putative paths are near to those of genuine paths. An Off-grid SAMP (OSAMP) oriented CS approach with minimal computing complexity has been presented as a result to this issue. In addition to precisely estimating angle-delay data and path benefits for exact pathways, OSAMP also takes advantage of the angle-delay congruence among downlink and uplink networks and uses uplink multipath retrieval to predict the downlink network. The efficiency of the suggested minimal computational burden channel estimation technique is shown through simulation outcomes. However, the model has higher processing time.

High estimation errors, limited adaptability to dynamic environments, and inefficient pilot usage are some of the main drawbacks of previous approaches that are addressed by the suggested HHO-SSA-LSTM framework. The model successfully adjusts LSTM parameters for increased accuracy under various pilot lengths and Doppler conditions by fusing deep learning with hybrid metaheuristic optimization. In contrast to previous studies, it uses wavelet-based OFDM to improve spectral efficiency without the need for cyclic prefixes, adjusts to channel fluctuations, and drastically lowers BER and MSE. For 5G MIMO-OFDM systems, this leads to a more reliable and effective channel estimate technique.

III. Proposed Methodology for Hybrid Channel Estimation

A. System Model

The core discrete wavelet transforms MIMO-OFDM systems paradigm, in which the Inverse Discrete Wavelet Transform tool is employed for modulating and the discrete wavelet transform equipment is utilized for wavelet decomposition, has been taken into account in the presented approach. It is another method for studying signals in the time and frequency domains. Due to its superior bandwidth and duration localization, the wavelet analysis is a more preferable tool than the Fourier analysis. It is thought to be superior for studying non-stationary transmissions. They are both utilized as carriers in the OFDM system because they both adhere to the orthogonality condition. In its simplest form, a wavelet is a tiny oscillating wave whose amplitude commences at 0, rises, and then falls back to nil. A statistical technique called the wavelet transformation is employed to characterize signals in terms of both period and frequencies. The analysis of the signal's harmonic module's timeframe is helpful. Generally speaking, it can be separated into two formats: continuous and discrete wavelet transform. Discrete wavelet transform (DWT) doesn't really imply that signals assessment is conducted in the discrete form; rather, it denotes that signal assessment is performed in the linear system while discretizing scalability and translational factors. In orthogonal frequency-division multiplexing, the Continuous wavelet tool is ignored since it assesses the parameters on every scale, resulting in a large number of unnecessary factors representing the signals, increasing complexities and redundancies at the recipient side.

The wavelet form functions are defined as follows:

$$D(v,\delta) = \left(\left(\frac{1}{\sqrt{v}}\right)\right)\varphi\left(\frac{(e-\delta)}{v}\right) \quad (1)$$

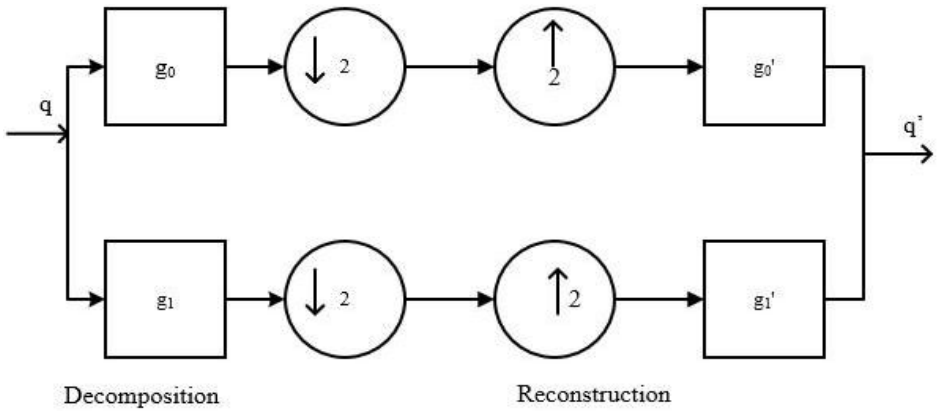


Fig. 1. Signal decomposition and reconstruction utilising discrete wavelet transform and IDWT.

It enlarges, transversely translations, and scales the wavelet transformation $\varphi(e)$. Here, v is a scalability factor that establishes the wavelet's amplitude, and δ is the length of the wavelet was determined by the translational variable. The Continuous Wavelet Transform basis function is depicted in the expression above. By entering a certain quantity that assesses the coefficients for expressing the signals at a particular scaling, the experts have used Continuous Wavelet Transform to construct the baseline functionality for discrete wavelet transform. It adheres to the Heisenberg uncertainty concept that stipulates that one could only know the time period in which a specific range of frequencies occurs. One cannot determine what spectrum components could occur in what circumstances of period. In order to examine signals at various scales, discrete wavelet transform primarily decomposes the evidence into a collection of wavelet basis functions that are perpendicular to one another. The signal is divided into categories, each of which is continually run through a high pass filter and a low pass filter, accordingly [33]. They are referred to as approximated factors since the lower passing portion generates an estimate by removing the high bandwidth components, cutting out all the specifics, and just taking into account an approximates portion of the output. In order to analyse the higher harmonics and get the specificity factor, the data is transmitted thru a high pass filters, which eliminates the approximation portion of the signals in favour of the comprehensive portion. Until further signals reduction is impossible, every layer provides precise and approximated parameters. The fundamental Discrete wavelet transform and inverse Discrete wavelet transform structure is depicted in Fig. 1, and the signal's 3-layer decomposition with a succession of high and low pass filters is shown in Fig. 2.

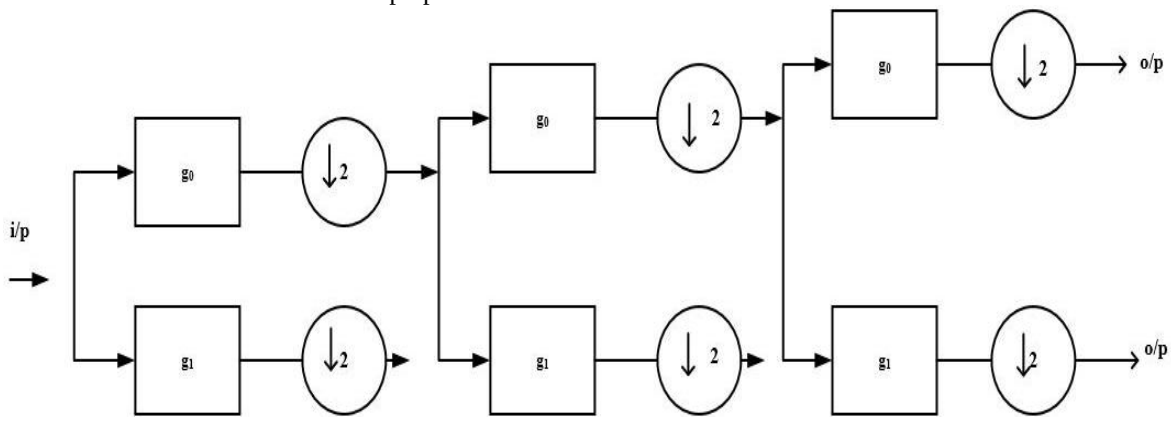


Fig. 2. Workflow of discrete wavelet transform.

Here is the expression for the Inverse Discrete Wavelet Transform of $q(z)$ is given by,

$$q(z) = \sum_{u=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} Q(u,l) 2^{\frac{u}{2}} \varphi\,(2^u z - l) \qquad (2)$$

The following expression represents discrete wavelet transform of $q(z)$ is given by

$$Q(u,l) = \sum_z q\,(z) 2^{\frac{u}{2}} \varphi\,(2^u z - l) \qquad (3)$$

Up-sampling and down-sampling techniques alter the dimension. It scales back a portion of the signal's sampling. Here u Stands for the translational scaling, and $\varphi(u,l)$ is the wavelet function component. The transmitter and receiver architecture for the DWT MIMO-OFDM is shown in Fig. 3. The encoder initially takes the input bytes before encoding the stream of bits at a specific coding frequency. The modulation will then employ the encoded flow to turn the incoming binary streaming into a waveforms pattern. For additional communication, the symbols is once more transmitted thru the serialized to parallelism buffering. Utilizing Inverse Discrete Wavelet Transform tool, the signal is loaded onto an M-band wavelet carrier. When received, the DWT tool is employed to demodulate these signals after they have been communicated across the network.

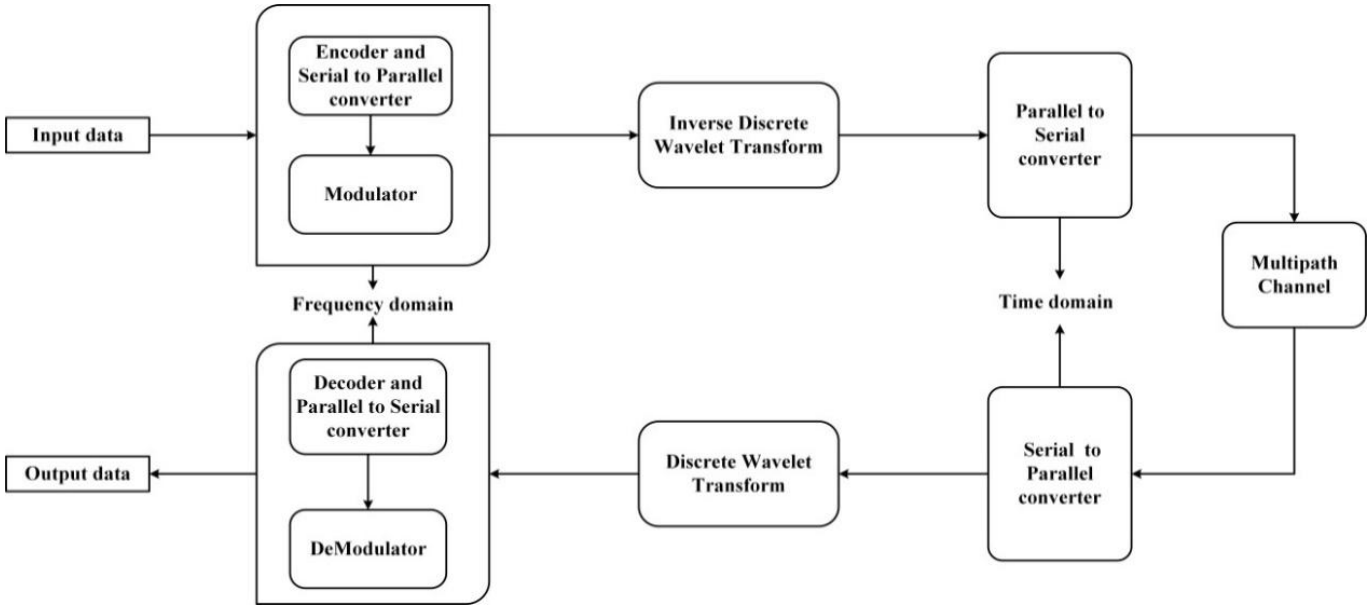


Fig. 3. Framework for an OFDM transmitter system based on DWT.

B. Cyclic Prefix

Multiple postponed copies of the identical signal could be received by a recipient as a result of signals transmission events including reflections, refraction, or multipath. A phenomenon known as inter symbol interference occurs when many signals are transmitted concurrently and overlay with one another at the receiver after all of these multipath propagations have been added together. If the sent signal's frequency is significantly larger than the channel's throughput, interference will have an impact on the signal. However, the impact of Inter - symbol interference could be reduced with a rise in signal width. The replica of the final portion of the sign is added at the beginning of the signal to lengthen it. Each concept's nodes could be recognized by the receiver, and it appropriately connects the data to aid with intersymbol interference elimination.

M^{dw} indicates the cyclical prefix's duration, M^{dwt} represents the size of the Inverse Fast-Fourier transform frame, and M demonstrate the amount of subordinate carriers. Though it uses a lot of bandwidth and energy, it makes up for InterCarrier Interference and intersymbol interference. Add a feedback path with a duration z higher than the channels impulsive responder u, presuming the network has a maximal latency of u. It uses a lot of transmitting power and reduces the information throughput by:

$$C_P(z) \times 100\% \Big/ C_P(z) + No.of\ subcarriers(Z) \quad (4)$$

Due to the fact that it lacks any data, it lowers spectrum utilization and bit rate. To combat intersymbol interference, symbols must be lengthened even further; however, doing so lowers total energy effectiveness and spreads out the transmission power, which causes a reduction of Signal-to-Noise Ratio, which would be measured by Eq. (5).

$$-10\log_{10}\frac{Overall\ OFDM\ symbol\ length - C_pLength}{Overall\ OFDM\ symbol\ length} \quad (5)$$

Since wavelet transforms are utilized, which have lengthier waveforms than standard transforms, discrete wavelet transform -OFDM signals overlap in period, which improves frequencies localization. As a consequence, extending the signal period to reduce intersymbol interference does not need to use a cyclic prefix. Additionally, whereas discrete wavelet transform-OFDM avoids this, Fourier transform signals pass via a rectangular waveguide, resulting in wider side lobes. Additionally, it lowers the system's difficulty from fast Fourier transforms $A\,(Zlog_2 Z)\ to\ A\,(Z)$. In addition to saving frequency by just not inserting Cyclic prefix to the Orthogonal Frequency Division multiplexing system signal, DWT-OFDM also reduces transmitting energy usage and increases system capacity [34].

1) Hybrid Pre-coding: It provides a quick overview of Fourier transforms and the CP decomposed to keep the article self-contained. Users can get additional information on the annotation and the fundamentals of Fourier transforms. A vector sometimes referred to as a manner or modal, is just a generalization of matrices to higher-order dimensions. It is possible to think of Fourier transforms including one phase and 2 modes as being represented by both vectors and matrices, accordingly. The Kronecker, outer, and Khatri-Rao products are each indicated in °,⊙,⊗ this article by the characters, and, correspondingly.

Let $Y \in D_{l1 \times l2 \times \ldots \times lu}$ indicate a vector of u−th rank with its $(l_1, l_2, \ldots l_u)$th admittance represented by$Y_{l1}, \ldots l_n$. The amount of dimensionality is represented by a tensor's order u in this case. The column and row of matrices are analogous to fibres at a higher form. By setting all indexes other than in, the $mode - u$ fibres of Y are produced as In-dimensional matrices. A vector could be divided into 2-dimensional segments by setting all but 2 of its indexes. The process that converts a vector into matrices is known as unfolding or matricization. The column of the resultant matrices is organized by the phase unfolding of a

vector$Y$, designated as $Y(n)$ and it is given in Eq. (6) and Eq. (7).

$$Y = \sum_{h=1}^{s} \mu_h d_h^{(1)} \circ d_h^{(2)} \circ \ldots \circ d_h^{(u)} \quad (6)$$

$$Y_{l1,l2,..lu} = \sum_{h=1}^{s} \mu_h d_{hl1}^{(1)} \circ d_{hl2}^{(2)} \circ \ldots \circ d_{hlu}^{(u)} \quad (7)$$

Following Eq. (8) is an expression for Y's mode-u unfolding:

$$Y_u = D_{(u)} \propto (D_{(u)} \odot \ldots . D_{(u+1)} \odot D_{(u-1)} \ldots . D_{(1)})^q \quad (8)$$

Where $\propto \cong diag(\mu_1, \ldots \mu_h)$. According to [35], the hybrid pre-coding method was implemented and the process is shown in algorithm.1. In the process, BS = Base Station, and MS = Mobile Station.

C. Hybrid Sparrow-Harris Hawks Optimization

The suggested HHO was based on a simulation of Harris hawks' foraging habits. The HHO replicates its cooperative foraging in two phases, i.e., exploration and exploitation phase—using a variety of strategies. Based on the parameter H, which represents the prey's escape energy, HHO carries out the change from the exploration to the exploitation phase. The mathematical formation is given in Eq. (9):

$$H = 2H_0(1 - i/I) \quad (9)$$

The early energy state of the prey is denoted as H_0, fluctuates at random among (1,1) for every iteration. The computation is $H_0 = 2 * ran - 1$ where I and i are the maximal and current iterations, and rand is the random integer of (0,1).

1) Searching phase: In this research, the searching phase is accomplished by the sparrow search optimization algorithm. The sparrow is highly clever and has a good memory, unlike numerous other small birds. Keep in mind that there are two different breeds of house sparrows which are kept as pets: the producer and the scrounger. While the scroungers rely on the producers to provide them with food, the producers constantly pursue for sources of food. The findings also demonstrate that the birds frequently transition between generating and scrounging behaviour patterns. Additionally, it may be noted that sparrows typically combine their producer and scrounger strategies to get food. The computational formula is developed to create the sparrow search method based on the earlier description of the sparrows. The study created equivalent rules based on the following idealized behaviour of the sparrows for convenience.

To search for food in the simulation study, virtual sparrows must be developed. The following Eq. (10) may be used to illustrate where sparrows are located:

$$A = \begin{bmatrix} A_{11} & A_{12} \ldots & A_{1m} \\ A_{21} & A_{22\ldots} & A_{2m} \\ A_{n1} & A_{n2} & A_{nm} \end{bmatrix} \quad (10)$$

Here, m is the dimensionality of the parameters that need to be optimized and n denotes the number of sparrows. The corresponding vector may therefore be used to indicate the fitness value of every sparrows as given in Eq. (11):

$$F_A = \begin{bmatrix} F([A_{11} & A_{12} \ldots & A_{1m}]) \\ F([A_{21} & A_{22\ldots} & A_{2m}]) \\ F([A_{n1} & A_{n2\ldots} & A_{nm}]) \end{bmatrix} \quad (11)$$

"Here, the amount of every row in F_A denotes the individual's fitness value and n is the number of sparrows. In the SSA, food is prioritised for producers with higher fitness ratings during the procedure of search. The producers are also responsible for charge of locating food and guiding the entire community's circulation. The producers may thus look for food in a wider variety of locations than the scavengers. The mathematic formulation is given in Eq. (12).

$$A_{i+1}^{xy} = \begin{cases} A_i^{xy} . exp\left(\frac{-t}{\beta . i_{max}}\right) \\ A_i^{xy} + R.M \end{cases} \begin{matrix} if\ D < S \\ if\ D \geq S \end{matrix} \quad (12)$$

Here, y = 1, 2,..., m, and i denotes the current iteration. The frequency of the y th dimensions of the ith sparrow at iteration i is represented by A_{xy}^{i}. The constant with the most iteration is called i_{max}. 0 and 1 are two random numbers. The alert value and the safety threshold are denoted by D and S respectively. R is a chance quantity that follows the normal distribution. M displays a $1 * m$ matrix with 1 as the value of each member. The producer switches to broad search mode at $D < S$, that indicates that there are no nearby predators. Many sparrows must immediately fly to other safe regions if $D \geq S$, which indicates that certain sparrows have identified the predator."

2) Exploitation phase: The procedure moves into the stage of development when$|H| < 1$. The prey that were followed and monitored during the exploration phase would be the subject of a raid hunt by the hawks throughout this stage. Nevertheless, when hunting in nature, the prey would often attempt to flee from the chase, hence hawks would also employ varied pursuing strategies for the various prey escape behaviour. HHO suggested four methods, which are all outlined below, to mimic this pursuing and hunting behaviour. HHO chooses a strategy based on the escape energy H and the probability of escape e.

a) Soft besiege: When $|H| \geq 0.5$ and ≥ 0.5, the prey had sufficient energy to make a random jump to safety, but hawks have encircled it. The hawk then decides to effectively hunt by using gentle besiege to deplete the prey's physical ability. Eq. (13) to Eq. (15) illustrate the computational equations.

$$A_{i+1}^{xy} = \Delta X_i^y - H * |J A_i^{p,y} - A_i^{xy}| \quad (13)$$

$$\Delta X_i^y = A_i^{p,y} - A_i^{xy} \quad (14)$$

$$J = 2(1 - e_5) \quad (15)$$

Here, e_5 is a random value between (0,1) and y is the distance between the prey's current location in the y th dimension as well as the current location of the xh hawk; where Jump (J) denotes a randomly changing jump intensity that alternates between (0,2) for each repetition.

b) Hard besiege: When the prey lacks the energy to flee and the hawks have encircled it, $|H| < 0.5$ and e <0.5, the hawks would decide to engage in a hard besiege swift raid hunt. Eq. (16) displays the behaviour model in question.

$$A_{i+1}^{xy} = A_i^{p,y} - H * |\Delta X_i^y| \quad (16)$$

c) Soft besiege with progressive rapid dives: The prey has had enough energy to break free of the siege as well as perform a zigzag motion in the situation of $|H| \geq 0.5$and $e < 0.5$, and the hawks haven't yet fully organized an surrounding denounce on the prey. The hawks now decide to concurrent using prey energy as well as eventually create a full surround. Eq. (17) to Eq. (20) illustrate the steps of this method.

$$A_{i+1}^{xy} \begin{cases} B_{i+1}^{xy} \; if \; f(B_{i+1}^{xy}) < f(B_{i+1}^{xy}) \\ C_{i+1}^{xy} if \; f(C_{i+1}^{xy}) < f(C_{i+1}^{xy}) \end{cases} \quad (17)$$

$$B_{i+1}^{xy} = A_i^{p,y} - H * |JA_i^{p,y} - A_i^{xy}| \quad (18)$$

$$C_{i+1}^{xy} = B_{i+1}^{xy} + T_y * L_y \quad (19)$$

$$L_y = L(r_y, s_y, \alpha_y) \quad (20)$$

Here T_y represents random number; L indicates Lévy flight function.

d) Hard besiege with progressive rapid dives: Hawks utilize this tactic to speed up and reduce the mean location distance among themselves and the prey in order to build a hard encircling circle earlier the raid in the situation of $|H| < 0.5$ and $e < 0.5$, when the prey lacks the energy to flee but is not entirely surrounded. Eq. (21) through Eq. (23) provide the models for this tactic.

$$\begin{cases} B_{i+1}^{xy} \; if \; f(B_{i+1}^{xy}) < f(B_{i+1}^{xy}) \\ C_{i+1}^{xy} if \; f(C_{i+1}^{xy}) < f(C_{i+1}^{xy}) \end{cases} \quad (21)$$

$$B_{i+1}^{xy} = A_i^{p,y} - H * |JA_i^{p,y} - A_i^{xy}| \quad (22)$$

$$C_{i+1}^{xy} = B_{i+1}^{xy} + T_y * L_y \quad (23)$$

The Sparrow Search Optimization (SSO) algorithm emulates the foraging and anti-predator actions of sparrows to solve optimization issues. In this algorithm, population members are separated into two primary roles: producers and scroungers. Producers are the sparrows with more energy, usually the ones that get assigned higher fitness values. They are responsible for actively searching the search space for regions with high resources (i.e., improved solutions). As soon as they find such a region, they alert other members of the group. In the event that there is danger (e.g., in the presence of a predator)—as expressed by a risk level above a particular threshold value—these producers lead the others to a safe area in the search space directly. Conversely, scroungers are less fit sparrows. Rather than searching on their own, they tail producers with the expectation of gaining access to good resources. Certain scroungers tactically monitor producers, and if they observe a producer discovering a high fitness solution, they move rapidly to that spot in a bid to upgrade their own position. If they succeed, they directly take advantage of the producer's solution; otherwise, they fall back to a rule-based updating mechanism. Population behavior demonstrates a balance natural to the system: while sparrows towards the center of the group meander randomly to preserve close proximity, those at the edges are more sensitive to threats and move quickly to less risky or more advantageous areas. The scrounger positions get updated according to certain mathematical laws in Eq. (24), that determine how they move compared to producers so as to gain opportunities for better solutions.

$$A_{i+1}^{xy} = \begin{cases} R.ep\left(\frac{A_i^{worst} - A_i^{xy}}{t^2}\right) \\ A_o^{i+1} + \left|A_i^{xy} - A_{i+1}^{xy}\right| \end{cases} \quad if \; T > n/2 \quad (24)$$

Here A_o denotes the ideal position. A_{worst} designates the world's worst location at the moment. Every member in the matrix R is given a random number between 1 and 1. The tth scrounger with the lowest fitness score is most certain to be starved when $T > n/2$.

These sparrows, which are aware of the threat, make about 10% to 20% of the overall population for the simulation study. Those sparrows were distributed at random in their starting places. The mathematical model may be depicted as follows using rules as given in Eq. (25):

$$A_{i+1}^{xy} = \begin{cases} A_i^{best}.\alpha\left|A_i^{xy} - A_i^{best}\right| & if \; f_t > f_g \\ A_{xy}^{i} + K.\left(\frac{A_i^{xy} - A_i^{worst}}{(f_t - f_g) + \varepsilon}\right) & if \; f_t = f_g \end{cases} \quad (25)$$

Here, the current global ideal location is A_i^{best}. With an average value and a variance of 0 and 1, respectively, the step size process parameters is a normal distribution of random integers. The random integer K is between [1, 1]. f_t is the current sparrow's fitness value in this instance. "The top and worst fitness values on the planet right now are f_g and f_t, respectively. To prevent a zero-division mistake, is the lowest constant.

For ease of use, when the sparrow was at the group's edge if $f_t > f_g$. A_i^{best} stands for where the population's centre is safe in that area. The sparrows inside the centre of the community are conscious of the threat and need to migrate closer to the others, as shown by the equation $f_t = f_g$. The step-size control coefficient and the sparrow's movement direction are both indicated by the letter K.

D. LSTM Based Channel Estimation

"The dispersion channels between the receiver and the transmitter in wireless communications networks must be known in order to perform coherent identification; these channels could be estimated by applying traditional estimation methods. This section introduces LSTM-based channel estimation strategies that encourage us to use DL techniques to reduce channel estimation errors. A recurrent neural network or LSTM neural network was used to examine the behaviours of the channel correlations (RNN). Fig. 4 shows the one-layer RNN's basic construction. This illustration shows that the RNN cell's input for the present time-step is its output for the previous time-step. By functioning in this way, the RNN has been able to recall previous input statistics. The computing unit, or fundamental RNN cell, executes the following computation in Eq. (26) and Eq. (27)."

$$d_t = f(R_{id}s_t + R_{dd}d_{t-1} + g_{id} + g_{dd}) \quad (26)$$

$$U_t = f(g_{do} + d_t R_{do}) \quad (27)$$

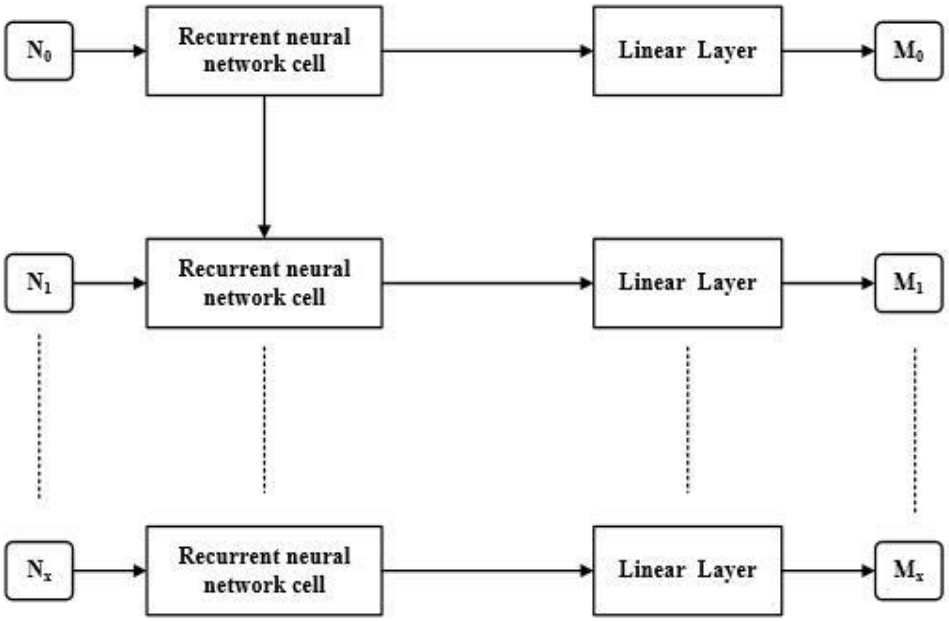


Fig. 4. Basic Structure of RNN layer.

where $f(\cdot)$ represents activation function; s_t is the input and U_t is the output; $d_t$and d_{t-1} are the hidden states at the time step t and $t-1$, correspondingly, R_{id}, R_{dd}, and R_{do} are the weights for input-hidden layer; hidden-next hidden layer, and hidden-output layer, respectively; and g_{id}, g_{dd}, and g_{do} were the corresponding biases.

"However, the basic RNN cell has some shortcomings. For starters, even if the channel at time step t has a connection to both the past and the future, it is unable to exploit the data's future information. In this case, the bidirectional network must be used to achieve higher efficiency. Second, a disadvantage of a simple RNN cell is that it cannot record long-term data. One solution to this problem is to use LSTM as an alternative. In this study, we propose a bidirectional LSTM (bi-LSTM) network for 5G channel estimation in order to overcome these problems."

The outcome of the LSTM cell's computation is indicated in the following formulae from Eq. (28) to Eq. (33).

$$f_t = f\left(V_f S_t + g_f + R_f d_{t-1}\right) \quad (28)$$

$$j_t = f\left(V_j S_t + g_j + R_j d_{t-1}\right) \quad (29)$$

$$b'_t = tand(V_b S_t + g_b + R_b d_{t-1}) \quad (30)$$

$$b_t = f_t \odot b_{t-1} + j_t \odot b'_t \quad (31)$$

$$p_t = f(V_o S_t + g_o + R_o d_{t-1}) \quad (32)$$

$$d_t = o_t \odot tand(b_t) \quad (33)$$

Where, R_f, R_j, R_b, R_o, V_f, V_j, V_b, V_o, g_f, g_j, g_b, and g_o are the respective weights of matrices and biases, as well as $tand$ is the hyperbolic tangent functional. The cell state b_t holds the critical data from the past, the forget function f_t specifies which statistics would be elapsed by the LSTM cell, and the new candidate values b'_t specifies which data would be upgraded to the cell state. The LSTM cell's hidden state function is represented by b_t and d_t. The LSTM cell has a stronger capacity to capture information than a basic RNN cell since it may avoid duplicate information by operating in this manner, allowing it to catch the vital data from the past. The bi-LSTM network's structure is depicted at the lowest part of Fig. 8. Moreover, the bi-LSTM technique combines two LSTM networks with two distinct orientations. The bi-LSTM's output considers the two LSTM cell's outputs via the linear layer as given in Eq. (34).

$$U_t = RD_t + g \quad (34)$$

Where, R and g are really the linear layer's weights and biases, and D_t is the hidden phase created by concatenating the forward-hidden phase ht with the backward-hidden phase d'_t. As a result, the bi-LSTM technique may take advantage of the relationship between the information in the present time step and both the past and the future. This research first collects the LS-estimated channels from across all antennas in order to apply the bi-LSTM framework to the network. "Next, it specifies a realization of the given data (input) for the procedure of training in Eq. (35).

$$X_{n-bi-LSTM} = \left\{\left[Re\{\hat{d}^n_{LS}(0)\}; Im\{\hat{d}^n_{LS}(0)\}\right], \ldots\ldots, \left[Re\{\hat{d}^n_{LS}(l-1)\}; Im\{\hat{d}^n_{LS}(l-1)\}\right]\right\} \quad (35)$$

l represents the sequence length taken into account for the bi-LSTM network. The number of characteristics for the input is 2NT×NR since the bi-LSTM's input, $\hat{d}^n_{LS}$ indicates the LS-estimated channel for entire NT×NR channel streams. The equivalent true channel is the bi-LSTM network's output as shown in Eq. (36).

$$O_{n-bi-LSTM} = \left\{\left[Re\{\hat{d}^n(0)\}; Im\{\hat{d}^n(0)\}\right], \ldots\ldots, \left[Re\{\hat{d}^n(l-1)\}; Im\{\hat{d}^n(l-1)\}\right]\right\} \quad (36)$$

The MSE loss function is taken into account since the goal of utilising a bi-LSTM network was to reduce the MSE between

the anticipated channel and the real channel. The bi-LSTM network's goal function is Eq. (37).

$$F_{bi-LSTM}(\mathcal{R}, A) = {}^{1}/_{Nl} \sum_{n=1}^{N} \sum_{j=0}^{l-1} \left\| \hat{d}^{n}(i) - d^{n}(i) \right\|_{2}^{2} \quad (37)$$

Where, N represents the training sample's total number, $\hat{d}^{n}(i)$ represents the actual channel conforming to $\hat{d}^{n}(i)$; $\mathcal{R}$ and A were all the bi-LSTM weights and biases, and the superscript n signifies the nth trained model. By employing gradient descent techniques to update $\mathcal{R}$ and A, the loss function could be reduced. The inadequate channel state data as a potential future expansion of the approach because this research assumes that the ideal instantaneous channels are accessible for the training stage.

The novelty of the proposed approach is integrating a hybrid metaheuristic optimization framework with deep learning overcoming the channel estimation of 5G MIMO-OFDM systems with great success. This method differs with traditional ones which involve only Deep learning or traditional estimators like LS, LMMSE and picks out the best of the two HHO and Sparrow Search Algorithm technologies to tune the parameters of a Long Short-Term Memory (LSTM). This HHO-SSA-LSTM hybrid model has been successful in adjusting itself to channel conditions dynamically in response to environment high-mobility and Doppler effects. Furthermore, OFDM modulation which is based on discrete wavelet transform (DWT) is used to enhance spectral efficiency and eliminate the effects of inter-symbol interference which needs not to use any cyclic prefixes. The metaheuristic layer globally optimizes the learning parameters in LSTM, resulting at such, the parameters capture better temporal relationships of the wireless channel, as well as nonlinear attributes of the wireless channel. This coupled architecture enables more efficient and robust channel estimation mechanism to have a lower BER, MSE, and SER than the existing approaches. In general, the suggested hybrid structure overcomes the weakness of the current estimation methods due to the smart process of the parameter's optimization and dynamic learning in a real-time communication case."

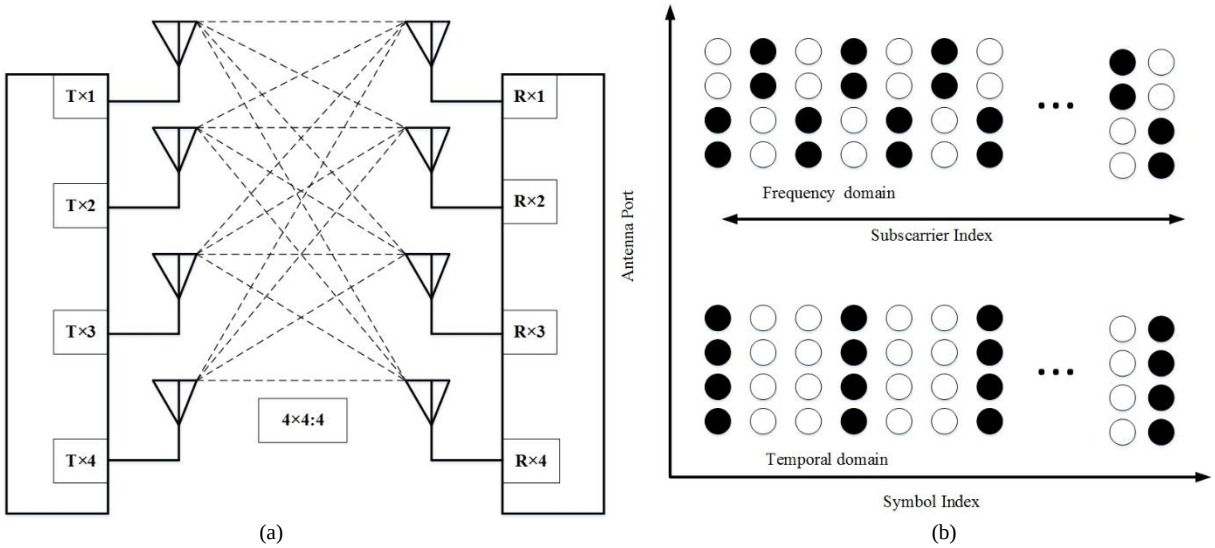
IV. Result and Discussion

This section evaluates the suggested hybrid optimization with DL estimates over the 5G channel profile's performance and compares with conventional approaches, such as LS, CNN, LSTM, FDNN, and LMMSE. Additionally, give a justification for each received outcome. The simulation's settings have been stated first, and the outcomes have been then shown and discussed. The parameters employed for MIMO-OFDM system is shown in Table I.

TABLE I. Parameter Setup

Parameters	Values
Modulation type	16-QAM
Model of noise	Gaussian Noise
FFT size	256
Sample frequency	3.84MHz
Cyclic prefix	24
Subcarrier spacing	15kHz
MIMO	4 × 4

Four antennas are used in 4x4 MIMO, also referred as 4T4R, to connect up to 4-data streams to the recipient device. 4x4 networks offer up to a 400% throughput boost over standard single antenna (SISO) networks. Construction complexity significantly rises as compared to earlier 2x2 antennas. Since the majority of manufacturers cannot offer four unique polarizations, nominal MIMO arrangements are the most typical. Additionally, 4x4 MIMO is up to 160% quicker than 2x2 MIMO in weak signal settings and roughly 90% faster in good signal circumstances. The employed 4x4 MIMO and the used pilot structure for the presented MIMO-OFDM construction are indicated in Fig. 5 (a) and (b), correspondingly. In Fig. 5 (a), T×1 and T×4 seems to be the transmitted antenna-indices and R×1 and R×4 seems to be the receiver antenna-indices.



(a) (b)

Fig. 5. Presented MIMO-OFDM (a) 4×4 MIMO (b) Pilot structure.

The scenario relating to the mobile velocity has been employed to evaluate the effectiveness of all the possible channel estimations. The outcomes of the simulation have been achieved for different models using different performance measures in order to evaluate the suggested DL model-based estimation's effectiveness. The suggested technique has been applied in MATLAB, and the experimental outcomes were obtained. The suggested algorithm evaluates the efficiency of the channel by adjusting the network's SNR in relation to the quantity of sender and recipient arrays. Moreover, the simulation was being run to assess the effectiveness of the HHO-SSA-LSTM-aided estimation, and the outcomes have been contrasted with the traditional LS, CNN, LSTM, FDNN, and LMMSE-based estimation using the BER, SER, and MSE vs SNR, efficiency, and execution time.

A. Performance Analysis

The presented channel estimation strategy's performance is assessed in regards to BER, SER, and MSE for SNR, efficiency, and execution time. The HHO-SSA-LSTM's measures have been contrasted with those of the conventional LS, CNN, LSTM, FDNN, and LMMSE. Additionally, the effectiveness of alternative channel estimation algorithms is assessed across a range of pilot lengths. The description of SNR, BER, SER, and MSE are described as follows:

- SNR: The noise effectiveness, and consequently the susceptibility of radio receivers, could be examined in a variety of methods. Comparing the signal as well as noise rates for a particular signal level, or SNR, seems to be the most straightforward technique. It goes without saying that the bigger the SNR—the difference among the signals and the undesired noise—the greater the radio recipient's sensitivity effectiveness.
- BER and SER: The chances of obtaining a symbol and bit in errors were denoted by SER and BER, correspondingly. By recreating a whole system with a large number of bits as well as comparing the proportion of bits or symbols obtained in error to the overall number-of-bits, BER and SER may be estimated. Hence SER is computed by Eq. (38).

$$BER = {N_S(E)}/{T_N(TS)} \quad (38)$$

Where, $T_N(TS)$ indicates total no. of transferred symbols and $N_B(E)$ indicates no. of symbols in error.

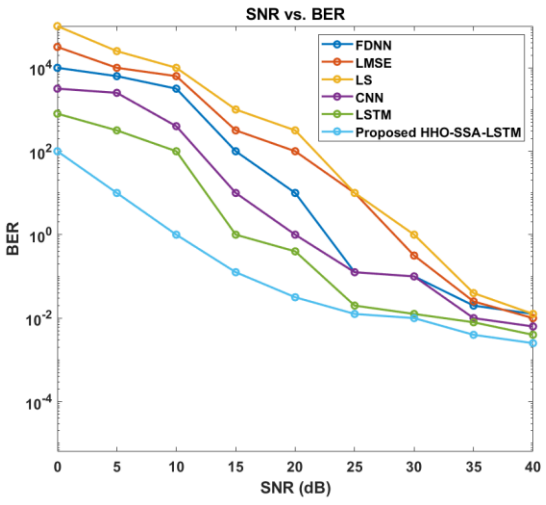
Moreover, BER is computed by Eq. (39).

$$BER = {N_B(E)}/{T_N(TB)} \quad (39)$$

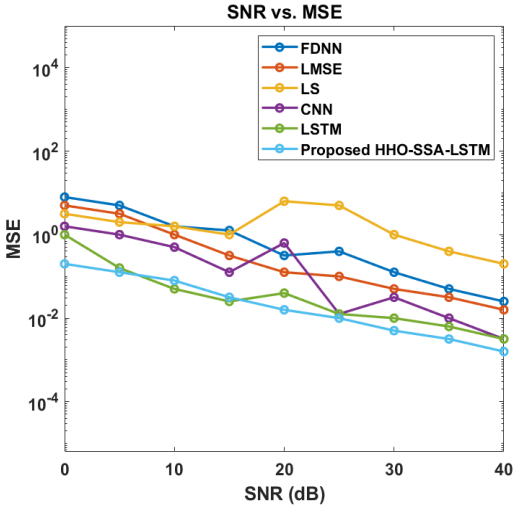
Where, $T_N(TB)$ indicates total no. of transferred bits and $N_B(E)$ indicates no. of bits in error.

- MSE: In statistics, MSE seems to be a measurement of the mean square error. It is an expected value-corresponding risk function. It is random, hence it could hasn't ever be null. It has been always employed in a non-negative context when assessing estimator performance. MSE seems to be a second moment that considers the sources of error as well as estimation variation. An unbiased estimation with invariance of estimate is the MSE. The mean squared variances between the predicted and real values have been used to compute it.

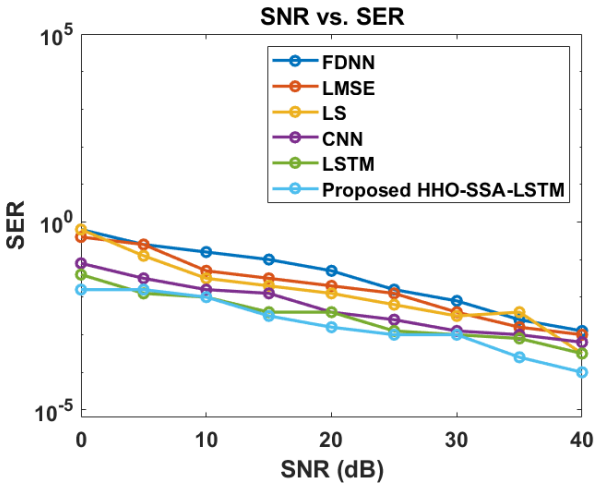
1) Performance evaluation using different pilot lengths: Under several pilot lengths (PL) (i.e., PL =128, 136, and160), the assessment metrics of several channel estimation algorithms are analyzed. Fig. 6 compares the BER, SER, and MSE of several channel estimate algorithms when the PL has been set to 128. As may be seen in Fig. 6 (a), BER (Bit Error Rate) decreases as SNR (Signal to Noise Ration) rises. The LSTM's BER is lowered to 30% less compared to the traditional CNN as long as the weight variables have been chosen for it in the best way possible. However, FDNN has been computationally more efficient than LMSE, and as a result, its BER was 50% lower contrasted to LMSE. In this method, HHO-SSA has been introduced to adjust the weight variables of LSTM in order to produce improved prediction outcome. As a result, the HHO-SSA-LSTM's BER is lower than those of the FDNN, LMSE, LS, CNN, and LSTM by 50%, 75%, 90%, 88%, and 90%, correspondingly.



(a)



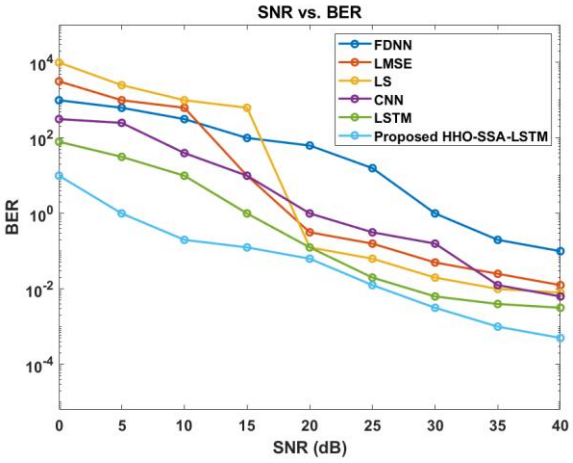
(b)



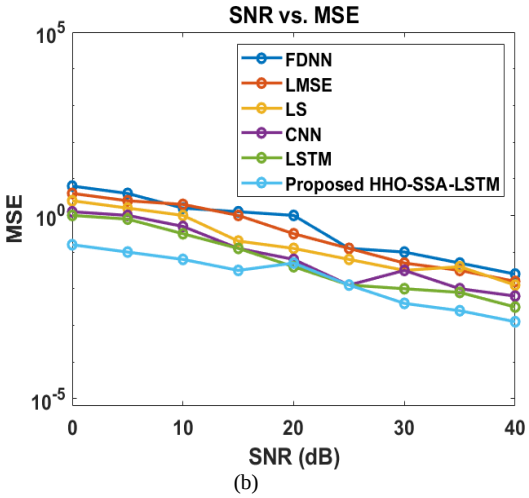
(c)

Fig. 6. "Performance of different channel estimation algorithms when PL=128 (a) BER vs SNR (b) MSE vs SNR (c) SER vs SNR"

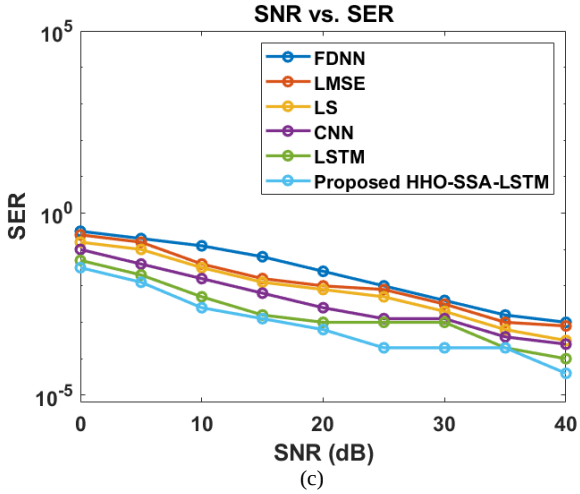
The HHO-SSA-LSTM's MSE (Mean Square Error) is lowered to 40%, 55%, 70%, 85%, and 92% less than that of the FDNN, LMSE, LS, CNN, and LSTM, correspondingly, when PL=128 as demonstrated in Fig. 6 (b). Additionally, as illustrated in Fig. 6 (c), the SER (Symbol Error Rate) of the HHO-SSA-LSTM was lower at PL=128 than it is for the FDNN, LMSE, LS, CNN, and LSTM at 50%, 53%, 64%, 79%, and 82%, correspondingly.



(a)



(b)



(c)

Fig. 7. "Performance of different channel estimation algorithms when PL=136 (a) BER vs. SNR (b) MSE vs. SNR (c) SER vs. SNR"

Fig.7 compares the BER, SER, and MSE of several channel estimate algorithms when PL =136. As may be seen in Fig. 7 (a), BER decreases as SNR rises. The CNN's BER is lowered to 75% less compared to the FDNN; moreover, LMSE's BER is mitigated to 60% than that of CNN. However, LSTM has been computationally more efficient than LS, and as a result, its BER was 25% lower contrasted to LS. In this method, HHO-SSA-LSTM has been introduced to offer improved prediction outcome. As a result, the HHO-SSA-LSTM's BER is lower than those of the all-other methods like FDNN, LMSE, LS, CNN, and LSTM by 20%, 53%, 72%, 80%, and 88%, correspondingly. In addition, HHO-SSA-LSTM's MSE is lowered to 53%, 65%, 72%, 78%, and 85% less than that of the FDNN, LMSE, LS, CNN, and LSTM, respectively, when PL=136 as depicted in Fig. 7 (b). Additionally, as represented in Fig. 7 (c), the SER of the HHO-SSA-LSTM was lower at PL=136 than it is for the FDNN, LMSE, LS, CNN, and LSTM at 65%, 70%, 75%, 89%, and 94%, respectively.

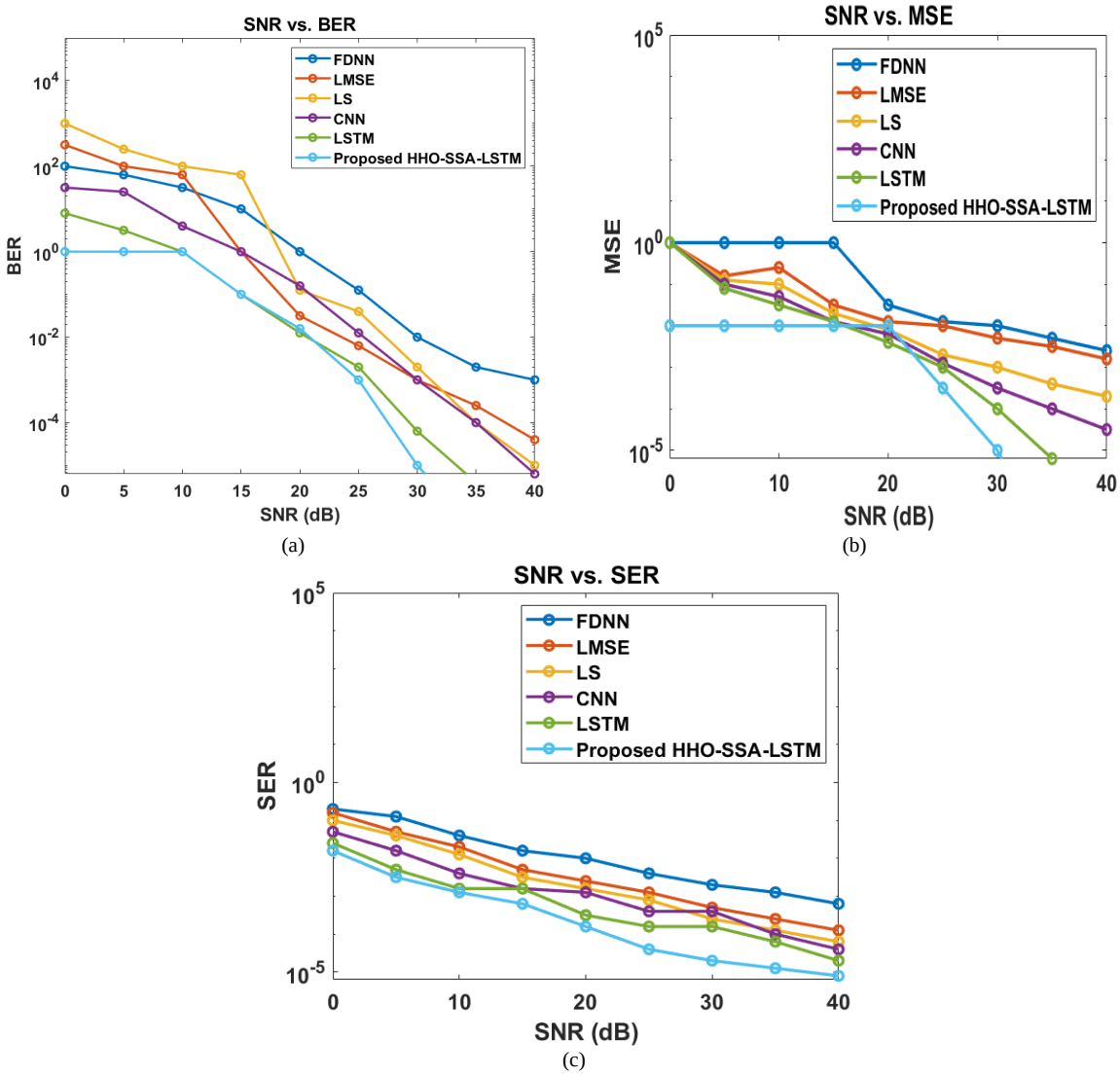


Fig. 8. "Performance of different channel estimation algorithms when PL=160 (a) BER vs. SNR (b) MSE vs. SNR (c) SER vs. SNR".

Fig. 8 compares the BER, SER, and MSE of several channel estimate techniques under the PL=160 condition. As shown in Fig. 8 (a), the BER of HHO-SSA-LSTM was lower than that of FDNN, LS, CNN, LMSE, and LSTM, falling to 76%, 80%, 42%, 83%, and 91%, accordingly. Additionally, as shown in Fig. 8 (b), the MSE of the HHO-SSA-LSTM has been lowered to 73%, 78%, 75%, 80%, and 86%, as opposed to the FDNN, LS, CNN, LMSE, and LSTM, correspondingly. Additionally, as shown in Fig. 8 (c), the SER of the HHO-SSA-LSTM significantly lowered to 58%, 69%, 76%, 84%, and 92%, as opposed to the FDNN, LS, CNN, LMSE, and LSTM, respectively. In addition, the efficiency of the different channel estimation algorithms is also computed for different pilot lengths. Efficiency seems to be the maximum degree of effectiveness that needs the fewest inputs and creates the greatest number of outputs. Efficiency appeals for cutting back on the amount of resources, particularly energy, that are utilized inefficiently to generate a given outcome. Additionally, efficiency seems to be a quantifiable notion that may be assessed by comparing the usable outcome to the entire input.

TABLE II. EFFICIENCY OF HHO-SSA-LSTM OVER FDNN, LSME, LS, CNN, AND LSTM WHEN NUMBER OF PL VARIED

Methods	Efficiency (%)		
	PL = 128	PL = 136	PL = 160
FDNN [36]	85.8	87.64	88.5
LSME [37]	87.7	88.9	90.2
LS [38]	81.1	84.7	87.9
CNN [39]	79.5	84.1	85
LSTM [40]	92.3	94.1	94.91
Proposed HHO-SSA-LSTM	96.41	96.75	96.92

The obtained efficiency is represented in Table II and Fig. 9. The result indicated that the presented HHO-SSA-LSTM has attained higher efficiency under PL = 128, 136, and 160 over other methods. The efficiency of LSTM is increased to 9%, 10%, and 11% than CNN under PL = 128, 136, and 160, respectively. Moreover, when PL = 128, the presented model is enhanced to 11%, 11%, 10%, 9%, and 4% as compared to FDNN, LSME, LS, CNN, and LSTM, respectively. Also, when PL = 136, the presented model is enhanced to 11%, 12%, 10%, 11%, and 2% as compared to FDNN, LSME, LS, CNN, and LSTM, respectively. In addition, when PL = 160, the presented model is enhanced to 12%, 15%, 12%, 11%, and 2% as compared to FDNN, LSME, LS, CNN, and LSTM, correspondingly.

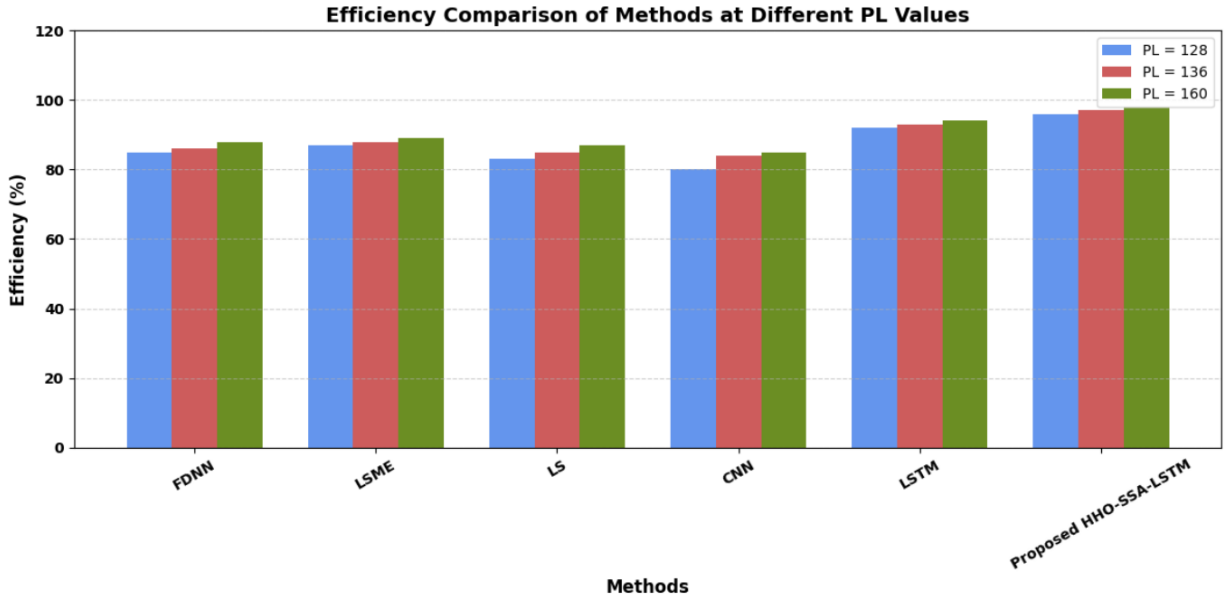


Fig. 9. Efficiency of HHO-SSA-LSTM over other methods.

2) Complexity analysis: The computational assessment of various channel estimate strategies in regards to execution time was shown in Fig. 10 and Table III. Every simulation and execution time measurement was carried out on a system that has an NVIDIA GeForce GTX 1660 Ti GPU, 16 GB of RAM, and an Intel Core i7-10750H CPU running at 2.60GHz. MATLAB R2021b was used for the implementation, and Windows 10 Pro (64-bit) was used. In order to ensure uniformity across all compared models, no GPU acceleration was used during execution time testing. As seen in the table, traditional channel estimation algorithms take longer to execute than DL-based ones due of their computational complexity. However, contrasted to FDNN, LSME, LS, CNN, and LSTM, the given HHO-SSA-LSTM takes a long time to execute. MATLAB's tic and toc functions were used to measure the execution time on a Windows 10 computer running an Intel Core i7-11700 CPU with 32 GB of RAM. After ten runs of each method, the average execution time was noted. Because of its hybrid optimization structure, the suggested HHO-SSA-LSTM model took 22 seconds, which is a little longer than other models. However, this is justified by the fact that it estimates channels with much greater accuracy and efficiency. The computational complexity has grown as a result of the integrated HHO and SSA approaches.

TABLE III. COMPLEXITY ASSESSMENT IN REGARDS TO EXECUTION TIME

Methods	Execution time (sec)
FDNN	21
LSME	18
LS	20
CNN	16
LSTM	10
Proposed HHO-SSA-LSTM	22

Contrasted to the proposed method, the other methods like FDNN, LMSE, LS, CNN, and LSTM has attained less execution time as 21s, 18s, 20s, 16s, and 10s, respectively. Those methods have less complexity due to implementation of single models. However, the presented model was a hybridized method; thus, the execution time was increased than other methods. Moreover, the execution time of the presented HHO-SSA-LSTM is 22s.

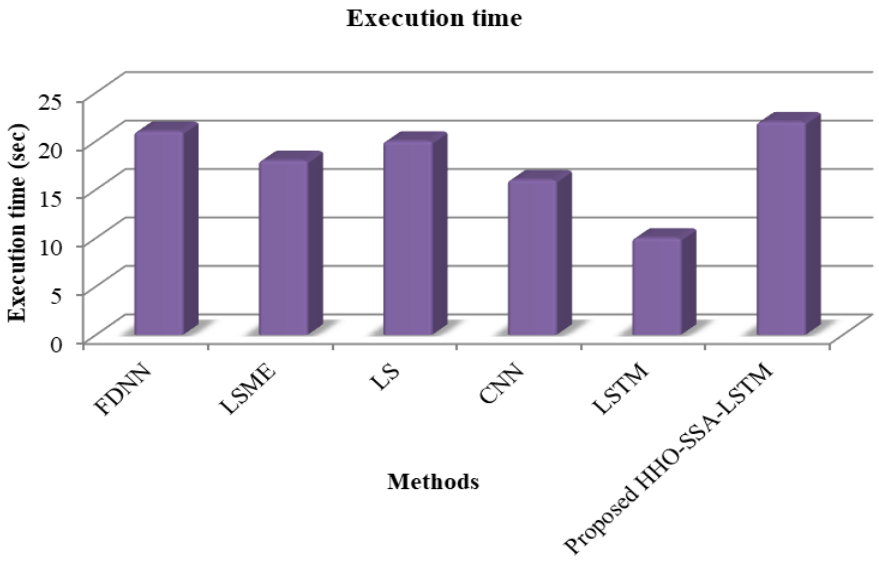


Fig. 10. Execution time.

B. Execution Time Comparison

"The comparative study of the suggested technique with the methods in the prior literature is illustrated in Table IV and Fig. 11. The presented HHO-SSA-LSTM method was compared with other existing methods like Compressive Sensing (CS) [24], OCEAN [41], and PSO-Adam-LSTM [42] in terms of execution time.

TABLE IV. COMPARATIVE ASSESSMENT

Reference	Method	Execution time (s)
[24]	CS	9
[41]	OCEAN	4
[42]	PSO-Adam-LSTM	20
Proposed	HHO-SSA-LSTM	22

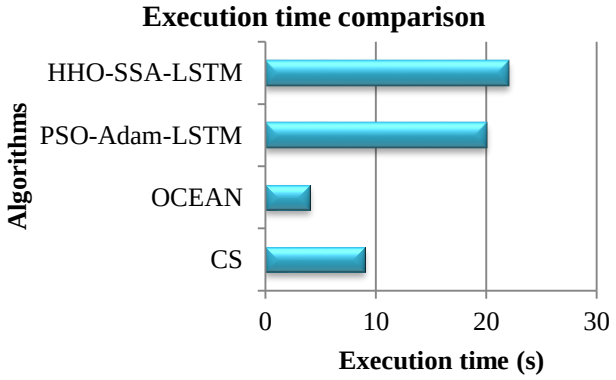


Fig. 11. Execution time comparison.

Haythem Bany Salameh et al. [24] achieved an execution time of 9s, as shown in the table. The researchers obtained highest MSE although execution time was less than the recommended technique. The method put out by Changqing Luo et al. [41] has attained execution time as 4s. Unfortunately, this takes the least amount of time to execute contrasted to the other techniques in the chart. In contrast to the suggested plan, Lipsa and Anand [42] were able to achieve an execution time of 20s. Furthermore, the provided work seems to have a longer execution time than these studies. However, the presented model has attained maximum BER and MSE as less, which is less than the other works.

C. Discussion

The proposed HHO-SSA-LSTM model outperformed traditional and deep learning-based channel estimation techniques—such as LS, LMMSE, CNN, LSTM, and FDNN—on various pilot lengths and signal-to-noise ratio scenarios consistently. The experimental findings showed considerable decreases in BER, SER, and MSE, especially at larger pilot lengths (PL=160), in which the suggested model attained as much as 91% decreased BER and 86% decreased MSE against isolated LSTM. Additionally, the model attained the most efficient estimation, with 96.92%, outperforming even sophisticated DL approaches. While the hybrid model added computational complexity, resulting in the longest execution time (22s), this comes at the expense of the significant improvements in estimation reliability and accuracy [43]. These results have been supported by Mahdi and Deniz [26], who showed that deep learning-based channel estimation models typically have higher execution times because of architectural complexity, comparable to the trade-offs seen in our suggested HHO-SSA-LSTM model, even though they improve accuracy over traditional methods.

The suggested approach is superior despite having a longer execution time because it provides significantly higher accuracy and efficiency, both of which are essential in high-reliability 5G MIMO-OFDM situations. The research presents a compelling argument for the necessity and justification of the trade-off, particularly for applications where accuracy is more important than speed. In comparison with other current benchmark solutions such as CS, OCEAN, and PSO-Adam-LSTM, the proposed scheme represents a reliable solution for channel estimation in 5G MIMO-OFDM systems [44], ideally applicable

for high-reliability and high-capacity communication scenarios. There are several restrictions to take into account, even with the HHO-SSA-LSTM framework's greater accuracy and efficiency. When compared to single-model approaches, the hybrid structure results in longer execution times due to increased computational overhead. Its direct application in real-time or energy-constrained applications may be limited as a result. Additionally, in low-pilot or rapidly changing settings, the model may not always have access to the full channel state information needed for training. The following studies will concentrate on lowering computational complexity to optimize the model for real-time deployment. Additionally, transfer learning and adaptive techniques will be explored to enhance performance in dynamic or low-pilot environments. There will also be consideration of integration with new 6G technologies including THz-band systems and RIS.

V. Conclusion and Future Works

This paper suggested a hybrid deep learning-based channel estimation scheme, HHO-SSA-LSTM, for 5G MIMO-OFDM systems under multi-path fading and Doppler shift. By combining Harris Hawks Optimization and Sparrow Search Algorithm to adjust LSTM parameters, the model showed notably reduced BER, SER, and MSE for various pilot lengths. The performance measurement showed up to 91% BER improvement and 86% MSE reduction in comparison with traditional and deep learning benchmarks like LS, LMMSE, CNN, FDNN, and LSTM. In addition, the novel approach attained maximum efficiency (96.92%) at a pilot size of 160, outperforming the available methods in all instances. The results prove the accuracy and resilience of the HHO-SSA-LSTM technique for practical application scenarios in real-world 5G channel estimation.

Although it has higher accuracy, the proposed model has two primary drawbacks: (1) higher computational complexity leading to longer runtime time (22s) than basic models, and (2) its requirement of a completely known channel profile at training time, which could restrict usability in fast-changing or low-pilot environments. Its computational overhead should be minimized in future studies by developing lightweight variants of the hybrid model for real-time embedded applications. Moreover, being able to extend the model to enable transfer learning and adaptive generalization across various channel conditions will make it more scalable for the next-generation wireless systems, such as 6G and beyond."

References

[1] J. G. Andrews et al., "What Will 5G Be?," IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1065–1082, June 2014, doi: 10.1109/JSAC.2014.2328098.

[2] X. Wang et al., "Millimeter Wave Communication: A Comprehensive Survey," IEEE Commun. Surv. Tutor., vol. 20, no. 3, pp. 1616–1653, 2018, doi: 10.1109/COMST.2018.2844322.

[3] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, and B. Ottersten, "Reconfigurable Intelligent Surface-Assisted Cell-Free Massive MIMO Systems Over Spatially-Correlated Channels," 2021, doi: 10.48550/ARXIV.2104.08648.

[4] J. Tanveer, A. Haider, R. Ali, and A. Kim, "Machine learning for physical layer in 5G and beyond wireless networks: A survey," Electronics, vol. 11, no. 1, p. 121, 2022.

[5] A. L. Ha, T. Van Chien, T. H. Nguyen, W. Choi, and V. D. Nguyen, "Deep Learning-Aided 5G Channel Estimation," in 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea (South): IEEE, Jan. 2021, pp. 1–7. doi: 10.1109/IMCOM51814.2021.9377351.

[6] C. Tong, "5G Technology Components and Material Solutions for Hardware System Integration," in Advanced Materials and Components for 5G and Beyond, C. Tong, Ed., in Springer Series in Materials Science. , Cham: Springer Nature Switzerland, 2022, pp. 1–32. doi: 10.1007/978-3-031-17207-6_1.

[7] R. Chataut and R. Akl, "Massive MIMO Systems for 5G and beyond Networks—Overview, Recent Trends, Challenges, and Future Research Direction," Sensors, vol. 20, no. 10, Art. no. 10, Jan. 2020, doi: 10.3390/s20102753.

[8] A. R. Mishra, Fundamentals of network planning and optimisation 2G/3G/4G: evolution to 5G. John Wiley & Sons, 2018.

[9] S. Hossain, "5G wireless communication systems," Am. J. Eng. Res. AJER, vol. 2, no. 10, pp. 344–353, 2013.

[10] K. S. Kim et al., "Ultrareliable and Low-Latency Communication Techniques for Tactile Internet Services," Proc. IEEE, vol. 107, no. 2, pp. 376–393, Feb. 2019, doi: 10.1109/JPROC.2018.2868995.

[11] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, "5G Deployment: Standalone vs. Non-Standalone from the Operator Perspective," IEEE Commun. Mag., vol. 58, no. 11, pp. 83–89, Nov. 2020, doi: 10.1109/MCOM.001.2000230.

[12] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," J. Ambient Intell. Humaniz. Comput., Feb. 2021, doi: 10.1007/s12652-020-02521-x.

[13] Z. Zhang, J. Chen, and J. Hou, "Analysis of the Evolution Scheme of NSA RAN Sharing to SA RAN Sharing," in 2021 International Wireless Communications and Mobile Computing (IWCMC), June 2021, pp. 204–211. doi: 10.1109/IWCMC51323.2021.9498959.

[14] I. Makhdoom, M. Abolhasan, and J. Lipman, "A comprehensive survey of covert communication techniques, limitations and future challenges," Comput. Secur., vol. 120, p. 102784, Sept. 2022, doi: 10.1016/j.cose.2022.102784.

[15] Aida Zaier and R. Bouallegue, "Channel Estimation Study for Block - Pilot Insertion in OFDM Systems Under Slowly Time Varying Conditions," Int. J. Comput. Netw. Commun., vol. 3, no. 6, pp. 39–54, Nov. 2011, doi: 10.5121/ijcnc.2011.3603.

[16] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," Found. Trends® Signal Process., vol. 11, no. 3–4, pp. 154–655, 2017, doi: 10.1561/2000000093.

[17] R. Dai, Y. Liu, Q. Wang, Y. Yu, and X. Guo, "Channel estimation by reduced dimension decomposition for millimeter wave massive MIMO system," Phys. Commun., vol. 44, p. 101241, Feb. 2021, doi: 10.1016/j.phycom.2020.101241.

[18] N. Shalavi, M. Atashbar, and M. Mohassel Feghhi, "Downlink channel estimation of FDD based massive MIMO using spatial partial-common sparsity modeling," Phys. Commun., vol. 42, p. 101138, Oct. 2020, doi: 10.1016/j.phycom.2020.101138.

[19] Z. Liu, L. Zhang, and Z. Ding, "Overcoming the Channel Estimation Barrier in Massive MIMO Communication Systems," 2019, doi: 10.48550/ARXIV.1912.10573.

[20] M. Boloursaz Mashhadi and D. Gündüz, "Deep Learning for Massive MIMO Channel State Acquisition and Feedback," J. Indian Inst. Sci., vol. 100, no. 2, pp. 369–382, Apr. 2020, doi: 10.1007/s41745-020-00169-2.

[21] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both?," IEEE Trans. Commun., vol. 67, no. 10, pp. 7331–7376, Oct. 2019, doi: 10.1109/TCOMM.2019.2924010.

[22] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," China Commun., vol. 14, no. 11, pp. 92–111, Nov. 2017, doi: 10.1109/CC.2017.8233654.

[23] H. Huang et al., "Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions," IEEE Wirel. Commun., vol. 27, no. 1, pp. 214–222, Feb. 2020, doi: 10.1109/MWC.2019.1900027.

[24] Z. Albataineh, K. Hayajneh, H. Bany Salameh, C. Dang, and A. Dagmseh, "Robust massive MIMO channel estimation for 5G networks using compressive sensing technique," AEU - Int. J. Electron. Commun., vol. 120, p. 153197, June 2020, doi: 10.1016/j.aeue.2020.153197.

[25] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-Driven Deep Learning Based Channel Estimation and Feedback for Millimeter-Wave Massive Hybrid MIMO Systems," IEEE J. Sel. Areas Commun., vol. 39, no. 8, pp. 2388–2406, Aug. 2021, doi: 10.1109/JSAC.2021.3087269.

[26] M. B. Mashhadi and D. Gunduz, "Pruning the Pilots: Deep Learning-Based Pilot Design and Channel Estimation for MIMO-OFDM Systems," IEEE Trans. Wirel. Commun., vol. 20, no. 10, pp. 6315–6328, Oct. 2021, doi: 10.1109/TWC.2021.3073309.

[27] O. A. Saraereh, I. Khan, Q. Alsafasfeh, S. Alemaishat, and S. Kim, "Low-Complexity Channel Estimation in 5G Massive MIMO-OFDM Systems," Symmetry, vol. 11, no. 5, p. 713, May 2019, doi: 10.3390/sym11050713.

[28] L. Ge et al., "Deep Neural Network Based Channel Estimation for Massive MIMO-OFDM Systems With Imperfect Channel State Information," IEEE Syst. J., vol. 16, no. 3, pp. 4675–4685, Sept. 2022, doi: 10.1109/JSYST.2021.3114229.

[29] N. Anughna and M. Ramesha, "Performance Analysis on various Diversity Schemes with Channel Equalization and Estimation Techniques in MIMO OFDM system," in 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India: IEEE, May 2022, pp. 1–5. doi: 10.1109/INCET54531.2022.9825175.

[30] L. Ge, Y. Zhang, G. Chen, and J. Tong, "Compression-Based LMMSE Channel Estimation With Adaptive Sparsity for Massive MIMO in 5G Systems," IEEE Syst. J., vol. 13, no. 4, pp. 3847–3857, Dec. 2019, doi: 10.1109/JSYST.2019.2897862.

[31] Y. Zhang, X. Zhu, Y. Liu, Y. Jiang, Y. L. Guan, and V. K. N. Lau, "Hierarchical BEM Based Channel Estimation With Very Low Pilot Overhead for High Mobility MIMO-OFDM Systems," IEEE Trans. Veh. Technol., vol. 71, no. 10, pp. 10543–10558, Oct. 2022, doi: 10.1109/TVT.2022.3184361.

[32] W. Ji, F. Zhang, and L. Qiu, "Multipath Extraction Based UL/DL Channel Estimation for FDD Massive MIMO-OFDM Systems," IEEE Access, vol. 9, pp. 75349–75361, 2021, doi: 10.1109/ACCESS.2021.3081497.

[33] M. A. Gupta, M. P. Nigam, and V. Chaurasia, "A Review on wavelet transform as a substitute to cyclic prefix removal in FFT in OFDM," Int. J. Eng. Sci. Res. Technol., 2014.

[34] A. Al-Jzari and K. Iviva, "Cyclic prefix length determination for orthogonal frequency division multiplexing system over different wireless channel models based on the maximum excess delay spread," Am. J. Eng. Appl. Sci., vol. 8, no. 1, p. 82, 2015.

[35] M. G. Kallur, "Hybrid PSO-GSA Algorithm for Channel Estimation in Massive MIMO System," J. Netw. Commun. Syst., vol. 4, no. 3, Art. no. 3, July 2021, Accessed: Feb. 17, 2023. [Online]. Available: https://publisher.resbee.org/admin/index.php/jnacs/article/view/17

[36] L. Ge et al., "Deep neural network-based channel estimation for massive MIMO-OFDM systems with imperfect channel state information," IEEE Syst. J., vol. 16, no. 3, pp. 4675–4685, 2021.

[37] L. Ge, Y. Zhang, G. Chen, and J. Tong, "Compression-based LMMSE channel estimation with adaptive sparsity for massive MIMO in 5G systems," IEEE Syst. J., vol. 13, no. 4, pp. 3847–3857, 2019.

[38] E. Björnson, J. Hoydis, L. Sanguinetti, and others, "Massive MIMO networks: Spectral, energy, and hardware efficiency," Found. Trends® Signal Process., vol. 11, no. 3–4, pp. 154–655, 2017.

[39] M. B. Mashhadi and D. Gündüz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems," IEEE Trans. Wirel. Commun., vol. 20, no. 10, pp. 6315–6328, 2021.

[40] O. A. Saraereh, I. Khan, Q. Alsafasfeh, S. Alemaishat, and S. Kim, "Low-complexity channel estimation in 5G massive MIMO-OFDM systems," Symmetry, vol. 11, no. 5, p. 713, 2019.

[41] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel State Information Prediction for 5G Wireless Communications: A Deep Learning Approach," IEEE Trans. Netw. Sci. Eng., vol. 7, no. 1, pp. 227–236, Jan. 2020, doi: 10.1109/TNSE.2018.2848960.

[42] L. Dash and A. S. Thampy, "Channel estimation using hybrid optimizer based recurrent neural network long short-term memory for MIMO communications in 5G network," SN Appl. Sci., vol. 5, no. 2, p. 60, Feb. 2023, doi: 10.1007/s42452-022-05253-z.

[43] C. A. ul Hassan et al., "Optimizing Deep Learning Model for Software Cost Estimation Using Hybrid Meta-Heuristic Algorithmic Approach," Comput. Intell. Neurosci., vol. 2022, no. 1, p. 3145956, 2022.

[44] L. Dash and A. S. Thampy, "Channel estimation using hybrid optimizer based recurrent neural network long short term memory for MIMO communications in 5G network," SN Appl. Sci., vol. 5, no. 2, p. 60, 2023.

Explainable Approach Using Semantic-Guided Alignment for Radiology Imaging Diagnosis

Fatima Cheddi[1], Ahmed Habbani[2], Hammadi Nait-Charif[3]
Smart Systems Laboratory ENSIAS, Mohammed V University in Rabat, Rabat, Morocco[1, 2]
National Center for Computer Animation, Bournemouth University, United Kingdom[3]

***Abstract*—The increased success of deep learning in the radiology imaging domain has significantly advanced automated diagnosis and report generation, aiming to enhance diagnostic precision and clinical decision-making. However, existing methods often struggle to achieve detailed morphological description, resulting in reports that provide only general information without precise clinical specifics and thus fail to meet the stringent interpretability requirements of medical diagnosis. Also, the critical need for transparency in clinical automated systems has catalyzed the emergence of explainable artificial intelligence (XAI) as an essential research frontier. To address these limitations, we propose an explainable system for report generation that leverages semantic-guided alignment and interpretable multimodal deep learning. Our model combines hierarchical semantic feature extraction from medical reports with fine-grained features that guide the model to focus on lesion-relevant visual features and use Concept Activation Vectors (CAVs) to explain how radiological concepts affect report generation. A contrastive multimodal fusion module aligning textual and visual modalities through hierarchical attention and contrastive learning. Finally, an integrated concept activation system that provides transparent explanations by quantifying how radiological concepts influence generated reports. Validation of our approach in comparisons with existing methods indicates a corresponding boost in report quality in terms of clinical accuracy of the description, localization of the lesion, and contextual consistency, positioning our framework as a robust tool for generating more accurate and reliable medical reports.**

Keywords*—Automated report generation; explainable AI; cross-modal fusion; contrastive learning; semantic-guided alignment

I. Introduction

The rapid success of deep learning in the medical domain has created transformative possibilities for improving diagnostic accuracy and clinical decision-making, particularly through AI-powered automated report generation. This solution enhances workflow efficiency, reduces the workload on radiologists, minimizes diagnostic errors, and improves the efficiency of patient treatment [1]. However, the integration of artificial intelligence in radiology has reached a critical inflection point, where diagnostic accuracy must now be matched by clinical interpretability to enable real-world adoption [2].

Recently, the increasing availability of the use of multimodal data in image analysis has created new opportunities for automating the generation of medical reports. Current models rely on multi-modal learning techniques that jointly process both imaging data and textual information [3]. These models perform a comprehensive analysis of medical imaging features alongside their associated clinical narratives, enabling a more thorough interpretation of radiological findings. Through the simultaneous processing of visual patterns and language representations. The combination of these modalities enables the models to effectively leverage both the nuanced features of the medical images and the diagnostic context from the associated text.

Transforming detailed medical images into descriptive text necessitates the accurate detection of anatomical structures, alongside precise identification of lesions based on their morphology and location. Although automated report generation systems have shown significant progress in analyzing chest radiographs [4], their widespread clinical adoption remains limited due to several challenges. Among these, a major barrier is the semantic and granularity gap between imaging modalities and textual descriptions, which leads to imperfect alignment between visual data and generated reports. Additionally, these systems suffer from data bias, as common conditions are often overrepresented in training datasets, whereas rare pathologies are underrepresented, hindering the robustness and generalizability of the models. Furthermore, interpretability continues to be a major challenge in medical applications, as AI models often operate as 'black boxes.' The doctors emphasize the need for transparent models that can provide clear, understandable explanations for their predictions, an essential requirement in medical practice.

Clinical automated systems require transparency, which has led to the creation of explainable artificial intelligence (XAI) becoming an essential research topic [5]. The ability to interpret AI outcomes remains essential for validating diagnostic accuracy while building clinician trust and meeting regulatory requirements [6]. The semantic gap between modalities, the "black-box" nature of deep learning models, and the need for high-fidelity descriptions—motivate the central research question of this work: How can we design a radiology report generation framework that not only achieves high clinical accuracy through fine-grained, semantically-guided alignment between images and text, but also provides transparent, concept-based explanations for its diagnostic outputs to foster clinical trust and adoption? This question guides our development of a novel architecture that simultaneously optimizes for descriptive precision and true model interpretability.

To answer this question, we introduce a new model based on semantic-guided alignment and interpretable multimodal deep learning named Explainable Report Generation Semantic-guided Alignment—Ex-RGSA—which directly uses important

signals and features to advance automated medical reporting. Unlike traditional methods focusing on global visual and textual features, this model emphasizes sensitivity to crucial lesion locations and fine-grained alignment, thereby improving the quality and reliability of generated reports.

Our approach addresses the persistent challenges of aligning visual and textual data, detecting small and rare pathologies, and ensuring the precision of diagnostic terminology. It uses structured semantic knowledge from medical reports. At the start of our framework lies the Explainable Semantic Feature Extractor (X-SFE) module, which consists of two subparts: the Multi-level Text Feature Extraction, capturing multi-level text features at the report, sentence, and word levels. Semantic Guided Knowledge extraction is a sub-part that identifies and encodes structured medical knowledge from historical reports, aiding the model in contextualizing lesion descriptions and refining diagnostic terminology. At the core of our framework lies the Semantic-Aligned Visual Extractor (X-SAVE) module. This mechanism leverages structured semantic knowledge extracted from medical reports to guide the visual extractor to focus on fine-grained lesion details related to the text. Additionally, the Contrastive Multimodal Fusion module (CMF) links visual-textual features anchored by concepts through three stages (global anatomy to regional findings to lesion-specific details). To further refine the generated reports, we introduce a semantic self-refining mechanism in the Decoder module that uses CAV confidence scores to correct terminology errors.

Experimental results demonstrate significant improvements in lesion localization, diagnostic precision, and contextual coherence, underscoring the effectiveness of our framework in generating high-quality medical reports. This positions our approach as a valuable tool for enhancing diagnostic workflows and improving patient outcomes. In summary, this paper presents three main contributions:

1) First, our primary contribution: is a novel trainable architecture for radiology report generation, introducing a new paradigm that significantly enhances multimodal alignment and semantic consistency. A key aspect of this architecture is its ability to improve the detection of small and rare lesions, further ensuring the diagnostic precision achieved by the Explainable Semantic Feature Extractor (X-SFE), the Semantic-Aligned Visual Extractor (X-SAVE), and the Multimodal Fusion (CMF) modules.

2) In the second contribution: we introduce an explainable architecture that fundamentally transforms how multimodal systems process medical data, combining concept activation vectors with hierarchical semantic alignment to produce reports clinicians can trust—unlike current "black box" systems. The framework's unique integration of visual-textual grounding with CAV-based explanations yields not just better performance, but more importantly, delivers the transparent diagnostic reasoning demanded in clinical practice.

3) The last contribution: Demonstrating superior performance on benchmark datasets: experimental results indicate that our framework achieves superior performance on publicly available datasets, with significant improvements in lesion localization, diagnostic precision, and contextual coherence. For instance, our model achieves a BLEU-4 score of 0.189, 0.221 in METHEOR and 0.412 in ROUGE-L on the IU X-RAY dataset, demonstrating a notable improvement compared to the baseline and delivering superior results on this dataset, particularly in handling complex cases with multiple lesions.

The remainder of this paper is organized as follows. A review of related work in explainable AI and multimodal report generation is given in Section II. The architecture of our proposed Ex-ReGSA model, including its main modules, is described in detail in Section III. The experimental setup, including the dataset, evaluation metrics, and implementation specifics, is covered in Section IV. The findings of our comparative and ablation studies are shown in Section V. Section VI, which discusses the results with suggestions for further research, addresses the limitations and wider ramifications of our work.

II. Related Work

The automatic generation of medical reports based on multimodal data has recently attracted much attention. In this field, diverse works are introduced to tackle key challenges, including cross-modal alignment, reducing data bias, and improving clinical relevance. This section shows an overview of related work along four key areas, including (1) multi-modal fusion and (2) explainable artificial intelligence-based report generation.

A. Multimodal Fusion-Based Method for Report Generation

In recent years, several studies have explored the use of multimodal data fusion to improve diagnostic accuracy and report coherence. The method proposed by Tang et al. [7] solves medical report generation challenges by proposing the "locate then generate" CAT pattern. The framework uses a multi-modality encoder combined with a dual-stream decoder to dynamically incorporate retrieved terminologies with preceding sentences. Similarly, the approach proposed by Iqbal et al. [8] improves radiology reporting by utilizing an adaptive multi-modal approach that encodes clinical data elements. The system uses "wisdom learning" to extract medical insights from radiology reports through textual embeddings while applying cross-modal coherence that employs clinical embeddings to direct visual feature learning and enhance semantic alignment between images, disease labels, and generated reports. Li etl. [9] continued to explore diverse data integration. The authors developed a context-enhanced framework to address the shortcomings of image-only dependent systems. The report generation process is improved through their method, which combines multiple contextual elements such as clinical texts, structured medical knowledge, and previous diagnostic outcomes with primary medical images. Zeng et al. [10] ensures that generated reports maintain internal coherence while preserving their contextual integrity. The model utilizes a relational memory module that updates itself with previously generated words to enhance word correlations, as well as a double LSTM network with an interaction module to preserve contextual information during text generation. Tsaniya etl. [11] conduct research on architectural advancements along with

input data enhancements. The authors developed a medical report generator that operates on a transformer-based model. The system integrates text feature embedding through BERT with visual feature extraction from a pre-trained CheXNet model enhanced by multi-head attention to evaluate how contrast-based image enhancement positively influences report quality.

B. XAI-Based Report Generation

Several recent approaches have focused on embedding explainability directly into the architecture of report generation models. Zhang et al. [12] proposed the Attribute Prototype-guided Iterative Scene Graph framework, which utilizes an autoregressive approach for structural edge reasoning to handle common attribute and region feature limitations and improve interpretability. Similarly, Tanida et al. [13] developed an interactive region-guided framework that begins by detecting anatomical regions and then describes key regions to generate medical reports. Chen et al. [14] introduced AdaMatch as a model to establish detailed connections between CXR image segments and medical report terminology; then they used their AdaMatch-Cyclic model to utilize these connections for producing understandable CXR reports, which showed excellent results on public datasets. The development of specialized datasets and models focused on aligning with clinical reasoning is also crucial. One such study [15] introduced the FG-CXR dataset, which provides fine-grained pairings between radiologists' generated captions and corresponding gaze attention heatmaps for specific anatomies. The same study proposed a Gen-XAI network, which simulates diagnostic activities by matching its outputs with radiologist gaze patterns and transcript data. The research by Taleb et al. [16] addresses broader topics beyond report generation. The researchers introduced ContIG as a self-supervised learning method that enables the alignment of medical images and genetic data through contrastive loss and utilizes gradient-based XAI for interpreting cross-modal connections to demonstrate XAI effectiveness on diverse medical data sources. XAI techniques are currently used to improve transparency across multiple diagnostic tasks that involve different data types. For example, Sangnark et al.[17] created an explainable deep learning model to diagnose dyssynergic defecation from abdominal X-rays and questionnaires utilizing cross-modal attention mechanisms with Grad-CAM and DeepSHAP for image and symptom data interpretation. Transparent COVID-19 interpretation research [18] implemented deep neural networks like EfficientNet and DenseNet together with XAI techniques including LIME and Grad-CAM as well as a new "Modified Grad-CAM++."

While existing methods have achieved remarkable performance on multimodal fusion and started to integrate explainability, from our literature review, we identify the following gaps in the existing methods: 1) Many methods use global feature alignment, which neglects the fine-grained, lesion-specific evidence that doctors often need to pay attention to for a detailed and accurate clinical description. 2) Explainability in these methods is often restricted to post-hoc visualizations, such as heatmaps that indicate where the model is looking. While they reveal the localization of model attention, these heatmaps do not offer explanations that are interpretable in terms of well-defined and clinically meaningful concepts, resulting in a "black box" problem that can undermine clinical trust. To address these problems, we employ hierarchical feature extraction and semantic-guided alignment to attend to fine-grained features. Moreover, we introduce Concept Activation Vectors (CAVs) to our model to enable quantifiable and interpretable concept-based explanations of the model prediction.

III. Proposed Method

The overview of our proposed model for enhancing the generation of radiology reports Ex-ReGSA is illustrated in Fig. 1 and comprises four major modules: Explainable Semantic Feature Extractor (X-SFE), which includes Multi-Level Text Feature Extraction and Semantic Guided Knowledge extraction. The second module is the Semantic-Aligned Visual Extractor (X-SAVE), the third module is the Contrastive Multimodal Fusion module (CMF); and the end module is a Dynamic Decoder (DD).

First, the input medical report is processed by the Explainable Semantic Feature Extractor module to extract rich-grained features from the radiology reports through hierarchical attention networks. Subsequently, the Semantic Guided Knowledge Extraction Module is designed to extract relevant semantic knowledge from the fine-grained features obtained by Multi-Level Text Feature Extraction by categorizing similar information into groups that have the same characteristics. Moreover, the Semantic-Aligned Visual Extractor serves to extract visual features from radiology images, guided by semantic output obtained by X-SFE, to ensure that only the important text-based features are captured. The features extracted by X-SFE and X-SAVE then feed into the Contrastive Multimodal Fusion module (CMF), which focuses on aligning semantic association between the extracted fine-grained visual and corresponding multi-level textual features, ensuring cohesive multi-modal representations. Finally, the Dynamic Decoder (DD) module integrates the aligned multi-modal features to generate, in a hierarchical process, a coherent and contextualized radiology report.

A. Explainable Semantic Feature Extraction (X-SFE)

1) Multi-level text feature Extraction: To extract multi-level features from medical report data, hierarchical attention mechanisms based on RadBERT[19], a Bidirectional Encoder Representations from Transformers (BERT) model specifically pre-trained on a large corpus of clinical radiology text, operate at three levels: report, sentence, and word. This approach captures both global diagnostic context and fine-grained clinical details while leveraging RadBERT’s improved pretraining for better representation learning.

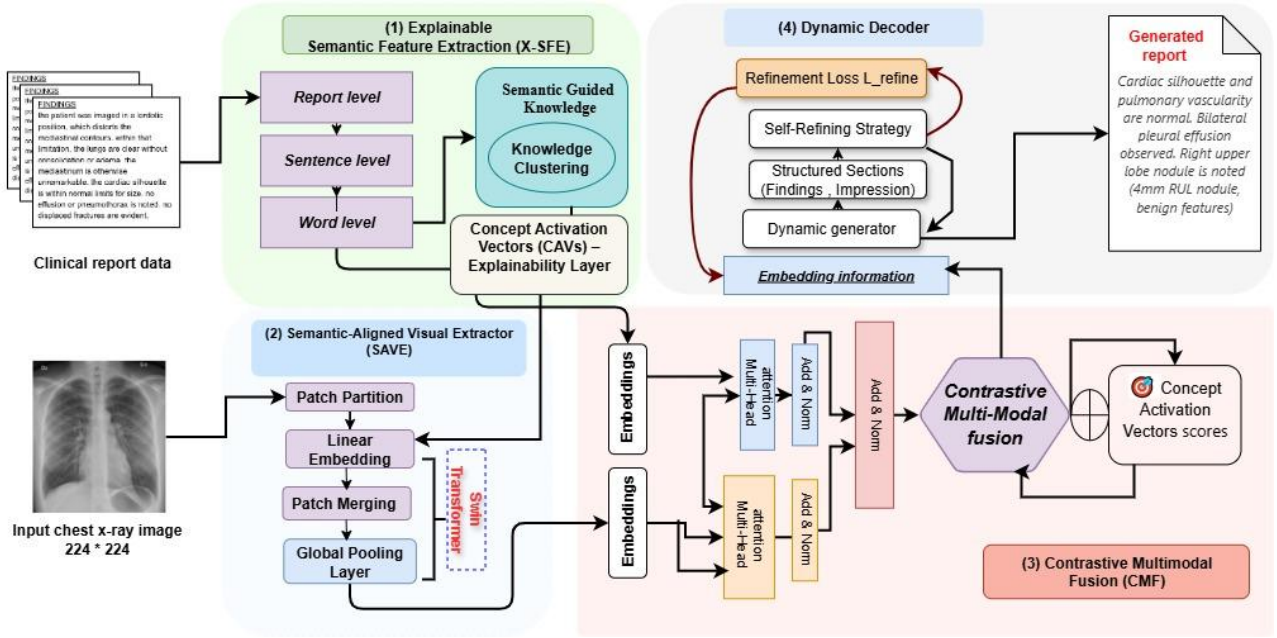


Fig. 1. Overview of the proposed Ex-ReGSA model.

Before robust hierarchical feature extraction can be applied to radiology reports data. They must first pass through an essential preprocessing stage. The process begins with extracting diagnostic sections such as "Findings" and "Impression," followed by text cleaning steps that include lowercasing text while removing unnecessary characters and artifacts. Crucially, the cleaned text, which we can denote as Rclean_text.

The prepared text undergoes processing by RadBERT, which conducts sub-word tokenization and creates foundational embeddings with deep context for all text tokens. The foundational token embeddings enable systematic creation of specialized representations covering all parts of the report down to the individual words and sentences. Converting the text into a sequence of M tokens, $T=\{t_1,t_2,...,t_M\}$. This token sequence T is then comprehensively processed by RadBERT to generate foundational contextual embeddings.

$$H = RadBERT(T) \quad (1)$$

where $H=\{h_1,h_2,...,h_M\}$ is the sequence of final hidden layer states from RadBERT. Each $hj \in R^{dh}$ represents a deeply contextualized embedding for token t_j, where dh is the dimensionality of RadBERT's hidden states. This output H serves as the rich, contextual basis from which all hierarchical features are subsequently derived. For Sentence-Level Feature Representation (Semb), Individual sentences within a clinical report often convey distinct findings or observations. To capture these, sentence-level embeddings, Semb, are derived for each of the N sentences, $S=\{S_1,S_2,...,S_N\}$, in the preprocessed report. For a given sentence Si, its representation $Semb(Si) \in R^{dh}$ is constructed by aggregating the contextual token embeddings $\{h_j \in H\}$ that correspond to the Qi tokens within that specific sentence (denoted $H_{Si}=\{h_{Si},1,...,h_{Si},Q_i\}$). A common aggregation method is mean pooling:

$$S_{emb}(S_i) = \frac{1}{Q_i} = \sum_{q=1}^{Q_i} h_{S_{i,q}} \quad (2)$$

Word embeddings for individual words make up the finest textual feature level Wemb. Word embeddings serve as essential tools for recognizing particular clinical entities as well as their complex meanings in medical reports. The word embedding $W_{emb}(W_k)$ for the k-th word W_k is derived from the RadBERT output H. Since RadBERT's tokenizer may split W_k into L_k sub-word tokens ($T(W_k)=\{t_{k,1},...,t_{k,Lk}\}$), the corresponding contextual token embeddings ($h_{k,1},...,h_{k,Lk}$ from H) are typically aggregated. The most common method is averaging:

$$W_{emb}(W_k) = \frac{1}{L_k} = \sum_{l=1}^{L_k} h_{K,l} \quad (3)$$

2) Semantic guided knowledge extraction: The Semantic Knowledge Module is designed to extract relevant knowledge from chest X-ray reports by categorizing similar information into groups that have the same characteristics. To achieve this, we employ a clustering algorithm on report-embedded representation vectors to group similar information together into several clusters. To quantify the presence and strength of these representations within each report and to guide the clustering process towards forming semantically interpretable groups, we leverage Concept Activation Vectors (CAV) [21]. In particular, given a set of embedding vectors $V = \{v_1, v_2, ..., v_n\}$, the K-Means algorithm [20] is used to group these vectors into groups (clusters), denoted Tk, based on their similarities. This process is defined by minimizing the Euclidean distance between each report's embedding Vi and the centroid Cj of its corresponding cluster. Initialize K centroids randomly, where each centroid mj (for j=1,2,…,K) represents the initial mean of a cluster. This can be expressed as Eq. (4).

$$T_k = \underset{j}{arg,min} \sum_{d=1}^{D} (v_{i,d} - m_{j,d})^2 + \alpha \cdot CAVsim(\text{vi, cj}) \quad (4)$$

Where T_k is the index of the cluster assigned to report v_i and m_j is the centroid, D is the dimensionality of each embedding vector y_i and $y_{i,d}$ and $m_{j,d}$ represent the d-th component of vectors vi and m_j, respectively. After this step, the centroids continue to get refined and updated as the meaning of the embeddings associated to each cluster. This iterative process executes until labeled reports do not change anymore, resulting in k groups of reports, where each cluster encapsulates specific features, like similar writing style or diagnostic terminology. And the function $CAVsim(\text{vi}, \text{cj})$ measuring the semantic or conceptual similarity between report vi and cluster centroid mj, based on their alignment with predefined concepts quantified by CAVs. While the α is a weighting parameter that balances the influence of the Euclidean distance against the CAV-guided conceptual similarity term.

B. Explainable Semantic-Aligned Visual Extractor (X-SAVE)

To reduce the problem of data bias and extract precise spatial and semantic features of disease lesions, we employ a multi-level feature extraction technique that relies heavily on prior knowledge obtained from semantic clusters found within clinical text reports. This guidance aims to align visual regions with relevant textual features and to direct the model's attention towards focal regions based on their semantic relevance, determined through cluster-specific importance signals. The X-SAVE module receives a collection of chest X-ray (CXR) images $X=\{x_1, x_2, \ldots, x_n\}$, where each image x_i is a tensor in $RC\times H\times W$, with C representing the number of channels and H and W representing spatial height and width, respectively. While H = W = 224 for CXR images. The feature extraction process starts by passing the input images through a SwinTransformer [22], which serves as the visual backbone. The Swin Transformer is selected for its effectiveness in computer vision tasks and its inherent hierarchical architecture, which captures features at multiple scales by progressively downsampling the input image. Enabling both localized fine-grained features and global contextual information. The process is illustrated in Eq. (5) to Eq. (7).

*1) Macro-level features (**F1**):* Capturing coarse, global anatomical structures, the image is downsampled by a factor of 4:

$$F1 = \text{SwinTransformer}(X) \tag{5}$$

were, $\mathbf{F1} \in \mathbf{R}^{(C_2\times H/4\times W/4)} = \mathbf{R}^{(C_1\times 56\times 56)}$, the resolution becomes 56 x 56 with C_1 channels.

*2) Intermediate level **(F2)**:* At the second level, highlighting regional characteristics and intermediate structures, the CXR-image is downsampled by another factor of 2:

$$F2 = \text{SwinTransformer}(F1) \tag{6}$$

Where $\mathbf{F2} \in \mathbf{R}^{(C_3\times H/8\times W/8)} = \mathbf{R}^{(C2\times 28\times 28)}$ and the resolution is reduced to 28 × 28, with C_2 channels.

*3) Micro-level features (**F3**):* Detailing fine-grained visual patterns crucial for localizing subtle lesions., The image is downsampled by a factor of 2:

$$F3 = \text{SwinTransformer}(F2) \tag{7}$$

Here $F3 \in R^{(C_4\times H/16\times W/16)} = \mathbf{R}^{(C3\times 14\times 14)}$ and the resolution becomes 14 × 14, and C_3 channels are used.

The key innovation lies in our semantic alignment integrated with the Concept Activation Vectors (CAVs) mechanism, where text-derived clinical concepts from radiology reports are clustered ($T_1,\ldots,T_k$) using RadBERT embeddings, and each cluster's importance weight w_k is computed based on feature dispersion from its centroid m_k to reflect semantic coherence, computed as Eq. (8):

$$wk = \frac{1}{|Tk|}\sum_{i\in Tk}^{n} ||vi - mk||2 \tag{8}$$

Where |Tk| is the number of reports in cluster T_k.

The visual features extracted for each input image x are then aligned with the clusters as follows in Eq. (9):

$$F_{Aligned} = \sum_{k=1}^{K}(wk.\,\text{TCAV}_k).\,Fk \tag{9}$$

where $Fk = Wk \cdot [F2 \oplus F3]$ represents the visual features associated with cluster T_k, and Wk is a learnable projection matrix.

C. Contrastive Multimodal Fusion (CMF)

To apply a hierarchical alignment and fusion, we use a transformer-based model. Current multi-modal fusion approaches rely on global feature fusion. We design multimodal fusion based on self-supervised attention, which comprises two key stages: cross-modal attention and self-supervised contrastive learning. The CMF module learns aligned representations by combining detailed feature interactions with a comprehensive semantic understanding from visual and textual data sources. The system accomplishes this through a unified process that uses hierarchical cross-modal attention mechanisms to map visual and textual features at various scales into detailed correspondences. The features are attentively fused and aligned and then projected into a shared embedding space, which undergoes refinement with a self-supervised contrastive learning objective to ensure robust end-to-end semantic alignment. The system achieves explainability by leveraging Concept Activation Vectors (CAVs) for understanding the semantic characteristics of the shared embedding space.

This initial step within CMF focuses on establishing detailed correspondence between the hierarchically extracted visual features and report features. As previously detailed, hierarchical visual features (F1 for global, F2 for regional, and F3 for fine-grained details) are obtained from the input CXR image. Similarly, hierarchical textual features (Remb for report-level, Semb for sentence-level, and wemb(j) for word-level) are extracted from the medical report.

To bridge representational differences, both feature sets undergo modality-specific normalization and are augmented with positional embeddings. The cross-modal attention then aligns these features.

The normalized and enriched features are then used to compute the queries (Qs), keys (Ks), and values (Vs) directly within the attention mechanism:

$$A_s = Softmax\left(\frac{Q_s K_s^T}{\sqrt{d}} + \log(\text{TCAV} + 1)\right) \tag{10}$$

Where $Q_s = F_{text}^{pos}$: The query matrix is derived from position-enriched textual features. And $K_s = F_{visual}^{pos}$ Is the key matrix derived from the position-enriched visual features. The d is the dimensionality of the key vectors, and serves as a scaling factor to ensure gradient stability in training. By using the attention scores As, the aligned features FS aligned are computed by applying As to the value matrix(Vs), which is also derived from the position-enriched visual features. $F_{aligned}^{s} = A_s V_s$ and $V_s = F_{visual}^{pos}$

The aligned features from all scales are concatenated into a unified representation:

$$F_{fusion}=Concat(F_{1aligned},F_{2aligned},F_{3aligned}) \quad (11)$$

After generating a unified representation (Ffusion), visual features will be projected into a shared latent space through a learnable matrix (Wv): F'visual=Wv.Ffusion and textual features with another one (Wt): F'text=Wtr. These embeddings are refined using multimodal contrastive learning, which optimizes a contrastive loss (Lcontrastive) to align positive pairs, such as a CXR image and its corresponding report, while distancing unrelated pairs. The loss is calculated in Eq. (12) as:

$$L_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(F'_{visual,i}, F'_{text,i})/\tau}{\sum_{j=1}^{N} \exp(sim(F'_{visual,i}, F'_{text,j})/\tau} \quad (12)$$

Where sim(x,y) is the cosine similarity, and τ is a temperature parameter controlling the sharpness of the similarity distribution. This end-to-end, self-supervised training paradigm leverages inherent data pairings without requiring manual annotations, ensuring scalability and robust semantic alignment for downstream tasks.

D. Dynamic Decoder (DD)

To overcome critical limitations in existing report generation decoders—such as static templates, neglect of hierarchical features, and lack of dynamic refinement. We propose a novel Dynamic Decoder (DD) that processes multi-level fused visual features (F_{fusion}) in a hierarchical structure. The decoder process uses corresponding report, sentence, and word-level textual embeddings together with fused visual features to construct final reports. A key innovation is its Dynamic Sentence-Type Adaptation mechanism, which uses learned, Concept Activation Vector (CAV)-influenced attention weights (αT,αA) where, for instance, αT is determined by Eq. (13):

$$\boldsymbol{\alpha}T = \sigma(W_T F_{fusion} + \beta \cdot CAVsim(F_{fusion}, c_k)\,,\ \alpha_A = 1 - \alpha_T) \quad (13)$$

Where CAV_{sim}: Measures alignment between fused features and concept c_k. These weights ensure critical abnormalities are prioritized in complex cases, while template sentences dominate in normal cases. The final representation is presented in Eq. (14):

$$Rt=\alpha_T T+\alpha_A A \quad (14)$$

where T represents normal features and A represents abnormal features.

Furthermore, a Feedback-Driven Refinement loop iteratively improves semantic consistency by re-encoding the generated report Y′ and minimizing a refinement loss (L_{refine}) against the input features F_{fusion}, defined as $L_{refine}=\|F_{fusion}-Encoder(Y')\|^2_2$. The DD is trained end-to-end with a hybrid loss function, combining objectives for token accuracy (L_{xe}), feature alignment (L_{refine}), and overall semantic coherence (via $L_{contrastive}$), enabling it to generate clinically accurate, contextually relevant, and professionally structured medical reports:

$$Lrefine : Ltotal=\lambda 1\ Lxe+\lambda 2\ Lcontrastive+\lambda 3\ Lrefine \quad (15)$$

IV. Experimental Settings

A. Dataset

IU X-ray [23]: is a publicly accessible repository of chest X-rays, featuring 7,470 images and their 3,955 associated radiological reports. The nature of this data makes it a valuable asset for developing and training models in tasks such as automated radiology report generation and computer-aided diagnostic systems. To maintain consistency and allow for comparison with prior research, this collection is conventionally divided into standard proportions: 70% of the data for model training, 10% for validation during development, and the remaining 20% for final model testing.

B. Evaluation Metrics

The quality of our model was evaluated using four NLG Natural Language Generation metrics. To measure precision in the generated report compared to ground truth clinical report, we used BLEU [24], which scores n-gram overlap. For an evaluation more aligned with human intuition, we used METEOR [25], as it considers variations like word stems and synonyms from WordNet. To specifically assess the fluency and coherence of the generated content, we employed ROUGE-L [26], which measures similarity based on the longest common subsequence of words found in both the generated and reference texts. Additionally, CIDEr [27] was utilized to assess how well the generated text captures salient concepts present across a set of reference texts, by weighing n-grams based on their Term Frequency-Inverse Document Frequency (TF-IDF) scores, thereby emphasizing consensus and information richness.

C. Implementation Details

Our medical report generation system is built on a hierarchical, multimodal architecture. For visual input, we process 224x224 resolution chest X-rays using a Swin Transformer backbone. This visual encoder uses a patch size of 4, an embedding dimension of 96, and four feature extraction stages with 3, 6, 12, and 24 attention heads, respectively. The text processing component utilizes RadBERT for hierarchical feature extraction at the word, sentence, and document levels, each with a 768-dimensional hidden state for attention. To integrate these, a multimodal fusion module projects both visual and textual features into a shared 128-dimensional space via linear transformations, followed by cross-modal attention with 4 parallel heads. Report generation is handled by a 6-layer transformer decoder (512 hidden units, 8 attention heads) that dynamically regularizes sentence types between templates and abnormal findings. The model was trained for 40 epochs using

a composite loss function—combining cross-entropy (weight 1.0), contrastive loss (weight 0.5, temperature 0.07), and a learned features refinement loss (weight 0.5)—optimized with Adam at a learning rate of 5×10^{-4}. For generating reports (inference), we employ a beam search with a width of 3, using hyperparameters (contrastive temperature τ between 0.05-0.2, loss balancing λ between (0.1-0.9) selected based on ROUGE scores on a validation set.

V. Experiments Results

A. Comparison Experiment

We evaluated our model using standard metrics: we employed BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr metrics to measure both the quality and relevance of the reports. As shown in Table I and Fig. 2, comparative analysis is performed against state-of-the-art models: The evaluation includes comparisons with advanced models HRGR-Agent[28], R2GEN[29], PPKED [30], AERMNet[10], C-Enhanced-F [9], and work by SAEED et al. [8]. These benchmarks provide thorough assessments of fluency quality along with informative content and clinical consistency.

Analyzing the BLEU scores, which measure n-gram precision against reference texts, our model demonstrates strong performance, particularly for lower-order n-grams. It achieves the highest BLEU-1 score of 0.517, outperforming all listed competitors, including notable models like SAEED et al. (0.499) and C-Enhanced-F (0.491). For higher-order n-grams, our model remains highly competitive: its BLEU-2 score is 0.351, slightly behind C-Enhanced-F, which scored 0.359. Similarly, its BLEU-3 score is 0.251, with C-Enhanced-F at 0.263, and its BLEU-4 score is 0.189, compared to C-Enhanced-F's 0.209. While C-Enhanced-F exhibits an edge in these higher-order BLEU metrics, indicating a closer match in longer contiguous word sequences, our model consistently surpasses other methods like R2GEN and PPKED across all BLEU categories. This suggests a strong grasp of lexical choice and phraseology.

In terms of the METEOR metric, which evaluates generated text quality by considering stemming, synonyms, and word order to better align with human judgment, our model scores 0.221. This is the highest METEOR score among all models for which this metric is reported in the table, surpassing AERMNet (0.219), C-Enhanced-F (0.212), and R2GEN (0.185). This leading performance in METEOR suggests that our model generates reports that are not only accurate in terms of word choice but also capture semantic similarity effectively, aligning well with human assessment criteria.

TABLE I. Performance Comparison

Model	B-1	B-2	B-3	B-4	M.	R-L	CIDEr
[28]	0.438	0.298	0.208	0.151	--	0.322	0.343
[29]	0.451	0.293	0.209	0.159	0.185	0.381	0.406
[30]	0.483	0.315	0.224	0.168	--	0.376	0.365
[10]	0.486	0.321	0.236	0.183	0.219	0.398	0.560
[9]	0.491	0.359	0.263	0.209	0.212	0.408	0.396
[8]	0.499	0.349	0.229	0.170	--	0.401	0.411
(Ours)	0.517	0.351	0.251	0.189	0.221	0.412	0.463

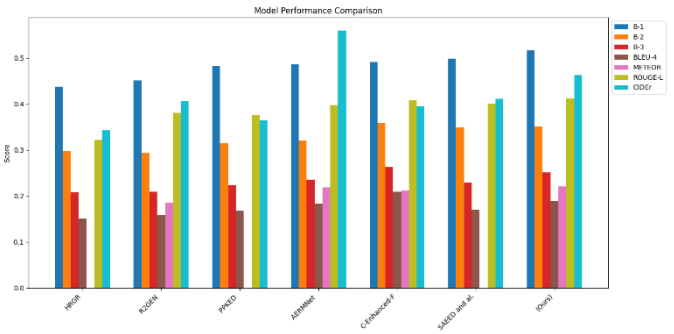


Fig. 2. Performance comparison of models on IU-CXR.

TABLE II. Ablation Experiment

Model	B-1	B-2	B-3	B-4	M.	R-L	CIDEr
Baseline Model	0.405	0.301	0.219	0.168	0.181	0.350	0.392
+ X-SFE	0.413	0.314	0.222	0.172	0.188	0.358	0.409
+ X-SAVE	0.422	0.332	0.239	0.181	0.192	0.373	0.441
+CMF	0.451	0.339	0.243	0.182	0.202	0.398	0.453
+Self-Refining mechanism	0.470	0.349	0.249	0.189	0.214	0.408	0.461
Full Model	0.517	0.351	0.251	0.189	0.221	0.412	0.463

Our model particularly excels in the ROUGE-L score, achieving 0.412. This metric, which measures the longest common subsequence between generated and reference texts, reflects fluency and the recall of important content. Attaining the top score in ROUGE-L indicates our model's superior ability to capture the structural similarity and main ideas present in the ground truth reports when compared to all other listed models, including strong performers like C-Enhanced-F (0.408), SAEED et al. (0.401), and AERMNet (0.398). This suggests that the reports generated by our model are coherent and effectively convey the necessary information.

The CIDEr metric, designed to measure consensus and human-like quality in image descriptions by weighting n-grams using Term Frequency-Inverse Document Frequency (TF-IDF), offers further insights into the model's performance. Our model achieves a CIDEr score of 0.413. This score is competitive and robust, surpassing or closely aligning with several models such as SAEED et al. (0.411), R2GEN (0.406), and C-Enhanced-F (0.396). However, it is noteworthy that AERMNet demonstrates a significantly higher CIDEr score of 0.560. This indicates AERMNet's strong capability in generating text that effectively captures n-grams frequently found across multiple reference

reports, a key aspect of consensus. While our model performs well against many peers in CIDEr, AERMNet sets a high benchmark for this specific measure of quality.

In summary, this comparative analysis reveals that our proposed model exhibits a strong and highly competitive performance profile. It achieves leading results in BLEU-1, METEOR, and ROUGE-L, signifying its strengths in generating lexically precise, semantically rich, and structurally coherent radiology reports. While C-Enhanced-F shows an advantage in higher-order BLEU scores, and AERMNet leads substantially in CIDEr, our model consistently outperforms many established baselines across multiple critical metrics. The particularly strong ROUGE-L and METEOR scores suggest that the generated reports are fluent and capture clinical information effectively, which is paramount for clinical utility. The CIDEr score of 0.413, while not the highest, is solid and points towards good consensus capture. These results collectively affirm the Efficacy of our model as a significant contribution to the field of automated radiology report generation, though future work could explore strategies to further enhance performance on metrics like CIDEr to match the top specialist models.

B. Ablation Experiment

In this section, we will conduct an ablation experiment to indicate the influence of each module on our model's overall performance. The experiments demonstrate the influence of the various components when they are integrated in the baseline model to quantify the gain of each module (X-SFE, X-SAVE, CMF, and DD) and the combination on final accuracy. Table II provides a detailed comparison over multiple evaluation metrics, including B-1, B-2, B-3, B-4, METEOR, and ROUGE-L, for the Indiana University dataset.

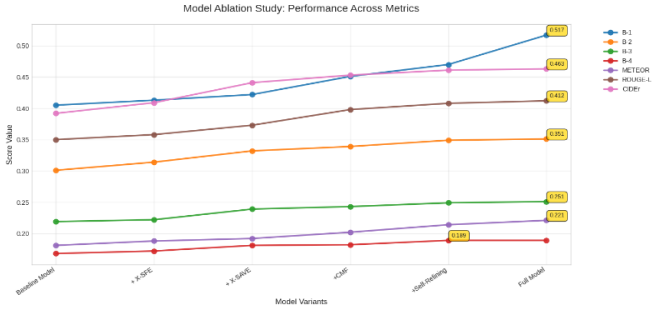


Fig. 3. Performance metrics of ablation study.

Table II and Fig. 3 illustrate that the full model yields substantial performance gains when compared to the baseline model, thereby underscoring the crucial role of every proposed module. Quantitatively, the Full Model surpassed the Baseline by 11.2% on BLEU-1, 5.0% on BLEU-2, 3.2% on BLEU-3, 2.1% on BLEU-4, 4.0% on METEOR, 6.2% on ROUGE-L, and 7.1% on CIDEr. These results indicate a marked improvement in report generation quality due to the additional components. A systematic evaluation of each component's individual effect was subsequently undertaken by comparing the various configurations.

BASELINE: The BASELINE model is a standard Transformer composite of three layers, eight attention heads, and 512 hidden units. This model serves as the starting point to evaluate the impact of additional modules.

+X-SFE (Explainable Semantic Feature Extraction): This configuration adds the explainable text feature extraction module to the baseline model. This component leads to discernible gains across all metrics. For instance, BLEU-1 rises to 0.413, ROUGE-L to 0.358, and notably, CIDEr improves by 1.7%. This suggests that enhancing the model's ability to capture hierarchical relations in the text provides better context and structure for report generation.

+X-SAVE (Explainable Semantic-Aligned Visual Extractor): This module focuses on fine-grained lesion detection. The X-SAVE module guides visual feature extraction using semantic knowledge derived from medical reports. This ensures that the model focuses on clinically significant regions of the image, significantly improving the model's ability to generate precise and accurate medical reports. This module yields more significant boosts, with BLEU-1 reaching 0.422, BLEU-2 improving from 0.314 to 0.332, ROUGE-L increasing to 0.373 (+0.015), and CIDEr showing a strong jump to 0.441 (+0.032). These results underscore the importance of semantic-aligned visual extractors in focusing the model on diagnostically relevant image regions, thereby enhancing the quality and relevance of the generated text, particularly reflected in the CIDEr score.

+CMF Multimodal Fusion module: In this setup, the CMF module is integrated into the model. As shown in Table II, this module improves visual-textual alignment by learning relationships between the two modalities. This leads to better cross-modal understanding, ensuring that both the visual and textual data are appropriately aligned, resulting in more accurate and coherent reports. This addition results in BLEU-1 increasing to 0.451, METEOR to 0.202 (+0.010), and ROUGE-L to 0.398 (+0.025). The CIDEr score also sees a healthy increase to 0.453. These results highlight the importance of multi-level cross-modal alignment in improving the integration of visual and textual data, leading to more accurate and coherent medical report generation.

+Self-refining mechanism: We add a self-refining mechanism strategy in our decoder. As shown in Table II, this configuration substantially outperforms BASE; BLEU-1 reaches 0.470, METEOR improves to 0.214, ROUGE-L to 0.408, and CIDEr to 0.461. This indicates the importance of this configuration in improving the semantic quality of reports that are generated by maintaining consistent and context-dependent terminology.

The full model, which merges all modules into an optimized layout, demonstrates superior performance. The BLEU-1 score reached 0.517, which represents a considerable increase from the former performance at 0.470. The METEOR score achieves 0.221, while ROUGE-L reaches a level of 0.412 and CIDEr advances to 0.463. Transitioning from the "+Self-Refining mechanism" configuration to the "Full Model" generates substantial BLEU-1 improvements and consistent metric gains, which demonstrate that combined optimization of all modules creates a powerful synergy or enables the self-refining process to achieve its maximum potential in this final configuration. The BLEU-4 score maintains a value of 0.189 across the final two

stages because advancements have been concentrated on the text structure and semantic meaning along with lower-order n-gram precision.

C. Visual Analysis

We demonstrated the clinical insight and effectiveness of our Ex-ReGSA model by choosing illustrative cases for qualitative analysis, which appear in Table III. The table presents a side-by-side analysis of the original radiological report against Ex-ReGSA-produced output and the model-derived explanations for its results, which appear within the context of the input radiological image. This visual experiment surpasses quantitative metrics by providing richer insight into Ex-ReGSA's ability to interpret complex clinical narratives and demonstrate its decision-making transparency.

As shown in Table III, the generated report shows a high degree of resemblance in terms of organs and lesion description, signifying the model's capability in encapsulating the intricate relationship between multi-level features in both images and reports. This enables its hierarchical attention mechanisms, which focus on both global and fine-grained details. Additionally, the output report of the case also contains detailed diagnostic details, such as specific characteristics of the referred lesions, which can be attributed to the X-SFE and X-SAVE during hierarchical feature extraction. This guidance enables the model to offer importance on critical areas, improving the clinical relevance of the generated reports. Furthermore, the "Interpretable Explanations" column provides important understanding of how the model arrives at its decisions and processes information internally. These examples illustrate how these explanations combine model-assigned confidence scores of specific findings (e.g., "cardiomegaly score 0.92," "pleural effusion score 0.93") with mentions of key visual features detected by internal components alongside potential semantic knowledge or quantifiers used by the model. Through detailed case-by-case explanations of its operational logic, Ex-ReGSA proves transparent because it shows both its pathologies and normality reports and explains the reasoning behind those conclusions, which builds confidence in its clinical usefulness and reliability.

TABLE III. THE VISUALIZATION OF OUR EX-RGSA MODEL

image	Ground truth	Generated report	Interpretable Explanations
	The heart is normal in size. The lungs are clear without evidence of pleural effusion, pneumothorax, or focal lesions. Bony structures are unremarkable.	Heart size is within normal limits. No pulmonary abnormalities detected lungs are clear bilaterally, with no signs of effusion, pneumothorax, or consolidation.	- The diagnosis of "clear lung fields" achieved a high confidence level of 0.97 because GGO and consolidation features were not present. - The "Normal cardiac size" score of 0.95 reflects that the cardiothoracic ratio falls within the established normal range. - Diffuse attention across both lung parenchyma (avg. intensity ;0.2 per zone), no focal hotspots. - The sharp and clear appearance of the costophrenic angles in visual features confirms the absence of pleural effusion with a high confidence level of 0.98, showing a contrast to typical fluid accumulation characteristics.
	Enlarged cardiac silhouette suggestive of cardiomegaly. Lungs are clear of consolidation, but mild pulmonary vascular congestion is noted, likely due to early congestive changes."	The cardiac silhouette is enlarged. Lungs appear clear . No acute infiltrates observed. Mild vascular congestion inferred from hilar vessel prominence."	- High score for "cardiomegaly" (0.92) as calculated cardiothoracic ratio > 0.55. - "Vascular congestion" score (0.75) from subtle hilar vessel prominence. - Strong focus on lateral cardiac borders (left avg. attention=0.88, right avg. attention=0.85). - Term "mild" for "vascular congestion" selected from "severity quantifiers" cluster - "No acute infiltrates" (confidence 0.90) as lung fields lack dense opacification features, distinguishing from decompensated heart failure.
	Moderate right pleural effusion with associated compressive atelectasis. Left lung parenchyma is normal. No cardiomegaly.	Right-sided pleural effusion is noted (moderate volume) causing obscuration of the costophrenic angle. Left lung remains clear. Cardiac size unremarkable.	- High score for "pleural effusion" (0.93) due to detection of costophrenic angle blunting & meniscus sign on the right. - Focus on right costophrenic angle (attention=0.94) and lateral right pleural lining. - Term "moderate" pulled from cluster "effusion quantifiers", correlated with estimated fluid volume - "Left lung clear" (confidence 0.96) confirmed by sharp left CPA and lack of opacity, ruling out bilateral effusion.

Collectively, Ex-ReGSA demonstrated through representative cases Table III its ability to produce accurate radiology reports that thoroughly include necessary clinical terms and findings from ground truth while maintaining high coherence. The "interpretable and explainable" evidence links generated statements with visual features and advanced analytical components to boost our model's clinical trustworthiness and usefulness.

The necessary degree of transparency supports radiologists by potentially decreasing their reporting workload and enhancing diagnostic speed while functioning as a dependable clinical tool. Ex-ReGSA's architecture demonstrates strong performance across various cases, which highlights its ability to manage different radiological presentations effectively.

VI. Discussion

Our research shows that the Ex-ReGSA model demonstrates robust performance while achieving competitive results in automated radiology report generation, as confirmed by quantitative data and qualitative evaluations along with an insightful ablation study. Ex-ReGSA achieves state-of-the-art or similar scores when compared to existing methods across essential metrics with exceptional performance in all metrics, which indicates high lexical accuracy and semantic depth along with strong structural coherence. Furthermore, the ablation study systematically confirmed the individual and collective contributions of its core architectural modules, namely the X-Semantic Feature Extractor (X-SFE), X-Semantic-Aligned Visual Extractor (X-SAVE), Cross-Modal Fusion (CMF), and the Self-Refining mechanism—each incrementally enhancing report quality towards the Full Model's performance. Complementing these objective scores, our qualitative analysis, showcased in Table III, illustrates Ex-ReGSA's ability to generate clinically relevant and accurate narratives for diverse cases, from identifying chronic conditions to recognizing acute findings such as pleural effusions and even subtle incidentalomas like calcified granulomas.

Ex-ReGSA achieves its strength through a well-designed architecture specifically created to tackle multimodal medical report generation challenges. The collaboration between advanced visual feature extraction methods (X-SFE and X-SAVE) delivers precise identification and interpretation of clinically important image regions. The Cross-Modal Fusion (CMF) module demonstrates essential functionality by combining visual cues with available textual information to create a comprehensive understanding. The output labeled "Interpretable and Explainable" stands out as a major advancement for achieving transparent AI within radiology. The model's decision-making process becomes clear through an analysis of system-detected features along with attention focal points and confidence levels. Ex-ReGSA produces reports while delivering a foundation for interpreting its conclusions. The model's inherent interpretability remains essential for developing clinical confidence and assisting with debugging and refining the model.

When compared to the broader landscape of radiology report generation research, Ex-ReGSA positions itself favorably. Its leading ROUGE-L score, for instance, suggests a superior capability in capturing the essential content and narrative flow of reference reports compared to many existing models. The METEOR score further supports its ability to generate semantically coherent and fluent text. While Ex-ReGSA achieves a solid CIDEr score (0.463 as per the ablation study's Full Model), some specialized models, such as AERMNet (which reported a CIDEr of 0.560 in comparative settings), demonstrate higher performance on this specific consensus-based metric. This suggests that while Ex-ReGSA demonstrates superior performance in report structure and semantic accuracy, it must improve n-gram overlap to maximize CIDEr rewards. While other models focus on single metrics and lack transparency, Ex-ReGSA demonstrates a distinct advantage through its comprehensive capabilities and built-in interpretability features. Therefore, our model, Ex-ReGSA, offers three key advantages over current methods. First, it achieves superior fine-grained accuracy by using a semantic-guided visual extractor (X-SAVE) to focus on specific, subtle lesions that other models miss. Second, it provides truly explainable and trustworthy AI by using Concept Activation Vectors (CAVs) to explain why it makes a diagnosis in quantifiable terms, rather than just showing a heatmap of where it looks. Finally, it ensures enhanced coherence and reliability through a dynamic, self-refining decoder that corrects its output, leading to more logical and professionally structured reports than static generators.

The study achieved promising results but remains limited by certain constraints. Although the IU X-ray dataset serves as a standard benchmark, its limited size and variety fail to capture the full spectrum of clinical scenarios found in worldwide medical practice. Further research is required to understand how the model handles fewer common conditions and complex cases that involve multiple subtle interacting findings. Our CIDEr score competes well but remains behind the field leaders, which demonstrates room for growth in producing reports that match multiple radiologists' preferred terminology. The existing interpretability methods provide valuable insights but represent only one phase in the continuous effort to achieve completely transparent and interactive artificial intelligence.

Future work should focus on several key directions. Addressing the CIDEr gap could involve exploring novel decoding strategies or incorporating more fine-grained common-sense and clinical knowledge into the generation process. Training and validating Ex-ReGSA on larger, more diverse, and multi-institutional datasets would be crucial for assessing its generalizability and robustness. Prospective clinical studies are also essential to evaluate Ex-ReGSA's real-world impact on radiologists' workflow, reporting efficiency, and diagnostic accuracy. Further development of the interpretability framework could involve more interactive explanation modalities, allowing clinicians to query the model's reasoning.

VII. Conclusion

Our research presents Ex-ReGSA as a novel method for automated radiology report generation which combines semantic-guided visual alignment, concept-based explainability via CAVs, and a dynamic self-refining decoder. Our model outperforms the state-of-the-art systems on major evaluation metrics such as ROUGE-L and METEOR, generating medically

accurate and fluent reports that are also interpretable. By explaining its diagnostic decisions in terms that can be quantified, Ex-ReGSA aims to promote transparency and trust in clinical AI, providing a powerful approach to improving diagnostic workflow, reducing radiologist workload, and helping interpretability in clinical decision-making. Although validation is successful, specific limitations are its testing on a single benchmark dataset and the potential to improve consensus-based metrics. Future work will focus on training with larger, multi-institutional datasets and conducting prospective clinical studies to affirm its real-world utility.

REFERENCES

[1] James H. Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, James Brink, Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology, Volume 15, Issue 3, Part B, 2018, Pages 504-508.

[2] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019).

[3] Wang, Xinyi, Grazziela Figueredo, Ruizhe Li, Wei E. Zhang, Weitong Chen, and Xin Chen."A Survey of Deep Learning-based Radiology Report Generation Using Multimodal Data."

[4] Tang, Yuhao & Wang, Dacheng & Zhang, Liyan & Yuan, Ye. (2024). An efficient but effective writer: Diffusion-based semi-autoregressive transformer for automated radiology report.

[5] Tjoa, Erico & Guan, Cuntai. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. PP. 10.1109/TNNLS.2020.3027314.

[6] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making, 20(1), 310.

[7] Y. Tang, Y. Yuan, F. Tao and M. Tang, "Cross-Modal Augmented Transformer for Automated Medical Report Generation," in *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 33-48, 2025.

[8] S. Iqbal, A. N. Qureshi, F. Khan, K. Aurangzeb and M. Azeem Akbar, "From Data to Diagnosis: Enhancing Radiology Reporting With Clinical Features Encoding and Cross-Modal Coherence," in *IEEE Access*, vol. 12, pp. 127341-127356, 2024.

[9] Hongzhao Li, Hongyu Wang, Xia Sun, Hua He, Jun Feng, Context-enhanced framework for medical image report generation using multimodal contexts, Knowledge-Based Systems, Volume 310, 2025,112913.

[10] ianhua Zeng, Tianxing Liao, Liming Xu, Zhiqiang Wang, AERMNet: Attention-enhanced relational memory network for medical image report generation,Computer Methods and Programs in Biomedicine,Volume 244,2024, 107979

[11] H. Tsaniya, C. Fatichah and N. Suciati, "Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement," in *IEEE Access*, vol. 12, pp. 25429-25442, 2024.

[12] K. Zhang *et al*., "Attribute Prototype-Guided Iterative Scene Graph for Explainable Radiology Report Generation," in *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4470-4482, Dec. 2024.

[13] T. Tanida, P. Müller, G. Kaissis and D. Rueckert, "Interactive and Explainable Region-guided Radiology Report Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 7433-7442.

[14] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509, Bangkok, Thailand. Association for Computational Linguistics.

[15] Pham, T.T. et al. (2025). FG-CXR: A Radiologist-Aligned Gaze Dataset for Enhancing Interpretability in Chest X-Ray Report Generation. In: Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H. (eds) Computer Vision – ACCV 2024. ACCV 2024. Lecture Notes in Computer Science, vol 15477. Springer, Singapore.

[16] A. Taleb, M. Kirchler, R. Monti and C. Lippert, "ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 20876-20889.

[17] S. Sangnark, P. Rattanachaisit, T. Patcharatrakul and P. Vateekul, "Explainable Multi-Modal Deep Learning With Cross-Modal Attention for Diagnosis of Dyssynergic Defecation Using Abdominal X-Ray Images and Symptom Questionnaire," in *IEEE Access*, vol. 12, pp. 78132-78147, 2024.

[18] Kinger, S., Kulkarni, V. Transparent and trustworthy interpretation of COVID-19 features in chest X-rays using explainable AI. Multimed Tools Appl 84, 19853–19881 (2025).

[19] Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A. and Hsu, C.N., 2022. RadBERT: adapting transformer-based language models to radiology. Radiology: Artificial Intelligence, 4(4), p.e210258.

[20] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020.

[21] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J. and Viegas, F., 2018, July. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning (pp. 2668-2677). PMLR.

[22] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[23] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[25] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

[26] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[27] Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[28] Li, Y., Liang, X., Hu, Z. and Xing, E.P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems, 31.

[29] Chen, Z., Song, Y., Chang, T.H. and Wan, X., 2020. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.

[30] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 13748-13757.

3D Reconstruction from JPG Images

Youssif Mohamed Mostafa, Maryam N. Al-Berry, Howida A. Shedeed
Scientific Computing Department-Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

***Abstract*—Three-dimensional (3D) reconstruction from two-dimensional (2D) images is a fundamental challenge in computer vision and photogrammetry, with applications in medical imaging, robotics, and augmented reality. This research introduces an image-based modeling pipeline designed to overcome the inherent limitations of Joint Photographic Experts Group (JPEG) images, such as lossy compression and reduced structural fidelity. The proposed hybrid framework integrates photogrammetric methods specifically Structure-from-Motion (SFM) and Dense Stereo Matching with advanced point cloud generation and surface reconstruction techniques. Initially, Marching Cubes was utilized to generate dense point clouds from sequential JPEG slices, followed by Poisson Surface Reconstruction to produce watertight 3D models. Structural details are further enhanced using Structural Similarity index (SSIM) guided texture refinement. Evaluated on the Kaggle Chest CT Segmentation dataset, the method achieves an SSIM score of 0.725, outperforming the JPEG-based reconstruction baseline of 0.675 by 7.4%. In addition to improved accuracy, the study explores the balance between computational cost and reconstruction quality, offering insights relevant to real time and resource constrained applications. By bridging photogrammetry with computer vision, this work advances practical 3D reconstruction from compressed medical images, enabling efficient digitization in low-bandwidth environments.**

Keywords*—*3D reconstruction; photogrammetry; computer vision; image-based modeling; point cloud generation; JPEG images

I. Introduction

Three-dimensional (3D) reconstruction of structures from two-dimensional (2D) images remains a persistent and challenging problem in computer vision, particularly in scenarios where direct 3D acquisition via specialized sensors is either impractical or unavailable. In the medical domain, particularly in diagnostic imaging and surgical planning, the ability to infer 3D anatomy from 2D image data is of critical importance. Classical approaches to this problem are grounded in multi-view geometry and utilize geometric relationships between multiple views to estimate depth, determine camera motion, and recover the underlying spatial structure [1], [2].

The vital components needed to create the 3D reconstructed model are the Red-Green-Blue (RGB) values and the depth map which is considered as the crucial value to reconstruct a watertight model [3]. The depth map contains the distance between the camera viewpoint and the reconstructed surface of the model.

In many clinical workflows, available datasets are stored as JPEG slices in compressed format from Computed Tomography (CT) scans. Although these images offer wide accessibility and storage efficiency, the JPEG format introduces several limitations, including lossy compression, quantization noise, structural artifacts and lack of depth information. Because the more information of the depth of each point in the point cloud, the closer the reconstructed model will be realistic and non-manifold. These issues can adversely affect key stages of the reconstruction pipeline, such as feature detection and matching, depth estimation, and surface extraction, thereby complicating the task of accurate anatomical reconstruction.

In this work, a pipeline for reconstructing 3D anatomical surfaces from JPEG-compressed CT image slices using classical computer vision techniques is proposed. The pipeline begins with robust feature detection and matching across sequential CT slices and masks, by extracting the global bounding box from the masks to extract the point of interest organ across all the CTs and crop them to the exact dimension, followed by camera pose approximation and sparse point cloud generation via SFM [4]. The sparse cloud is then densified using Multi-View Stereo (MVS) to generate a detailed volumetric representation of the anatomical region [5]. To extract a continuous and topologically coherent surface from this volumetric data, the Marching Cubes algorithm is applied [6], which constructs an initial isosurface by interpolating scalar field values across a 3D grid. For further refinement, Poisson Surface Reconstruction is employed [7], which formulates surface recovery as a spatial Poisson problem, enabling the generation of smooth, watertight surfaces that preserve geometric detail and continuity. Finally, face reduction technique is applied to reduce the number of faces and polygons in the model while maintaining a visually similar model according to the SSIM factor.

This study demonstrates that high-quality 3D reconstructions of anatomical structures can be achieved using classical methods, even when relying solely on JPEG-compressed CT images and the estimation of the depth value. The proposed pipeline is evaluated on clinical datasets, highlighting its effectiveness and the trade-offs involved in balancing reconstruction accuracy, computational efficiency, and the limitations of compressed medical imaging data.

The rest of the paper is organized as follows: Section II provides a survey on related works of 3D reconstruction. Section III explains the background and methods used. Section IV details the proposed pipeline, Section V presents the used datasets, experiments, and achieved results. Finally, section VI presents our conclusions and future work directions.

II. Related Work

Many researchers proposed several techniques to contribute to the 3D reconstruction problem either by using deep learning-based methodologies [8], [9], or classical computer vision

techniques [10], [11], [12], [13]. The following introduces a summary of the state-of-the-art methodologies that use classical and deep learning computer vision techniques, since they are more related to the proposed method.

In study [10], the authors started to collect images from the medical field, like CT images, and then improved the quality of the images. Two different methods were used to create the 3D model. The first method, called "volume rendering", was used to scan the CT and MR images. It started from one pixel and produced a steady light beam and scanned each pixel in the image. If the light source is different, the resulting image will be different. The second method, called "surface rendering", used a series of related surfaces from volumed data, and started to extract the outer lines of each slice image. The results demonstrated that the statistical model improved alignment precision and robustness and provided valuable insights for enhancing image-guided diagnosis and interventions in clinical settings.

G. Caravaca *et al.* [11] used 3D reconstruction to simulate the Kimberley outcrop on Mars. They used multiple cameras with different perspectives and camera lenses in different calculated locations to collect images of Gale crater, which contains a lot of geographical terrains. Firstly, they used the cameras that were mounted on the Mars Science Laboratory Rover, known as *Curiosity,* to capture images from Mars' surface. Secondly, they used the captured and collected images from Mars Hans Lens Imager (MAHLI). Finally, they combined the outputs from those two steps in one database. They used 2175 images from the collected images to be passed to a 3D mesh generator to generate a 3D mesh from the overlapped images. The final step was to use those mesh points as an input for a 3D reconstruction software (Agisoft)[1].

In study [12], Siyu Ren and Junhui Hou started with real-world scans like: ScanNet which were fed to Local Geometry representation to get a dense point cloud with un-oriented normal vectors. This dense point cloud was fed to geometry-aware Unsigned Distance Field (UDF) to predict the unsigned distance field, and the last step was to use Edge-based Marching Cubes to generate the output mesh.

In study [8], a method to improve the quality and resolution of CT images was proposed. The method aimed to improve diagnostic precision within the Internet of Health Things (IoHT) framework. They used Convolutional Neural Network (CNN) and 3D reconstruction and super-resolution (SR) to overcome the limitations of low-resolution CT images and make it possible to retrieve finer anatomical information for precise medical analysis. Specifically, the authors integrated a Generative Adversarial Network (GAN) with a Residual Network (ResNet) architecture to effectively upsample and refine CT images while preserving critical structural features. Experimental results demonstrate that the proposed method significantly outperforms traditional interpolation techniques, achieving superior image clarity and resolution as validated by quantitative metrics such as peak signal-to-noise ratio (PSNR) and SSIM.

[1] https://www.agisoft.com/

A. M. I. Mahmoud et al. [9] utilized endoscopic video footage capturing the arteries that supply blood from the heart to the prostate. The video frames were first preprocessed by removing distorted or low-quality images, after which a subset of 30 key frames was selected for analysis. These frames were then enhanced using Adobe Lightroom to reduce noise and improve clarity. Next, a 3D point cloud was generated from the processed frames and converted into a polygonal mesh, forming a detailed 3D model of the arterial structure. This model was used to quantitatively assess the precise volume of prostate tissue that must be removed to optimize blood flow dynamics, to improve vascular stability, and reduce blood pressure.

In study [13], the authors discussed the advancement of 3D CT imaging techniques in thoracic surgery, emphasizing the development of the Resection Process Map (RPM). Traditional 3D CT imaging provided static images, which may not accurately represent the dynamic nature of lung tissues during surgery. To address this, the RPM was developed to generate patient-specific, dynamic 3D images that simulate intraoperative anatomical changes, such as lung deformation due to traction. This semi-automatic system allowed thoracic surgeons to visualize different resection paths and high-quality surgical maps efficiently, enhancing preoperative planning and intraoperative navigation. The RPM demonstrated high accuracy in delineating vascular branches and bronchi, with a median image acquisition time of only 2 minutes, making it a valuable tool in modern thoracic surgical practices.

Recent advances in 3D reconstruction have leveraged various techniques, such as Deep Learning, to improve the quality and the reliability of the reconstructed model, but challenges persist. CNN-based method [8] Have achieved notable results using GANs to generate super-resolution images but often suffer from the lack of ground truth of the reconstructed images and rely on the GAN result image. Additionally, several works including [10] explored to use the CT images to reconstruct the 3D model regards the effect of the background of the CT on the reconstructed model. These limitations highlighted the need for more accurate approaches. In this work, an efficient pipeline that uses the segmented CT images to reconstruct a 3D model from JPG images is proposed.

III. Background and Methods

In this section, the techniques used in the proposed method are discussed.

A. Background

In this section, marching cubes, Poisson surface reconstruction, and face reduction are discussed.

1) Marching cubes algorithm: is a widely used algorithm to extract the polygonal mesh to represent the surface of 3D scalar field. The algorithm processes the voxel grid by comparing the scaler value of the each voxel's corner and the values of user-defined isovalues using a predefined table that represents the possible 256 values of all the configurations of triangle inside the cube, and the triangle of the voxel is generated when the intersection of the voxel and the

isosurface which lead to approximate the isosurface. For accurate rendering, vertex positions are interpolated along voxel edges, and surface normals are computed via gradient estimation. The key parameter which affect the result of marching cubes algorithm is the isovalue **τ,** as it affects the size of the model and the details generated. It decides the isosurface of the 3D scalar field. As the value of this level decreases (eg. 30 %), more details will be generated, and if the value of the level increases (eg.70%), fewer details will be generated and a less harsh isosurface is obtained.

2) Poisson surface reconstruction: is a powerful technique in 3D graphics that reconstructs smooth, watertight surfaces from oriented point clouds commonly acquired from 3D scanners or depth sensors. Unlike mesh-based approaches, it treats surface reconstruction as a spatial Poisson problem, interpreting input points and their normals as samples of an indicator function's gradient field. By solving this Poisson equation over an adaptive octree structure, the method derives an implicit function whose zero level set robustly defines the output surface. Key advantages include intrinsic noise resilience, efficient handling of large datasets, and topologically guaranteed results (e.g., no holes or self-intersections) [7]. These properties make it indispensable for applications demanding high fidelity, such as reverse engineering, cultural heritage digitization, and medical imaging, where accurate and complete surface models are critical. The key parameter of Poisson surface reconstruction is the depth. As the value of the depth increases, more details will be generated, but it will take more time to be rendered. If the value of the depth decreases, less details will be generated and the rendering will require less time.

3) Face reduction: is a fundamental technique in 3D modeling and computer graphics aimed at reducing the number of polygonal faces in a mesh while retaining its essential shape, visual fidelity, and structural integrity. This process, also called mesh simplification or decimation, is critical for optimizing 3D models in performance-sensitive applications such as video games, virtual reality, and real-time web rendering, where lower computational overhead and efficient memory usage are paramount. Advanced algorithms achieve this by strategically collapsing edges, merging vertices, or clustering polygons, prioritizing geometric features that contribute most to the model's silhouette and surface detail. Popular methods, including edge collapse, vertex clustering, and Quadric Error Metrics (QEM), offer distinct trade-offs: QEM preserves curvature with high accuracy, while clustering enables faster, coarser simplification. The choice of the technique depends on the target balance between computational efficiency, geometric precision, and artistic control over the final mesh [14].

IV. Proposed Method

Fig. 1 shows the architecture of the proposed method. First, pre-processing operations are applied to the input raw CT images. The next two consecutive stages to extract the surface and point cloud: a Marching Cubes stage, which is responsible for edge extraction for each image, and a Poisson Surface Reconstruction stage, which works to extract the mesh and faces from the provided edges. Finally, a face reduction step is used to remove redundant vertices and edges. The following sub-sections highlight the detailed procedure for each phase.

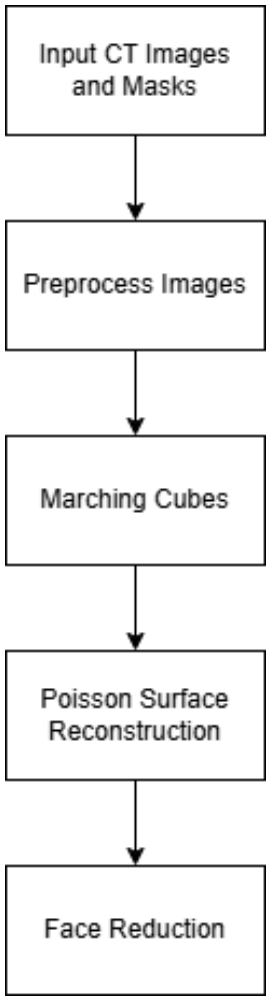
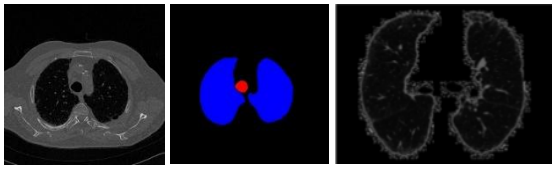


Fig. 1. The proposed annotation pipeline.

A. Preprocessing

The dataset CT images appear in the shape of (W_i, H_i, 3) Where W_i represents the width of the CT slice i, H_i represent the height of CT Slice i and 3 for the three channels of each image (red, green, blue). These input images go through a sequence of pre-processing operations as follows: as shown in Fig. 2, first each slice is processed by applying pixel-wise product with its corresponding mask to extract the image with the organ of interest (in our case the blue labeled organ in Fig. 2). Second, the bounding box is extracted for each slice by calculating its local minimum x-axis value and local maximum x-axis value and the same is done in the y-axis direction as well. The third step is to crop each slice to extract the organ image without the distortion of the background. Finally, all the images are stacked in one 3D array to be used as one unit, and this stack is in the shape of (W_i, H_i, n), where n is the number of slices.



(a) (b) (c)

Fig. 2. Steps for (a) Original image, (b) Mask of the image, (c) Result image after cropping and masking.

B. Point Cloud Extraction

In this step, two algorithms are used consecutively:

1) The first is the marching cubes algorithm: The Marching Cubes algorithm is used to convert the volumetric data provided by the stacked slices into a mesh by interpolating the scalar value. It uses an iso-surface thresholding to separate the tissue of the patient organ from the background to create the surface of the organ.

2) The second algorithm is poisson surface reconstruction: The point cloud provided by the marching cubes algorithm is used to create a smooth, watertight mesh to refine the resulting surface and improve the mesh.

C. Face Reduction Step

In this step the number of the faces generated by the Poisson surface reconstruction algorithm is reduced, as shown in the results section, the number of the vertices and faces in the initial 3D model are too much to be visualize, so the number of the faces of the result model are reduced by removing the non-manifold edges, removing degenerate triangles and finally removing duplicated vertices.

V. Experimental Results

The proposed model is evaluated against a benchmark dataset, which is available on Kaggle platform.

A. Dataset

Chest-ct-segmentation[2]: The dataset consists of 16078 CT Slices and 16078 corresponding masks for a total of 114 patients. Each CT image is represented in GrayScale format while its corresponding mask is represented in RGB format. Each channel in the mask is responsible for representing a segment area for its corresponding organ. Red was used for the Trachea, blue for the Lung, and green for the Heart. In the experiments performed, only the blue channel is used as the Lungs of the patient contain a lot of details and also will be easier to visualize.

Each patient has their own number and the CTs for each patient has it's own number like: ID00323637202285211956970-1.

Three different patients were selected, and for each patient, two different organs were reconstructed from the CT scans. The selected organs were the Lung and the Heart. The IDs of the patients selected in the experiments are listed below, along with the number of CT images available in the dataset:

1) ID00323637202285211956970: 258 CT images and the corresponding 258 CT masks.

2) ID00411637202309374271828: 268 CT images and corresponding 268 CT masks.

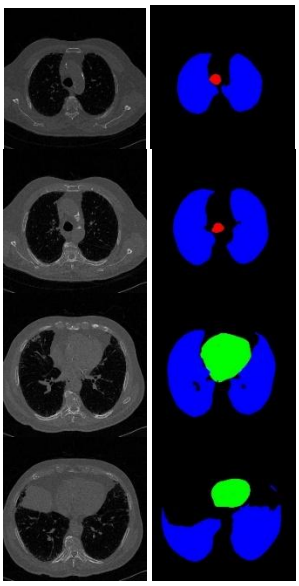
3) ID00329637202285906759848: 260 CT images and corresponding 260 CT masks.

Each CT image is of 512×512 size, and the corresponding mask is 512×512×3 with each channel having the mask of a spacefic Organ. Fig. 3 shows sample snips from the used dataset.

B. Environmental Settings

The proposed model was implemented using Python 3.8[3] with TensorFlow 2.4 [15] backend and some helper libraries like OpenCV[4], NumPy[5], Open3D[6], scikit-image[7], scikit-learn[8], trimesh[9] and joblib[10].

The Marching Cubes algorithm is used with a level equal to 40 to capture more details and produce a richer model. Depth in the Poisson Surface Reconstruction algorithm is set to 11 to create a more detailed model, but it affects the reconstruction time. Regarding the hardware used, the system was run on Kaggle[11], which has a GPU Tesla P100, and 2 CPUs, 30 GB RAM, and 5 GB as hard drive. Finally, MeshLab[12] open-source software is used to visualize the resulting 3D model.



(a) (b)

Fig. 3. Samples from the Chest-CT-segmentation dataset, (a) CT images and (b) Corresponding masks.

[2] https://www.kaggle.com/datasets/polomarco/chest-ct-segmentation

[3] https://keras.io/
[4] https://opencv.org/
[5] https://numpy.org/
[6] https://www.open3d.org/
[7] https://scikit-image.org/
[8] https://scikit-learn.org/stable/
[9] https://trimesh.org/
[10] https://joblib.readthedocs.io/en/stable/
[11] https://www.kaggle.com/
[12] https://www.meshlab.net/

C. Results and Discussions

1) Experiments: In this section, the model performance is discussed. In the first experiment, we use the Marching Cubes Algorithm only and try to reconstruct the model. In the second experiment we combined the Marching cubes algorithm and Poisson Surface Reconstruction. After that we experimented the whole proposed pipeline in which we get the bounding box of the organ to reduce the size of the input image and its corresponding mask, after that we combine face reduction technique to reduce the redundant vertices and faces. Table I compares the memory cost, time cost, visual evaluation and size of the reconstructed model for the of all the experiments.

As shown in Table I, the proposed method has achieved the minimum size of the reconstructed model, as outlined in the proposed steps. First, the global bounding box of the organ from the CT image masks was extracted using its corresponding channel mask. Second, the CT bounding boxes were stacked in one stack. Third, the projection of the masks on the original CT images in the stack were used to separate the organ from the background and the resulting stack of images utilized as an input for the Marching Cubes algorithm.

TABLE I. COMPARISON RESULTS OF 3D RECONSTRUCTION

Method	Memory Cost (MB)	Time (minutes)	Size of model (MB)	Visual Evaluation
Marching Cubes	680	3	3.87	sparse representation
Marching Cubes + Poisson Surface Reconstruction	1056	7	1200	sparse representation with slice problems
Marching Cubes + Poisson surface reconstruction (with bounding box)	1214	8	400	More dense and detailed, but with a big size and redundant point problem
Marching Cubes + Poisson surface reconstruction with bounding box+ faces reduction (**Proposed**)	2680	15	12.8	Dense and detailed representation

To test the generalization of the proposed model, two different sets of experiments are performed. The first experiment is based on an EDLF-CGAN model in the data set used to extract our model, and in the second experiment the proposed model was applied on the dataset, and the results were evaluated based on peak signal-to-noise ratio (PSNR) in Eq. (1) and structural similarity (SSIM) in Eq. (3).

The PSNR is a statistical analysis indicator that is based on the gray level of image pixels, which is defined by the mean square error (MSE) in Eq. (2) between the original image I (i, j) and the reconstructed image projection from the reconstructed model K (i, j). Generally, the higher the PSNR value is, the better the image restoration is

$$PSNR = 10 \times log_{10}\left(\frac{(2^n - 1)^2}{MSE}\right) \quad (1)$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m}\sum_{j=0}^{n}\|I(i,j) - K(i,j)\|^2 \quad (2)$$

SSIM is a similarity indicator of two images. The first image is the original image, and the second is the reconstructed image projection from the reconstructed model.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

μ: Average intensity, σ: Standard deviation, σ_{xy}: Cross-correlation. C_1, C_2C_1, C_2: Stability constants.

2) Results: Table II shows the achieved results of those experiments for each patient on the lung segment against EDLF-CGAN [8]. Table III shows the achieved results of the heart segment.

TABLE II. PSNR AND SSIM ON LUNG

Model	PSNR	SSIM
EDLF-CGAN[8] (patient 1)	34.7458	0.541
EDLF-CGAN[8] (patient 2)	34.8461	0.551
EDLF-CGAN[8] (patient 3)	34.7470	0.543
Proposed (Patient 1)	36.6978	0.710
Proposed (Patient 2)	**37.1298**	**0.725**
Proposed (Patient 3)	36.6989	0.715

TABLE III. PSNR AND SSIM ON HEART

Model	PSNR	SSIM
EDLF-CGAN[8] (patient 1)	32.0021	0.422
[8] (patient 2)	32.8991	0.466
EDLF-CGAN[8] (patient 3)	32.0374	0.430
Proposed (Patient 1)	35.1347	0.632
Proposed (Patient 2)	**35.7804**	**0.675**
Proposed (Patient 3)	35.1487	0.638

As shown in Tables II and Table III, the performance of the proposed pipeline shows a significant improvement in SSIM when compared with EDLF-CGAN [8]. This is especially particularly in patient 2 as a slightly higher number of CTs and masks compared to the other patients. On the other hand, the PSNR was slightly improved with no statistically significant change. Additionally, the results of the lung show a slight increase over the heart. This is probably because the lung has a bigger reconstruction surface area and fewer structural discontinuities (such as voids or holes) in the segmentation masks.

In addition to the fact that the proposed method shows a slight enhancement in the time needed to reconstruct the model by about 27%, it also reduces the storage space used by the model by 60%. Fig. 4 shows a sample of the reconstructed models.

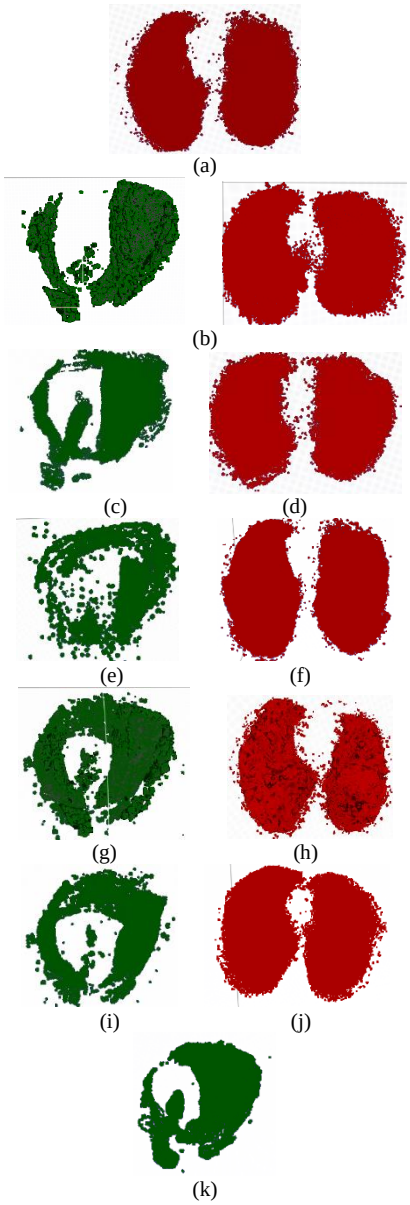


Fig. 4. Examples of models (a) Proposed patient 1 lung, (b) Proposed patient 1 heart, (c) Proposed patient 2 lung, (c) Proposed patient 2 heart, (d) Proposed patient 3 lung, (e) Proposed patient 3 heart, (f) EDLF-CGAN[8] patient 1 lung, (g) EDLF-CGAN[8] patient 1 heart, (h) EDLF-CGAN[8] patient 2 lung, (i) EDLF-CGAN[8] patient 2 heart, (j) EDLF-CGAN[8] patient 3 lung, (k) EDLF-CGAN[8] patient 3 heart.

VI. Conclusion

3D reconstruction is one of the most used fields that has become popular in recent decades. It is used in many applications like medical field applications, civil engineering field applications, games, space discovery, and urban cities reconstruction field. To avoid using learning-based models, this work proposed a traditional pipeline for three-dimensional reconstruction from JPEG-compressed computed tomography (CT) images. The proposed pipeline successfully creates precise and smooth 3D anatomical surfaces by combining Marching Cubes, Multi-View Stereo, Structure-from-Motion, and Poisson Surface Reconstruction. The suggested method shows that classical multi-view geometry and volumetric processing may produce high-quality reconstruction despite the difficulties caused by lossy compression and low depth information in JPEG photos. Results from experiments verify the pipeline's resilience and applicability for medical imaging and visualization applications, particularly when there is only compressed image data available. Even while the suggested pipeline shows encouraging outcomes when it comes to reconstructing 3D anatomical structures from JPEG-compressed CT data, there are still a number of unexplored regions. Future directions may include enhancements in the posture estimation and surface continuity accuracy by advances in feature identification and matching across very comparable CT slices. Second, the negative impacts of JPEG compression may be further reduced by including image preprocessing methods like denoising, contrast enhancement, or artifact correction. Furthermore, the method's generalizability would be enhanced by expanding its support to include additional medical imaging modalities like MRI or PET. Finally, a balance between interpretability and adaptation to intricate anatomical changes may be provided by using hybrid techniques that integrate lightweight, explainable machine learning modules with classical geometry.

References

[1] H. Ham, J. Wesley, and Hendra, “Computer vision based 3D reconstruction : A review,” International Journal of Electrical and Computer Engineering, vol. 9, no. 4, pp. 2394–2402, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2394-2402.

[2] S. Speidel, S. Bodenstedt, F. Vasconcelos, and D. Stoyanov, “Interventional imaging: Vision,” Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 721–745, Jan. 2019, doi: 10.1016/B978-0-12-816176-0.00034-X.

[3] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” IEEE Trans Pattern Anal Mach Intell, vol. 33, no. 5, pp. 898–916, 2011, doi: 10.1109/TPAMI.2010.161.

[4] A. Eltner and G. Sofia, “Structure from motion photogrammetric technique,” Developments in Earth Surface Processes, vol. 23, pp. 1–24, Jan. 2020, doi: 10.1016/B978-0-444-64177-9.00001-1.

[5] Y. Furukawa and C. Hernández, “Multi-view stereo: A tutorial,” Foundations and Trends in Computer Graphics and Vision, vol. 9, no. 1–2, pp. 1–148, Jun. 2015, doi: 10.1561/0600000052.

[6] J. W. Suh and Y. Kim, “Example in Computer Graphics,” Accelerating Matlab with GPUs, pp. 157–191, Jan. 2014, doi: 10.1016/B978-0-12-408080-5.00007-9.

[7] M. Kazhdan, M. Bolitho, and H. Hoppe, “Poisson Surface Reconstruction,” Eurographics Symposium on Geometry Processing, 2006.

[8] J. Zhang et al., “3D Reconstruction for Super-Resolution CT Images in the Internet of Health Things Using Deep Learning,” IEEE Access, vol. 8, pp. 121513–121525, 2020, doi: 10.1109/ACCESS.2020.3007024.

[9] A. M. I. Mahmoud, E. N. Mohamed, A. M. Tawfeek, and A. Y. Elbanhawy, “A computational approach to standardise prostatic resection interventions by virtual reconstruction,” 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 413–418, Dec. 2021, doi: 10.1109/ICICIS52592.2021.9694165.

[10] J. Yao and R. Taylor, “Assessing Accuracy Factors in Deformable 2D/3D Medical Image Registration Using a Statistical Pelvis Model,” 2003.

[11] G. Caravaca, S. Le Mouélic, N. Mangold, J. L’Haridon, L. Le Deit, and M. Massé, “3D digital outcrop model reconstruction of the Kimberley outcrop (Gale crater, Mars) and its integration into Virtual Reality for simulated geological analysis,” Planet Space Sci, vol. 182, Mar. 2020, doi: 10.1016/j.pss.2019.104808.

[12] S. Ren, J. Hou, X. Chen, Y. He, and W. Wang, "GeoUDF: Surface Reconstruction from 3D Point Clouds via Geometry-guided Distance Representation," Proceedings of the IEEE International Conference on Computer Vision, pp. 14168–14178, Nov. 2022, doi: 10.1109/ICCV51070.2023.01307.

[13] T. F. Chen-Yoshikawa, "Evolution of Three-Dimensional Computed Tomography Imaging in Thoracic Surgery," Cancers 2024, Vol. 16, Page 2161, vol. 16, no. 11, p. 2161, Jun. 2024, doi: 10.3390/CANCERS16112161.

[14] D. P. Luebke, "A developer's survey of polygonal simplification algorithms," IEEE Comput Graph Appl, vol. 21, no. 3, pp. 24–35, 2001, doi: 10.1109/38.920624.

[15] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, "Sample size selection in optimization methods for machine learning," Math Program, vol. 134, no. 1, pp. 127–155, Aug. 2012, doi: 10.1007/S10107-012-0572-5.

A Focused Survey of ECG Datasets for Artificial Intelligence-Based Atrial Fibrillation Detection

ASSALHI Imane[1], Bybi Abdelmajid[2], Oulad Hamdaoui Hanaa[3],
Ebobisse Djene Yves Frederic[4], Drissi Lahssini Hilal[5]
MEAT Laboratory, EST Sale, EMI Rabat, Mohammed V University in Rabat, Sale, Rabat, Morocco[1, 2, 3]
SmartiLab Laboratory, EMSI Rabat, Rabat, Morocco[1, 4]
LASTIMI Laboratory, EST Sale, EMI Rabat, Mohammed V University in Rabat, Sale, Rabat, Morocco[5]

***Abstract*—Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia and increases the risk of stroke, heart failure, and mortality. Electrocardiography (ECG) is the most important technology for AF detection because it is inexpensive, non-invasive, and provides clinically useful information. However, the variability of ECG patterns, particularly during paroxysmal AF creates challenges in detecting AF. Artificial Intelligence (AI) offers a promising opportunity to improve AF recognition. However, AI performance is contingent on obtaining high-quality and diverse ECG datasets. This paper presents a focused survey of 15 publicly available and clinical ECG datasets used in AI-driven AF detection research between 2023 and 2025. We analyze the datasets based on acquisition methods, ECG type, format, lead configurations, annotation richness, and their application in AI models. Our comparative analysis reveals major trends, challenges such as data imbalance and motion artifacts, and gaps in current datasets including limited demographic diversity and underrepresentation of wearable ECG data. This study aims to guide future research toward more robust, interpretable, and inclusive AF detection models.**

***Keywords*—Atrial fibrillation; ECG datasets; Artificial Intelligence; AI-ECG; dataset survey; AF detection**

I. Introduction

Atrial fibrillation (AF) is a serious and prevalent cardiac condition characterized by irregular electrical activity in the atria, which can lead to blood clots, stroke, and heart failure [1]. As the global burden of AF rises, early and accurate detection has become critical for timely treatment and risk management. Electrocardiograms (ECGs) serve as the gold standard for detecting AF due to their ability to capture cardiac electrical activity in a fast, non-invasive, and cost-effective manner [2].

In recent years, Artificial Intelligence (AI), particularly machine learning (ML) and deep learning (DL), has shown remarkable performance in analyzing ECG signals to automatically detect AF. However, these AI models depend heavily on the quality, structure, and diversity of the ECG datasets used for training and validation. While previous surveys often focus on AI architectures, little attention has been paid to the ECG datasets themselves.

This paper addresses that gap by providing a focused survey of ECG datasets used in AI-based AF detection. The survey is structured as follows: First, a technical background discusses ECG types for AF diagnosis and highlights commonly used AI technologies. Second, the related work section emphasizes the importance of our survey and outlines our added value compared to existing reviews, which often overlook the role of ECG datasets in AI-based AF diagnosis. Third, we present our methodology for assembling recent studies from 2023-2025 and extracting the datasets used. The survey results follow, organized in two tables: one summarizing the extracted datasets, and the second providing an analysis grid of 12 selected studies detailing preprocessing techniques, AI models, architectures, and performance metrics. Finally, we discuss insights from these results, highlighting trends, challenges, and future research directions.

II. Background

Electrocardiograms (ECGs) record the heart's electrical activity and can be captured using different lead configurations, such as the standard 12-lead or single-lead used in wearable devices. The duration of ECG recordings can vary from seconds to days, with longer recordings, known as Ambulatory ECGs, often obtained via wearable devices. These continuous recordings are particularly useful for detecting paroxysmal atrial fibrillation (AF), which occurs intermittently and may not be captured in shorter ECG sessions. To address this, AI models like Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Convolutional Neural Networks (CNNs), and Long Short-Term Memory Networks (LSTMs) are trained on diverse ECG data, including ambulatory recordings, with recent models incorporating explainable AI (XAI) techniques to improve interpretability and clinical trust [3]. The data can be stored as raw signals or transformed into images for visual or AI-based analysis [4].

III. Related Work

Recent studies have explored AI-based tools for assessing atrial fibrillation (AF) burden, showing high correlation with manual physician assessments. Notably, AI demonstrated strong agreement and minimal bias, offering an efficient and accurate alternative for AF burden evaluation [5]. However, when it comes to AI-driven AF detection, the dataset plays a crucial role in the precision of the diagnosis. The quality, variety, and characteristics of ECG datasets significantly impact the performance of AI models. Despite the growing importance of these datasets, there is a noticeable gap in existing reviews and surveys. Many reviews discuss ECGs from a biometric perspective [6] or focus on ECGs for arrhythmia detection in general, but few address the specifics of AI-based AF detection[7], [8]. My previous review focused on the latest AI

technologies for AF detection [9], but this current work expands by compiling and analyzing the ECG datasets specifically used for AI-driven AF diagnosis. This review aims to fill the gap by highlighting the diverse datasets used in AF detection, which will aid future research in understanding the strengths and limitations of various data sources and model implementations in the field.

IV. METHODOLOGY

To conduct this dataset-oriented survey on atrial fibrillation (AF) detection using artificial intelligence (AI) models applied to electrocardiogram (ECG) data, we followed systematic review principles, inspired by PRISMA guidelines as shown in Fig. 1, to ensure transparency and reproducibility in the selection and analysis of studies. Bibliographic references were collected using the following keywords: "Artificial Intelligence", "AI", "Machine Learning", "ML", "Deep Learning", "DL", "Atrial Fibrillation", "AF", "AFib", "Electrocardiogram", and "ECG". These were queried in several major scientific databases, including Scopus, Science Direct, Web of Science, and IEEE Xplore. As part of the inclusion criteria, we considered only papers that had the core terms in the title, ensuring a strict focus on AF detection using AI applied to ECG data with a total of 304 papers. Additionally, the time frame was restricted to 2023–2025 and limited to 25 June 2025, and only articles published in English were considered. We included only original research articles, conference papers, and technical studies, while review and survey papers were examined separately to identify potential overlaps or gaps but were ultimately excluded from the analysis. All retrieved references were exported into Mendeley for management and screening. After removing duplicates, 81 unique entries remained. During the screening process, papers were excluded if they:

- Focused on physiological signals other than ECG (e.g., PPG or ECHO).
- Studied diseases related to AF rather than AF itself (e.g., stroke outcomes or AF recurrence after catheter ablation).
- Did not report the use of AI techniques for AF detection.

Following the title and abstract screening, 48 papers met the inclusion criteria. In the full-text screening phase, 33 full-text papers were successfully retrieved. After reviewing the literature, it became clear that a lack of standardization in ECG datasets used across AI models was a significant limitation. This prompted a focus on compiling a dataset-oriented survey of ECG datasets for AI-based AF detection. We've selected 33 studies from 2023 to 2025, extracting key information about 15 different ECG datasets. The dataset attributes included ECG type, acquisition method, sample size, format, annotation quality, and availability. To gather this information, we conducted online searches for dataset sources and details, including any risk factors or clinical labels associated with each dataset. The extracted data were organized in Table I, summarizing each dataset's key characteristics and their relevance to AI-based AF detection. This compilation provides a comprehensive overview of the datasets, filling a gap in the current literature.

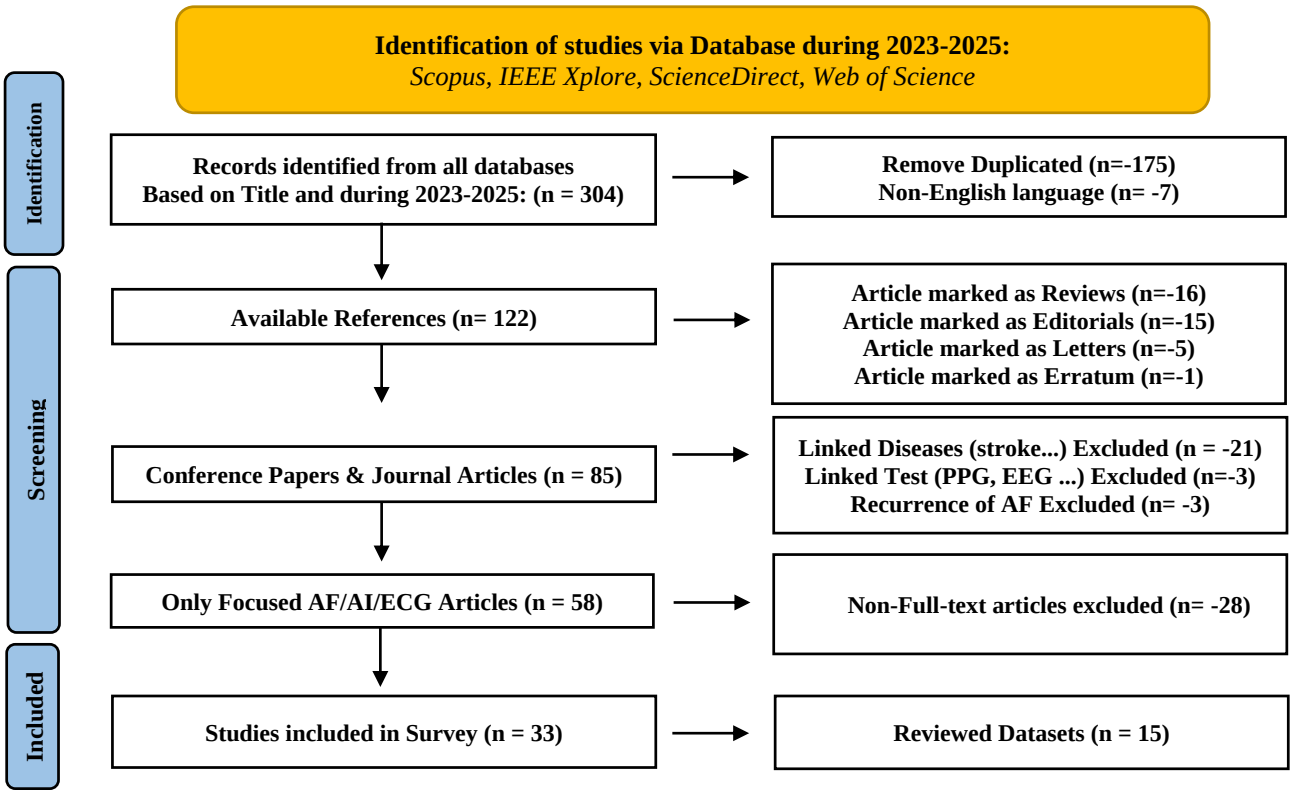


Fig. 1. PRISMA flow diagram representing the paper selection process for the dataset-oriented survey on AI-based Atrial Fibrillation detection.

TABLE I. SUMMARY OF ECG DATASETS USED FOR ATRIAL FIBRILLATION DETECTION

Ref	ABR	Name of Dataset	Year	Source type	Acquisition device	Sample Size	Duration	ECG Type	ECG Format	Number of ECG leads	Annotation	Link	Papers
[10]	CPSC 2018	- The China Physiological Signal Challenge Database - CPSC-Extra Database	2018	Public	Clinical devices from 11 hospitals	10 330 records: - Set 1: 6877 Records - Set 2: 3453 Records	6s to 60s	Stand-ard ECG	Signal	12	- ECG only - Age & Sex only	The China Physiological Signal Challenge 2018	[11] [12] [13]
[14]	INCART	St. Petersburg Institute of Cardiological Technics 12lead Arrhythmia Database	2008	Public	Holter Monitor	75 Records from 32 Patient	30 min	Ambu-latory ECG	Signal	12	Over 175,000 beat annotations; auto-detected and manually corrected; includes age, sex, diagnosis.	St Petersburg INCART 12-lead Arrhythmia Database v1.0.0	[12] [15]
[16]	PTB	PTB Diagnostic ECG Database	2004	Public	PTB custum prototype Recorder	549 Records from 290 Patient	Varies up to minutes	Rest-ing ECG	Signal	15 leads: •12 •3 Frank	Diagnosis, age, gender, history, medications, interventions, and clinical summaries *(some records missing Annotations).*	PTB Diagnostic ECG Database v1.0.0	[11] [12] [13] [17] [18]
[19]	PTB-XL	PTB-XL, a large publicly available ECG dataset	2019	Public	Device from Achiller AG	21 799 Records from 18 689 Patients	10s	Res-ting ECG	Signal	12	SCP-ECG codes (71 statements), multi-label: diagnostic, form, rhythm; metadata includes age, sex, height, weight, heart axis, infarction stage, signal noise info.	PTB-XL, a large publicly available electrocardiography dataset v1.0.3	
X	Georgia	The Georgia 12-lead ECG Challenge (G12EC) Database	2000	Public	Not specified *(Assuming the clinical devices)*	10 344 Records	10s	Rest-ing ECG	Signal	12	Labels not clearly defined in brief, Includes diagnosis/class	Georgia 12-Lead ECG Challenge Database	[11] [12]
[20]	2017 PhysioNet Challenge	The PhysioNet/Computing in Cardiology Challenge 2017	2017	Public	Kardia Mobile: AliveCor	8528 Records	9s to 60s (Varies)	Mini-mal ECG	Signal	Single lead	Rhythm Labels : • Normal • Atrial • Fibrillation • Other Rhythms • Noisy	AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017 v1.0.0	[21] [22] [23] [24] [25] [26] [27] [28]
X	MUSE	the MUSE cardiology information system (GE,Healthcare, Chicago, IL, USA)	X	Clinical Not Avail-able	Clinical	Diverse and Over 300 000 Records	Usially 10s	Rest-ing ECG	Signal	12	• Raw waveform signals • Demographics (age, sex, etc.) • Clinical diagnoses and physician interpretations • ECG measurements (HR, PR interval, QT, etc.) • Not publicly available, • Access typically requires collaboration with a hospital using the MUSE system.	MUSE Cardiology Information System | GE HealthCare (United Kingdom)	[3] [29]

Ref	ABR	Name of Dataset	Year	Source type	Acquisition device	Sample Size	Duration	ECG Type	ECG Format	Number of ECG leads	Annotation	Link	Papers
X	PHYJ	Hospital of Yangjiang (PHYJ),Yangjiang, China, Database	X	Clinical Not Available	Clinical	100 Records from 100 Patients	15s	Resting ECG	Array	12	• Labels confirmed by Holter monitoring and cardiologist review • Not publicly available.	X	[30]
X	EHR	the ECG and electronic health record (EHR) databases from AZ Delta	2022	Clinical Not Available	GE MUSE Cardiology System	173 537 ECGs from 68 880 Patients	10s	Resting ECG	MUSE Format	• 12 • Single	• AF diagnosis from structured EHR + MUSE ECG labels • Not publicly available	Best Electronic Health Record (EHR) Datasets & Databases 2025 | Datarade	[31]
[32]	Chapman	Chapman University and Shaoxing People's Hospital database,	2020	Clinical Available	GE MUSE Cardiology System	10 646 Patients	10s	Resting ECG	Digital	12	• Expert-labeled: - 11 rhythms + 67 conditions	A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients	[13] [33] [34]
X	CNUH	Chonnam National University Hospital databases	2022	Clinical Not Available	MobiCARE-MC100	3059 ECGs from 6720 patients	60s	Mobile ECG	Digital	Single	• Labeled as "AF" (masked AF) or "Healthy" based on 12-lead ECG history • Not publicly available	X	[35]
[36]	AFDB	MIT-BIH Atrial Fibrillation Database	2000	Public	Clinical Ambulatory ECG recorders	25 records	10h	Ambulatory ECG	Digital	2 leads	• Rhythm: AFIB, AFL, J, N (manual) • Beat: .qrs (auto), .qrsc (manually corrected, some records)	MIT-BIH Atrial Fibrillation Database v1.0.0	[37] [28] [38] [39]
[40]	LTAF	The Long Term AF Database	2000	Public	Clinical Ambulatory ECG recorders	86 Records from 80 Patients	21h to 24h	Ambulatory ECG	Digital	2 lead or 3 leads	• ST episode annotations • beat annotations • ST level measurements	Long Term ST Database v1.0.0	[38]
[41]	MITDB	The MIT-BIH Arrhythmia Database	2001	Public	Clinical Ambulatory ECG recorders	48 Records from 47 Patients	30min	Ambulatory ECG	Binary	2 leads	• Beat-by-beat annotations by cardiologists (~110,000 beats annotated)	MIT-BIH Arrhythmia Database v1.0.0	[26] [28] [38]
[42]	NSRDB	MIT-BIH Normal Sinus Rhythm Database	2000	Public	Clinical Ambulatory ECG recorders	18 Subjects	24h	Ambulatory ECG	Binary	2 leads	• No significant arrhythmias; • Rhythm annotations only • Healthy volunteers	The MIT-BIH Normal Sinus Rhythm Database	[38]

V. ECG DATASETS LANDSCAPE

Over the past decade, a wide range of ECG datasets have been utilized to develop AI models for Atrial Fibrillation (AF) detection. Table I presents a comparative analysis of 15 ECG datasets that were extracted following the used methodology in section before. These datasets differ significantly in terms of source (public vs. clinical), ECG type (resting vs. ambulatory), number of leads (from single-lead to 15-lead), signal duration (6 seconds to 24 hours), and data format (digital signals, WFDB, MUSE, or image-based ECG scans).

Publicly accessible datasets like PTB-XL, AFDB, MITDB, and the PhysioNet 2017 Challenge have been widely adopted due to their availability and rich rhythm annotations [19], [20], [36], [41]. Clinical datasets such as MUSE, PHYJ, and CNUH, while offering more realistic clinical scenarios, are often inaccessible, limiting reproducibility and external validation[44]. There is growing interest in single-lead and ambulatory ECG datasets, especially for wearable and real-time monitoring applications [36].

In contrast, 12-lead ECGs remain the most common in clinical datasets due to their diagnostic richness [19], [32]. Regarding duration, some datasets contain short snapshots (typically 10 seconds) while others, such as LTAF and MITDB, include long-term ambulatory ECGs suitable for detecting paroxysmal or transient AF episodes. Datasets vary considerably in annotation granularity. Some, like the PhysioNet 2017 Challenge and AFDB, provide beat-level or rhythm-level labels, which are essential for supervised learning. Others, like PTB and CNUH, contain diagnostic summaries but lack detailed temporal annotation, which restricts their use in time-series AF detection tasks. A few datasets offer clinical metadata such as

age, sex, and comorbidities, but this information remains inconsistent and often incomplete, hindering the study of model fairness and demographic generalization. A variety of acquisition devices are represented, including hospital-grade machines, Holter monitors, and portable solutions like KardiaMobile [5], [45]. Additionally, several large-scale, non-public clinical datasets (e.g., from hospitals in the US, Israel, Japan, and the UK Biobank) have been used in recent studies to validate generalization performance across diverse patient populations [46], [47]. One notable example involved over 320,000 ECGs from 130,000+ patients using both Philips and GE systems [48]. These real-world datasets, though inaccessible to the broader community, are essential for robust AF detection, especially in challenging cases such as paroxysmal or asymptomatic AF [40], [41], [49], [50], [51]. Despite recent advancements, the lack of standardized annotations, inconsistent demographic diversity, and limited access to prospective real-world datasets remain key barriers to model development and evaluation

VI. AI Models Landscape for AF Diagnosis

Table II provides a comprehensive overview of 12 representative studies from 2024–2025 that apply AI techniques to ECG datasets for atrial fibrillation (AF) detection. Each entry outlines the dataset used, preprocessing pipeline, model architecture, performance, and key innovations. Together, they illustrate the current state of applied AI in this domain and support a synthesis of prevailing trends, methods, and research directions.

Key Observations:

- **Dominance of Deep Learning (DL):** CNNs remain the most widely adopted models, often serving as backbone architectures in hybrid setups (e.g., CNN-LSTM, CNN-RNN). Lightweight architectures (e.g., MobileNet, ultra-compact CNNs) are increasingly favored in resource-constrained environments like wearables and mobile devices [3][11][21].
- **Rise of hybrid and transformer models:** Hybrid models combining CNNs with LSTMs or GRUs enhance temporal pattern recognition. Transformer-based models are also emerging, particularly in multimodal settings (e.g., ECG + HRV + demographics), demonstrating high AUROC scores (e.g., 0.9668 with VGG-16) [24] [33] [38].
- **Preprocessing pipelines vary widely:** Signal preprocessing includes traditional filtering (e.g., Butterworth, Pan-Tompkins), wavelet transforms (e.g., CWT), and advanced time–frequency representations (e.g., STFT, spectrograms). Some models integrate domain-specific techniques like ECG segmentation (PQRST) or RR interval extraction for feature fusion. [22] [30] [33] [35] [37].
- **Explainability and interpretability:** Explainable AI (XAI) is increasingly used through techniques like Layer-wise Relevance Propagation (LRP), anomaly scoring, and saliency maps. Studies applying LRP or segment-wise analysis provide insights into model behavior and potential diagnostic biomarkers, such as T-wave/ST-segment deviations in paroxysmal AF[3] [33] [34].
- **Dataset-specific adaptations:** Several studies tailor their models to specific datasets public or clinical accounting for noise, class imbalance, and sampling frequency. For example, the use of DBSCAN-GAN for denoising and synthetic augmentation on noisy wearable ECG datasets shows promise for generalization [24].
- **Clinical integration and risk factors:** A subset of studies incorporates demographic features (e.g., age, sex) and electronic health record (EHR)-derived risk scores. This multimodal fusion enhances AF prediction, particularly in identifying paroxysmal AF from normal sinus rhythm, and aligns with real-world deployment needs [31].
- **Performance benchmarks:** Reported metrics span AUROC (up to 0.98), F1-scores (up to 0.99), and accuracy (>96% in some cases). However, performance varies across datasets and is often affected by noise, signal duration, or sampling frequency. Lightweight models maintain high performance with reduced computational cost, enabling deployment in real-time monitoring systems [18].
- **Cross-dataset and external validation:** Some studies emphasize generalization by testing across multiple public datasets (e.g., PTB-XL, MITDB, PhysioNet 2017). However, only a few employs external validation on private clinical data an essential step for real-world readiness [11].

This comparative summary not only emphasizes the practical applications of each dataset, but also reveals methodological trends, current performance benchmarks, and emerging best practices in AI-driven AF detection.

TABLE II. OVERVIEW OF AI MODELS FOR ATRIAL FIBRILLATION DETECTION FROM ECG SIGNALS

Papers	Year	Dataset	Preprocessing Techniques	AI Models Type	Architecture details	Performance Metrics	Key Contribution	Notes
[11]	2025	• CPSC2018 • PTB-XL • Georgia	• Raw ECG; • Feature extraction; • Demographic data (age, sex); • Saliency maps	• Deep Learning: ○ CNNs, ○ RNNs (GRU), ○ Transformers	• AlexNet, • VGG-16, • LeNet, • ResNet, • Inception, • FCN, • GRU, • Transformers	• AUROC: ○ VGG-16 (0.9668), ○ AlexNet (0.9617); • Sensitivity: ~0.9225 with HRV+demo	• Multimodal input (ECG + HRV + demographics) • Improves AF detection; • Simple models outperform complex ones	• Final best performers: AlexNet and VGG-16 due to efficiency, performance, and interpretability • Best performance on PTB-XL; generalizability and label quality vary across datasets
[21]	2025	Physionet challenge 2017	• Bandpass filtering • Short-Time Fourier Transform (STFT) • Reverse polar transform • PanTompkins (P-T) algorithm	• CNN • Ensemble Voting	• Pretrained MobileNet, ResNet50, DenseNet121 • 5-fold cross-validation • Soft and hard voting ensemble	• Accuracy, • Precision, • Recall, • F1-score	• Introduced reverse polar-transformed spectrograms for ECG • Improved AFib detection with better visual representation • Effective use of compact square matrices for CNN input	• Sensitive to signal amplitude and filter type • Well-suited for real-time and wearable ECG analysis
[3]	2025	MUSE	• Feature extraction; • Demographic data (age, sex);	• Deep Learning (CNN)	Not fully disclosed • CNN • Layer wise Relevance Propagation	• AUROC: 0.905 ± 0.00	• Demonstrated ability to predict paroxysmal AF onset from normal sinus rhythm ECGs using deep learning. • Used LRP to identify T-wave/ST abnormalities as key predictors	• Focused on paroxysmal AF prediction. • Applied LRP (Layer-wise Relevance Propagation) for XAI.
[22]	2025	Physionet challenge 2017	• Bnechmark Algorithm • Expert cardiologist review • removal of artifacted ECGs	• Deep Learnig	• Cloud-based AI platform: "The Willem Artificial Intelligence platform" [43] • Trained on >520,000 patients' ECGs; • detects 23+ arrhythmias; • Uses data from: ○ PTB-XL, ○ MIT-BIH, ○ Georgia DB, ○ ESC DB, ○ AHA DB, etc.	• Accuracy: 96.4% • Sensitivity: 84.2% • Specificity: 97.6% • PPV: 78.0% • NPV: 98.4%	• High-performance AF detection from 1-lead ECGs, • Outperforming rule-based algorithms • Real-world deployability	• Detected additional arrhythmias (PVC, PAC, AV block); • Platform was not trained on test ECGs (external validation) • Performance degraded with noise/artifacts
[23]	2025	Physionet challenge 2017	• Raw ECG used directly; • Class imbalance addressed using Neurokit, BioSPPy, and EngZeeMod filters for synthetic data augmentation	• Hybrid DL ○ CNN ○ LSTM	• Ultra-lightweight CNN for feature extraction • LSTM for temporal patterns; • 64.9K parameters; • 0.48 ms inference time	• F1 Score: 99.56%, • Size gain: 102.25 dB, • 1.06% Gain over SOTA	• Introduces a compact, real-time architecture with high accuracy on imbalanced ECG data	• Optimized for speed and embedded applications; • Avoids complex preprocessing

Papers	Year	Dataset	Preprocessing Techniques	AI Models Type	Architecture details	Performance Metrics	Key Contribution	Notes
[17]	2025	PTB-XL	• Coarse graining (1s/0s) • Complexity feature extraction from all 12 leads; • Evaluated each lead and combinations.	• Traditional ML	• Multiple models optimized using Bayesian optimization. • KNN was tuned over 1–99 neighbors (odd values) with 10 distance functions; • SVM models tested with linear, • Quadratic polynomial, • Gaussian kernels. • Leave-one-out cross-validation was used for robust evaluation.	• Peak accuracy ~0.69 at 125 Hz sampling (using lead V6).	• First human application of complexity analysis for detecting a history of PAF from normal sinus rhythm recordings.	Additional experiments at 500 Hz were conducted using various coarse graining techniques: • LZ76 (A), • LZ78 (B), • Titchener (C); also reporting results with configurations: o BD (Beat Detection), o FD (Feature Detection), KM (K-means), o TC (Threshold Crossing) across individual leads and their combination. o KNN outperformed SVM overall.
[30]	2024	PHYJ	• Signal Preprocessing (Filter) • Signal Dimensionality Reduction (SDR); • Kernal Principal Component Analysis (KPCA) for dimensionality reduction; • Wavelet Transform (CWT) for time–frequency domain;	• Deep Learning (CNN)	• AlexNet , • VGG19, • ResNet152 , • Inception-v3, • Inception ResNet-v2	• Accuracy (ACC), • Sensitivity (SEN), • Specificity (SPF) • Cross Validation	• First approach to detect PAF from sinus rhythm ECGs using full 12-lead CWT + KPCA + deep network fusion	• Highlights potential to detect PAF from non-diagnostic ECGs; • Future work aims to scale validation and involve open-access datasets for pretraining
[31]	2024	EHR MUSE	• ECG converted to OMOP-CDM; • Risk factor extraction from ICD codes, Rx, and measurements	• Multiple residual neural network	• ResNet • Random Forest	• AUC = 0.74 (ECG only), • AUC = 0.76 (ECG + 6 RFs); • Stable across age and sex	• Demonstrates that AF can be predicted during sinus rhythm using 1-lead ECG + 6 key risk factors • Model matches 12-lead performance.	• Designed for **real-world deployment** • Reduces **age bias** seen in prior studies Prospective clinical validation planned
[33]	2024	• Chapman • PTB-XL	• ECGs segmented into PreQ, QRS, and PostS • Only Lead II used	• Unsupervised DL: o LSTM Autoencoder o XGBoost	• LSTM-based autoencoder trained on normal ECG segments • Anomaly detection via MSE per segmenT • Postprocessing with XGBoost classifier on anomaly scores	• AUROC (PreQ): o Experiment A: 0.96, o Experiment B: 0.90, o Experiment C: 0.95 • AUROC (XGBoost): 0.98 • F1 Score: 0.94	• Demonstrated explainable, segment-wise anomaly detection of AF without supervision • PreQ segment most predictive of AF	• Cross-dataset validation (Germany/China) • Highlights clinical interpretability by segment-level scores • Addresses physician concerns about DL "black-boxing"

Papers	Year	Dataset	Preprocessing Techniques	AI Models Type	Architecture details	Performance Metrics	Key Contribution	Notes
[24]	2024	Physionet challenge 2017	• Two-stage ECG denoising and filtering • ECG-SQE (signal quality evaluation using Pan-Tompkins based ML) • 9-sec segmentation • Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for outlier filtering	• Deep learning (hybrid)	• MuDANet: Multi-stream CNN-RNN with Dense Attention o Dual-stream Conv-RNN (enhances signal representation o Attention mechanism improves focus on discriminative features o Final classification into AFR, NSR, and Other • DBSCAN-GAN: o Density-based clustering to isolate clean data o GAN trained only on clean samples for synthetic augmentation	• F1 Score: 0.876 (baseline) • F1 Score: 0.962 (with DBSCAN-GAN and 10-fold CV)	• Introduced MuDANet: dual-stream CNN-RNN with dense attention for AF detection. • Proposed DBSCAN-GAN for noise-aware, class balanced synthetic ECG generation. • Demonstrated strong generalization using synthetic data in a noisy real-world dataset.	• Real-time capable and wearable-device friendly • Combines traditional ECG signal cleaning (Pan-Tompkins) with advanced DL and GAN-based augmentation • Overcomes common data imbalance and outlier noise issues in wearable ECG datasets
[35]	2024	CNUH	• PQRST detection and trimming • Baseline correction • 0.5 Hz high-pass Butterworth filter • Segmentation into 10-second intervals • Random under-sampling to reduce class imbalance	• Deep Learning o CNN o LSTM o RNN	• ResNet50: deep CNN architecture with skip connections • RNN and LSTM models for comparison • Training with AdamW optimizer, early stopping, batch size = 32	• ResNet50: o F1 = 71.9% o Recall = 79.3% o Precision = 65.8% o Accuracy = 70.5% o AUC = 0.79 • External set" o F1 = 64.1%, o AUC = 0.68	• Demonstrated that deep learning can detect masked AF from NSR in single-lead mobile ECG	• Study used a private clinical dataset; • High potential for real-world wearable device applications
[38]	2024	• AFDB • LTAF • MITDB • NSRDB	• R-peak detection to compute RR intervals (RRIs) • Removal of poor-quality or noisy data • ECG augmentation via vertical flipping	• Hybrid CNN-LSTM	• Multi-input fusion network combining ECG and RRI features • Res-CNN for morphological features • Bi-LSTM for temporal sequence modeling of RRI • Low complexity design for real-time application	• High accuracy across 4 external datasets • Ablation study validated robustness • Maintains performance with ectopic beats	• Proposed MIF-AFNet: a multi-input fusion model integrating RR interval + ECG features for robust AF detection • High generalization capability across datasets and rhythms	• Limitations: requires ≥30s input; R-peak detection accuracy affects performance; lacks diverse arrhythmias • Future work: clinical validation, enhanced R-peak detection, training on multi-arrhythmia data, improved noise handling

VII. Analysis and Critical Evaluation

Building on the comparative review of ECG datasets and applied AI models, this section critically evaluates key trends, limitations, and methodological patterns observed across recent studies on AF detection.

- **Patterns and bias in dataset usage:** Despite the growing number of ECG datasets, the majority of recent studies rely heavily on PTB-XL, PhysioNet 2017, and AFDB. This over-reliance introduces dataset bias, limiting the generalizability of models to broader clinical populations. Less common datasets, especially clinical ones like MUSE or EHR, are underutilized due to accessibility barriers, despite offering more realistic and varied signals.
- **Dataset limitations affecting model performance:** Several datasets exhibit class imbalance (e.g., fewer AF vs. non-AF samples), requiring oversampling, Synthetic Minority Over-sampling Technique (SMOTE), or weighted loss functions. Wearable ECG datasets often contain motion artifacts and noise, while many collections lack demographic metadata (e.g., age, sex, comorbidities), which hampers model personalization. Only a few datasets integrate clinical risk factors, despite their value in enhancing prediction.
- **Clinical vs. Wearable datasets:** Clinical datasets (e.g., PTB-XL, Chapman) offer clean, high-resolution ECGs from controlled environments. In contrast, wearable datasets enable long-term monitoring but introduce higher noise levels and data variability. Many models perform well on clean datasets but degrade on wearable ones unless robust preprocessing or augmentation is applied. This creates a gap between lab performance and real-world application [33]..
- **Single-lead vs. Multi-lead ECGs:** 12-lead ECGs remain the standard for model development and benchmarking due to their diagnostic richness. However, single-lead ECGs, especially from wearables, are gaining traction for screening. While multi-lead ECGs generally outperform in accuracy, well-designed models using single-lead inputs (e.g., AZ Delta study) demonstrate that simpler signals combined with clinical features can achieve comparable performance, especially in ambulatory settings.
- **Reproducibility and transparency gaps:** Many studies omit details about signal preprocessing, segmentation, or lead selection. Some apply noise filters or augmentation without reproducible descriptions. Additionally, few publish code or preprocessing pipelines, making it difficult to replicate findings or benchmark new approaches. This lack of transparency reduces trust and slows progress.
- **Standout approaches:** A few studies stand out for their innovation. For example, the AZ Delta AF-SR study combined a single-lead ECG during sinus rhythm with six clinical risk factors, achieving strong predictive performance while reducing age-related bias ideal for wearable screening [31]. Another notable work used LSTM autoencoders to detect anomalies in specific ECG segments (PreQ), enhancing interpretability and requiring no labeled AF samples.

Overall, while current approaches show promise, overcoming dataset biases, improving reproducibility, and integrating richer clinical context remain essential for advancing reliable and generalizable AF detection models.

VIII. Discussion and Feature Directions

A. Key Takeaways:

Our analysis of recent AI-based Atrial Fibrillation (AF) detection studies reveals consistent patterns across dataset usage and model development. A key concern is the heavy dependence on a limited set of datasets particularly PTB-XL, PhysioNet 2017, and AFDB while more clinically representative datasets remain underutilized. This trend narrows the scope of model validation and hinders generalizability to real-world scenarios. Moreover, datasets with limited demographic diversity and missing metadata restrict the ability to evaluate models for fairness, particularly across age groups, sex, or comorbidities. Additionally, wearable ECGs, while increasingly relevant for remote monitoring and early AF screening, are still underexplored due to their noisy nature and signal variability. As a result, models often perform well in controlled, clinical conditions but falter when deployed in real-world or ambulatory environments.

B. Gaps and Limitations in Current Approaches:

1) Overfitting to benchmarks: Many models are optimized for performance on a few benchmark datasets, which inflates accuracy metrics but limits cross-dataset robustness.

2) Lack of dataset standardization: Inconsistent data formats, label definitions (e.g., AF types), and missing preprocessing documentation make reproducibility difficult and comparisons unreliable.

3) Fairness blind spots: Absence of demographic and clinical metadata (e.g., sex, age, comorbidities) prevents analysis of bias, reducing the trustworthiness of AI models for sensitive applications.

4) Underutilization of wearable data: Despite their value for long-term monitoring, wearable ECG datasets are rarely used in training or evaluation due to signal quality issues.

5) Reproducibility challenges: Very few studies share their code, training configurations, or preprocessing pipelines, making model replication and clinical translation problematic.

C. Future Research Directions

To address these challenges and unlock the full potential of AI in AF detection, we recommend the following research directions:

1) Standardization and benchmarking needed: Most public datasets vary in signal format, label structure, and metadata availability. Establishing a benchmark framework with unified preprocessing and labeling would allow for fair model comparisons and reproducibility.

2) Underuse of wearable and real-time ECGs: Wearable devices are central to future AF screening strategies, yet few studies utilize real-world data from such sources. Research should focus on noise-robust algorithms and domain adaptation techniques to bridge the performance gap.

3) Explainability still lacking: Explainable AI remains an exception rather than the rule. Future models should incorporate interpretability methods (e.g., attention maps, ECG segment analysis, SHAP values) to support clinical trust and decision-making.

4) Missing demographics and clinical metadata: Datasets should include age, sex, risk factors, and comorbidities to support personalized predictions and fairness assessment. This metadata is crucial for clinical deployment and regulatory validation.

5) Hybrid modeling with clinical context: Few models integrate non-ECG features like EHR data, medication history, or symptom logs. Incorporating these modalities can improve predictive performance and reduce overreliance on limited signal features.

D. Proposed Dataset Guidelines for Future Research

Based on our review, the following dataset practices are strongly encouraged to support future innovation:

- Provide at least 3 leads when possible (e.g., lead I, II, V1) for improved feature richness.
- Share both raw and preprocessed signal versions.
- Use consistent and annotated labels, including AF types (e.g., paroxysmal, persistent).
- Include demographic and clinical metadata (age, sex, comorbidities).
- Ensure public access under a research license to foster collaboration and reproducibility.

IX. CONCLUSION

This survey examined 15 ECG datasets from 33 recent AI models for Atrial Fibrillation (AF) detection studies, revealing a strong reliance on PTB-XL, PhysioNet 2017, and AFDB. While these resources have advanced the field, their overuse, coupled with missing demographic and clinical metadata, limits generalizability and fairness. Our analysis highlights the growing gap between model performance on clean benchmark datasets and the noisy reality of wearable ECGs. We emphasized the importance of dataset quality, transparency in preprocessing, and the inclusion of patient context for real-world readiness. This work represents the first survey focused specifically on datasets in AI-based AF detection. By mapping current usage patterns and limitations, we provide a guide for future research toward more inclusive, robust, and clinically meaningful AI tools. A collaborative push for standardized, annotated, and diverse datasets including wearable ECGs and multimodal data is essential to bridge the gap between lab success and clinical impact.

REFERENCES

[1] "The Clinical Profile and Pathophysiology of Atrial Fibrillation | Circulation Research." Accessed: Jun. 25, 2025. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.114.303211

[2] S. S. Chugh et al., "Worldwide Epidemiology of Atrial Fibrillation: A Global Burden of Disease 2010 Study," Circulation, vol. 129, no. 8, pp. 837–847, Feb. 2014, doi: 10.1161/CIRCULATIONAHA.113.005119.

[3] Y. Jin, W. Chang, B. Ko, K.-H. Choi, and K. H. Lee, "Explainable paroxysmal atrial fibrillation diagnosis using an artificial intelligence-enabled electrocardiogram," Korean Journal of Internal Medicine, vol. 40, no. 2, pp. 251–261, 2025, doi: 10.3904/kjim.2024.130.

[4] F. M. Dias et al., "Artificial Intelligence-Driven Screening System for Rapid Image-Based Classification of 12-Lead ECG Exams: A Promising Solution for Emergency Room Prioritization," IEEE Access, vol. 11, pp. 121739–121752, 2023, doi: 10.1109/ACCESS.2023.3328538.

[5] E. Hennings et al., "Assessment of the atrial fibrillation burden in Holter electrocardiogram recordings using artificial intelligence," Cardiovascular Digital Health Journal, vol. 4, no. 2, pp. 41–47, Apr. 2023, doi: 10.1016/j.cvdhj.2023.01.003.

[6] D. Meltzer and D. Luengo, "ECG-Based Biometric Recognition: A Survey of Methods and Databases," Sensors, vol. 25, no. 6, Mar. 2025, doi: 10.3390/s25061864.

[7] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, "Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017–2023," Frontiers in Physiology, vol. 14, 2023, doi: 10.3389/fphys.2023.1246746.

[8] J. Rahul and L. D. Sharma, "Advancements in AI for cardiac arrhythmia detection: A comprehensive overview," Computer Science Review, vol. 56, May 2025, doi: 10.1016/j.cosrev.2024.100719.

[9] ASSALHI Imane and BYBI Abdelmajid, "Artificial Intelligence for Atrial Fibrillation Detection: A Systematic Review of Recent Advances in ECG-Based Deep Learning Models," in 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, 2025.

[10] F. Liu et al., "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection," j med imaging hlth inform, vol. 8, no. 7, pp. 1368–1373, Sep. 2018, doi: 10.1166/jmihi.2018.2442.

[11] A. Rawshani et al., "Integrating deep learning with ECG, heart rate variability and demographic data for improved detection of atrial fibrillation," Open Heart, vol. 12, no. 1, 2025, doi: 10.1136/openhrt-2025-003185.

[12] B. B.-S. Chuang and A. C. Yang, "Optimization of using multiple machine learning approaches in atrial fibrillation detection based on a large-scale data set of 12-lead electrocardiograms: Cross-sectional study," Jmir Formative Research, vol. 8, 2024, doi: 10.2196/47803.

[13] A. Petroni, F. Cuomo, G. Scarano, S. Colonnese, P. Francia, and M. Pediconi, "MUSE: MUlti-lead Sub-beat ECG for remote AI based atrial fibrillation detection," Journal of Network and Computer Applications, vol. 212, 2023, doi: 10.1016/j.jnca.2022.103544.

[14] V. Tihonenko, A. Khaustov, S. Ivanov, and A. Rivin, "St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database." physionet.org, 2007. doi: 10.13026/C2V88N.

[15] A. Makhir, M. H. E. Alaoui, A. Jilbab, and L. Bellarbi, "Classification of Atrial Fibrillation and Cardiac Arrhythmias by a CNN-BiLSTM Hybrid Model with DWT Preprocessing," presented at the 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2024, 2024. doi: 10.1109/IRASET60544.2024.10549460.

[16] R.-D. Bousseljot, D. Kreiseler, and A. Schnabel, "The PTB Diagnostic ECG Database." physionet.org, 2004. doi: 10.13026/C28C71.

[17] S. Creasy, V. Alexeenko, K. Jeevaratnam, P. J. Aston, G. Y. H. Lip, and G. Tse, "Electrocardiogram sampling frequency for the optimal performance of complexity analysis and machine learning models: Discrimination between patients with and without paroxysmal atrial fibrillation using sinus rhythm electrocardiograms," Heart Rhythm O2, vol. 6, no. 1, pp. 48–57, 2025, doi: 10.1016/j.hroo.2024.11.002.

[18] T. Bender, P. Gemke, E. Idrobo-Avila, H. Dathe, D. Krefting, and N. Spicher, Benchmarking the impact of noise on deep learning-based classification of atrial fibrillation in 12-lead ECG, vol. 302. 2023, pp. 977–981. doi: 10.3233/SHTI230321.

[19] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," Sci Data, vol. 7, no. 1, p. 154, May 2020, doi: 10.1038/s41597-020-0495-6.

[20] G. Clifford et al., "AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017," presented at the 2017 Computing in Cardiology Conference, Sep. 2017. doi: 10.22489/CinC.2017.065-469.

[21] D. Kwon, H. Kang, D. Lee, and Y.-C. Kim, "Deep learning-based prediction of atrial fibrillation from polar transformed time-frequency electrocardiogram," Plos One, vol. 20, no. 3 March, 2025, doi: 10.1371/journal.pone.0317630.

[22] F. D. Guio et al., "Enhanced detection of atrial fibrillation in single-lead electrocardiograms using a Cloud-based artificial intelligence platform," Heart Rhythm, 2025, doi: 10.1016/j.hrthm.2024.12.048.

[23] A. Khan et al., "A deep learning-based ultra-lightweight architecture for atrial fibrillation detection using single-lead ECG recordings," IEEE Access, vol. 13, pp. 86474–86486, 2025, doi: 10.1109/ACCESS.2025.3567093.

[24] G. Msigwa, E. Ntambala, and J. Yun, "Enhancing atrial fibrillation classification from single-lead electrocardiogram signals using attention-based networks and generative adversarial networks with density-based clustering," Engineering Applications of Artificial Intelligence, vol. 133, p. 108607, Jul. 2024, doi: 10.1016/j.engappai.2024.108607.

[25] H. Zhang, H. Zhao, and Z. Guo, "Artificial Intelligence-Based Atrial Fibrillation Recognition Method for Motion Artifact-Contaminated Electrocardiogram Signals Preprocessed by Adaptive Filtering Algorithm," Sensors, vol. 24, no. 12, p. 3789, Jun. 2024, doi: 10.3390/s24123789.

[26] L. Xia, S. He, Y. F. Huang, and H. Ma, "Multiscale Dilated Convolutional Neural Network for Atrial Fibrillation Detection," Plos one, vol. 19, no. 6, p. e0301691, 2023.

[27] M. E. A. Bourkha, A. Hatim, D. Nasir, and S. El Beid, "Enhanced Atrial Fibrillation Detection-based Wavelet Scattering Transform with Time Window Selection and Neural Network Integration," International Journal of Advanced Computer Science and Applications, vol. 14, no. 12, pp. 494–503, 2023, doi: 10.14569/IJACSA.2023.0141252.

[28] Y. Hu, M. Wang, H. Tang, T. Feng, and C. Liu, "Detection of paroxysmal atrial fibrillation from dynamic ECG recordings based on a deep learning model," Journal of Personalized Medicine, vol. 13, no. 5, 2023, doi: 10.3390/jpm13050820.

[29] H. Gruwez et al., "Detecting paroxysmal atrial fibrillation from an electrocardiogram in sinus rhythm: External validation of the AI approach," Jacc Clinical Electrophysiology, vol. 9, no. 8P3, pp. 1771–1782, 2023, doi: 10.1016/j.jacep.2023.04.008.

[30] M. Deng et al., "An intelligent computer-aided diagnosis method for paroxysmal atrial fibrillation patients with nondiagnostic ECG signals," Biomedical Signal Processing and Control, vol. 88, p. 105683, Feb. 2024, doi: 10.1016/j.bspc.2023.105683.

[31] S. Dupulthys et al., "Single-lead electrocardiogram Artificial Intelligence model with risk factors detects atrial fibrillation during sinus rhythm," EUROPACE, vol. 26, no. 2, 2024, doi: 10.1093/europace/euad354.

[32] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," Sci Data, vol. 7, no. 1, p. 48, Feb. 2020, doi: 10.1038/s41597-020-0386-x.

[33] S. Choi et al., "Diagnosis of atrial fibrillation based on AI-detected anomalies of ECG segments," Heliyon, vol. 10, no. 1, p. e23597, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23597.

[34] A. Mohan, D. Elbers, O. Zilbershot, F. Afghah, and D. Vorchheimer, "Deciphering Heartbeat Signatures: A Vision Transformer Approach to Explainable Atrial Fibrillation Detection from ECG Signals," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.09474

[35] J. Kim, B. Ko, M. Lee, K. H. Lee, S. J. Lee, and Y.-S. Lee, "Identification of atrial fibrillation with single-lead mobile ECG during normal sinus rhythm using deep learning," Journal of Korean Medical Science, vol. 39, no. 5, 2024, doi: 10.3346/jkms.2024.39.e56.

[36] G. B. Moody and R. G. Mark, "MIT-BIH Atrial Fibrillation Database." physionet.org, 1992. doi: 10.13026/C2MW2D.

[37] S. Mandala, A. R. P. Wibowo, Adiwijaya, Suyanto, M. S. M. Zahid, and A. Rizal, "The effects of daubechies wavelet basis function (DWBF) and decomposition level on the performance of artificial intelligence-based atrial fibrillation (AF) detection based on electrocardiogram (ECG) signals," Applied Sciences (Switzerland), vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053036.

[38] Y. Zou et al., "A generalizable and robust deep learning method for atrial fibrillation detection from long-term electrocardiogram," Biomedical Signal Processing and Control, vol. 90, 2024, doi: 10.1016/j.bspc.2023.105797.

[39] M. M. Rahman, M. W. Rivolta, R. Sassi, and F. Badilini, "Quantifying uncertainty of a deep learning model for atrial fibrillation detection from ECG signals," in Computing in cardiology, 2023. doi: 10.22489/CinC.2023.340.

[40] F. Jager et al., "The Long-Term ST Database." physionet.org, 1995. doi: 10.13026/C2G01T.

[41] G. B. Moody and R. G. Mark, "MIT-BIH Arrhythmia Database." physionet.org, 1992. doi: 10.13026/C2F305.

[42] T. A. L. The Beth Israel Deaconess Medical Center, "The MIT-BIH Normal Sinus Rhythm Database." physionet.org, 1990. doi: 10.13026/C2NK5R.

[43] Idoven 1903 S.L., "Evaluation of Electrocardiographic Data From High-risk Cardiac Patients Using WillemTM Cardiologist-level Artificial Intelligence Software. WILLEM Trial.," clinicaltrials.gov, Clinical trial registration NCT05890716, May 2025. Accessed: Jun. 25, 2025. [Online]. Available: https://clinicaltrials.gov/study/NCT05890716

[44] "MUSE Cardiology Information System." Accessed: Jun. 25, 2025. [Online]. Available: https://www.gehealthcare.com/en-gb/products/diagnostic-cardiology/muse-v9?npclid=CjwKCAjw0a-SBhBkEiwApljU0nguvZYAZ1DaaIftQ4qgx-ruc99u6H5Q2Rc63w7lgTIGvqD964bblRoC8TkQAvD_BwE

[45] D. Mannhart et al., "Clinical validation of an artificial intelligence algorithm offering cross-platform detection of atrial fibrillation using smart device electrocardiograms," Archives of Cardiovascular Diseases, vol. 116, no. 5, pp. 249–257, May 2023, doi: 10.1016/j.acvd.2023.04.003.

[46] X. Wang et al., "Genetic Susceptibility to Atrial Fibrillation Identified via Deep Learning of 12-Lead Electrocardiograms," Circulation Genomic and Precision Medicine, vol. 16, no. 4, pp. 340–349, 2023, doi: 10.1161/CIRCGEN.122.003808.

[47] N. Ben-Moshe, S. B. Brimer, E. Zvuloni, J. A. Behar, K. Tsutsui, and L. Sornmo, "RawECGNet: Deep Learning Generalization for Atrial Fibrillation Detection from the Raw ECG," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 9, pp. 5180–5188, 2024, doi: 10.1109/JBHI.2024.3404877.

[48] Y. Kim et al., "Predicting Future Incidences of Cardiac Arrhythmias Using Discrete Heartbeats from Normal Sinus Rhythm ECG Signals via Deep Learning Methods," Diagnostics, vol. 13, no. 17, p. 2849, Sep. 2023, doi: 10.3390/diagnostics13172849.

[49] A. Raghunath et al., "Artificial intelligence–enabled mobile electrocardiograms for event prediction in paroxysmal atrial fibrillation," Cardiovascular Digital Health Journal, vol. 4, no. 1, pp. 21–28, Feb. 2023, doi: 10.1016/j.cvdhj.2023.01.002.

[50] G. Baj et al., "Comparison of discrimination and calibration performance of ECG-based machine learning models for prediction of new-onset atrial fibrillation," BMC Medical Research Methodology, vol. 23, no. 1, 2023, doi: 10.1186/s12874-023-01989-3.

[51] T. Hygrell et al., "An artificial intelligence–based model for prediction of atrial fibrillation from single-lead sinus rhythm electrocardiograms facilitating screening," EP Europace, vol. 25, no. 4, pp. 1332–1338, Apr. 2023, doi: 10.1093/europace/euad036.

DLCA-CapsNet: Dual-Lane CDH Atrous CapsNet for the Detection of Plant Diseases

Steve Okyere-Gyamfi[1]*, Michael Asante[2], Yaw Marfo Missah[3], Kwame Ofosuhene Peasah[4], Vivian Akoto-Adjepong[5]
Kwame Nkrumah University of Science and Technology, Kumasi, Ghana[1, 2, 3, 4]
Catholic University of Ghana, Sunyani, Ghana[1]
University of Energy and Natural Resources, Sunyani, Ghana[5]

***Abstract*—Humanity's survival, development, and existence are deeply intertwined with agriculture, the source of most of our food. Plant disease detection helps in securing food, but manual plant disease detection is error-prone and labor-intensive. Convolutional Neural Networks (CNNs) are highly effective for automated plant disease classification, but their difficulty in recognizing differently oriented images means they need large datasets with many variations to work best. Capsule Networks (CapsNets) were developed to overcome the shortcomings of CNNs and can function effectively with smaller datasets. However, CapsNets process every part of an input image, so their performance can suffer when dealing with complex visuals. To tackle this challenge, DLCA-CapsNet was introduced. DLCA-CapsNet integrates a Color Difference Histogram (CDH) layer for key feature extraction, atrous convolution layers to enlarge receptive fields while maintaining spatial details, along with max-pooling, standard convolutional layers, and a dropout layer. The proposed DLCA-CapsNet method was evaluated on datasets including apple, banana, grape, maize, mango, pepper, potato, rice, tomato, as well as CIFAR-10 and Fashion-MNIST. The model demonstrated strong performance with high test accuracies in plant disease detection and on CIFAR-10 and Fashion-MNIST. It improved test accuracies by 6.78%, 14.82%, 6.14%, 5.07%, 21.12%, 40.32%, 4.64%, 0.76%, 10.23%, 13.73%, and 2.03%, while also reducing the number of parameters in millions by 6.16M, 6.16M, 6.16M, 6.16M, 7.14M, 5.68M, 5.92M, 7.62M, 7.62M, and 6.54M respectively when compared with the original CapsNet. On sensitivity, F1-Score, precision, specificity, Receiver Operating Characteristics, Precision-Recall values, accuracy, disk size, and parameters generated, etc., the DLCA-CapsNet achieved better performance compared to the original CapsNet and other advanced CapsNets reported in the literature. The findings suggest that this efficient and computationally less demanding method can significantly enhance plant disease classification and contribute incrementally to efforts aligned with the SDG 2 goal by offering a lightweight, scalable solution that can be adapted for field use in resource-constrained settings.**

***Keywords*—Color Difference Histogram (CDH); Convolutional Neural Network (CNN); atrous Convolution; Capsule Neural Network; plant disease detection; dynamic routing; AI in agriculture**

I. INTRODUCTION

Agriculture has consistently been a crucial social and economic sector for humanity. The production of food is particularly critical, with high demand from every household. As a result, employing innovative technologies to improve the sector is essential for the agri-food industry. Today, artificial intelligence stands out as a significant technological tool extensively utilized in contemporary society. Specifically, Deep Learning (DL) has numerous applications owing to its capability to learn strong representations from images [1] [2] [3].

CNNs are the primary DL architecture for image classification, primarily due to the significant attention they have received in recent years. Multiple CNN architectures have been developed to enhance their performance. However, traditional CNNs still have many limitations. They do not emphasize the arrangement or spatial connections among the components of the image, so they require extensive datasets in various variations to achieve high performance [4], which results in the augmentation of data [5] [6].

To address this issue, a novel architecture named CapsNet, which mimics the human brain, has been introduced for extracting information from images [7]. Despite their numerous advantages, such as performing well with fewer data, CapsNets try to identify every object in an image, which reduces their ability to generalize effectively, especially on unseen and complex images [8][9][10]. To address the issues with CapsNet, researchers sometimes deepen and widen their models, resulting in longer training times and more parameters. Moreover, increasing the CapsNet model depth does not necessarily improve performance. Selecting efficient feature extractors for the encoder network is crucial for enhancing the CapsNet model's performance. Therefore, this study seeks to answer the following research questions:

- Can a well-performing CapsNet be designed to perform well on plant diseases and complex images?
- Can good feature extractors be incorporated into the CapsNet to improve the encoder network for better feature extraction of the CapsNet Model?
- How can we explain the "black box" phenomenon of this AI model to enhance comprehension of the model and confidence in the model's output?
- Can the CapsNet model be optimized to make it deployable on resource-constrained devices?

This study suggests a low-parameterized shallow CapsNet with efficient feature extraction abilities to help classify plant diseases efficiently using CDH [11] and atrous convolutions [12] [13]. Results from experiments conducted on 11 publicly accessible datasets demonstrate that the proposed DLCA-CapsNet model performs on par with the most advanced models available for detecting complex images and other plant diseases.

The visualization of class capsule clusters, activation maps, and reconstructed images demonstrates the superior efficiency of the feature extractors of DLCA-CapsNet compared to the feature extraction capabilities found in the traditional CapsNet model.

This study's contributions are:

- A novel and well-performing CapsNet design named DLCA-CapsNet is introduced.
- CDH is utilized to extract significant features in CapsNet models. Additionally, the DLCA-CapsNet employs atrous convolution to enhance spatial representation.
- We conducted thorough visualizations of plant diseases and other datasets to enhance the explainable artificial intelligence (XAI) field.
- The DLCA-CapsNet model generated fewer parameters and had a smaller size, enhancing its suitability for deployment on devices with limited resources.

The remainder of the paper is structured as follows: Section II reviews related literature, Section III outlines the methodology employed, Section IV discusses the experimental results, and Section V provides the conclusion and suggestions for future research.

II. Related Works

A. Introduction

In recent years, plant disease detection has emerged as a crucial area of research in agriculture, meeting the demand for timely and precise diagnosis to ensure crop health and productivity. CNNs have been widely used in this domain, demonstrating notable success in image classification tasks compared to CapsNet. This section reviews the advancements and applications of Capsule networks for plant disease detection, highlighting their effectiveness in improving detection accuracy and generalisation.

B. Review of CapsNet for Plant Disease Detection in Literature

Mensah et al. proposed combining CapsNet and Gabor to detect deformed and unclear diseases in citrus and tomatoes, utilising the PlantVillage dataset. Their system attained 93.33% validation accuracy for classifying citrus diseases and 98.13% with 12 million parameters for classifying tomato diseases [14]. Verma and their team created an optimized technique for computing features by leveraging Squeeze and Excitation (SE) Networks. These networks are applied before the original CapsNet in the classification process to assess the severity of plant diseases. They incorporated AlexNet and ResNet into CapsNet to get two SE CapsNets. They used a tomato dataset from PlantVillage and classified Late Blight diseases into late, middle, early, and healthy stages. The SE-ALexNet CapsNet had 9 million parameters with an accuracy of 92.1%, and SE-Res CapsNet had 19 million parameters with an accuracy of 93.75% [15]. Vasudevan and Karthick proposed a composite technique for detecting grape diseases. They obtain a leaf area by utilizing a method based on a graph. To augment the dataset, they employ a Generative Adversarial Network. Although the model was computationally complex, it attained a validation accuracy of 97.63% on diseases in grapes using captured data and data from the PlantVillage dataset [16]. Xu et al. proposed a model, combining CapsNet with inception modules to amplify receptive fields using diverse dilation rates. This helps to improve the multi-level characteristics of diseases in apple leaves. The model attained 93.16% validation accuracy [17]. Mensah et al. introduced a model by replacing sigmoid with SoftMax, CNN with Local Binary Pattern, and dynamic routing with k-means routing. Experimentally on CIFAR-10, fashion-MNIST, MNIST, citrus, maize, and tomato, 75.80%, 92.72%, 99.68%, 99.41%, 96.79%, and 98.06% validation accuracies were attained, generating 5.2 million for CIFAR-10, 2.8 million parameters for fashion MNIST and MNIST, and 8.4 million for tomato and citrus [18]. Verma and co-authors used the conventional CapsNet to classify potato leaf images from the PlantVillage dataset. The experimental results reveal 9,856,768 parameters produced by the model and 91.83% validation accuracy attained [19]. Oladejo and Ademola classified banana diseases by proposing a CapsNet model, in which they changed the neuron number in the fully connected layer of the original CapsNet to two 512 and 1024 instead of 3, hence reducing the time for training the network. Also, they used an optimizer for momentum to upgrade the network processing speed. On banana leaf disease detection, the model attained a validation accuracy of 95% [20]. Atlan changed the traditional CapsNet model by adjusting the three fully connected layers to 960, 768, and 4096. On assessing the model, it gained a sensitivity of 96.37%, an accuracy of 95.76%, and a specificity of 97.49% when used to classify bell pepper diseases [21]. Mensah and colleagues chose to use the Gabor filter rather than CNNs. This change allowed the initial layers to capture spatial and texture relationships, improving overall performance. Max Pooling was used to select the most important features after the Gabor filter convolution, reducing the feature vectors' dimensionality. Tests conducted on tomato datasets from plant villages showed a validation accuracy of 97.98%, with a total parameter count of 8,708,128 [22]. Anant suggested a technique called AppleCaps that overcomes the spatial invariance issue discovered in CNNs. The model gained 87.06% accuracy on an augmented dataset containing diseased apple leaves [23]. Mensah et al. proposed a CapsNet model that uses dual input. The output from these inputs is merged and submitted to the layers of the original CapsNet. Upon assessing the system on the tomato and CIFAR-10 datasets, it generated 5.48 million (M) and 6.04 M parameters and 76.58% and 93.03% validation accuracies, respectively [24]. Peker suggested a CapsNet that uses multiple ensemble channels by implementing different CapsNet models on images with various pre-processing methods. By merging the networks to learn diverse features from data, the model increases in performance and attains 98.15% on 10 disease classes of the tomato dataset [25]. Abouelmagd and colleagues developed a computer vision approach using an improved CapsNet to recognize and categorize ten distinct diseases affecting tomato leaves from standard datasets. To minimize overfitting, they employed pre-processing and data augmentation approaches when training. Their CapsNet method achieved an accuracy of 96.39% [26]. Zang and colleagues proposed a method for detecting plant leaf diseases by combining residual networks (ResNet) and CapsNet. They improved the traditional ResNet by replacing the kernel with a set of 3×3 convolutional kernels and

incorporating an attention mechanism to guide the model in prioritizing important features. The upgraded ResNet was then integrated with CapsNet. This combined model, SE-SK-CapResNet, achieved accuracy rates of 98.58%, 97.19%, and 95.08% on the PlantVillage, Tomato Leaf Disease, and AI Challenger 2018 datasets, respectively [27]. Mensah and co-authors proposed Shallow and Multi-input CapsNets. The shallow CapsNet used LBP, the squash function, and a normalizer in the standard CapsNet model. The model was assessed on datasets, CIFAR-10, fashion-MNIST, and tomato, and achieved validation accuracies of 75.75%, 92.70%, and 97.33% with 4.6M, 2.5 M, and 4.1M parameters, respectively. Furthermore, the multi-input CapsNet merged the results of three (3) convolution layers before feeding them to the standard CapsNet model. The model was also assessed on the same dataset as shallow CapsNet and attained validation accuracies of 63.95 %, 91.45%, and 94.04% with 4.3 M, 2.2 M, and 4.0 M parameters, respectively [28]. Andrushia and co-authors proposed a technique for classifying grape leaf diseases by inserting convolution layers before the primary capsules. This resulted in speeding up the dynamic process by reducing the number of capsules. On both the non-augmented and augmented grape leaf disease datasets from PlantVillage, the model achieved 99.12% accuracy [29].

C. Summary

These models have performed well on different plant disease datasets. Still, these current CapsNet models for plant disease identification are often limited by scalability, processing speed, many parameters, large size, and robustness to complex backgrounds. Developing a more advanced CapsNet model could enhance accuracy, improve generalization, reduce parameter count and size on disk, and ultimately aid in more effective crop management.

III. Methodology

A. Capsule Network

Unlike CNNs, CapsNets [7] can identify orientation, texture, and pose. In CapsNets, neurons are organised into capsules, each of which has an activity vector that encodes different instantiation parameters for detecting a specific object type. These capsules provide a probability of the object's presence and its generalised pose. Capsules receive these activity vectors from the capsules in the preceding layer, and the connections between these layers, called coupling coefficients, have varying values. If the present capsule identifies a dense cluster of earlier predictions, strongly suggesting the object's presence, it generates a high probability, a process referred to as routing by agreement. Therefore, if the lower layer capsule's prediction aligns with the current capsule's real output, the coefficient among them rises, calculated using the softmax function as illustrated in Eq. (1).

$$\hat{u}_{j|i} = W_{ij} u_i \qquad (1)$$

Where $\hat{u}_{j|i}$ represents the j^{th} capsule output or prediction vector, W_{ij} and u_i denote the weight matrix and output vector of the capsule i in the lower-level layer, respectively. Coupling coefficients are determined through the softmax function in Eq. (2), reflecting the level of alignment between capsules in adjacent layers.

$$C_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \qquad (2)$$

Here, b_{ij} represents the probability amongst the two (2) capsules based on logarithmic values, set to zero. The j^{th} capsules input vector s_j is determined as in Eq. (3),

$$s_j = \sum_{i=1}^{N} c_{ij} \,.\, \hat{u}_{j|i} \qquad (3)$$

Ultimately, the subsequent squash function in Eq. (4) is utilised to confine the output within the range of 0 to 1.

$$v_j = \frac{||s_j||^2}{1+||s_j||^2} \frac{s_j}{||s_j||} \qquad (4)$$

Eq. (5) calculates the loss function for the capsules in the final layer. Here, T_k equals 1 if class k is active, and 0 if otherwise. The values of λ, m^-, and m^+ are determined during the learning process.

$$L_k = T_k \, max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k) \, max(0, ||v_k|| - m^-)^2 \qquad (5)$$

B. UnitsColour Difference Histogram (CDH) Feature Map Detection

The CDH method prioritizes color, how edges line up, and color changes that look natural to us. It captures these features in a way that mimics how our eyes and brain understand them. This method introduces a novel visual descriptor that combines color, edge direction, and how we visually perceive color differences, all while taking into account how these elements are arranged in space. With edge orientation identified and colors quantized via CDH, the next stage is to find small structural elements. This is done by looking at each pixel and then using a 3x3 filter to apply the CDH framework for color and edge quantization [11]. Examining a filter's central value in relation to its eight surrounding values yields a set of edge and color maps. From these maps, we can derive features representing color and edge characteristics by computing the difference (Δ) in color intensity and edge direction for each *L*a*b** component. In the final step, we integrate these two properties into one histogram. In a quantized image, the brightness or color levels of each pixel, denoted as $C(x, y)$, are limited to a specific set of discrete values. These values span from a minimum of 0 up to a maximum of $W - 1$. Neighbouring pixels, identified by their coordinates (x, y) and $(x'y')$ have associated color index values as $C(x, y) = w_1$ and $C(x'y') = w_2$. The orientation image for edge by $\theta(x, y)$ stores orientation information as discrete values v (ranging from 0 to $V - 1$) for each pixel (x, y). At specific coordinates (x, y) and $(x'y')$, the orientation angles are as $\theta(x, y) = v_1$ and $\theta(x'y')=v_2$.

Color difference histograms for neighboring pixels separated by distance D, with color quantization W and edge orientation quantization V, are defined using Eq. (6) and Eq. (7).

$$H_{color}(\, C(x, y)\,) = \begin{cases} \sum \sqrt{(\Delta L^2) + (\Delta a^2) + (\Delta b^2)} \\ where\ \theta(x, y) = \theta(x'y'); max(|x - x'|), (|y - y'| = D) \end{cases} \qquad (6)$$

$$H_{ori}(\, \theta(x, y)\,) = \begin{cases} \sum \sqrt{(\Delta L^2) + (\Delta a^2) + (\Delta b^2)} \\ where\ C(x, y) = C(x'y'); max(|x - x'|), (|y - y'| = D) \end{cases} \qquad (7)$$

Here, ΔL, Δa, and Δb denote the differences between two color pixels, with D set to a value of 1. When the edge orientation is W and the color quantization level is V, the CDH feature is computed as shown in Eq. (8):

$$H_{CDH} = H_{color}(0), H_{color}(1) \dots H_{color}(W-1), H_{ori}(0), H_{ori}(1) \dots H_{ori}(V-1) \quad (8)$$

Image retrieval features can be created by combining a color histogram H_{color} with a histogram of edge orientations H_{ori}. If, for instance, we divide the color space into 72 bins and the edge orientations into 18 bins, the combined feature vector, represented as H, will have a total of 90 dimensions (72 + 18).

C. Atrous Convolutions

Atrous convolution or dilated convolution is crafted to see a wider input area without needing more computing power and parameter count. They are primarily used in tasks like semantic segmentation [12] [13]. Regular deep CNN combines convolutional layers with max-pooling. The downside is that the feature maps shrink in size by 50% every time a max-pooling operation happens. As a result, projecting the processed feature information back onto the initial image leads to less detailed feature extraction as the neural network goes deeper. Atrous (Dilated) convolution tackles this problem by allowing for more thorough feature extraction. It introduces a new parameter known as the rate (r). Atrous convolution functions much like standard convolution, with the key difference being that its kernel weights are distributed with a spacing of r positions, creating sparsely connected convolution layers. Atrous convolution expands its field of view by changing how spread out its sampling points are. Different levels of this spread can capture information at various scales without more computational resources. For example, a standard 3x3 convolution can be modified to see the same area as a 5x5 or 7x7 convolution, allowing it to pick up image features of different sizes. By applying dilated convolutions with varying dilation rates, receptive fields at multiple scales can be effectively captured, and dilated convolution functions can serve as a mechanism for multi-scale convolutional processing. Eq. (9) and Eq. (10) present the formulations for computing the dilated convolution kernel and its corresponding receptive field, respectively.

$$n = k + (k-1)(r-1) \quad (9)$$

$$l_m = l_{m-1} + [(f_m - 1) \prod_{i=1}^{m-1} S_i] \quad (10)$$

In this context n and k denote the sizes of the dilated and standard convolution kernels, respectively. The term l_{m-1} refers to the receptive field size of the $l_{m-1}{}^{\text{th}}$ layer, while the receptive field at the m^{th} layer is determined after applying convolution. f_m represents the convolution kernel size at the m^{th} layer, and S_i indicates the stride of layer l. Fig. 1 illustrates the concept of atrous (dilated) convolutions utilizing a 3×3 kernel with dilation rates of 1, 2, and 4.

D. Proposed Architecture

The proposed DLCA-CapsNet model found in Fig. 2 comprises a CDH layer, two astrous convolution layers, two traditional convolution layers, max-pooling layers, batch normalizers, and a dropout layer. The input image, rescaled to 32x32x3, is subjected to the CDH layer, and does not add any parameters. The 32x32x3 feature map from the CDH layer is sent through two different lanes (Lane1(L1) and Lane2(L2)), each starting with atrous convolutions (Atrous_Conv1 and Atrous_Conv2) with a dilation rate of 2, which processes the feature map subjected to them to produce a 32x32x32 feature map. Each feature map from these astrous convolutions from L1 and L2 is sent to a max-pooling layer (MP1 and MP3), producing feature maps of dimensions 16x16x32. The outputs from the max-pooling layers from L1 and L2 are forwarded to the convolution layers (Conv1 and Conv2) of 64 filters, a 3×3-sized kernel operating with a stride length of 1, respectively. They produce 14x14x64 feature maps sent to max pooling layers (MP2 and MP4) in these separate lanes to produce 7x7x64 feature maps each for L1 and L2. Feature maps from L1 and L2 are then merged or concatenated to produce a 7x7x128 feature map. This merging allows the model to gain the ability to learn integrated features. The merged output is then sent to a dropout layer, then to a Primary Capsule (PC) layer with eight dimensions and 16 channels, a 3×3-sized kernel operating with a stride length 2. Output from the PC layer is sent to the class capsule (Plant_DiseaseCaps), which produces its output for reconstruction. This class capsule layer consists of 16D capsules and considers every class within a dataset. The decoder layer, comprised of three fully connected layers encompassing 512, 1024, and 3072 neurons, respectively, gets the feature map and decodes the entity's properties. The batch normalization layers incorporated in the model helped in the consistent data distribution and improved and simplified the training. The atrous, CDH, dropout, and max-pooling layers help retrieve very salient features submitted to the model from the image for better categorization.

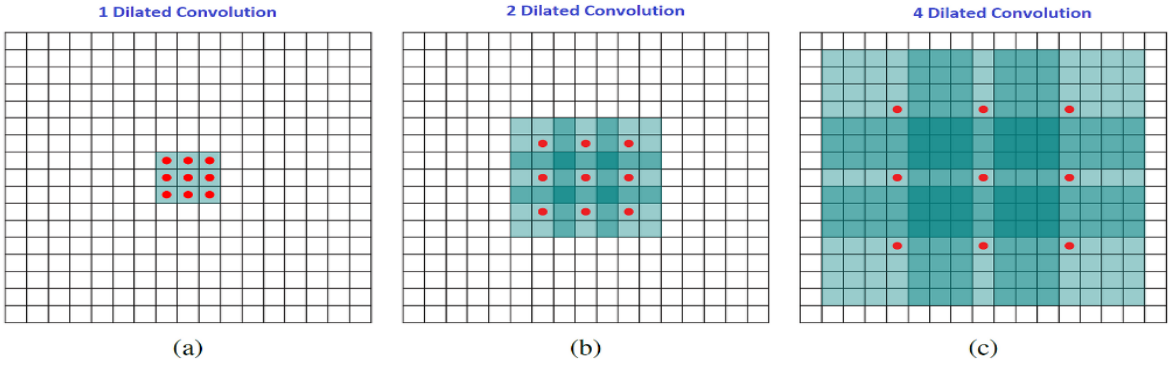


Fig. 1. Demonstrating the idea of astrous convolution. Dilation significantly increases the receptive fields while maintaining full resolution. (a) 1-rate dilated convolution (b) 2-rate dilated convolution and (c) 4-rate dilated convolution.

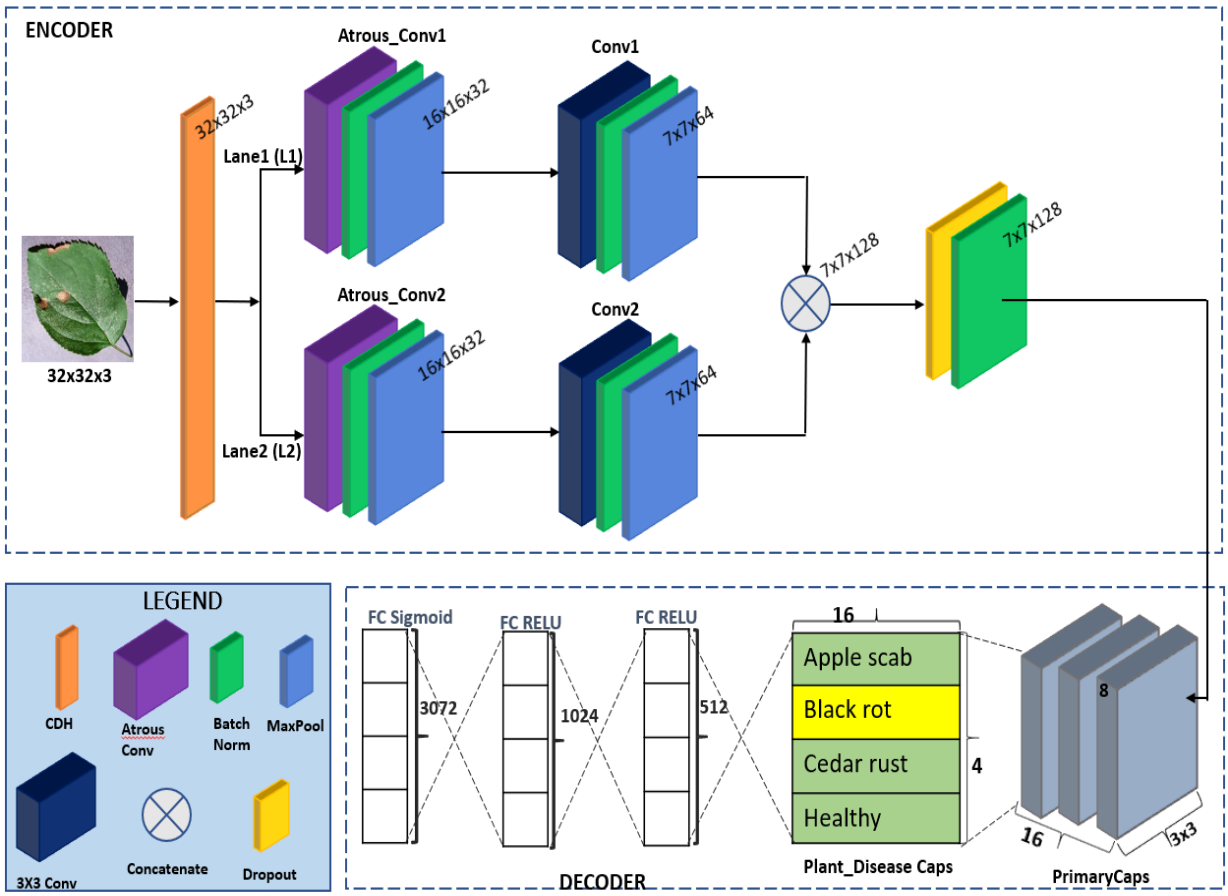


Fig. 2. Architecture of the suggested DLCA-CapsNet model.

E. Datasets

All datasets used are publicly available datasets.

The apple, grape, corn, pepper, potato, and tomato datasets, scaled at 256x256, are a portion of the Plant Village [30].

Apple: It comprises 3171 images sorted into four groups: three infected classes and one healthy class. 0,1,2, and 3 represents Apple_scab, Black_rot, Cedar_apple_rust, and healthy respectively.

Grape: It contains 4062 images in four groups; 0,1,2, and 3 represents Black_rot, Esca_(Black_Measles), Leaf_blight_(Isariopsis_Leaf_Spot, and healthy respectively.

Corn: It is made up of 3852 images sorted into four groups; three infected classes and one healthy class. 0,1,2, and 3 represents Cercospora_leaf_spot Gray_leaf_spot, Common_rust, Northern_Leaf_Blight, and healthy respectively.

Pepper: It is made up of 2475 images sorted into two groups; one infected class and one healthy class. 0 and 1 represent healthy and Bacterial_spot, respectively.

Potato: comprises 2152 images sorted into three groups: two infected classes and one healthy class. 0,1, and 2 represent Healthy, Early_blight, and Late_blight, respectively.

Tomato: It comprises 18160 images sorted into ten groups; nine infected and one healthy class. 0,1,2,3,4,5,6,7,8 and 9 represents Bacterial_spot, Early_blight, Late_blight, Leaf_Mold, Septoria_leaf_spot, Spider_mites Two-spotted_spider_mite, Target_Spot, Tomato_mosaic_virus, Tomato_Yellow_Leaf_Curl_Virus, and Healthy respectively.

Banana comprises 937 images sorted into four groups: three infected classes and one healthy class. 0, 1, 2, and 3 represent Cordana, Pestalotiopsis, Sigatoka, and Healthy [31].

Mango comprises 4000 images sorted into eight groups: seven infected classes and one healthy class. 0,1,2, 3, 4, 5, 6, and 7 represent Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Powdery Mildew, Sooty Mould, and Healthy, respectively [32].

Rice: It comprises 5932 images sorted into four infected classes. 0, 1, 2, and 3 represent Bacterial_blight, Blast, Brown_spot, and Tungro, respectively [36].

Fashion-MNIST: contains 70,000 grayscale fashion product images, each with dimensions of 28x28 pixels. It is divided into ten categories, each containing 7,000 images. The training set includes 60,000 images, while the remaining 10,000 images form the test set. This dataset is more complex than MNIST [33].

CIFAR-10: The dataset contains 50,000 images for training and 10,000 images for testing. Each image has a resolution of 32x32x3 and includes diverse backgrounds and objects, making it more complex than the Fashion-MNIST dataset [34].

To standardize input across datasets, images were resized to 32 × 32 × 3 and split into training and testing sets in an 80:20 ratio. Data augmentation was not applied, as CapsNet is suitable for limited data, and the DLCA-CapsNet is optimized for extracting critical features. Given that most of the ten datasets are imbalanced, high performance would underscore the model's effectiveness in dealing with real-world data imbalance.

F. Implementation Details

The setup involved utilizing a Windows machine with an RTX 2080 SUPER GPU featuring 8GB of dedicated memory and 32GB of RAM. The implementation utilized Keras and Python through Anaconda, with TensorFlow as the backend. For training. The optimization process used Adam, operating with a learning rate of 0.001 and a decay rate of 0.9, with training conducted in batches of 100 samples. To ensure optimal progress during training, the model with the best performance was saved at each iteration. The authors referenced the architecture of the original CapsNet, which is available at the following GitHub repository: https://github.com/XifengGuo/CapsNet-Keras.

IV. Results and Discussion

In this part of the article, we assessed the proposed DLCA-CapsNet model by conducting a comparative analysis with the traditional CapsNet and other cutting-edge CapsNet techniques for classifying plant disease. This comparison aims to determine the most effective model for classifying diseases from plant images.

A. Performance Evaluation

The outcomes of the DLCA-CapsNet architecture, which was trained on images from the plants, CIFAR-10, and fashion-MNIST datasets, are presented here. Confusion matrices are also p, loss and validation curves that depict accuracy (ACC), Area Under the Curve (AUC) for both Precision-Recall (PR) and Receiver Operating Characteristics (ROC) curves, sensitivity (SEN), precision (PRE), specificity (SPE), and F1-Score (FS) are presented to assess the model's performance. Eq. (11)-(15) represent the various performance metrics calculated by considering the confusion matrix's TP, FP, TN, FN, representing True Positive, False Positive, True Negative, False Negative.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (11)$$

$$Precision(P) = \frac{TP}{TP+FP} \quad (12)$$

$$Recall(R)/Sensitivity = \frac{TP}{TP+FN} \quad (13)$$

$$Specificity = \frac{TN}{TN+FP} \quad (14)$$

$$F1-Score = 2\left(\frac{P*R}{P+R}\right) \quad (15)$$

Additionally, discussions cover visual representations of reconstructed images, clusters of class capsules, and layer activation maps, which illustrate the model's internal mechanisms and help users have confidence in the model. Again, an ablation study demonstrates the model's flexibility and robustness, highlighting the components that significantly impact its performance. Moreover, the quantity of parameters and the model's disk size are also discussed.

Based on the confusion matrices in Fig. 3 for the tomato and mango datasets, it is evident that the DLCA-CapsNet model categorised the images into their respective classes more accurately than the traditional model. From the confusion matrices, accuracy per class, sensitivity, precision, F1-score, and specificity values can be calculated. Such metrics ensure a well-rounded analysis of model performance, which is crucial for small and unevenly distributed dataset scenarios, where relying solely on accuracy might overlook critical details. Based on Table I and Table II, it can be noticed that the proposed DLCA-CapsNet model achieved higher accuracy, sensitivity, precision, F1-Score, and specificity, surpassing values attained by the traditional model when the tomato and mango datasets are considered respectively. This suggests that the proposed model outperforms the traditional model and effectively generalizes unseen data.

Also, Fig. 4 depicts the PR and ROC curves of the tomato dataset, providing insights into the model's resilience and effectiveness on imbalanced datasets. The suggested model outperformed the traditional model, as indicated by superior curve shapes and values. The DLCA-CapsNet significantly outperformed the original CapsNet on the Tomato dataset, achieving an overall ROC of 100% and a PR of 99.5%, compared to the original CapsNet's lower ROC of 97.7% and PR of 88%.

The training and validation loss and accuracy for both the DLCA-CapsNet and traditional CapsNet are shown in Fig. 5 for tomato, mango, and CIFAR-10 datasets. The DLCA-CapsNet model converged faster and achieved higher validation accuracies than the traditional model on the various datasets. It can be seen that, the DLCA-CapsNet significantly outperformed the original CapsNet in validation accuracy across eleven diverse datasets (apple, banana, grape, corn, mango, pepper, potato, rice, tomato, CIFAR-10, and Fashion-MNIST), achieving consistently higher validation accuracy scores of 99.69%, 95.77%, 99.63%, 97.66%, 99.50%, 100%, 100%, 100%, 98.82%, 77.31%, and 93.01% compared to that of the traditional CapsNet model that achieved lower validation accuracies of 92.91%, 80.95%, 93.49%, 92.59%, 78.38%, 59.68%, 95.36%, 99.24%, 88.59%, 63.58%, and 90.98% for the same dataset. This suggests the proposed model generalizes better on unseen data than the traditional model across all the datasets. Examining the original or traditional CapsNets validation accuracy for the CIFAR-10 dataset reveals its initial increase but decline around the 20th epoch. In contrast, the proposed model maintained its peak accuracy until the end.

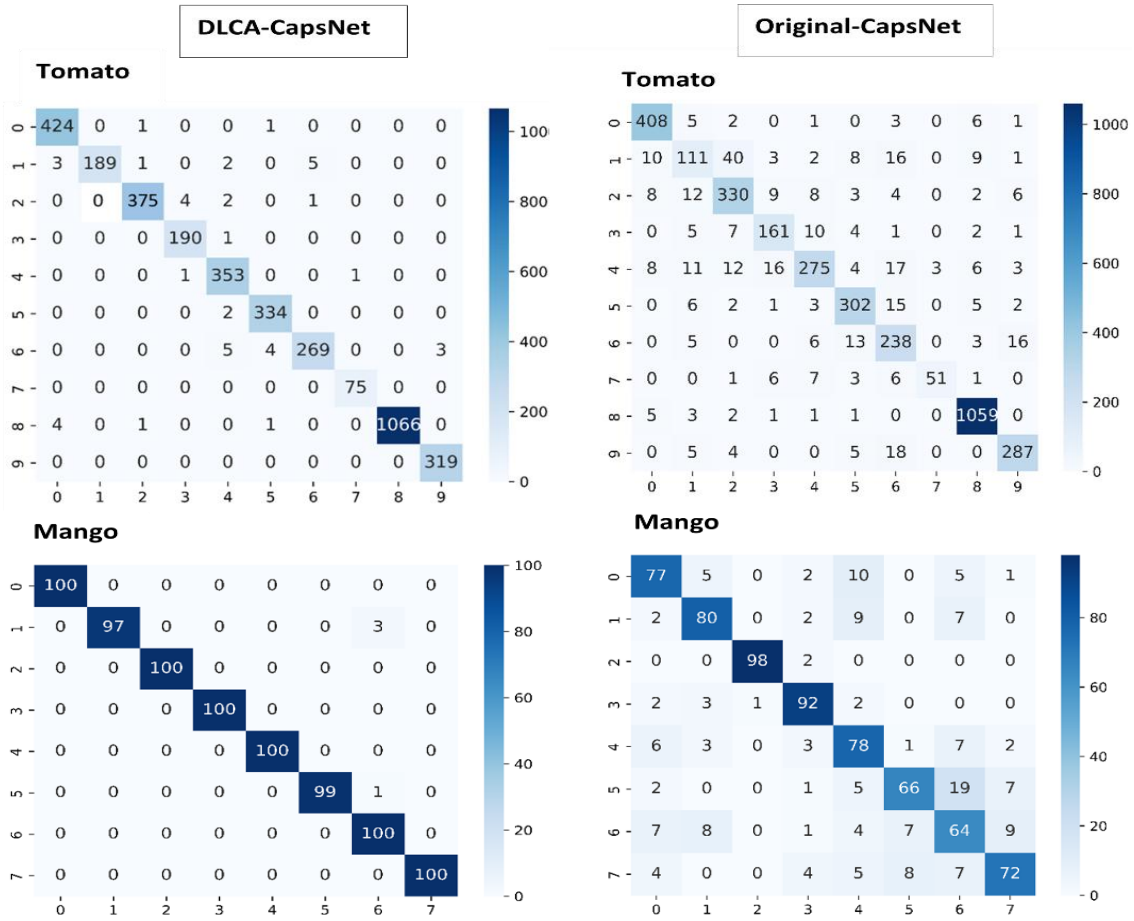


Fig. 3. Confusion matrices comparing the performance of the DLCA-CapsNet and the original CapsNet models on the tomato and mango datasets.

TABLE I. PERFORMANCE RESULTS OF DLCA-CAPSNET AND THE ORIGINAL CAPSNET ON THE TOMATO DATASET

Model (Dataset)	Class	TP	FP	FN	TN	ACC (%)	PRE (%)	SEN (%)	SPE (%)	FS (%)	Data Size
Original-CapsNet (Tomato)	0	408	31	18	3180	98.65	92.94	95.78	99.04	94.34	426
	1	111	52	89	3385	96.12	68.10	55.50	98.49	61.16	200
	2	330	70	52	3185	96.65	82.50	86.39	97.85	84.40	382
	3	161	36	30	3410	98.19	81.73	84.29	98.96	82.99	191
	4	275	38	80	3244	96.76	87.12	77.47	98.84	82.01	355
	5	302	41	34	3260	97.94	88.05	89.88	98.76	88.96	336
	6	238	80	43	3276	96.62	74.84	84.70	97.62	79.47	281
	7	51	3	24	3559	99.26	94.44	68.00	99.92	79.07	75
	8	1059	34	13	2531	98.71	96.89	98.79	98.68	97.83	1072
	9	287	30	32	3288	98.30	90.54	89.97	99.10	90.25	319
DLCA-CapsNet (Tomato)	0	424	7	2	3204	99.75	98.38	99.53	99.78	98.95	426
	1	189	0	11	3437	99.70	100	94.50	100	97.17	200
	2	375	3	7	3252	99.73	99.21	98.17	99.91	98.69	382
	3	190	5	1	3441	99.84	97.44	99.48	99.86	98.45	191
	4	353	12	2	3270	99.62	96.71	99.44	99.63	98.06	355
	5	334	6	2	3295	99.78	98.24	99.41	99.82	98.82	336
	6	269	6	12	3350	99.51	97.82	95.73	99.82	96.76	281
	7	75	1	0	3561	99.97	98.68	100	99.97	99.34	75
	8	1066	0	6	2565	99.84	100	99.44	100	99.72	1072
	9	319	3	0	3315	99.92	99.07	100	99.91	99.53	319

TABLE II. PERFORMANCE RESULTS OF DLCA-CAPSNET AND THE ORIGINAL CAPSNET ON THE MANGO DATASET

Model (Dataset)	Class	TP	FP	FN	TN	ACC (%)	PRE (%)	SEN (%)	SPE (%)	FS (%)	Data Size
Original-CapsNet (Mango)	0	77	23	23	677	94.25	77.00	77.00	96.71	77.00	100
	1	80	19	20	681	95.13	80.81	80.00	97.29	80.40	100
	2	98	1	2	699	99.63	98.99	98.00	99.86	98.49	100
	3	92	15	8	685	97.13	85.98	92.00	97.86	88.89	100
	4	78	35	22	665	92.88	69.03	78.00	95.00	73.24	100
	5	66	16	34	684	93.75	80.49	66.00	97.71	72.53	100
	6	64	45	36	655	89.88	58.72	64.00	93.57	61.25	100
	7	72	19	28	681	94.13	79.12	72.00	97.29	75.39	100
DLCA-CapsNet (Mango)	0	100	0	0	700	100	100	100	100	100	100
	1	97	0	3	700	99.63	100	97.00	100	98.48	100
	2	100	0	0	700	100	100	100	100	100	100
	3	100	0	0	700	100	100	100	100	100	100
	4	100	0	0	700	100	100	100	100	100	100
	5	99	0	1	700	99.88	100	99.00	100	99.50	100
	6	100	4	0	696	99.50	96.15	100	99.43	98.04	100
	7	100	0	0	700	100	100	100	100	100	100

B. Number of Parameters and Size of Disk

Increasing model complexity (by adding layers or increasing the size of layers) often improves performance on intricate images. However, this makes the models larger and computationally demanding, hindering their use on devices with limited resources like phones and embedded systems. DLCA-CapsNet is smaller in size and has fewer parameters than the original CapsNet and other top models (see Table III). Additionally, the CDH component we used does not add any extra parameters. The DLCA-CapsNet achieves a notable decrease in parameter count (in millions (M)) by 6.16M, 6.16M, 6.16M, 6.16M, 7.14M, 5.68M, 5.92M, 7.62M, 7.62M, and 6.54M respectively, when compared to the conventional CapsNet considering apple, banana, grape, corn, mango, pepper, potato, rice, tomato, CIFAR-10, and Fashion-MNIST datasets. Furthermore, it results in a reduction in disk space usage by 23.5MB, 23.5MB, 23.5MB, 23.5MB, 27.2MB, 21.6MB, 22.6MB, 23.5MB, 29.1MB, 29.1MB, and 24.8MB, respectively, for the same datasets.

TABLE III. COMPARISON OF DISK SIZE (S) IN MB AND NUMBER OF PARAMETERS (P) IN MILLIONS (M)

CapsNet Models/Reference	P(M) and S(MB)											
		Apple	Banana	Grape	Maize	Mango	Pepper	Potato	Rice	Tomato	CIFAR-10	Fashion-MNIST
Gabor CapsNet [14]	P	-	-	-	-	-	-	-	-	12.00	-	-
CapsNet[18]	P	-	-	-	8.40	-	-	-	-	8.40	5.20	2.80
K-Means CapsNet[37]	P	-	-	-	-	-	-	-	-	5.12	-	-
CapsNet [19]	P	-	-	-	-	-	-	9.86	-	-	-	-
Gabor-Maxpooled CapsNet [22]	P	-	-	-	-	-	-	-	-	8.71	-	-
Shallow/Multi-Input CapsNet [28]	P	-	-	-	-	-	-	-	-	4.10/ 4.00	4.60/ 4.30	2.50/ 2.20
Dual-Input CapsNet [24]	P	-	-	-	-	-	-	-	-	6.04	5.48	-
Original CapsNet [7]	P	10.13	10.13	10.13	10.13	11.21	9.59	9.86	10.13	11.75	11.75	8.22
	S	38.6	38.6	38.6	38.6	42.7	36.5	37.6	38.6	44.8	44.8	31.3
DCLA-CapsNet proposed	P	3.97	3.97	3.97	3.97	4.07	3.91	3.94	3.97	4.13	4.13	1.68
	S	15.1	15.1	15.1	15.1	15.5	14.9	15.0	15.1	15.7	15.7	6.5

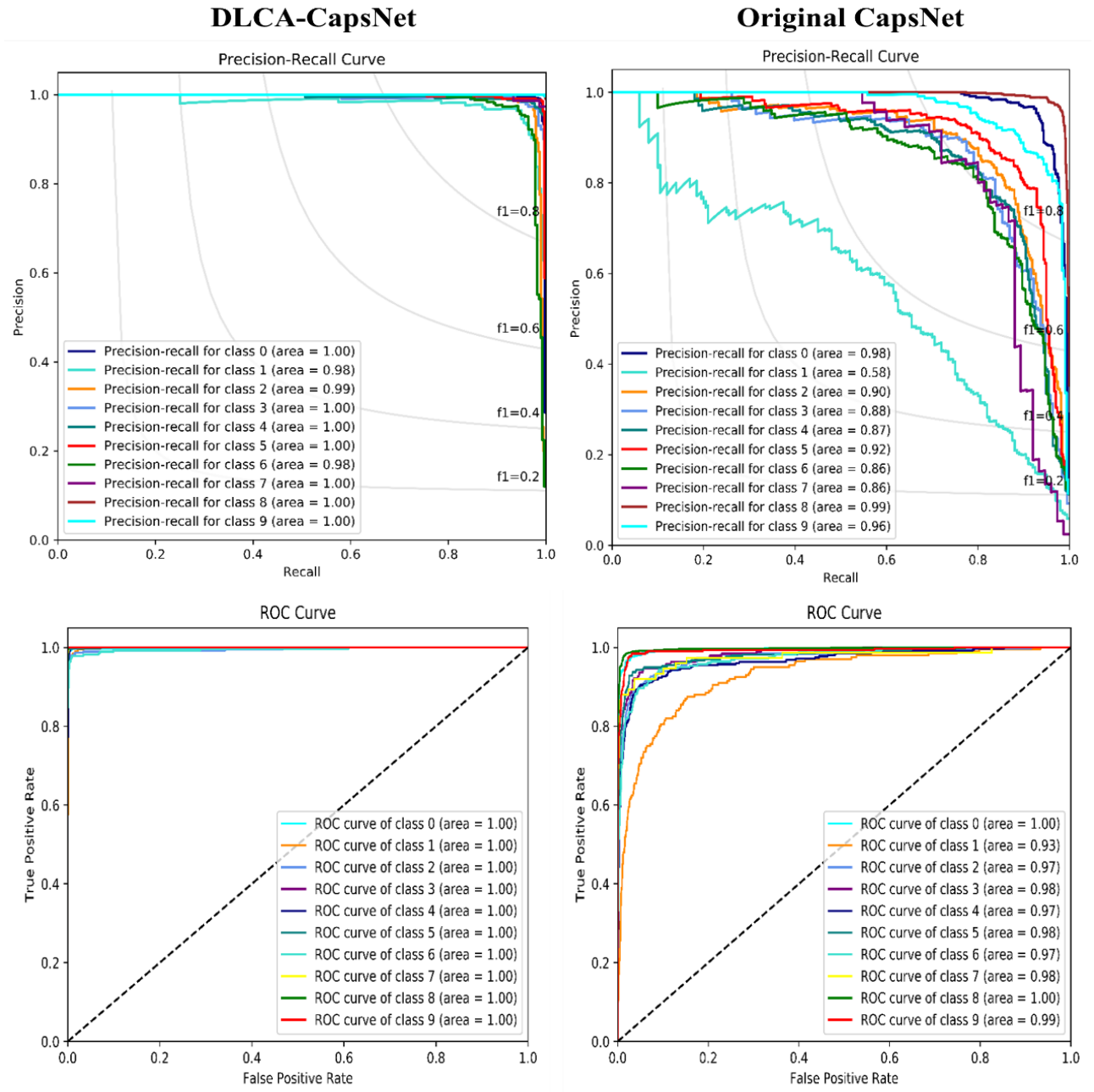


Fig. 4. PR and ROC curves comparing the performance of DLCA-CapsNet and Original CapsNet models on the Tomato dataset.

C. Ablation Study

To assess the model parts that affect its performance, an ablation study was performed [35]. The layers are removed one by one to check the layers of the model that significantly affect the performance of the model. Considering the tomato and potato leave diseases, it can be seen in Table IV that the model's performance is highly impacted by the CDH layer.

TABLE IV. ABLATION RESULTS

Layers	Validation accuracy (%)	
	Tomato	Potato
-CDH	82.12	90.17
-Atrous_Conv1/ Atrou_Conv2	97.75	95.95
-MP1 & MP3	98.62	98.88
-Conv1 & Conv2	97.88	97.17
-MP2 & MP4	98.75	98.28
-Dropout	98.88	99.14
+All Layers	98.82	100

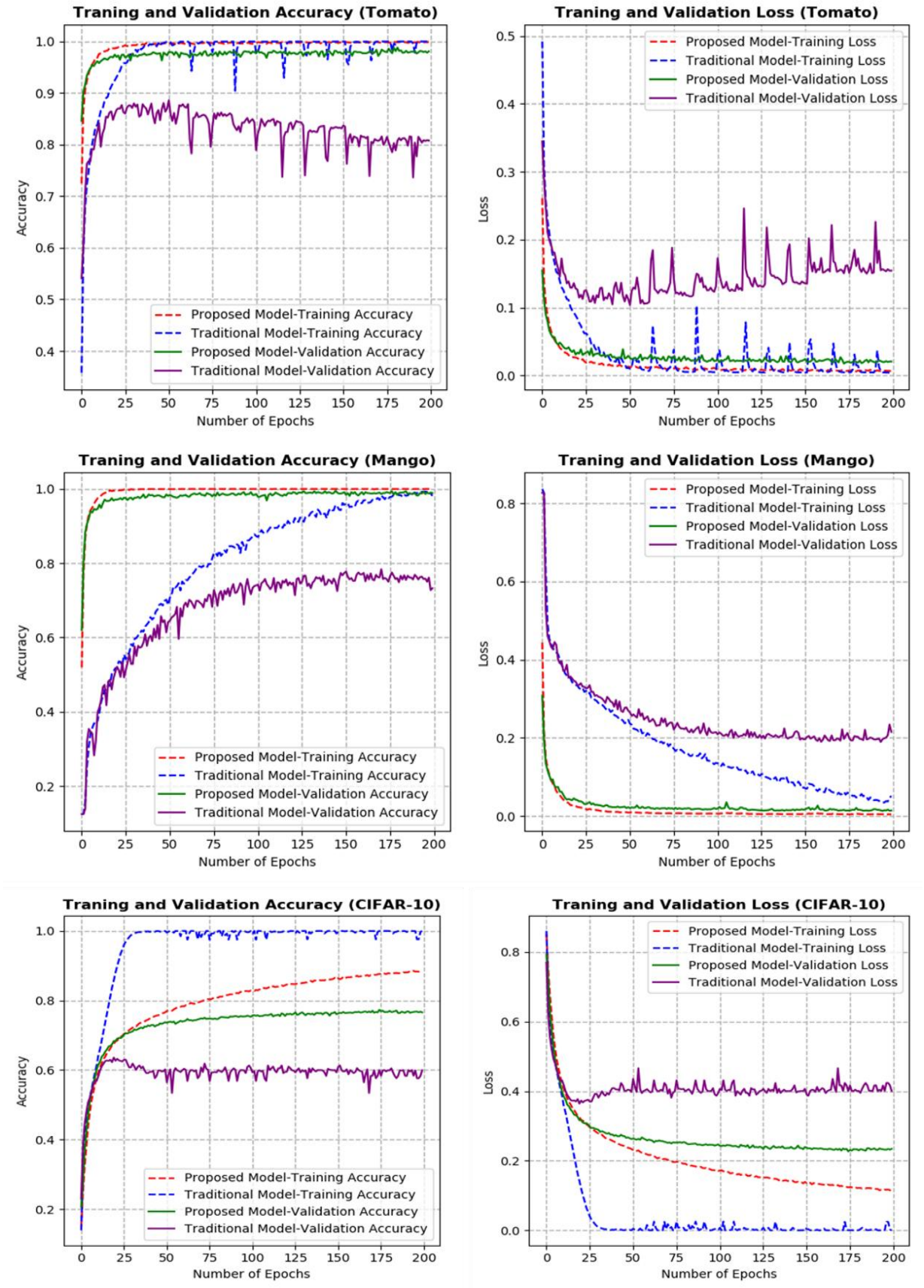


Fig. 5. Accuracy and loss performance for the proposed DLCA-CapsNet and the original CapsNet models across three datasets: tomato, mango, and CIFAR-10.

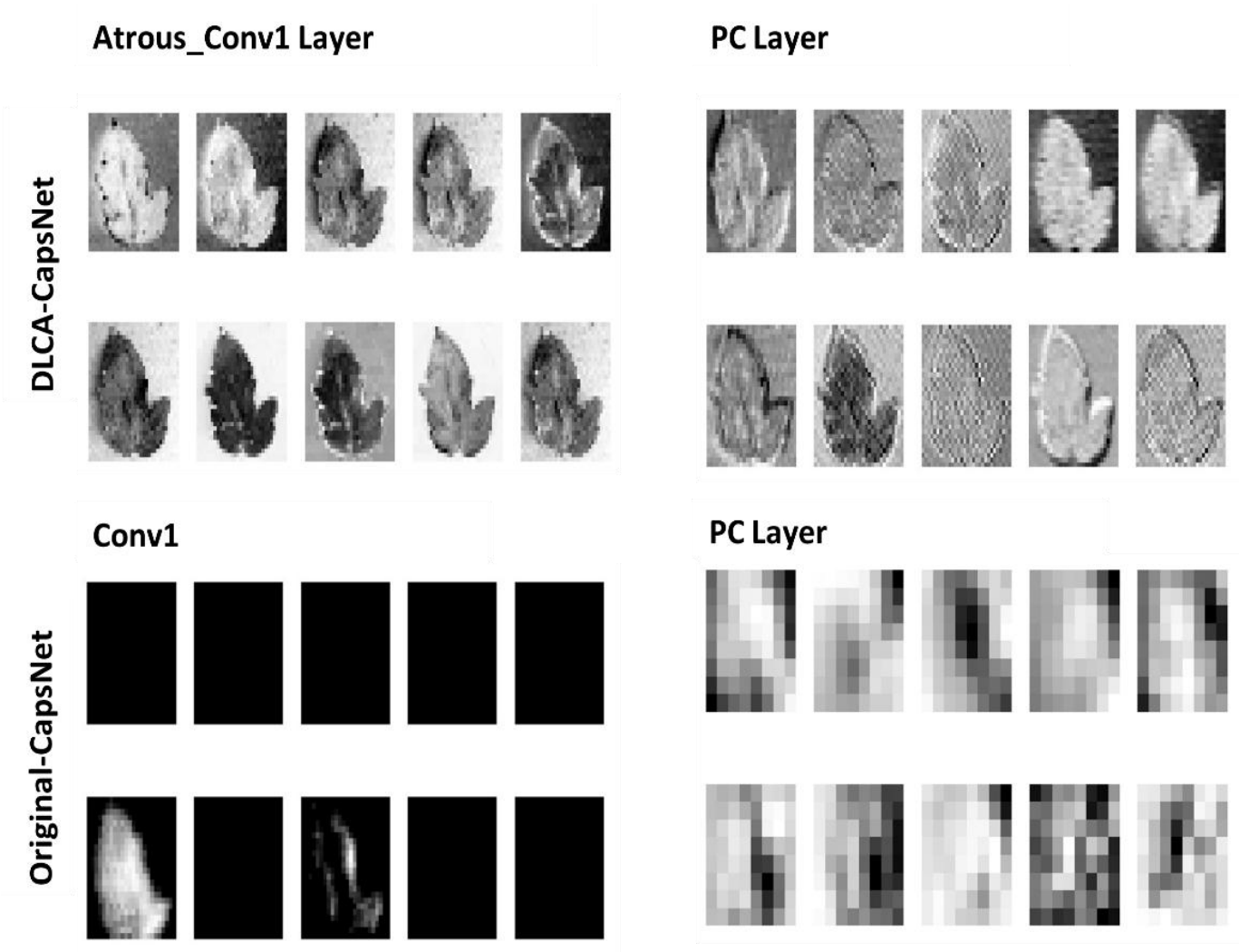


Fig. 6. Activation maps of the DLCA-CapsNet and the original CapsNet models on the Tomato dataset.

D. Model Interpretability

Fig. 6 visually compares activation maps from one of the atrous convolution layers in the proposed DLCA-CapsNet model (which takes input from the CDH layer using the tomato dataset) to the activation map from the convolution layer of the traditional CapsNet. A comparison of these maps reveals that the atrous convolution in the DLCA-CapsNet captures more detailed features, suggesting that the CDH layer plays a crucial role in extracting significant information, something the convolution layer of the traditional CapsNet alone fails to achieve. Furthermore, comparing the activation maps from the primary capsule layer of the proposed DLCA-CapsNet and the original model reveals that DLCA-CapsNet captures more relevant features. This improvement is attributed to the proposed model's ability to extract significant features earlier in the network, which enhances the quality of the features passed to its primary capsule layer, unlike the traditional CapsNet, whose convolutional layer was less effective in feature extraction.

The clusters generated at the class capsule layers were also visualized using t-distributed stochastic neighbor embedding (t-SNE). As shown in Fig. 7, considering the DLCA-CapsNet model, the clusters at the class capsule layer are more distinctly grouped by class, with fewer outliers compared to the traditional model, particularly for the tomato, mango, and pepper datasets. This indicates that the DLCA-CapsNet model demonstrates a stronger ability to distinguish between different classes in the dataset than the traditional model. Also, the reconstruction technique aids in identifying the predicted class of an image and the confidence of that prediction. Referring to Fig. 8, which presents three rows of reconstructed images from the tomato and banana datasets for both the proposed and traditional CapsNet models, it is evident that the proposed model generates images of slightly better quality with higher class probabilities. These visual outputs from the model layers enhance interpretability and support the goals of Explainable Artificial Intelligence (XAI).

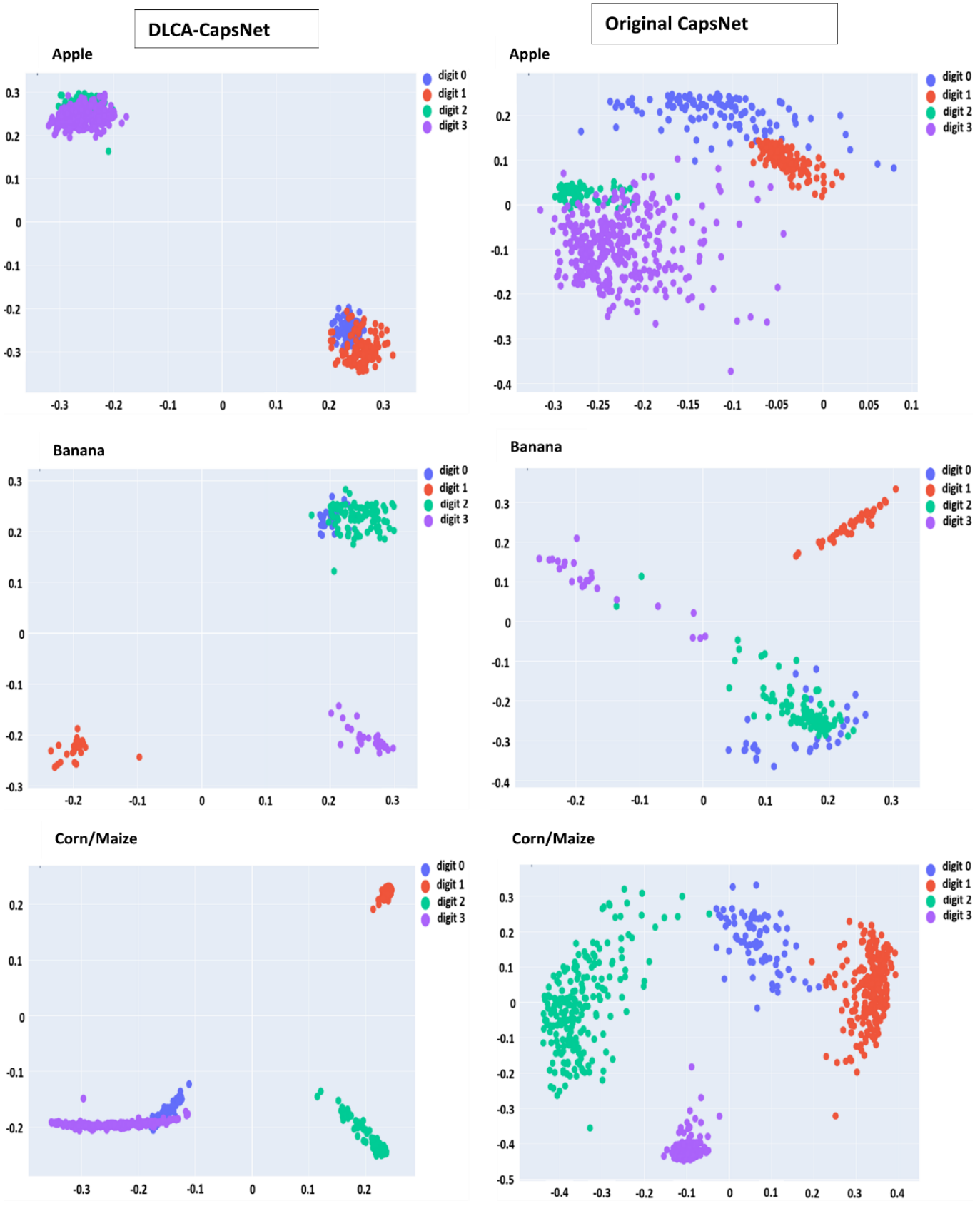


Fig. 7. Class capsule clusters for Apple, Banana, and Corn in DLCA-CapsNet and the original CapsNet models.

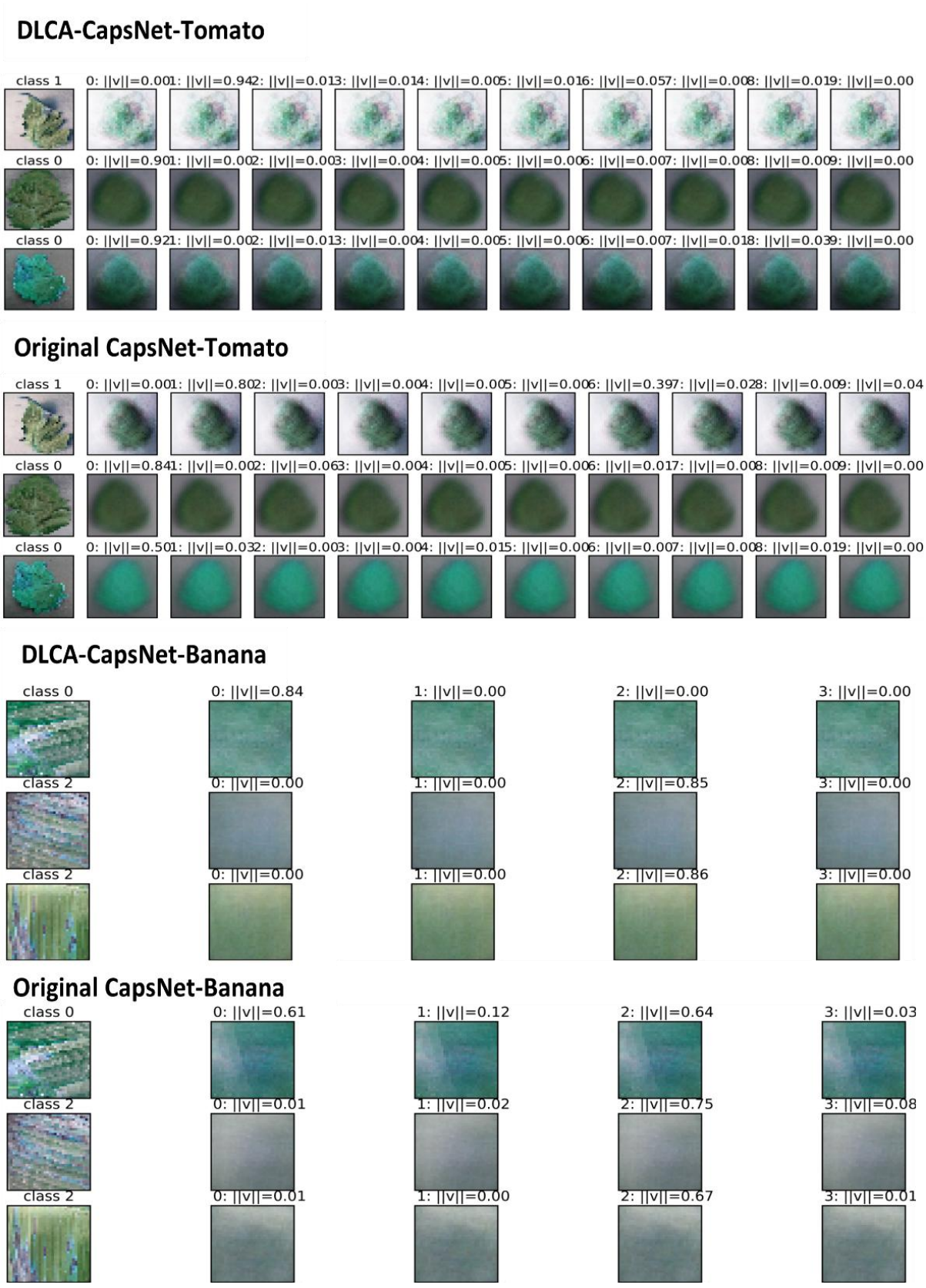


Fig. 8. Reconstructed images generated by the DLCA-CapsNet and Original CapsNet models when applied to the Tomato and Banana datasets.

E. Comparison of Results

Table V compares the proposed DLCA-CapsNet model and state-of-the-art models applied to CIFAR-10, Fashion-MNIST, and the nine plant disease datasets from the literature. The comparison also includes various routing algorithms, despite the DLCA-CapsNet model utilizing a dynamic routing technique. Table V shows that the less parameterized DLCA-CapsNet model, with enhanced feature extraction capabilities, achieved higher validation accuracies across various datasets, surpassing the original CapsNet by 6.78%, 14.82%, 6.14%, 5.07%, 21.12%, 40.32%, 4.64%, 0.76%, 10.23%, 13.73%, and 2.03% for the apple, banana, grape, corn, mango, pepper, potato, rice, tomato, CIFAR-10, and Fashion-MNIST datasets, respectively. It also resulted in a reduction of 6.16M, 6.16M, 6.16M, 6.16M, 7.14M, 5.68M, 5.92M, 7.62M, 7.62M, and 6.54M in model size (in millions) compared to the traditional CapsNet. Furthermore, the DLCA-CapsNet model led to a decrease in disk size by 23.5MB, 23.5MB, 23.5MB, 23.5MB, 27.2MB, 21.6MB, 22.6MB, 23.5MB, 29.1MB, 29.1MB, and 24.8MB for the same datasets compared to the traditional CapsNet, as shown in Table III. Also, comparing the proposed DLCA-CapsNet with existing models found in the literature for plant disease detection, the DLCA-CapsNet also outperformed them, as shown in Table V and Table III. These results demonstrate the superior generalization ability and lower computational complexity of the DLCA-CapsNet over the original CapsNet. This exceptional performance, necessary for complex images, can be credited to the superior feature extraction methods we used to isolate only the most relevant features from the images.

TABLE V. COMPARISON BETWEEN PREVIOUS STUDIES AND THE PROPOSED DCLA-CAPSNET MODELS

CapsNet Models/Reference	Validation Accuracy (%)										
	Apple	*Banana*	*Grape*	*Corn*	*Mango*	*Pepper*	*Potato*	*Rice*	*Tomato*	*CIFAR-10*	*Fashion-MNIST*
Gabor CapsNet [14]	-	-	-	-	-	-	-	-	98.13	-	-
E-GAN CapsNet [16]	-	-	97.63	-	-	-	-	-	-	-	-
CapsNet [18]	-	-	-	96.79	-	-	-	-	98.06	75.80	92.72
Dilated CapsNet [17]	93.16	-	-	-	-	-	-	-	-	-	-
K-Means CapsNet [37]	-	-	-	97.99	-	-	-	-	98.80	-	-
ConvCapsNet [29]	-	-	99.12	-	-	-	-	-	-	-	-
CapsNet [19]	-	-	-	-	-	-	91.83	-	-	-	-
CapsNet [21]	-	-	-	-	-	95.76	-	-	-	-	-
CapsNet [20]	-	95.00	-	-	-	-	-	-	-	-	-
Gabor-Maxpooled CapsNet [22]	-	-	-	-	-	-	-	-	97.98	-	-
Shallow/Multi-Input CapsNet [28]	-	-	-	-	-	-	-	-	97.33/ 94.04	75.75/ 63.95	92.7/ 91.45
Dual-Input CapsNet [24]	-	-	-	-	-	-	-	-	93.03	76.58	-
Multi-Channel CapsNet [25]	-	-	-	-	-	-	-	-	98.15	-	-
CapsNet [26]	-	-	-	-	-	-	-	-	96.39	-	-
SE-SK CapsNet [27]	-	-	-	-	-	-	-	-	97.19	-	-
Original CapsNet [7]	92.91	80.95	93.49	92.59	78.38	59.68	95.36	99.24	88.59	63.58	90.98
DCLA-CapsNet Proposed	99.69	95.77	99.63	97.66	99.50	100	100	100	98.82	77.31	93.01

V. CONCLUSION

This study proposed an optimized DLCA-CapsNet model for classifying plant diseases, CIFAR-10, and fashion-MNIST. Modifications were made by adding a CDH, which does not add any parameters, atrous convolutions, max-pooling, or a dropout layer. All these layers contributed to the efficient feature extraction abilities of the proposed model. The DLCA-CapsNet model was assessed using evaluation metrics such as sensitivity, F1-score, precision, specificity, ROC and PR values, accuracy, disk size, and parameters generated. The DLCA-CapsNet model results were compared with those of the traditional CapsNet and other models found in the literature, and outperformed them as shown in Table III and Table V. The proposed DLCA-CapsNet model's fewer parameters make it usable on IoT and resource-constrained devices, and the better validation accuracies show its generalization ability on unseen data. An ablation study was performed to ascertain the layers of the model that influence its performance. Again, the model interpretability regarding visualizing clusters at the class capsule, activation maps, and reconstruction of images was discussed. The proposed model's better performance shows its effectiveness in detecting plant diseases from plant leaf images than those found in literature as analyzed in Table III and Table V. The findings suggest that this efficient and computationally less demanding method can significantly enhance plant disease classification and contributes incrementally to efforts aligned with the SDG 2 goal by offering a lightweight, scalable solution that can be adapted for field use in resource-constrained settings.

Nonetheless, environmental variables like uneven lighting and intricate backgrounds in real-world conditions can hinder or limit model performance. Subsequent research will aim to enhance generalizability and real-world applicability by evaluating the model across more diverse, challenging settings, incorporating additional datasets, and investigating real-time implementation on edge and mobile platforms for agricultural use.

REFERENCES

[1] T. Li et al, "Applications of Deep Learning in Fundus Images: A Review," *Medical Image Analysis*, 69, 101971, 2021.

[2] X. S. Wei et al, "Fine-Grained Image Analysis with Deep Learning: A Survey," *IEEE transactions on pattern analysis and machine intelligence*, 44(12), 8927-8948, Nov. 2021.

[3] X. Chen et al, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical image analysis*, 79: 102444, 2021.

[4] A. Farahat, F. Effenberger, and M. Vinck, "A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations," *Neural Networks*, vol. 160, pp. 400-414, Aug. 2023.

[5] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imaging*, vol. 6, no. 6, pp. 1–19, 2020, doi: 10.3390/JIMAGING6060052.

[6] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.

[7] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," *Appl. Biosaf.*, vol. 22, no. 4, pp. 185–186, 2017, doi: 10.1177/1535676017742133.

[8] M. Mitterreiter, M. Koch, J. Giesen, and S. Laue, "Why Capsule Neural Networks Do Not Scale: Challenging the Dynamic Parse-Tree Assumption," in *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, no. 8, 2023.

[9] V. Mazzia, F. Salvetti, and M. Chiaberge, "Efficient-CapsNet: Capsule Network with Self-Attention Routing," *Scientific reports*, 11.1: 14634, 2021.

[10] S. Cao, Y. Yao, and G. An, "E2-capsule neural networks for facial expression recognition using AU-aware attention," *IET Image Process.*, vol. 14, no. 11, pp. 2417–2424, 2020, doi: 10.1049/iet-ipr.2020.0063.

[11] G. H. Liu and J. Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, 2013, doi: 10.1016/j.patcog.2012.06.001.

[12] S. Qiao, L. C. Chen, and A. Yuille, "Detectors: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10213-10224, 2021.

[13] L. Zhang, J. Zhang, Z. Li, and Y. Song, "A multiple-channel and atrous convolution network for ultrasound image segmentation," *Medical Physics*, vol. 47, no. 12, pp. 6270–6285, Dec. 2020.

[14] P. K. Mensah, B. A. Weyori, and M. A. Ayidzoe, "Gabor Capsule Network for Plant Disease Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 388–395, 2020.

[15] S. Verma, A. Chug, R. P. Singh, A. P. Singh, and D. Singh, "SE-CapsNet : Automated evaluation of plant disease severity based on feature extraction through Squeeze and Excitation (SE) networks and Capsule networks University School of Information , Communication & Technology (USIC & T), Guru Gobind Singh Indra," *Kuwait J. Sci.*, vol. 49, no. 1, pp. 1–31, 2022.

[16] N. Vasudevan and T. Karthick, "A Hybrid Approach for Plant Disease Detection Using E-GAN and CapsNet," *Comput. Syst. Sci. Eng.*, vol. 46, no. 1, pp. 337–356, 2023, doi: 10.32604/csse.2023.034242.

[17] C. Xu, X. Wang, and S. Zhang, "Dilated convolution capsule network for apple leaf disease identification," *Front. Plant Sci.*, vol. 13, no. November, pp. 1–13, 2022, doi: 10.3389/fpls.2022.1002312.

[18] P. Mensah, B. Asubam, and A. Abra, "Exploring the performance of LBP-capsule networks with K-Means routing on complex images," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2574–2588, 2022, doi: 10.1016/j.jksuci.2020.10.006.

[19] S. Verma, A. Chug, and A. P. Singh, "Exploring capsule networks for disease classification in plants," *J. Stat. Manag. Syst.*, vol. 23, no. 2, pp. 307–315, 2020, doi: 10.1080/09720510.2020.1724628.

[20] B. F. Oladejo and O. O. Ademola, "Automated Classification of Banana Leaf Diseases using an Optimized Capsule Network Model," pp. 119–130, 2020, doi: 10.5121/csit.2020.100910.

[21] G. ALTAN, "Performance Evaluation of Capsule Networks for Classification of Plant Leaf Diseases," *Int. J. Appl. Math. Electron. Comput.*, vol. 8, no. 3, pp. 57–63, Sep. 2020, doi: 10.18100/ijamec.797392.

[22] P. K. Mensah, B. A. Weyori, and A. A. Mighty, "Max-pooled fast learning gabor capsule network," *2020 Int. Conf. Artif. Intell. Big Data, Comput. Data Commun. Syst. icABCD 2020 - Proc.*, 2020, doi: 10.1109/icABCD49160.2020.9183823.

[23] A. Anant, "AppleCaps : A Capsule Model for Classification of Foliar Diseases in Apple Leaves Rakhi Ashok Sonkusare National College of Ireland Supervisor :," *Natl. Coll. Irel.*, 2021.

[24] P. K. Mensah and M. A. Ayidzoe, "Overview of CapsNet Performance Evaluation Methods for Image Classification using a Dual Input Capsule Network as a Case Study," *Int. J. Comput. Digit. Syst.*, vol. 1, no. 1, 2022.

[25] M. Peker, "Multi-channel capsule network ensemble for plant disease detection," *SN Appl. Sci.*, vol. 3, no. 7, 2021, doi: 10.1007/s42452-021-04694-2.

[26] L. M. Abouelmagd, M. Y. Shams, H. S. Marie, and A. E. Hassanien, "An optimized capsule neural networks for tomato leaf disease classification," *Eurasip J. Image Video Process.*, vol. 2024, no. 1, 2024, doi: 10.1186/s13640-023-00618-9.

[27] X. Zhang, Y. Mao, Q. Yang, and X. Zhang, "A Plant Leaf Disease Image Classification Method Integrating Capsule Network and Residual Network," *IEEE Access*, vol. 12, no. February, pp. 44573–44585, 2024, doi: 10.1109/ACCESS.2024.3377230.

[28] P. K. Mensah, B. A. Weyori, and M. A. Ayidzoe, "Evaluating shallow capsule networks on complex images," *Int. J. Inf. Technol.*, vol. 13, no. 3, pp. 1047–1057, 2021, doi: 10.1007/s41870-021-00694-y.

[29] A. D. Andrushia, T. M. Neebha, A. T. Patricia, S.Umadevi, N.Anand, and A. Varshney, "Image Based Disease Classiication in Grape Leaves Using Convolutional Capsule Network Image based Disease Classification in Grape Leaves using Convolutional Capsule Network," *Soft Comput.*, 2022.

[30] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," 2015.

[31] S. E. Arman, M. A. B. Bhuiyan, H. M. Abdullah, S. Islam, T. T. Chowdhury, and M. A. Hossain, "BananaLSD: A banana leaf images dataset for classification of banana leaf diseases using machine learning," *Data Br.*, vol. 50, p. 109608, 2023, doi: 10.1016/j.dib.2023.109608.

[32] S. I. Ahmed *et al.*, "MangoLeafBD: A comprehensive image dataset to classify diseased and healthy mango leaves," *Data Br.*, vol. 47, p. 108941, 2023, doi: 10.1016/j.dib.2023.108941.

[33] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," *Mach. Learn.*, pp. 1–6, 2017.

[34] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *ASHA*, vol. 34, no. 4, 2009.

[35] M. Li and L. Janson, "Optimal Ablation for Interpretability," *Advances in Neural Information Processing Systems*, vol 37, pp.109233-109282, Sep. 16, 2024.

[36] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. K. Behera. "Deep feature-based rice leaf disease identification using support vector machine." *Computers and Electronics in Agriculture* 175 (2020):105527, doi: 10.1016/j.compag.2020.105527

[37] K. P. Mensah, B. A. Weyori, and A. M. Ayidzoe, "Capsule network with K-Means routing for plant disease recognition," *J. Intell. Fuzzy Syst.*, pp. 1–12, 2020, doi: 10.3233/JIFS-201226.

A Comparative Study of Machine Learning Techniques for AE-Based Corrosion Detection with Emphasis on Transformer Models

Osama Shahid Ali, Lukman B A Rahim
Center for Cyber Physical Systems, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

***Abstract*—Corrosion-induced damage poses a critical threat to the structural integrity of fluid transport pipelines, necessitating advanced detection strategies for early intervention. This study investigates the use of acoustic emission (AE) monitoring in conjunction with machine learning techniques to identify anomalies indicative of corrosion. A comprehensive analysis of supervised, unsupervised, semi-supervised, and self-supervised learning methods is presented, with emphasis on their suitability for AE-based anomaly detection. Building upon this foundation, we implement and evaluate multiple machine learning models—including K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN)—and compare them to a Transformer-based model integrated into a hybrid CNN-Transformer architecture. Experimental results demonstrate that the hybrid model outperforms all baselines, achieving R-squared values of 0.7037 for Acoustic Signal Level (ASL) and 0.6836 for Root Mean Square (RMS), thus confirming its superior ability to capture both local and long-range dependencies in acoustic emission data. A systematic review of recent Transformer-based corrosion detection models further contextualizes the results. This research highlights the promise of Transformer-based models in robust, real-time corrosion monitoring and offers a pathway toward more intelligent, machine learning-driven infrastructure maintenance systems.**

Keywords*—Acoustic emissions; transformer based models; machine learning

I. Introduction

The structural integrity of pipelines is critical to the safe and efficient transport of fluids across industrial infrastructure. Among the various threats to pipeline reliability, corrosion remains one of the most pervasive and damaging phenomena, often leading to catastrophic failures, economic losses, and environmental hazards. As such, early and accurate detection of corrosion-related anomalies is essential for preventive maintenance and risk mitigation. Traditional inspection methods, such as ultrasonic testing, visual inspection, or radiography, are often constrained by accessibility, resolution, and cost. These limitations have catalysed the shift toward non-destructive monitoring approaches, particularly those based on acoustic emission (AE) signals, which can detect internal damage mechanisms in real-time without interrupting operations.

Within this context, the field of anomaly detection has gained considerable traction. By identifying deviations from normal behaviour in sensor data, anomaly detection methods enable the recognition of subtle, early-stage corrosion signatures. These methods span from classical statistical techniques to advanced machine learning and deep learning models. More recently, the adoption of Transformer-based architectures—originally developed for natural language processing—has opened new avenues for capturing long-range dependencies and intricate temporal patterns in time-series data, including AE signals. Transformers, especially when combined with Convolutional Neural Networks (CNNs), have demonstrated superior performance in learning both localized and global representations, making them particularly well-suited for signal-based corrosion detection.

This study contributes to this evolving landscape by applying and evaluating a hybrid CNN-Transformer model for AE-based corrosion anomaly detection in pipelines. Unlike conventional approaches that often struggle with noisy, high-dimensional, and unlabelled signal data, the hybrid model leverages CNNs for localized feature extraction and Transformer encoders for contextual temporal modelling. A comprehensive comparison against classical machine learning algorithms and other deep learning models reveals the hybrid model's significant performance advantage in predicting key corrosion indicators such as Acoustic Signal Level (ASL) and Root Mean Square (RMS).

Furthermore, this work contextualizes its findings within the broader literature by examining prior studies on Transformer-based corrosion detection across various domains, including bridge inspection, UAV imagery, pipeline surveillance, and 3D point cloud segmentation. By bridging insights from both image-based and signal-based corrosion detection, this paper not only validates the effectiveness of Transformer models in AE applications but also lays the groundwork for future self-supervised learning strategies in industrial health monitoring systems.

II. Background of the Study

A. Anomaly Detection

Anomaly detection is the process of identifying data points, patterns, or events that significantly deviate from the normal or expected behaviour within a dataset [1]. These deviations, known as anomalies or outliers, can indicate critical incidents such as errors, defects, fraud, or other unusual activities. Anomaly detection is used across various fields like cybersecurity, finance, manufacturing, and healthcare to detect problems early and prevent larger issues. It typically involves

building a model of normal behaviour and then comparing new data to this model to find anything that stands out as abnormal.

Anomaly detection offers several significant advantages, especially in systems where early identification of unusual behaviour can prevent serious consequences. One of the primary benefits is its ability to uncover rare or previously unknown issues, such as hidden faults in machinery, undetected cyberattacks, or fraudulent financial transactions, which might not be identifiable using standard monitoring techniques [2]. It also supports automation in monitoring processes, reducing the need for manual oversight and allowing organizations to respond more quickly and efficiently to emerging problems. Furthermore, anomaly detection improves decision-making by providing insights into unexpected behaviour and enabling predictive maintenance, risk management, and enhanced security. Its flexibility across domains and ability to handle large volumes of data in real time make it a powerful tool for increasing operational reliability and safety.

B. Anomaly Detection Use Cases

Anomaly detection plays a pivotal role in cybersecurity, especially for identifying novel or stealthy attacks that signature-based systems often miss. For instance, a study on smart-city IoT networks applied federated and split learning methods to detect anomalies on the UNSW-NB15 dataset, achieving accuracy rates above 97% and reducing false positives in real-time intrusion detection. Another work introduced a PCC-CNN deep-learning model capable of detecting multi-class attacks in IoT environments using datasets like NSL-KDD and CICIDS2017, demonstrating robust performance in real-time classification [3], [4], [5]. These approaches exemplify how anomaly detection enhances threat detection in dynamic, decentralized network ecosystems.

In manufacturing, anomaly detection enables predictive maintenance by identifying early signs of equipment degradation. A machine-learning system using Random Forests and SVM, trained on sensor data like vibration, temperature, and pressure, successfully flagged anomalies indicative of impending motor and pump failures. This led to a remarkable 30% reduction in unplanned downtime and substantial cost savings. Additionally, a semi-supervised approach for pharmaceutical manufacturing equipment demonstrated its effectiveness in detecting failures in complex multivariate time series without requiring labelled fault data [6], [7]. These techniques shift maintenance strategies from reactive to proactive, improving operational efficiency and equipment lifespan.

Deep learning–based anomaly detection methods are transforming system reliability across industrial settings. One paper [8] explored autoencoders, RNNs, and CNNs to examine sensor streams in a manufacturing environment. The proposed deep-learning framework outperformed traditional techniques, enabling earlier detection of equipment faults. Another study [9] implemented a federated learning setup using 1D-CNN and bi-LSTM to process time-series sensor data across distributed devices, achieving 97.2% test accuracy and enabling edge-based anomaly detection that preserves data privacy. Together, these works illustrate how advanced neural architectures facilitate real-time fault prediction and distributed diagnostics in Industry 4.0 systems.

In civil and structural engineering, anomaly detection methods are increasingly used to identify corrosion on critical infrastructure—like bridges, towers, and building façades—using aerial imagery [10] collected by drones or unmanned aerial vehicles (UAVs). For instance, CorrDetector [11] is an ensemble deep learning model based on convolutional neural networks (CNNs) that analyses drone-captured images of structures (e.g., telecom towers) to locate and identify corroded regions [12]. This framework achieved significantly better classification accuracy compared to prior state-of-the-art models, enabling automated, remote, and scalable corrosion monitoring in hard-to-reach areas. Such anomaly-driven detection enhances safety, reduces inspection costs, and enables timely maintenance decisions [13].

C. Types of Anomalies

Point anomalies refer to individual data instances that deviate significantly from the majority of the data and can be identified as unusual regardless of the context in which they occur [14]. These anomalies are the simplest and most intuitive form, appearing as sharp spikes or dips in data distributions. A point anomaly stands alone as being inconsistent with the expected pattern—there is no need to examine its surrounding values or additional variables [15]. For example, in banking, if a customer typically spends less than $100 per transaction, but one day a transaction of $10,000 is recorded, that transaction is likely to be a point anomaly. Similarly, in a medical dataset, a sudden and extreme heart rate recorded in an otherwise stable trend could be flagged as anomalous. These anomalies are typically rare but can be highly impactful, often signalling errors, fraud, or critical failures. Because they occur in a variety of domains—from finance and healthcare to environmental monitoring and manufacturing—point anomalies are often the first type considered when building anomaly-aware systems. Despite their simplicity, understanding and classifying them accurately is crucial due to the potential severity of the events they represent.

Contextual anomalies occur when a data point is considered normal in one context but anomalous in another. Unlike point anomalies, these require additional contextual information—such as time, location, or season—to assess whether the data instance is truly unusual [16]. A contextual anomaly cannot be identified solely by looking at the value itself; its significance emerges only when the surrounding environment or condition is considered [17]. For example, a temperature of 25°C might be perfectly normal during the summer but highly unusual during the winter, making it a contextual anomaly in the latter case. Another example would be a sudden drop in stock prices that is abnormal only when viewed against the typical patterns of that specific market on a certain day of the week. These types of anomalies are especially prevalent in time-series and spatiotemporal data, where the same value can be either expected or unexpected depending on when and where it occurs. Recognizing contextual anomalies is essential in fields such as climate science, behavioural analytics, and industrial monitoring, where environmental variability plays a key role in defining what is normal.

Collective anomalies refer to a set or sequence of data instances that are anomalous when considered together, even though the individual data points within the group may appear normal on their own. The anomaly lies not in the data points themselves, but in the pattern or relationship between them. This type of anomaly typically appears in sequential or structured data, such as time-series or spatial data grids [18]. For example, in network traffic analysis, a single packet may appear normal, but a burst of similar packets over a short duration might indicate a denial-of-service attack—a collective anomaly. Similarly, in health monitoring systems, a slight elevation in blood pressure over one reading might not be a concern, but a sustained increase over several hours could signal the onset of a medical emergency. These anomalies are crucial in identifying coordinated or evolving patterns that could go unnoticed when data is analysed in isolation. They are especially important in domains like cybersecurity, fault detection, and system diagnostics, where anomalies emerge not from individual values, but from the way groups of values interact over time or space.

D. Anomaly Detection Techniques

1) Statistical techniques are among the earliest and most fundamental approaches to anomaly detection [19]. These methods rely on the assumption that normal data points follow a known statistical distribution (e.g., Gaussian, Poisson), and anomalies are observations that deviate significantly from this distribution. Common statistical measures used include mean, variance, standard deviation, and probability density functions.

A classic example is the Z-score method, where any data point that lies beyond a threshold (commonly 3 standard deviations from the mean) is considered anomalous[20]. This approach works well for data that is symmetrically distributed, such as in many industrial process control settings. Grubbs' Test is another statistical test used for identifying outliers in a univariate dataset that assumes normality; it tests the hypothesis that the extreme value in the dataset is an outlier [21].

Another foundational statistical method is the Interquartile Range (IQR) approach, which defines outliers as data points lying below $Q1-1.5\times IQR$ or above $Q3+1.5\times IQR$, where Q1 and Q3 are the first and third quartiles, respectively [19]. This is particularly useful for skewed distributions and boxplot-based anomaly visualization.

More advanced techniques include Kernel Density Estimation (KDE), a non-parametric way to estimate the probability density function of a random variable [22]. Anomalies are identified as points in regions of low estimated density. KDE is flexible and does not require assumptions about the distribution form.

In time-series applications, statistical methods like Moving Average and Exponential Smoothing are employed to detect deviations from expected trends. These are often combined with control limits to flag values that exceed expected variation. Similarly, Cumulative Sum Control Charts (CUSUM) and Exponentially Weighted Moving Average (EWMA) charts are widely used in quality control to detect small shifts in process mean [20].

These techniques are particularly valuable due to their interpretability, low computational cost, and ease of implementation. They are commonly used in industries such as finance (to detect fraud), manufacturing (to monitor production quality), and environmental science (to identify climate anomalies).

2) Time-series anomaly detection involves techniques that account for the temporal structure and order of the data [23]. Unlike traditional static datasets, time-series data includes observations collected over time, making the detection of anomalies more challenging due to seasonality, trends, and autocorrelation.

One of the most traditional techniques is the Autoregressive Integrated Moving Average (ARIMA) model. It predicts future values based on past observations. Anomalies are detected by comparing actual values to ARIMA-based forecasts; significant deviations signal potential anomalies. ARIMA is particularly effective when the data shows clear trends or periodic patterns.

Seasonal Decomposition of Time Series (STL) separates a time series into seasonal, trend, and residual components. Anomalies are typically identified in the residual component. This method helps isolate irregularities that cannot be explained by seasonality or trend.

Change-point detection techniques aim to identify points in time where the statistical properties of the series change [24]. These can include abrupt shifts in mean, variance, or autocorrelation structure. Algorithms like Bayesian Change Point Detection, Pruned Exact Linear Time (PELT), and Binary Segmentation are used in applications such as financial market analysis and equipment failure prediction [25].

More recent developments include the Matrix Profile method, which enables fast, scalable similarity search and anomaly detection in large time-series datasets [26]. By identifying the least similar sub sequences, Matrix Profile effectively highlights unusual patterns without requiring a model to be trained.

Time-series specific techniques are essential in fields like energy consumption forecasting, stock market analysis, industrial equipment monitoring, and healthcare analytics, where anomalies often follow complex temporal dynamics.

3) Signal-based anomaly detection techniques are widely used in domains where data is collected in the form of continuous waveforms, vibrations, or frequency-based signals. These techniques focus on identifying irregularities in the shape, amplitude, or frequency components of the signal data, which may indicate faults, noise, or other abnormal conditions. Applications include structural health monitoring, acoustic emission testing, vibration analysis in machinery, and biomedical signal processing such as ECG and EEG monitoring.

One common method is spectral analysis, where the frequency content of a signal is examined using tools like the Fourier Transform [27]. Anomalies may manifest as sudden spikes or drops in power at certain frequencies. In rotating machinery, for instance, deviations from baseline frequency signatures often suggest wear, imbalance, or bearing failure.

Another key method is wavelet transform analysis, which provides both frequency and time localization of signal features [28]. This is particularly useful for identifying transient anomalies, such as sudden impacts or short-duration faults. Wavelets can decompose signals at multiple scales, making them suitable for multi-resolution analysis.

Envelope analysis is another powerful technique used to detect impacts and modulated signals hidden within a complex waveform [29]. It's often used in gear and bearing diagnostics, especially where periodic impulses are masked by noise.

Cross-correlation and coherence techniques are used to compare signals across multiple channels or sensors to detect spatially coherent anomalies, such as propagating cracks in structures.

In AE-based corrosion monitoring, signal-based techniques can distinguish between noise, crack growth, and corrosion signals based on parameters like rise time, amplitude, duration, and energy [30]. These methods are critical in environments where traditional techniques may not detect internal flaws until catastrophic failure.

Signal-based methods are valued for their sensitivity and ability to capture minute signal changes that precede larger failures. However, they require high-fidelity data acquisition systems and advanced filtering to isolate meaningful features.

4) Machine learning techniques form a core component of modern anomaly detection systems. These techniques leverage data-driven learning models to discover complex and nonlinear relationships between features that can indicate abnormal patterns. Depending on the availability of labelled data and the structure of the problem, machine learning techniques are broadly categorized into four main types: supervised, unsupervised, semi-supervised, and self-supervised learning. Each of these categories provides different advantages and trade-offs, and their usage is dictated by the nature of the data and the application domain.

E. Machine Learning Anomaly Detection Techniques

1) Supervised learning involves training models on datasets that include labelled examples of both normal and anomalous data. These models learn to distinguish between the two classes based on the features present in the training data. Algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forests, Logistic Regression, and Neural Networks are widely used in this category.

This approach can be highly effective when ample labelled data is available. It allows for the creation of highly discriminative models that can achieve strong generalization on unseen data [31]. However, its major limitation lies in the difficulty of acquiring sufficient labelled examples of anomalies, which are typically rare and diverse.

Supervised anomaly detection has been successfully applied in spam detection, medical diagnostics, and quality assurance systems where historical labelled datasets are available. Techniques such as cross-validation and synthetic data generation (e.g., SMOTE) are often used to handle class imbalance [32].

2) Unsupervised learning techniques are particularly useful when labelled data is unavailable, which is a common scenario in anomaly detection where anomalies are rare and difficult to label [1]. These models attempt to learn the inherent structure of the dataset without predefined labels, assuming that normal patterns dominate the data distribution. Anomalies are identified as instances that do not conform to these learned patterns.

Common approaches include clustering algorithms like k-means and DBSCAN, where outliers are data points that do not belong to any cluster or are far from cluster centroids. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Isolation Forest also fall under this category [14]. PCA reduces the data into a lower-dimensional space where anomalies can be observed as deviations from the principal components. One-Class SVM is another widely used method that attempts to learn the boundary of normal data and flags deviations as anomalies.

Unsupervised techniques are widely used in cybersecurity for detecting unauthorized access, in finance for fraud detection, and in manufacturing for identifying process faults. They require minimal prior knowledge about the data and are effective for exploratory analysis.

3) Semi-supervised learning sits between supervised and unsupervised methods. It utilizes a large amount of unlabelled data and a small amount of labelled data—usually only normal instances [2]. The model is trained to capture the distribution of normal data and detect deviations as potential anomalies.

A prominent approach in this category includes One-Class SVM, which learns a decision function that identifies the region in the feature space where the majority of the data lies. Autoencoders, a type of neural network, are also commonly used; they are trained to reconstruct normal data with high fidelity [33]. At inference time, high reconstruction errors suggest that the input does not conform to the learned normal pattern, signalling an anomaly.

Semi-supervised methods are well-suited to industrial applications, such as predictive maintenance, where anomaly data is limited. They are also used in biomedical monitoring, where only healthy data may be available for training.

4) Self-supervised learning (SSL) is an emerging paradigm in anomaly detection that leverages large volumes of unlabelled data by creating pseudo-labels through pretext tasks [34]. These pretext tasks are designed to train models to learn general data representations, which can then be used for

downstream anomaly detection without requiring manually labelled anomalies.

In the context of anomaly detection, SSL often begins by defining surrogate tasks such as predicting masked input features (e.g., in tabular data), reconstructing corrupted signal segments (e.g., in time series), or predicting the temporal order of sequence segments [35]. Models trained on these tasks develop an internal representation of normal data distribution. During inference, deviations from expected patterns are flagged as anomalies.

In time-series anomaly detection, a common SSL strategy involves forecasting future values based on historical data using models like Transformers or Temporal Convolutional Networks. Large prediction errors are used as indicators of anomalous behaviour [36]. Similarly, contrastive learning—a subfield of SSL—maximizes agreement between different views or augmentations of the same data point while minimizing agreement with others, making it powerful for identifying rare deviations.

In industrial and biomedical applications, self-supervised pretraining has shown to improve anomaly detection performance [37]. For example, in acoustic emission-based corrosion monitoring, self-supervised pretraining allows the model to learn signal patterns unique to material degradation without needing explicitly labelled defect types.

SSL not only reduces the dependency on labelled data but also enhances robustness and generalizability across unseen data distributions. It has proven effective in fraud detection, predictive maintenance, healthcare monitoring, and cybersecurity, where anomalies evolve or emerge dynamically over time.

5) Deep learning techniques have revolutionized the field of anomaly detection by enabling systems to automatically extract hierarchical representations from raw, high-dimensional data. These techniques, primarily based on artificial neural networks, are capable of modelling complex and nonlinear data distributions, making them exceptionally powerful for detecting subtle and intricate anomalies.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are among the most commonly used deep learning models for anomaly detection. CNNs are particularly effective in applications involving image or spatial data, where anomalies may appear as localized irregularities. For instance, in industrial quality inspection, CNNs can identify manufacturing defects from high-resolution images [38]. RNNs, on the other hand, are designed to process sequential data and are extensively used in time-series anomaly detection, such as sensor data monitoring, by capturing temporal dependencies and flagging unexpected deviations.

Autoencoders are another widely used deep learning technique. These unsupervised neural networks are trained to reconstruct their input data. During testing, if the reconstruction error is high for certain inputs, those inputs are flagged as potential anomalies. Variants such as Variational Autoencoders (VAEs) and Denoising Autoencoders add probabilistic and noise-resilience components to improve detection robustness [39].

Generative Adversarial Networks (GANs) have also gained popularity for anomaly detection [40]. These models consist of a generator that creates fake data and a discriminator that tries to distinguish between real and generated data. The discriminator's inability to classify an instance properly can indicate that the data does not follow the expected distribution, thus signalling an anomaly. GAN-based methods have been applied in fields such as cybersecurity, medical imaging, and predictive maintenance.

Long Short-Term Memory (LSTM) networks, a type of RNN, are particularly effective for learning long-range dependencies in time-series data [41]. They have been employed in health monitoring systems, power grid analysis, and transportation networks to predict normal behaviour and detect anomalies through forecast errors.

Deep learning models benefit from large-scale datasets and computational resources, enabling them to outperform traditional methods in many scenarios [42]. However, they also present challenges such as model interpretability, high computational cost, and the need for careful hyperparameter tuning.

6) Transformers represent a significant advancement in machine learning models for anomaly detection due to their superior ability to model long-range dependencies and capture complex temporal patterns in data. Originally designed for natural language processing, Transformers have since been successfully adapted to various domains including vision, time-series analysis, and industrial monitoring [43]. Unlike traditional recurrent architectures like RNNs or LSTMs, Transformers rely on self-attention mechanisms that allow them to process all elements of a sequence simultaneously, leading to better scalability and improved performance, especially on large and noisy datasets.

One of the key advantages of Transformers in anomaly detection lies in their flexibility and robustness. They can be applied in both supervised and self-supervised settings and are capable of learning high-quality representations from unlabelled data. In comparison to CNNs and RNNs, Transformers offer superior performance in handling multivariate time-series data, where correlations across different features and timestamps must be considered jointly [44]. Moreover, self-attention helps in identifying subtle contextual anomalies that may span long sequences, which traditional models often miss.

This makes Transformers particularly valuable in industrial and infrastructure monitoring applications, such as in predictive maintenance and fault detection. One emerging application is anomaly detection in corrosion monitoring, especially using Acoustic Emission (AE) data, where annotated anomalies are difficult to obtain and the signals are often non-stationary and high-dimensional [45]. The ability of Transformers to learn from limited or unlabelled data and their capacity to model complex temporal dynamics makes them an ideal choice for such scenarios.

Furthermore, recent research has demonstrated that Transformer-based architectures can outperform conventional techniques in both accuracy and generalization when applied to corrosion-related anomaly detection tasks [46]. Their compatibility with self-supervised learning strategies enhances their effectiveness in low-resource settings, making them particularly relevant for pipeline health monitoring systems and structural integrity assessments.

The discussion of Transformer applications in corrosion detection, is expanded further in the next section, where we delve into their specific contributions, datasets, and models.

III. Prior Studies of Transformer Based Models for Corrosion Anomaly Detection

A. Comparative Analysis of Transformer Based Models for Corrosion Detection

The following paragraphs go into a deeper dive into each of the studies mentioned in Table I.

TABLE I. Comparative Study

Study	Transformer Model	Input Data	Learning Setup	Dataset	Accuracy/F1
Efficient Metal Corrosion Area Detection Model Combining Convolution and Transformer	Visual Transformer integrated in MCD-Net	Corrosion images (public dataset A and self-collected Dataset B)	Supervised Learning	Dataset A (AASHTO/BIRM corrosion levels), Dataset B (natural scene images)	F1 Score: 84.53%
CorFormer: a hybrid transformer-CNN architecture for corrosion segmentation on metallic surfaces	CorFormer (hybrid Transformer-CNN)	Corrosion images on metallic surfaces	Supervised Learning	Not stated (mentions 10 validation splits)	IoU improved by 2.7% over SOTA; FPS: 28
Civil Infrastructure Damage and Corrosion Detection: An Application of Machine Learning	Transformer block within CycleGAN (residual block-based Transformer)	UAV aerial images of bridges and steel infrastructure	Supervised Learning	1300 images (Bolte Bridge & Sky Rail, Victoria, AU)	F1 Score: 0.83; Accuracy: 0.989
A Multilevel Bridge Corrosion Detection Method by Transformer-Based Segmentation in a Stitched View	Transformer-based segmentation (e.g., SegFormer)	Stitched images of bridge surfaces	Supervised	Public dataset from Virginia Tech (440 images), augmented with Lundamo Bridge dataset (94 images)	F1 Score: 68.2% (improved from 60–61% with CNNs)
P-DETR: A Transformer-Based Algorithm for Pipeline Structure Detection	Pipe Detection Transformer (P-DETR, DETR-based with FNT module)	Aerial drone images of pipelines	Supervised	Custom dataset (2553 raw + augmented to 6127 images)	mAP: 55% (↑3 AP over DETR); Recall: 73.9%
Corrosion segmentation method of concrete drainage pipes based on point transformer	Point Transformer (U-Net-like)	3D point clouds from real + simulated concrete pipe data	Supervised Learning	200 real + 200 simulated 3D point cloud samples (Azure Kinect DK)	Accuracy: 94.36%; MIoU: 86.31%
SegFormer: Semantic Segmentation Based Transformers For Corrosion Detection	SegFormer	Bridge corrosion images from inspection reports (VDOT)	Supervised Learning	440 annotated images (AASHTO/BIRM guidelines)	Mean Acc: 81.39%; MIoU: 71.16%
Double-Attention YOLO: Vision Transformer Model for Transmission Line Fittings & Rust	Vision Transformer + Dual Attention YOLO	Hazy/defogged images of transmission line connection fittings	Supervised Learning	Not explicitly named; 7700 annotated images (Stage 2)	mAP@0.5:0.95 = 0.8674mAP@0.5 = 0.9948Macro-F1 = 0.661 (stage 1), 0.568 (stage 2)
Corrosion SAM: Adapting Segment Anything Model with Parameter-Efficient Fine-Tuning for Structural Corrosion Inspection	Segment Anything Model (SAM) + Vision Transformer (ViT-H) with Adapter modules (PEFT)	RGB corrosion images resized to 512×512; binarized	Supervised Learning	VDOT (440 images)KRDB (955 images)	VDOT: IoU 69.72%, Dice 81.04%KRDB: IoU 60.78%, Dice 72.97%
Pipeline defect recognition algorithm based on CNN-Transformer model for internal detection in PE gas pipelines	CNN + Transformer with Squeeze-and-Excitation (no positional encoding)	Internal pipeline defect images (1200), 1080×1080 px, collected via inspection robot	Supervised Learning	Custom dataset (1200 images) of PE pipeline defects	Accuracy: 97.8% (avg. over 10 runs)

The study titled “Efficient Metal Corrosion Area Detection Model Combining Convolution and Transformer” proposes MCD-Net, a hybrid deep learning architecture that integrates convolutional layers with a visual Transformer encoder to enhance corrosion detection on metallic surfaces [46]. By combining local feature extraction with global context modelling, the model effectively addresses challenges such as occlusions, lighting variability, and irregular corrosion patterns. It employs an attention-based multi-layer feature fusion mechanism and a multi-scale feature enhancement strategy to refine boundary segmentation. MCD-Net achieved an F1 score of 84.53% on a public dataset, demonstrating improved robustness to noise and lighting compared to conventional CNNs.

In "CorFormer: A Hybrid Transformer-CNN Architecture for Corrosion Segmentation on Metallic Surfaces", the authors present CorFormer, a real-time segmentation model that combines CNNs and Transformers within a unified framework [47]. Transformer layers are embedded within the CNN encoder, while a Semantic Gap Merger (SGM) bridges feature-level disparities. A hierarchical decoder further processes multi-scale features, enabling the model to capture both localized and broad defects. CorFormer improved Intersection-over-Union (IoU) by 2.7% over baseline methods and achieved real-time performance at 28 FPS, making it well-suited for industrial monitoring applications.

The study "Civil Infrastructure Damage and Corrosion Detection: An Application of Machine Learning" introduces a hybrid CycleGAN-based model incorporating Transformer-enhanced residual blocks for corrosion segmentation in UAV-acquired bridge imagery [48]. Positioned between the encoder and decoder of the generator, the Transformer blocks enable domain-adaptive learning, improving the model's robustness to texture and lighting variations. Trained on 1,300 annotated images from Melbourne's Bolte Bridge and Sky Rail, the model achieved an F1 score of 0.83 and an overall accuracy of 98.9%, outperforming traditional networks like PSPNet and SegNet, and demonstrating strong generalization capabilities across diverse aerial scenes.

The study "A Multilevel Bridge Corrosion Detection Method by Transformer-Based Segmentation in a Stitched View" addresses the limitations of CNNs when applied to wide, high-resolution structural images [49]. Utilizing a Transformer-based semantic segmentation model (e.g., SegFormer), the study incorporates stitched bridge surface imagery to enable large-area defect detection. Trained on 440 annotated images from Virginia Tech and additional data from Norway's Lundamo Bridge, the model achieved an F1 score of 68.2%, significantly outperforming U-Net and DeepLabV3+. Although performance was impacted by input resolution, the approach showed strong potential for deployment in field inspection workflows.

In "P-DETR: A Transformer-Based Algorithm for Pipeline Structure Detection", the authors propose an enhanced DETR architecture tailored for small-object detection in aerial pipeline imagery [50]. The model integrates a Feature Normalization and Transformation (FNT) module to improve spatial resolution and feature fusion. Trained on an augmented dataset of 6,127 drone-captured images, P-DETR achieved a mean Average Precision (mAP) of 55% and a recall of 73.9%, outperforming DETR, YOLOv3, and SSD baselines. While highly accurate in static image analysis, real-time drone-based deployment is identified as a future direction for further validation.

The study "Corrosion Segmentation Method of Concrete Drainage Pipes Based on Point Transformer" introduces a 3D point cloud segmentation method using a Point Transformer to identify corrosion in concrete drainage pipes [51]. A U-Net-like architecture is employed, trained on a combined dataset of real and simulated point clouds generated using an Azure Kinect DK. The model achieved 94.36% accuracy and a mean IoU of 86.31%, surpassing PointNet++ and PAConv. The best results were obtained with a 1:1 real-to-simulated data ratio. Despite its computational intensity, the model excels in precise 3D segmentation tasks for corrosion detection.

"SegFormer: Semantic Segmentation Based Transformers for Corrosion Detection" explores the use of SegFormer for corrosion classification based on annotated bridge inspection images following AASHTO and BIRM guidelines [52]. The dataset includes four corrosion severity levels, with class imbalance noted for the "severe" category. After fine-tuning, the model achieved a mean accuracy of 81.39% and a mean IoU of 71.16%. The study emphasizes the need for balanced datasets and tailored preprocessing to improve segmentation performance in underrepresented classes.

The paper "Double-Attention YOLO: Vision Transformer Model Based on Image Processing Technology in Complex Environment of Transmission Line Connection Fittings and Rust Detection" presents a two-stage framework for corrosion detection under visual degradation [53]. The approach integrates defogging preprocessing with a YOLO-based detector that combines Vision Transformers, a dual attention mechanism (CBAM), and GhostNet for efficient inference. Trained on 7,700 annotated images, the model achieved a mAP@0.5 of 99.48% and macro-F1 scores of 0.661 and 0.568. It outperformed multiple baselines—including ATSS, FCOS, and DETR variants—particularly in scenes with small objects, haze, and occlusion.

"Corrosion SAM: Adapting Segment Anything Model with Parameter-Efficient Fine-Tuning for Structural Corrosion Inspection" applies the SAM foundation model with ViT-H backbone to the corrosion domain using adapter-based Parameter-Efficient Fine-Tuning (PEFT) [54]. By inserting adapters into frozen Transformer layers, the model adapts with minimal computational overhead. Evaluated on the VDOT and KRDB datasets, Corrosion SAM achieved IoU scores of 69.72% and 60.78%, and Dice scores of 81.04% and 72.97%, respectively. It consistently outperformed DeepLabV3+, U-Net, and SegFormer, demonstrating the viability of efficient fine-tuning techniques in adapting foundation models to specialized inspection tasks.

Finally, "Pipeline Defect Recognition Using a CNN–Transformer Model" presents a lightweight architecture combining CNNs with Transformers and a Squeeze-and-Excitation module to classify defects in polyethylene (PE) gas pipelines [55]. The model omits positional encoding—leveraging zero-padding from CNNs—and uses Layer Normalization and multi-head attention for stability. Trained on 1,200 images captured by pipeline inspection robots, the model achieved an average accuracy of 97.8% over 10 randomized test runs. It consistently outperformed standard CNN and ViT architectures on small and medium-sized datasets, underscoring its potential for real-world pipeline monitoring.

As can be seen from these studies, the use of transformer-based models improved the accuracy of detecting anomalies caused by corrosion. Therefore, in this study, a transformer-based model was used on the experimental data set along with other models for a comparative analysis of the performances of each model.

IV. METHODOLOGY

A. Experimental Setup

The dataset used in this study was generated through a controlled laboratory experiment designed to simulate real-world corrosion in fluid transport pipelines. As shown in Fig. 1, ten acoustic emission (AE) sensors were strategically distributed along the pipeline—five on the left side (with girth welds between the source and sensors) and five on the right (without girth welds). Accelerated corrosion was induced at the centre of the pipeline over a three-hour period, with sensors spaced at 100 cm intervals to capture both near-field and far-field AE signal variations.



Fig. 1. Setup for the data acquisition.

This experimental dataset was chosen over public datasets to ensure high-fidelity, labelled AE signals under known and controlled corrosion conditions. The setup allowed us to evaluate the performance of machine learning models in detecting corrosion-induced anomalies while accounting for attenuation effects due to girth welds. Such a controlled environment also provided reliable ground truth for model validation, making the dataset particularly suitable for the supervised, unsupervised, and hybrid models evaluated in this study.

B. Data Analysis

After filtering and analysing the data, strong correlations was found between the parameters ABS-ENERGY and Signal Strength, along with distance (channel 3 being the closest and channel 8 being the furthest). The scatterplot in Fig. 2 shows the relationship. These relationships will be used to train the machine learning models that would detect the corrosion.

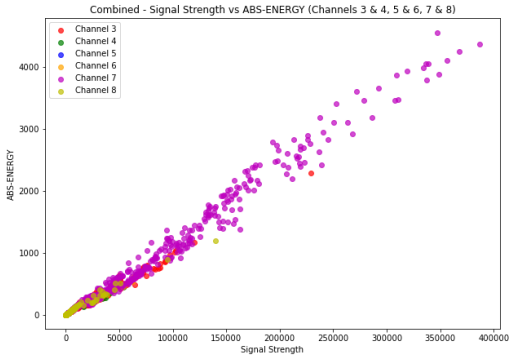


Fig. 2. Relationship between ABS-ENERGY and RMS.

V. RESULTS AND DISCUSSION

A. Machine Learning Models and Transformer Based Models Applied

In the following paragraphs, each ML model used, including the Hybrid CNN-Transformer, is described in detail along with its performance on the acquired data from the experiment. The bar graph in Fig. 3 provide a visual representation of the performances of each ML model.

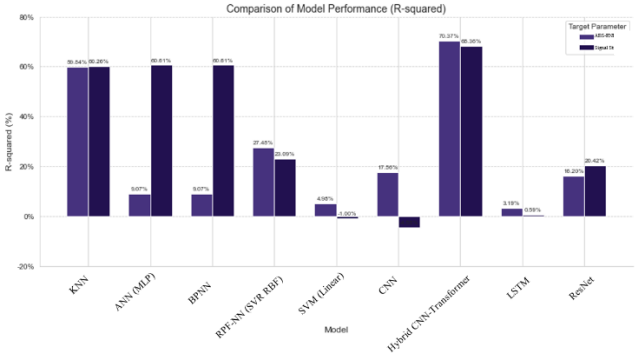


Fig. 3. Comparison of ML models.

1) K-Nearest Neighbours (KNN) is a non-parametric, instance-based learning algorithm that predicts the value of a target variable by averaging the values of its k-nearest neighbors in the feature space. For both 'ASL' and 'RMS', KNN demonstrated relatively strong performance, achieving R-squared values of 0.5984 and 0.6026, respectively. This suggests that the local patterns in the 'Distance' and 'CH' feature space are somewhat indicative of the target parameters. The moderate to high R-squared scores imply that a significant portion of the variance in both 'ASL' and 'RMS' can be explained by the values of their nearest neighbours. However, the performance is not perfect, indicating that other factors or more complex relationships might be at play.

2) The Artificial Neural Network (ANN) and Backpropagation Neural Network BPNN, both implemented as a Multi-Layer Perceptron with similar architectures (two hidden layers with 64 and 32 neurons), exhibited substantially lower performance for 'ASL' (R-squared of 0.0907) compared to KNN. For 'RMS', the performance was significantly better, reaching an R-squared of 0.6061, comparable to KNN. The discrepancy in performance for 'ASL' suggests that the non-linear relationships captured by the MLP were not as effective as the local averaging performed by KNN for this specific parameter. Conversely, for 'RMS', the MLP was able to model the underlying relationships as effectively as KNN. The fact that both ANN and BPNN yielded identical results is expected given their identical implementation and random initialization (controlled by the random state). This highlights the sensitivity of neural network performance to the specific target variable and the potential need for more extensive hyperparameter tuning.

3) The Support Vector Regression with Radial Basis Function kernel (RBF-NN), implemented using SVR with an RBF kernel, showed moderate predictive power for both 'ASL' (R-squared of 0.2748) and 'RMS' (R-squared of 0.2309). The RBF kernel allows the model to capture non-linear relationships through the use of radial basis functions. The performance was lower than KNN and the better performing neural network for 'RMS', indicating that the specific non-linear mapping learned by the RBF kernel was not as well-

suited to the underlying data distribution for these parameters compared to the other effective models.

4) The Support Vector Regression (SVM) with a linear kernel demonstrated the poorest performance among the dedicated regression models for 'ASL', achieving an R-squared of only 0.0498. For 'RMS', the R-squared was even slightly negative (-0.0100), indicating that the model performed worse than simply predicting the mean of the 'RMS' values. This suggests that the relationship between the features ('Distance' and 'CH') and the target parameters is likely non-linear, and a linear model is insufficient to capture the underlying patterns. The near-zero or negative R-squared values highlight the importance of selecting a model that can accommodate the complexity of the data relationships.

5) The Convolutional Neural Network (CNN) adapted for this task by treating each data point as a single timestep, also exhibited low performance, with an R-squared of 0.1756 for 'ASL' and a negative R-squared of -0.0455 for 'RMS'. CNNs are typically designed for sequential or spatial data, and their poor performance here likely stems from the lack of a meaningful sequential structure in the provided features in relation to the target variables. The convolutional layers, designed to detect local patterns across a sequence, may not have been effectively utilized in this context.

6) The Hybrid CNN-Transformer Model yielded the highest performance for both 'ASL' (R-squared of 0.7037) and 'RMS' (R-squared of 0.6836) among all the models evaluated. This model leverages the strengths of Convolutional Neural Networks (CNNs) for local feature extraction from the raw acoustic emission (AE) waveforms or their spectrograms, and Transformer encoders for capturing long-range temporal dependencies. The CNN layers are adept at identifying intricate, localized patterns and transient events within the AE signals, which are crucial for understanding the underlying physical processes. Subsequently, the Transformer's self-attention mechanism processes these extracted features, allowing the model to learn complex, non-linear relationships across extended segments of the AE data. The high R-squared scores suggest that the model effectively translates the rich information embedded in the AE signals (which implicitly capture 'Distance' and 'CH' effects, along with other nuanced characteristics) into highly accurate predictions for both 'ASL' and 'RMS'. The strong performance of this hybrid architecture indicates its superior ability to learn hierarchical and contextual representations from the sequential AE data, bridging the gap between local signal characteristics and their global impact on the target parameters.

7) The Long Short-Term Memory network (LSTM), a type of recurrent neural network designed for sequential data, showed very poor performance, with an R-squared of 0.0319 for 'ASL' and 0.0059 for 'RMS'. Similar to the CNN, the poor performance is likely due to the lack of a clear temporal dependency in the provided features. Treating the data points as a sequence of single timesteps did not allow the LSTM's recurrent architecture to effectively learn predictive patterns for these parameters.

8) The Deep Residual Network (ResNet) implemented also showed relatively low performance, with an R-squared of 0.1620 for 'ASL' and 0.2042 for 'RMS'. While ResNets are powerful for learning deep representations, their architecture, adapted here for non-sequential data, did not outperform simpler models like KNN or the Decision Tree. The skip connections, designed to ease the training of deep networks, might not have provided a significant advantage in this context with the given data structure.

VI. Limitations and Future Work

While the study demonstrates the potential of machine learning and Transformer-based models for corrosion detection using acoustic emission data, several limitations should be acknowledged. One notable limitation is that the effect of temperature on the pipeline and the acoustic signal propagation was not considered during the experiment. In real-world scenarios, pipelines operate under varying thermal conditions, which can alter AE signal characteristics such as amplitude, attenuation, and wave speed. Ignoring temperature effects may limit the generalizability of the trained models to operational environments with fluctuating temperatures.

Future work should incorporate temperature as a variable in both data collection and model training to improve robustness. This could involve using temperature-compensated sensors, collecting data under controlled thermal variations, or integrating temperature data as an additional input feature for model training. Additionally, further research could explore the deployment of the proposed hybrid CNN-Transformer model in real-world pipelines with dynamic environmental conditions and test its performance across different corrosion types and pipeline materials.

VII. Conclusion

The study conclusively demonstrates the significant advantage of employing a Hybrid CNN-Transformer Model for accurately predicting Acoustic Signal Level (ASL) and Root Mean Square (RMS) in acoustic emission-based corrosion detection. Achieving R-squared values of 0.7037 for ASL and 0.6836 for RMS, this transformer-based architecture outperformed all other evaluated machine learning models, including traditional methods like KNN, SVM, and even other deep learning models like ANNs, CNNs, LSTMs, and ResNets.

This success can be attributed to the inherent design of the Hybrid CNN-Transformer. It effectively harnesses the Convolutional Neural Network's ability to extract nuanced, local features and transient patterns from raw AE waveforms or their spectrograms, which are critical for characterizing the distinct acoustic signatures of corrosion. Subsequently, the Transformer encoder's self-attention mechanism excels at learning complex, long-range temporal dependencies and contextual relationships across these extracted features. This synergistic combination allows the model to process the high-dimensional, sequential AE data comprehensively, bridging the gap between subtle local signal changes and their broader

implications for corrosion-related parameters. The high R-squared values unequivocally indicate that this advanced, transformer-based approach can effectively decipher the intricate link between AE signals and the target parameters, representing a significant step forward in robust and accurate corrosion detection through acoustic emission monitoring.

ACKNOWLEDGMENT

This research is ongoing research supported by YUTP Research Grant Scheme.

REFERENCES

[1] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, "Anomaly detection," Computers, Materials and Continua, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.

[2] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," Jun. 2014. doi: 10.1016/j.sigpro.2013.12.026.

[3] D. Kim and T. Y. Heo, "Anomaly Detection with Feature Extraction Based on Machine Learning Using Hydraulic System IoT Sensor Data," Sensors, vol. 22, no. 7, Apr. 2022, doi: 10.3390/s22072479.

[4] R. W. Anwar, M. Abrar, A. Salam, and F. Ullah, "Federated learning with LSTM for intrusion detection in IoT-based wireless sensor networks: a multi-dataset analysis," PeerJ Comput Sci, vol. 11, pp. 1–30, 2025, doi: 10.7717/peerj-cs.2751.

[5] I. Priyadarshini, "Anomaly Detection of IoT Cyberattacks in Smart Cities Using Federated Learning and Split Learning," Big Data and Cognitive Computing, vol. 8, no. 3, Mar. 2024, doi: 10.3390/bdcc8030021.

[6] J. C. Quiroz, N. Mariun, M. R. Mehrjou, M. Izadi, N. Misron, and M. A. Mohd Radzi, "Fault detection of broken rotor bar in LS-PMSM using random forests," Measurement (Lond), vol. 116, pp. 273–280, Feb. 2018, doi: 10.1016/j.measurement.2017.11.004.

[7] F. Ferraz Júnior, R. A. F. Romero, and S. J. Hsieh, "Machine Learning for the Detection and Diagnosis of Anomalies in Applications Driven by Electric Motors," Sensors, vol. 23, no. 24, Dec. 2023, doi: 10.3390/s23249725.

[8] J. Garcia, L. Rios-Colque, A. Peña, and L. Rojas, "Condition Monitoring and Predictive Maintenance in Industrial Equipment: An NLP-Assisted Review of Signal Processing, Hybrid Models, and Implementation Challenges," May 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/app15105465.

[9] R. Abdalla, H. Samara, N. Perozo, C. P. Carvajal, and P. Jaeger, "Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs)," ACS Omega, vol. 7, no. 21, pp. 17641–17651, May 2022, doi: 10.1021/acsomega.1c05881.

[10] L. Yu, E. Yang, C. Luo, and P. Ren, "AMCD: an accurate deep learning-based metallic corrosion detector for MAV-based real-time visual inspection," J Ambient Intell Humaniz Comput, vol. 14, no. 7, pp. 8087–8098, Jul. 2023, doi: 10.1007/S12652-021-03580-4/FIGURES/12.

[11] D. J. Atha and M. R. Jahanshahi, "Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection," Struct Health Monit, vol. 17, no. 5, pp. 1110–1128, Sep. 2018, doi: 10.1177/1475921717737051;CTYPE:STRING:JOURNAL.

[12] W. Wu, T. Tao, J. Xiao, Y. Yao, and J. Yang, "Unsupervised Anomaly Detection on Metal Surfaces Based on Frequency Domain Information Fusion," Sensors 2025, Vol. 25, Page 2250, vol. 25, no. 7, p. 2250, Apr. 2025, doi: 10.3390/S25072250.

[13] T. Okura, N. Kasai, H. Minowa, and S. Okazaki, "Application of Machine Learning for Data with an Atmospheric Corrosion Monitoring Sensor Based on Strain Measurements," Metals 2022, Vol. 12, Page 1179, vol. 12, no. 7, p. 1179, Jul. 2022, doi: 10.3390/MET12071179.

[14] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104, 2000, doi: 10.1145/342009.335388.

[15] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation-based anomaly detection," ACM Trans Knowl Discov Data, vol. 6, no. 1, Mar. 2012, doi: 10.1145/2133360.2133363.

[16] E. Calikus, S. Nowaczyk, and O. Dikmen, "Context discovery for anomaly detection," Int J Data Sci Anal, vol. 19, no. 1, pp. 99–113, Jan. 2024, doi: 10.1007/S41060-024-00586-X/FIGURES/5.

[17] M. A. Hayes and M. A. Capretz, "Contextual anomaly detection framework for big sensor data," J Big Data, vol. 2, no. 1, pp. 1–22, Dec. 2015, doi: 10.1186/S40537-014-0011-Y/TABLES/9.

[18] L. Bontemps, V. L. Cao, J. McDermott, and N. A. Le-Khac, "Collective Anomaly Detection based on Long Short Term Memory Recurrent Neural Network," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10018 LNCS, pp. 141–152, Mar. 2017, doi: 10.1007/978-3-319-48057-2_9.

[19] D. M. Hawkins, "Identification of Outliers," Identification of Outliers, 1980, doi: 10.1007/978-94-015-3994-4/COVER.

[20] D. C. Montgomery, "Introduction to Statistical Quality Control, 8th Edition | Wiley," Wiley Online Library, p. 786, 2019, Accessed: Jun. 24, 2025. [Online]. Available: https://www.wiley.com/en-us/Introduction+to+Statistical+Quality+Control%2C+8th+Edition-p-9781119399308

[21] F. E. Grubbs, "AD-781 499 PROCEDURES FOR DETECTING OUTLYING OBSERVATIONS IN SAMPLES," 1974.

[22] F. Gustafsson, "Adaptive Filtering and Change Detection," 2000, Accessed: Jun. 24, 2025. [Online]. Available: http://www.wiley.co.uk

[23] "Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting Principles and Practice. OTexts. - References - Scientific Research Publishing." Accessed: Jun. 24, 2025. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=2849375

[24] M. Lavielle, "Using penalized contrasts for the change-point problem," Signal Processing, vol. 85, no. 8, pp. 1501–1510, Aug. 2005, doi: 10.1016/J.SIGPRO.2005.01.012.

[25] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," J Am Stat Assoc, vol. 107, no. 500, pp. 1590–1598, 2012, doi: 10.1080/01621459.2012.737745.

[26] C.-C. M. Yeh et al., "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," pp. 1317–1322, Feb. 2017, doi: 10.1109/ICDM.2016.0179.

[27] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," Mech Syst Signal Process, vol. 20, no. 2, pp. 282–307, Feb. 2006, doi: 10.1016/J.YMSSP.2004.09.001.

[28] "A Wavelet Tour of Signal Processing | ScienceDirect." Accessed: Jun. 24, 2025. [Online]. Available: https://www.sciencedirect.com/book/9780123743701/a-wavelet-tour-of-signal-processing

[29] R. B. Randall, "Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications," Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications, pp. 1–289, Dec. 2010, doi: 10.1002/9780470977668.

[30] C. U. Grosse and M. Ohtsu, "Acoustic emission testing: Basics for Research-Applications in Civil Engineering," Acoustic Emission Testing: Basics for Research-Applications in Civil Engineering, pp. 1–404, 2008, doi: 10.1007/978-3-540-69972-9/COVER.

[31] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," Journal of Network and Computer Applications, vol. 60, pp. 19–31, Jan. 2016, doi: 10.1016/J.JNCA.2015.11.016.

[32] J. Ma and S. Perkins, "Time-series Novelty Detection Using One-class Support Vector Machines," Proceedings of the International Joint Conference on Neural Networks, vol. 3, pp. 1741–1745, 2003, doi: 10.1109/IJCNN.2003.1223670.

[33] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," Pattern Recognit, vol. 58, pp. 121–134, Oct. 2016, doi: 10.1016/J.PATCOG.2016.03.028.

[34] L. Jing and Y. Tian, "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 11, pp. 4037–4058, Feb. 2019, doi: 10.1109/TPAMI.2020.2992393.

[35] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. Part F129685, pp. 665–674, Aug. 2017, doi: 10.1145/3097983.3098052/SUPPL_FILE/ZHOU_ANOMALY_DETECTION.MP4.

[36] J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi, "Unsupervised Scalable Representation Learning for Multivariate Time Series," Adv Neural Inf Process Syst, vol. 32, 2019.

[37] C. Ding and C. Wu, "Self-Supervised Learning for Biomedical Signal Processing: A Systematic Review on ECG and PPG Signals," medRxiv, p. 2024.09.30.24314588, Oct. 2024, doi: 10.1101/2024.09.30.24314588.

[38] "Long Short Term Memory Networks for Anomaly Detection in Time Series." Accessed: Jun. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/304782562_Long_Short_Term_Memory_Networks_for_Anomaly_Detection_in_Time_Series

[39] J. An and S. Cho, "SNU Data Mining Center 2015-2 Special Lecture on IE Variational Autoencoder based Anomaly Detection using Reconstruction Probability," 2015.

[40] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10265 LNCS, pp. 146–147, Mar. 2017, doi: 10.1007/978-3-319-59050-9_12.

[41] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 387–395, Jul. 2018, doi: 10.1145/3219819.3219845/SUPPL_FILE/HUNDMAN_DETECTING_SPACECRAFT_ANOMALIES.MP4.

[42] A. Vaswani et al., "Attention Is All You Need," Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[43] S. Li et al., "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting," Adv Neural Inf Process Syst, vol. 32, Jun. 2019, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/1907.00235

[44] Z. Rong, D. Shandong, N. Xin, and X. Shiguang, "Feedforward Neural Network for Time Series Anomaly Detection," Dec. 2018, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/1812.08389

[45] "Sensors | Special Issue : Structural Health Monitoring Using Acoustic Emission Signals: The Role of Machine Learning." Accessed: Jun. 24, 2025. [Online]. Available: https://www.mdpi.com/journal/sensors/special_issues/QE6G26RX6Y

[46] J. Guo, L. Wang, and L. Hua, "Efficient Metal Corrosion Area Detection Model Combining Convolution and Transformer," Applied Sciences 2024, Vol. 14, Page 9900, vol. 14, no. 21, p. 9900, Oct. 2024, doi: 10.3390/APP14219900.

[47] A. Subedi, C. Qian, R. Sadeghian, and M. R. Jahanshahi, "CorFormer: a hybrid transformer-CNN architecture for corrosion segmentation on metallic surfaces," Mach Vis Appl, vol. 36, no. 2, pp. 1–21, Mar. 2025, doi: 10.1007/S00138-025-01663-2/METRICS.

[48] H. S. Munawar, F. Ullah, D. Shahzad, A. Heravi, S. Qayyum, and J. Akram, "Civil Infrastructure Damage and Corrosion Detection: An Application of Machine Learning," Buildings 2022, Vol. 12, Page 156, vol. 12, no. 2, p. 156, Feb. 2022, doi: 10.3390/BUILDINGS12020156.

[49] Z. Lu, T. Jiang, J. Slavič, · Gunnstein, and T. Frøseth, "A multilevel bridge corrosion detection method by transformer-based segmentation in a stitched view," J Civ Struct Health Monit, 123AD, doi: 10.1007/s13349-025-00928-8.

[50] I. A. Aromoye, L. H. Hiung, and P. Sebastian, "P-DETR: A transformer-based algorithm for pipeline structure detection," Results in Engineering, vol. 26, p. 104652, Jun. 2025, doi: 10.1016/J.RINENG.2025.104652.

[51] N. Wang et al., "Corrosion segmentation method of concrete drainage pipes based on point transformer," Measurement, vol. 234, p. 114893, Jul. 2024, doi: 10.1016/J.MEASUREMENT.2024.114893.

[52] A. Safa, A. Mohamed, B. Issam, and H. Mohamed-Yassine, "SegFormer: Semantic Segmentation Based Tranformers for Corrosion Detection," 6th International Conference on Networking and Advanced

[53] Z. Song, X. Huang, C. Ji, and Y. Zhang, "Double-Attention YOLO: Vision Transformer Model Based on Image Processing Technology in Complex Environment of Transmission Line Connection Fittings and Rust Detection," Machines 2022, Vol. 10, Page 1002, vol. 10, no. 11, p. 1002, Oct. 2022, doi: 10.3390/MACHINES10111002.

[54] C. Chai, Y. Gao, H. Li, and X. Zhu, "Corrosion SAM: Adapting Segment Anything Model with Parameter-Efficient Fine-Tuning for Structural Corrosion Inspection".

[55] X. Niu, X. Guo, H. Song, W. Hu, G. Jiang, and J. Guan, "Pipeline Defect Recognition Algorithm Based on CNN Transformer Model for Internal Detection in PE Gas Pipelines," Proceedings of 2024 Academic Conference of China Instrument and Control Society, ACCIS 2024, pp. 531–536, 2024, doi: 10.1109/ACCIS62068.2024.10948683.

Smartphone-Integrated Sensor-Based DFU Risk Assessment Using CatBoost and Deep Neuro-Fuzzy Intelligence

Dr. Jayashree J, Dr. Vijayashree J*, Dr. Perepi Rajarajeswari, Saravanan S
Department of Database Systems-School of Computer Science and Engineering,
Vellore Institute of Technology, VIT, Vellore, 632014, India

***Abstract*—Diabetic Foot Ulcer (DFU) is a serious and common complication of diabetes mellitus, which can lead to lower limb amputation if not identified and treated in its early stages. This study introduces an integrated and intelligent system designed for the early detection and severity classification of DFUs by combining sensor-driven data collection with machine learning techniques in a mobile application. The research is based on a dataset comprising both clinical features (D-1 to D-16) and key sensor-based readings gathered from 316 participants. After preprocessing and normalization, the clinical data undergoes feature selection using CatBoost, which filters out the five least impactful features while preserving all sensor data due to its diagnostic relevance. The refined dataset is then processed using a Deep Neuro-Fuzzy Network (DN-FN) to deliver real-time DFU severity predictions, categorized into Low, Mid, and High-risk levels. The solution is deployed through an intuitive smartphone interface, enabling users to input clinical data once and conduct periodic sensor-based tests—including vibration, pressure, and temperature readings. The mobile application interfaces with embedded hardware via Bluetooth and performs offline inference using a compact version of the trained model. The system is designed to offer both patients and healthcare professionals a practical and interpretable tool for continuous monitoring of foot health, with the ultimate goal of reducing the risk and impact of DFU complications.**

Keywords*—Bayesian Optimization; CatBoost; Deep Neuro-Fuzzy Networks (DN-FN); Diabetic Foot Ulcer (DFU) prediction; sensor-based risk stratification

I. Introduction

Diabetes mellitus continues to rise as a critical global health issue, with around 537 million adults affected as of 2021—a number expected to surge to 783 million by 2045 (International Diabetes Federation, 2021). [1] One of the most serious complications associated with diabetes is Diabetic Foot Ulcers (DFUs), which impact an estimated 15% to 25% of diabetic patients during their lifetime. [2] DFUs contribute to over 85% of diabetes-related lower-limb amputations, placing a heavy burden on healthcare infrastructure and severely affecting patients' quality of life. [3]

India, with over 77 million diabetic adults, holds the second-largest diabetic population globally [4]. The country faces considerable difficulties in detecting DFUs early, especially in semi-urban and rural regions. Challenges such as limited diagnostic resources, irregular medical follow-ups, and a scarcity of trained healthcare providers often lead to delayed intervention in foot-related complications.

Although diagnostic technologies have progressed, DFU evaluation largely remains manual and reactive. Standard assessment techniques depend heavily on clinical expertise and are often unavailable or impractical in remote settings [5]. This highlights the pressing need for an affordable, scalable, and intelligent solution capable of facilitating both early diagnosis and ongoing risk monitoring.

This work proposes a smart and accessible system that merges clinical information with sensor-derived data, enhanced by advanced machine learning techniques, and implemented through a smartphone-based application. The system enables early and proactive DFU risk classification by employing CatBoost for efficient feature selection and a DN-FN for accurate prediction. It emphasizes interpretability, adaptability, and reliable performance, including functionality in offline scenarios—addressing the practical limitations in current DFU screening approaches.

The remainder of this paper is organized as follows: Section II presents an in-depth literature review of existing DFU prediction methodologies. Section III outlines the dataset, hardware setup, feature selection using CatBoost, and the development of the DN-FN model. Section IV discusses the results and model performance based on standard evaluation metrics. Section V concludes with a summary of key contributions and outcomes.

II. Literature Review

Chatratia et al. [6] present a smart home health monitoring system integrating machine learning models for predicting type 2 diabetes and hypertension based on blood pressure and glucose readings, with Support Vector Machine (SVM) achieving the highest accuracy (75%) for diabetes prediction. This system facilitates real-time alerts to healthcare providers, promoting early detection and intervention. Similarly, Agrawal et al. [7] compared statistical methods and machine learning models for plantar pressure-based diabetic foot classification. Using wireless insoles, they found that the AdaBoost model achieved the highest accuracy (0.85), outperforming traditional methods, suggesting that ML models significantly enhance diabetic foot prediction and classification.

This work was supported and funded by the Indian Council of Medical Research (ICMR) and facilitated by Vellore Institute of Technology, Vellore.

Reddie et al. [8] introduced a low-cost, purely mechanical plantar pressure evaluation device utilizing bistable compliant mechanisms, allowing non-specialist healthcare workers to assess foot pressure, though sensitivity improvements are needed. Ghazi et al. [9] developed a smart shoe system integrating temperature, humidity, plantar pressure, and oxygen saturation sensors, with efficient real-time monitoring and responsive data transmission, though future work is needed to improve sensor accuracy. Kularathne et al. [10] presented Dia-Shoe, a smart diabetic shoe with IoT-enabled sensors, transmitting real-time data to a mobile app and showing high calibration accuracy, emphasizing the importance of refining sensors and predictive modeling for DFU prevention.

Ming et al. [11] proposed a telemedical monitoring system using sensor-equipped insoles to detect plantar temperature asymmetry, where intervention groups showed no ulcer cases, demonstrating feasibility but limited by low ulceration incidence during the study. Bus et al. [12] evaluated the DIATEMP system involving daily foot temperature monitoring for DFU prevention. Although no significant reduction was seen at monitored sites, significant benefits were observed when adherence to activity reduction was high, stressing that patient behavior plays a critical role in telehealth interventions for DFU.

De Pascali et al. [13] developed a smart insole-based monitoring platform integrating microfabricated sensors for plantar pressure, temperature, and sweat glucose detection. This real-time wireless system proved capable of detecting early signs of ischemic damage and inflammation, with strong technical performance in piezoelectric and glucose sensing. Although promising, the system remains in early-stage development, requiring clinical validation and broader usability testing. This effort represents an important step in combining bio-sensing with wearable technologies for DFU prevention.

Reddie et al. introduced a purely mechanical plantar pressure evaluation device designed for low-resource settings. Utilizing bistable compliant mechanisms, the system provides binary feedback without electronics. Though sensitivity was relatively low in healthy subjects (25.6%), performance improved in heavier individuals, indicating potential with design enhancements. This approach offers a minimalistic and accessible solution for DFU screening where advanced electronics are impractical.

Nagarajan et al. [14] proposed a cryptographically secure data transmission method in the Internet of Medical Things (IoMT) using a hybrid RES-256 model, which combines RC6 encryption, ECDSA digital signatures, and SHA-256 hashing to safeguard sensitive health data. The system ensures end-to-end confidentiality, integrity, and authentication between implantable medical devices and clinical repositories. This architecture offers resilience against known attacks such as DoS and router-level intrusions, and is validated using ECG signals from the MIT-BIH dataset, showing efficient encryption/decryption performance. While their focus was on data protection rather than analytics, the architectural principles are directly applicable to DFU systems, particularly in securing patient sensor streams (pressure, vibration, and temperature) in mobile-based diagnostic applications.

Ramana et al. [15] developed an ambient intelligence-based intrusion detection system for IoT environments using a Reinforcement Learning-integrated Deep Q-Network (RL-DQN). The model performs binary attack classification at the edge and multi-class detection in the cloud, enhancing real-time threat identification in wireless sensor networks. Evaluated on datasets like UNSW-NB15 and CICIDS2017, it outperforms traditional ML methods in intrusion detection accuracy. Their design affirms the feasibility of deploying intelligent, lightweight inference models on resource-limited platforms—an essential requirement for smartphone-based DFU screening in rural or offline clinical scenarios.

Fakhar et al. [16] proposed a deep learning-based MLFNN model for early diabetes detection using the PIMA dataset. Their approach included data normalization, advanced activation functions (ELU, SELU), and robust validation, achieving superior performance over traditional classifiers. This study is reviewed for its methodological relevance—particularly in data preprocessing and neural network-based medical prediction—which aligns with our DFU system's clinical data handling and predictive modeling approach.

These studies collectively highlight the progress in applying machine learning and deep learning models to clinical and sensor-based data for effective DFU detection and risk stratification. Building on this foundation, the proposed system introduces a hybrid framework that integrates CatBoost for feature selection and a DN-FN for predictive modeling, with an emphasis on real-time DFU risk assessment and seamless deployment via smartphone-based applications for enhanced accessibility and scalability.

III. Methods and Methodology

Conventional diabetic foot ulcer (DFU) risk prediction systems predominantly rely on static clinical parameters and lack integration of real-time physiological data, resulting in suboptimal early detection and limited clinical utility. Furthermore, many existing approaches employ opaque machine learning models and manual feature selection techniques, which constrain model interpretability and hinder their deployment in patient-facing applications.

To address these limitations, the proposed study introduces a multi-modal, smartphone-integrated predictive framework that combines dynamic sensor inputs—pressure, temperature, and vibration sensitivity—with comprehensive clinical attributes. Feature selection is performed using the CatBoost algorithm, which offers high efficiency and interpretability in handling heterogeneous data types. For classification, a Deep Neuro-Fuzzy Network (DN-FN) is employed to model nonlinearities and uncertainty in physiological signals while maintaining interpretability through fuzzy logic layers.

This end-to-end methodology enables real-time DFU risk stratification (low/mid/high) and seamless deployment in a mobile environment, making it suitable for remote, home-based monitoring and early intervention.

A. Data Acquisition and Integration

This study is based on a dataset collected from 316 participants, including 192 individuals with diabetes and 124

healthy controls, gathered from several small clinics and healthcare centers across the Vellore district in Tamil Nadu, India. The dataset is divided into two main feature groups: clinical and sensor-based attributes. In Table I, the clinical data (D-1 to D-16) covers a broad range of demographic and medical details, such as diabetic condition, age, gender, BMI, HbA1c, blood pressure, family history, and key lifestyle factors like diet, physical activity, smoking, and alcohol use. It also includes anthropometric indicators like height and weight to improve model accuracy. The sensor-derived data is captured through dedicated hardware, with pressure readings from force-sensitive resistors (FSRs), temperature values from DS18B20 digital sensors, and vibration sensitivity measured manually using an eight-point tactile test on both feet to evaluate neuropathy. While users enter their clinical information once during initial registration via a smartphone app, the sensor readings are collected in real time during each assessment, enabling dynamic monitoring and responsive DFU risk prediction.

TABLE I. DATASET FEATURES AND CORRESPONDING LABELS

Data Attributes	Identifiers
Diabetic	D-1
Age	D-2
Gender	D-3
BMI	D-4
HbA1c	D-5
Blood Pressure	D-6
Family History	D-7
Diet	D-8
Physical Activity	D-9
Smoking	D-10
Alcohol	D-11
Education	D-12
Profession	D-13
Early Signs of DFU	D-14
Height	D-15
Weight	D-16
Pressure of Right Fore Foot	RP-1
Pressure of Right Heel	RP-2
Pressure of Left Fore Foot	LP-1
Pressure of Left Heel	LP-2
Temperature of Right Fore Foot	RT-1
Temperature of Right Heel	RT-2
Temperature of Left Fore Foot	LT-1
Temperature of Left Heel	LT-2
Vibration Sensitivity - Right Foot	RV-1 to RV-8
Vibration Sensitivity - Left Foot	LV-1 to LV-8

B. Data Preprocessing

The dataset comprises 316 individuals, each with 16 clinical attributes and 24 sensor readings, including plantar pressure, skin temperature, and vibration sensitivity. This heterogeneous data demands careful handling to maintain quality and compatibility for machine learning applications.

To address missing entries, different imputation strategies are used based on the feature type. Numerical clinical variables such as age, BMI, HbA1c, and blood pressure are filled using mean imputation, ensuring that the overall distribution is preserved. For categorical attributes like gender, diet, and smoking status, mode imputation is applied to replace missing values with the most common category in each field. Notably, sensor data—including pressure, temperature, and vibration sensitivity—is collected in real time during each user session, making it inherently complete and not requiring imputation.

Categorical features are then transformed into numerical formats using appropriate encoding methods. Binary categories such as gender, smoking, and alcohol usage are processed using label encoding. For multi-class variables like education level and profession, one-hot encoding is used to prevent the model from learning false ordinal relationships. This encoding ensures that the dataset remains interpretable and machine-friendly without introducing bias.

Normalization is applied to all continuous numerical features to bring them to a uniform scale, facilitating faster and more stable model training. Min-Max normalization is used to scale attributes such as age, BMI, HbA1c, blood pressure, height, and weight. Similarly, sensor data—including plantar pressure and temperature—is normalized to account for individual variations, while vibration sensitivity values, already in binary form, are directly incorporated without transformation.

C. Feature Selection Using CatBoost

To enhance the model's performance and generalization capability, the CatBoost algorithm [17] [18] —an advanced gradient boosting method developed by Yandex—is utilized for assessing the importance of input features. By training CatBoost on the clinical dataset, it generates importance scores for all 16 clinical variables (D-1 to D-16). One of the key strengths of CatBoost is its native support for categorical features, enabling efficient processing while minimizing overfitting through ordered boosting strategies. As shown in Fig. 1, the feature importance rankings indicate that, five clinical attributes—D-9 (Physical Activity), D-8 (Diet), D-10 (Smoking), D-12 (Education), and D-13 (Profession)—were identified as having minimal relevance to DFU risk prediction and were thus excluded. These variables showed weak associations with the target outcome, contributing little to the model's predictive capacity. The refined feature set, comprising the remaining 11 clinical features and all sensor-based inputs, serves as the input layer for the modeling phase, ensuring improved classification efficiency and model interpretability.

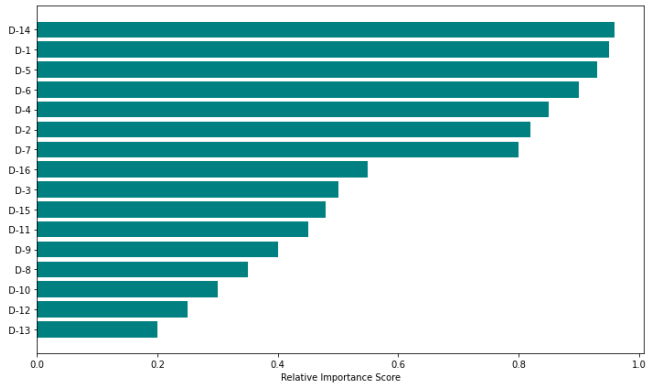


Fig. 1. Given dataset's clinical feature importance for DFU severity evaluation – comparison-ranking chart.

This targeted pruning reduced dimensional redundancy while maintaining the model's diagnostic robustness. The remaining high- and moderately ranked clinical predictors were then integrated with the sensor data for input into the DN-FN, ensuring an optimized and clinically meaningful feature set for risk classification.

D. Deep Neuro-Fuzzy Network (DN-FN)

The selected features are then passed to the DN-FN for the prediction phase. The DN-FN is a hybrid model that integrates deep learning with fuzzy logic to capture complex, nonlinear relationships in the data [19] [20]. As shown in Fig. 2, the DN-FN operates in a multi-phase architecture—fuzzification, fusion, and learning—to transform the input data into fuzzy sets, combine them into a unified representation, and learn from these patterns to predict the DFU risk level. This approach ensures the model not only performs accurate predictions but also remains interpretable, making it easier for clinicians to understand how specific inputs influence the risk classification.

Algorithm 1: DN-FN Prediction Algorithm

Input:

Sensor + clinical feature vector $\boldsymbol{z_i} \in R^d$ for patient i, with true label $y_i \in \{L, M, H\}$

1. Fuzzification Phase (Input → Fuzzy Layer)

Each input feature is fuzzified using Gaussian membership functions:

$$\mu_{ij}^{(k)} = \exp\left(-\frac{\left(z_{ij}-m_j^{(k)}\right)^2}{2\left(\sigma_j^{(k)}\right)^2}\right) \quad (1)$$

- $m_j^{(k)}$: center of k-th MF for feature j
- $\sigma_j^{(k)}$: width of the MF

This generates fuzzy activations for each input dimension.

2. Fusion Phase (Fuzzy → Fusion Layer)

Combine all fuzzy outputs using fuzzy rule-based logic to form rule strengths:

$$r_q = \prod_{j=1}^{d} \mu_{ij}^{(k_{qj})} \text{ (rule strength for rule } q) \quad (2)$$

Normalize all rule strengths:

$$\tilde{r_q} = \frac{r_q}{\sum_{q'=1}^{Q} r_{q'}} \quad (3)$$

Pass through a weighted linear combination to form the fused representation:

$$f_i = \sum_{q=1}^{Q} \tilde{r_q} \cdot \left(\boldsymbol{w_q}^{\top} \boldsymbol{z_i} + b_q\right) \quad (4)$$

3. Learning Phase (Fusion → Hidden Layer → Output Layer)

Step 3.1: Hidden Layer Processing

The fused scalar or vector f_i is input to a deep learning stack (MLP or BiLSTM). For simplicity, assuming MLP:

$$\boldsymbol{h_1} = \phi(\boldsymbol{W_1} f_i + \boldsymbol{b_1}) \quad (5)$$

$$\boldsymbol{h_2} = \phi(\boldsymbol{W_2 h_1} + \boldsymbol{b_2}) \quad \text{(and so on, as per hidden depth)} \quad (6)$$

- ϕ: non-linear activation (e.g., ReLU, tanh)

Step 3.2: Output Layer

Final dense layer maps hidden activations to class logits:

$$\boldsymbol{o_i} = \boldsymbol{W_o h_n} + \boldsymbol{b_o} \quad (7)$$

Softmax applied for class probabilities:

$$\hat{y}_i = \arg\max\left(\text{softmax}(\boldsymbol{o_i})\right) \quad (8)$$

Output:

Predicted class label $\hat{y}_i \in \{L, M, H\}$ for each patient/sample

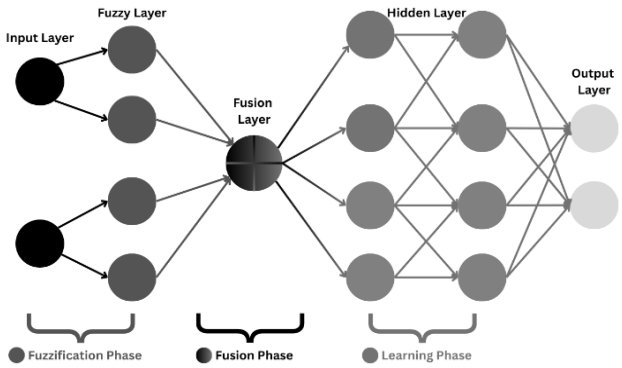


Fig. 2. DN-FN Prediction model – network architecture.

Furthermore, Bayesian Optimization is employed to fine-tune the hyperparameters of the DN-FN model. Unlike conventional methods such as grid search or random search, this approach leverages a probabilistic model to explore the hyperparameter space more intelligently, selecting optimal configurations based on prior evaluations. Key parameters—including learning rate, tree depth, and regularization factors—are optimized through this method, leading to improved model performance and better generalization to new data. The model's effectiveness is first validated using an 80:20 stratified train-test split, ensuring a consistent distribution of diabetic and healthy cases across both sets. To enhance the reliability of performance assessment, 5-fold cross-validation is also conducted, offering a more comprehensive evaluation across different subsets of the dataset.

E. Smartphone-Based Integration and Real-Time Inference

The Diabetic Foot Screening Companion (DFSC) mobile app, built using Android Studio with Java and Kotlin, integrates a TensorFlow Lite (TFLite)-based backend for fast, offline inference of DFU risk using the trained Deep Neuro-Fuzzy Network (DN-FN) model. Users input their clinical data once during registration—excluding the five least significant features removed via CatBoost—and the app stores this locally. During each assessment, the app connects via Bluetooth to an Arduino-based embedded system that collects real-time sensor data across three modalities: vibration sensitivity at 16 foot points (RV-1 to RV-8, LV-1 to LV-8), plantar pressure via Force Sensing Resistors (RP-1, RP-2, LP-1, LP-2), and localized temperature using DS18B20 sensors (RT-1, RT-2, LT-1, LT-2). These readings are processed on-device when the user taps "Calculate Severity," triggering the DN-FN model to classify DFU risk into Low, Mid, or High. Results are presented with color-coded severity indicators, health tips, and recommendations. Designed for accessibility in rural or resource-limited areas, the DFSC app enables proactive self-monitoring through its user-friendly interface, offline capability, and comprehensive multi-sensor integration. Fig. 3 represents the mobile application's navigations and functionalities and Fig. 4 shows the built model's prototype device.

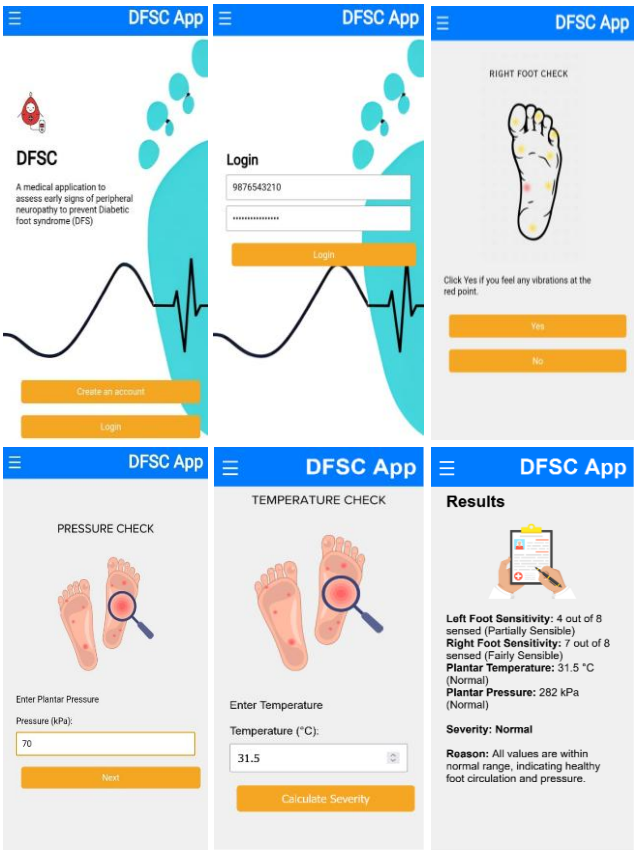


Fig. 3. Smartphone Application's step by step navigation, sensor-based foot inspection and results.

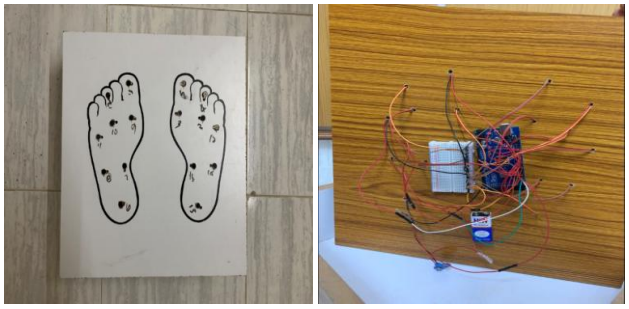


Fig. 4. DFU Sensor device – model prototype.

IV. Results and Discussion

A. Results

Table II presents the proposed model benchmarked against state-of-the-art algorithms.

TABLE II. Proposed System Performance Comparison Across State of the Art Models Using Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	MCC
Random Forest	86.2	84.9	82.3	83.5	88.4	76.1
Support Vector Machine (SVM)	85.1	82.7	80.2	81.4	87.2	72.9
XGBoost	88.6	87.1	84.3	85.7	89.3	79.6
ANN (3-layer MLP)	89	87.5	85.6	86.5	90.2	81.2
LightGBM-EANFIS	90.5	89.3	87.4	88.3	92.1	84.7
Proposed DN-FN - CatBoost	92.4	91.2	89.7	90.4	94.5	88

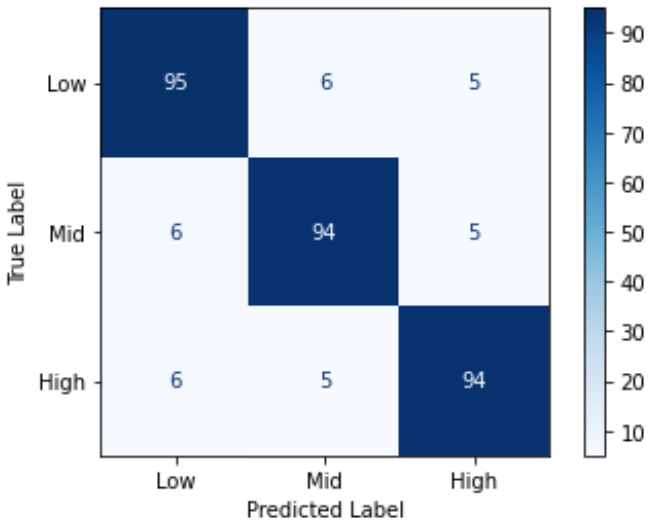


Fig. 5. Confusion matrix for three class DFU severity classification.

The performance of the proposed DFU risk prediction system, utilizing the CatBoost-selected features and the DN-FN model, demonstrates strong classification accuracy across all three severity levels—Low, Mid, and High. Fig. 5 visualizes the confusion matrix of the proposed model which correctly classified 95 Low-risk, 94 Mid-risk, and 94 High-risk instances, with only minimal misclassifications between adjacent classes. These results indicate the model's ability to effectively differentiate among DFU severity levels, particularly in borderline cases where physiological parameters might overlap. The near-diagonal dominance of the confusion matrix highlights the model's robustness and reliability in real-world risk stratification, affirming its suitability for integration into the DFSC mobile application for continuous, on-device diabetic foot monitoring.

B. Discussion

Compared to prior studies, the proposed smartphone-integrated system demonstrates notable advancements in both predictive accuracy and deployment feasibility. Agrawal et al. achieved 85% accuracy using AdaBoost for plantar pressure-based classification, while the proposed model surpassed this with improved precision and real-time DFU severity grading across multimodal sensor data. Similarly, Ghazi et al. and Kularathne et al. developed smart footwear systems incorporating temperature and pressure sensors, but lacked advanced feature selection or neuro-fuzzy logic for personalized risk profiling. Ming et al. and Bus et al. emphasized telemedical approaches, though their effectiveness was limited by adherence and simplistic data interpretation. In contrast, the proposed system leverages intelligent preprocessing, robust feature selection via CatBoost, and a deep neuro-fuzzy framework for explainable, real-time DFU detection, making it more adaptive and clinically relevant for at-home screening in resource-constrained settings.

C. Limitations and Future Work

While the proposed system demonstrates significant potential, it does face certain limitations that could impact the comprehensiveness of diabetic foot ulcer (DFU) risk assessment. A notable shortcoming is the exclusion of pulse oximeter data and microvascular flow indicators, which are critical for evaluating peripheral circulation and anticipating DFU development. Moreover, the current smartphone-based implementation lacks individualized interpretability tools such as SHAP or LIME, restricting users from understanding the underlying reasoning behind each prediction. Another limitation is the absence of longitudinal tracking, which is essential for monitoring changes in foot health over extended periods.

To overcome these gaps, future iterations of the system will integrate pulse oximeter sensors to support real-time vascular health evaluation. Additionally, efforts will be made to embed SHAP-driven explainability directly into the mobile application, offering both clinicians and users greater transparency into prediction logic. The dataset will also be diversified to include participants from various ethnic backgrounds, thereby improving the model's generalizability and relevance. Furthermore, the architecture will adopt federated learning techniques to facilitate on-device model training while preserving user privacy. Collectively, these enhancements aim to bolster the system's clinical utility, reliability, and adaptability in diverse, real-world healthcare environments.

V. CONCLUSION

This study validates the effectiveness of combining CatBoost-based feature selection with a Deep Neuro-Fuzzy Network (DN-FN) for early and interpretable prediction of Diabetic Foot Ulcer (DFU) risk. By leveraging a hybrid input space of clinical attributes and real-time sensory data—including plantar pressure, foot temperature, and vibration sensitivity—the proposed framework demonstrates superior performance over existing machine learning models in terms of accuracy, reliability, and semantic transparency. The smartphone-integrated deployment ensures offline operability and user accessibility, making it particularly suitable for rural and resource-constrained settings. The system allows patients to input clinical data once while performing periodic sensory tests, enabling continuous foot health monitoring. Comparative evaluation shows that the proposed approach outperforms state-of-the-art models in key evaluation metrics, underscoring its robustness and clinical relevance. Future enhancements will focus on incorporating vascular health sensors like pulse oximeters, advanced explainable AI techniques such as SHAP for instance-level interpretation, and privacy-aware training methods like federated learning to expand its utility and scalability in real-world clinical practice.

ACKNOWLEDGMENT

This work was supported by the Indian Council of Medical Research (ICMR). The authors express their gratitude to ICMR and Vellore Institute of Technology, Vellore for their support.

DISCLOSURE OF INTERESTS

The author(s) declare(s) that there is no conflict-of-interest regarding the publication of this paper.

REFERENCES

[1] Liu, Y., et al., Time trends and advances in the management of global, regional, and national diabetes in adolescents and young adults aged 10–24 years, 1990–2021: analysis for the global burden of disease study 2021. Diabetology & Metabolic Syndrome, 2024. 16(1): p. 252.

[2] Armstrong, D.G., et al., Diabetic foot ulcers: a review. Jama, 2023. 330(1): p. 62-75.

[3] Yachmaneni Jr, A., et al., A comprehensive review of the vascular consequences of diabetes in the lower extremities: current approaches to management and evaluation of clinical outcomes. Cureus, 2023. 15(10).

[4] Joseph, A., et al., Genetics of type 2 diabetes mellitus in Indian and Global Population: A Review. Egyptian Journal of Medical Human Genetics, 2022. 23(1): p. 135.

[5] Anbarasi, L.J., et al., An overview of current developments and methods for identifying diabetic foot ulcers: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2024. 14(6): p. e1562.

[6] Chatrati, S.P., et al., Smart home health monitoring system for predicting type 2 diabetes and hypertension. Journal of King Saud University-Computer and Information Sciences, 2022. 34(3): p. 862-870.

[7] Agrawal, D.K., et al., Smart Insole-Based Plantar Pressure Analysis for Healthy and Diabetic Feet Classification: Statistical vs. Machine Learning Approaches. Technologies, 2024. 12(11): p. 231.

[8] Reddie, M. and D. Frey, A resource-efficient plantar pressure evaluation system for diabetic foot risk assessment. Diabetology, 2024. 5(2): p. 206-222.

[9] Ghazi, R.F., J.S. Chiad, and F.M. Abdulghani, Design and manufacturing a smart shoe for diabetic foot ulcer monitoring and prediction system using internet-of-things technology. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024. 46(2): p. 83.

[10] Kularathne, N., et al. Dia-Shoe: a smart diabetic shoe to monitor and prevent diabetic foot ulcers. in 2019 International Conference on Advancements in Computing (ICAC). 2019. IEEE.

[11] Ming, A., et al., Telemedical Monitoring of Plantar Temperature in Diabetic Patients at Risk of Foot Ulcers: The Randomized Smart Prevent Diabetic Feet Trial. Deutsches Ärzteblatt international, 2024. 121(1): p. 9.

[12] Bus, S.A., et al., Effectiveness of at-home skin temperature monitoring in reducing the incidence of foot ulcer recurrence in people with diabetes: a multicenter randomized controlled trial (DIATEMP). BMJ Open Diabetes Research and Care, 2021. 9(1): p. e002392.

[13] De Pascali, C., et al., Modeling, fabrication and integration of wearable smart sensors in a monitoring platform for diabetic patients. Sensors, 2021. 21(5): p. 1847.

[14] Nagarajan, S.M., et al., Secure data transmission in internet of medical things using RES-256 algorithm. IEEE Transactions on Industrial Informatics, 2021. 18(12): p. 8876-8884.

[15] Ramana, T., et al., Ambient intelligence approach: Internet of Things based decision performance analysis for intrusion detection. Computer Communications, 2022. 195: p. 315-322.

[16] Fakhar, M.H., et al., A Deep Learning-based Architecture for Diabetes Detection, Prediction, and Classification. Engineering, Technology & Applied Science Research, 2024. 14(5): p. 17501-17506.

[17] Geeitha, S., et al., Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer. Scientific Reports, 2024. 14(1): p. 19828.

[18] Wang, H., et al., Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. Journal of Big Data, 2024. 11(1): p. 44.

[19] Rashed, B.M. and N. Popescu, Medical image-based diagnosis using a hybrid adaptive neuro-fuzzy inferences system (ANFIS) optimized by GA with a deep network model for features extraction. Mathematics, 2024. 12(5): p. 633.

[20] Shabu, S.J., et al., An Improved Adaptive neuro-fuzzy inference framework for lung cancer detection and prediction on Internet of Medical Things platform. International Journal of Computational Intelligence Systems, 2024. 17(1): p. 228.

Decoding Sales Order Anomalies: Advanced Predictive Modeling and Discrepancy Resolution Utilizing Machine Learning Algorithms

Amit Kumar Soni, Pooja Jain
IIIT Nagpur, Nagpur, India

***Abstract*—This study examines the accuracy of order prediction and determines the grounds for order block predictions. It sets order deviation by calculating forecasted variation using R^2 scores and mean absolute deviation. The blocks that are checked mainly include—business partner block, credit block, common block, and delivery block. Demand forecasts compare six months' worth of sales data against mean absolute deviation and coefficient of variation. This study puts forth a proposal for solving discrepancies between sales order forecasts and confirms credit management's system credit limits on sales orders. Parameters for evaluating orders are set relying on historical data. Machine Learning (ML) has been utilized in this study—which involves Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms to improve accuracy where they achieve 96% and 93% respectively.**

Keyword*—Block predictions; credit; machine learning; sales data

I. Introduction

In this study—forecast predictions and reasons behind sales order evaluation using Machine Learning (ML) have undergone analysis [1]–[5]. This study underscores prediction delays due to inconsistencies in orders. Business success hinges on the continuous prediction of demand concerning forecast quantity and actual sales quantity. This aids companies in boosting production efficiency by crafting an optimal supply chain and ensuring product availability. Generally, SAP—predictive maintenance and service—aids customers with a unified solution for identifying issues within large machine fleets. It enhances services and optimizes service planning for individual forecasts. SAP cloud edition stands as an innovative solution in predictive analytics. Big data in SAP surpasses the capacity of traditional databases for capturing product data. However, ML facilitates big data visualization and analysis to provide real-time outcomes. Data sensors can be used for obtaining real-time master data. SAP ensures enhanced security through encrypting sensor data and is achieved via user management and authentication. Therefore, the issues concerning the sales order are not addressed in this study.

Hypotheses can rectify the mismatch between anticipated and actual sales quantities. It is therefore useful for analyzing sales order confirmations and showcasing the role of ML in predicting delays. Demand varies across periods—making accurate forecasting highly critical for forecasting order quantities. A six month dataset is used to evaluate the forecast. This requires that demand can be predicted through forecasts per specific period because there is volatility across the horizon. For this matter, a promise check is available—which is employed in checking whether products are available in terms of quantity at the time of creating a sales order and—hence—confirms schedule lines at the line-item level. This ensures that the available promise check succeeds during order creation by giving quantity to the sales order. This forecasted order remains unutilized by any customer requirement within Supply Chain Management (SCM), Advanced Planning and Optimization (APO) as well as Demand Planning Module (DPM). The strategy of forecasted order may be missing, or its strategy has deviated from that defined on the product master's strategy tab. On average it will be 5% when comparing the projected forecast with the last three months' sales data. Data for modeling consists of past events while future occurrences are used as inputs for the next period's forecasting. Therefore, an 80% of train dataset was utilized to build models with 20% left out as a test dataset. The R-squared score is used to evaluate how well a given forecast is performing.

The study is as follows; the background will be seen in the following section. The related works are listed in Section III. The problem statement is given in Section IV. The implementation is carried out in Section V. The experimental analysis is presented in Section VI, and in Section VII, we propose a conclusion and future directions for the research.

II. Background

The study highlights a block in display mode and discusses credit blocks in sales orders. Sales orders having credit blocks appear on monitor stock. Such deliveries remain blocked if their credit limit is exceeded. Credit checks are done upon every sales orders and delivery documents. System-guided changes to new credit conditions—whenever—customers' credit positions demonstrate dynamic and static functionality respectively. Credit checks aim to mitigate the risk of customers exceeding their credit limits—necessitating re-evaluation of prior approved deliveries upon any change in delivery. Therefore, a careful analysis of credit checks and limits is essential for each sales order that results in an order block and thus—halts subsequent processing. When a sales order document is created for an account with an assigned credit limit—**it** changes the items automatically—set to the—delivery block in the document. This automatic setting occurs whenever document changes are made—as the credit limit re-

evaluates the document's value against the credit limit in the account master data. The credit limit is maintained in the financial tab while the delivery block credit limit is entered in the sales data tab. A delivery block is set upon account addition—since—a credit limit check does not occur before item addition. Item addition triggers credit limit check as per credit limit check configuration. The function of credit limit check must be turned on after feeding it with the necessary data.

However, un-invoiced orders will include items invoiced—but not cleared in financial accounting—needed for calculating the invoiced open amount. Verifying the credit limit proposed by the system on the sales order requires checking the limit amount in the account master data. Sales order confirmations are analyzed based on delivery block settings. If a delivery block is required and not cross-checked with the account financial data tab—different options like customer request or a customer-specific delivery block must be maintained. When the header-level delivery block and confirmation block are selected—confirmed quantities in schedule lines are deleted upon saving the order. Open sales order credit values—based on confirmed quantities at the line-item level. Transfer of requirement occurs even if a delivery block with confirmation block is set at the header level. Delivery or credit block settings in the order do not prevent requirement transfer. Setting the confirmation block for the delivery block results in zero confirmed sales order quantities. If the "confirmation block" indicator in customizing is set for the relevant delivery block—the system resets the confirmed quantity when saving the sales document. If the fixed date and quantity indicator is set—the system does not transfer to material requirements planning. Approved deliveries should remain 'approved'. Therefore, automatic assignment of delivery blocks occurs even if the document value hasn't increased. Sales orders should not be released in customer demand unless all line items have green color available to promise status—enabling the release button. Force confirmation in the actions tab releases customer demand without errors. Sales orders show completed status even with partial deliveries due to delivery rule configuration.

Delivery rule configuration for sales order line items specifies single delivery full quantity—aiming for one single delivery for all products in the sales order. Creating a single outbound delivery against such a sales order update its status to finish—marking the sales order as completed. The common issue of being unable to create service confirmation for sales order items arise from service items being tied to project tasks. Updating project messages changes the status of sales orders. If a sale order has “process” as its status—there is a line item that has not been released. The release order option will be enabled but due to the line order status cannot save follow-up service confirmation. It is important that sales orders are confirmed before they can be processed further. It is necessary to analyze fields affecting delivery blocks. An order must be released first before new service confirmation is created. Rejection reasons are chosen based on the sales document type and sales organization at the header level of a sales order. Values are displayed in drop-down lists for customizing for orders and rejection reason types based on requirements. When an item is cancelled—the sales order goes into process status—necessitating order release before service confirmation. Therefore, to solve these issues—we used the Machine Learning (ML) that facilitates order evaluation and classification using a hybrid model. ML can assign classes for order evaluation parameters. Order parameters in this study are selected for their impact on the ML order evaluation model analysis.

III. Related Work

Several studies had been conducted such as [6] has suggested in their paper assessing forecast model performance in SAP. The study [7] have evaluated the use of time series analysis for predicting the level. For sales order data optimization—[8] have suggested in their paper using statistical methods in SAP optimizer planning. The use of ML for forecasting is recommended by study [9]. Consumer demand during ramp-up and ramp-down periods is influenced by a combination of the holiday and the day of the week it falls on according to study [10]. The holiday effect and weekday combination will vary each year—causing affected product locations to exhibit variable sales patterns throughout the year. There are specific periods when consumer demand is generally difficult to forecast. Forecasting can operate on an aggregated level if no disaggregation rule is maintained for the forecast key figure. The study in [11]—discussed the use of four SVM implementations. They give the features for class probabilities output—cross-validation error estimation, and hyperparameter estimation for the Gaussian Radial Basis Function (RBF) kernel. The study in [12] have conducted a detailed analysis of ERP data and its visualization. ML algorithms can be applied to classification problems like order ranking. The study in [13] have suggested that data mining can be used along with predictive analytics. However, this study proposes forecasting in relation to quantity in the sales order. It highlights that ML techniques can be used for forecasting. The performance of predicting forecast sales order values are assessed using R-squared score and Mean Absolute Deviation (MAD). An R-squared score value of 73% and MAD of 100 is calculated based on sales orders training data. The requirements planning run—perform poorly because it takes a long time for the system to calculate usage probabilities. When calculations are too complex—the system terminates the Market Retail Price (MRP) list and displays the error message. In study [14]—they compared various regression estimators by using the Mean Squared Error (MSE) criterion and constructed an alternative to the least squares estimator in case of multi-collinearity. A business partner with roles of account and sales and service partner are required to continue with the account role and block the sales and service partner role. When the partner role is blocked—it blocks the account role and the complete business partner role. However, in this study—a delivery block at the header level with the confirmation block indicator deselected. Open sales order values are calculated with the confirmed quantity of the schedule lines. A reduction in order quantity can be a reason for rejection—when applied at the item level. This occurs whenever any changes in the saved order occurred—developing inconsistencies in the delivery block at the top level.

IV. PROBLEM STATEMENT

The analyzed problems in this study include the reasons for sales orders not aligning with the forecast—resulting in inconsistent consumption when sales orders are created before the forecast. Another problem is block in display mode of the order. The system issues an error indicating that the relevant trading contract is blocked and there is an error in customizing. Some other problems include:

1) Sales orders with a credit block are displayed on the requirements list app.

2) Sales order creation is blocked due to a business partner block and analysis of confirmation of sales orders for the repetitive fields "delivery block" and "description".

3) Calculation of order performance and analysis of order evaluation parameters.

4) An order evaluation model is designed in the SAP application using evaluation criteria such as order reason, delivery date, and quantity. The order evaluation score is calculated based on weighted scores.

5) Complex dependencies are assigned to individual activities in the routing lead to long runtime in lead time scheduling.

Improving the evaluation of individual dependencies is necessary as dependencies that lead to too many combinations are problematic in terms of performance. This is applicable to specified and non-specified conditions. The reasons for order blocks generally are—forecast issues are due to improper settings for requirement strategy. Inconsistencies exist between the allocation indicator of the sales order requirement class and the assignment mode of the requirement strategy due to missing settings of the check mode. Improving forecast accuracy for affected product locations during variable demand periods is required. External figures or data more indicative of the demand than the Demand Influencing Factors (DIFs) generated by the forecasting are needed. Disabling some system demand influencing factors can improve forecast accuracy. Another reason is association management requires editing when the sales order is in display mode. The error appears as an error in customizing. When the sales order is generated from GTM with credit management—it does not get released using transaction code VKM3. Sales orders created with credit status "not approved" must be released using transaction VKM4. Additionally, other reason is the manual pre-implementation for two data elements—"description" and "delivery"—is required. "Description" is replaced by "created by username" and "delivery block" is replaced by "item delivery block". Restricting order rejection reasons based on the sales document type and sales organization in the dropdown list during configuration is required. Also, the order performance with ML is based on parameters like order reason, delivery date, and quantity [15]–[20]. Weight assignments are calculated during the training of the classifier on training data. The system calculates usage probabilities for configurable materials with planned independent requirements. Maintaining dependent forecast requirements in the planning segment of the stock requirements list is necessary. The MRP creates more forecast planned orders that require usage probability calculations to cover these dependent requirements. The long runtime of materials or products entered in the order results from the usage probabilities calculated in the BOM explosion. The planning statistics for the total planning run issue a ranking list of ten products with the longest runtime. The information report is displayed using the transaction RMMDPERF. Running the program SAPLCUFE for complex calculations to calculate usage probability is required. The program's execution time can be checked on the planning process overview screen. Detailed scheduling for forecast planned orders can be deactivated, or a lower detailed scheduling horizon can be set. Relocating characteristics planning into long-term planning can be done. Changes in forecasting are carried out at longer regular intervals—for instance—monthly. A long-term planning run is executed after such changes, and the report RM60CH00 is run to copy the simulated dependent demands created to operative planning. Saving interactive forecasting results produces unexpected and unrelated messages when the forecast key figures remain unchanged—and the 'enable forecast BOD' flag is activated in general settings for the forecast. Various inconsistencies are checked and reported by running the check_table_consistency after upgrades, migration, crashes, or out-of-memory termination.

V. IMPLEMENTATION

The algorithms that are used in developing a forecast data model from the SAP application consider several different factors which influence the order forecast. These factors must have standard strategies with different category groups and sales order categories. There are combinations of PIR segment and assignment mode supported by standard strategies in SAP. The sales order category of a sales order is checked by transaction code /SAPAPO/RRP3—while categories assigned to a category group are checked by /SAPAPO/SNPCG. When requirement strategy 10 is used—no consumption occurs. If no category group is maintained—then the forecast tab will show the order should not consume the forecast. Products views help locate the requirement strategy field of the forecast orders. The requirement strategy must match with the strategy assigned to the product master otherwise delete and re-release forecasts. Category type = "3" (forecasts) for all forecast orders. All orders having forecast order categories can be consumed even if the category is not set in the requirement strategy. Algorithm 1 shows the steps that are required by the allocation indicator of the sales item's requirement class.

Algorithm 1: Allocation indicator

START

1. *Find the requirement type of the sales order item.*
2. *Find the corresponding requirement class from table VBBE – BDART.*
3. *Find the allocation indicator of requirement class.*

Ensure the allocation indicator is correctly mapped to the order status of sales order in the APO system.

END

It is impossible to manually reconcile forecast error calculations against system-generated calculations. The Coefficient of Variation (CV) is calculated based on MAD. When 'adjust bias of forecast' is selected:

- MAD = AVG (|bias adjusted forecast – sales|).
- Bias adjusted forecast = input forecast * (1 - bias used).

If adjust for bias is adjust positive or adjust all:

- CV = 1.25 * (MAD / MAX [AVG (bias adjusted forecast) or AVG (sales)]) * (1 - bias used).

Forecasting has a negative bias therefore not bias adjusted. The process of fixing the forecast is recommended when there is a negative bias—which indicates under-forecasting. If bias used = 1, CV = 0. Outlier periods are those where both Forecast and sales are 0 or are NULL—before determining bias removes them. When sales and forecast are either NULL or both 0 in the same period—then exclude that period from calculating AVG (forecast) and AVG (sales). The requirement class defines the check mode on the advanced planning side. On S/4HANA—the check mode must be maintained in table /SAPAPO/ATP06. Forecasts do not consume forecasts—when the assignment mode of check mode does not match the requirement strategy's assignment as shown in Table I.

TABLE I. FORECAST STRATEGY

Product Master Forecast Strategy	Sales Order Status
20	1
30	2
35	2
40	3

If a new product site has forecast consumption—then forecast orders are created but the result of forecasting consumption may be inconsistent at first. The discrepancy is fixed by any subsequent modification in orders related to it. When credit block is released—the workflow gets retriggered. Dynamic credit check sets credit limits for customers based on undelivered not yet billed sales orders, undelivered not yet invoiced deliveries, unbilled accounting document open amounts and unpaid bills. Every time there is a save of document or check credit selected—credit limit check automatically starts. In case there are changes in the credit master data of a business partner—this event can trigger workflow functions notifying recipients about them and these with follow-on processes in SAP credit management. Scoring models might self-tune themselves using economic variables predicting delays in business payments. For any partner being monitored under SAP credit management—there should be a master record in the system. To transfer business partners—assign role UKM000–SAP credit management, score recalculation from SAP easy access screen and define default procedure for score calculation along with rule checking. Beforehand two data elements must undergo manual pre-implementation steps where “description” is replaced by “created by username” and “delivery block” by “item delivery block”. These elements are implemented through Algorithms 2 and 3:

Algorithm 2: Establish a data component CREATED_BY_USER_NAME

START

1. *Begin Transaction SE11.*
2. *Select "CREATED_BY_USER_NAME" as the data type and input it into the input field.*
3. *Select 'data element'—then—'continue' and confirm it.*
4. *Write a user description in the description field.*
5. *On the data type tab—select an elementary type for this type of data or set a domain instead.*
6. *Under the field label tab—fill in created by user name, created by user.*
7. *Save your work on the data element now that it’s complete.*
8. *Check over everything again so that everything matches up perfectly within this new system before activating anything at all within this program.*
9. *Activate data elements.*

END

Algorithm 3: Generate data element DELIV_BLOCK_REASON_SCHEDLINE

START

1. *Open transaction SE11.*
2. *Choose 'data type' and enter 'DELIV_BLOCK_REASON_SCHEDLINE' in the input field. Then select 'create'.*
3. *Select 'data element'—then 'continue'. Confirm any warnings when prompted.*
4. *Fill the description field with ‘blocked for delivery at schedule line’.*
5. *In the ‘data type’ tab—choose ‘elementary type’. Set the data type as ‘domain’ and give the data type as 'LIFSP'. Click on enter.*
6. *In the ‘field label’ tab—provide values for item delivery block.*
7. *Save this data element.*
8. *Run consistency check on this data element.*
9. *Activate data elements.*

END

Order processing is assessed with a data model built using an extraction algorithm. Key fields like credit release and delivery date are identified (refer to Table II). Deviations in quantity, delivery date, and block status predict delays. Order performance is measured against reason, date, and quantity (refer to Table III). Each order is ranked. Orders with minimal delivery date and quantity deviations (<= 100) can be blocked based on a configuration guide. This analysis focuses on order reason, delivery date, and quantity per line item.

TABLE II. ORDER EVALUATION FEATURES

Feature	SAP Field Description
Order	(VBAK), Attribute-Nominal
Header level fields	
Sales Document Type	Table TVAK, Field-AUART
Order Reason (Reason for the Business Transaction)	Table TVAU, Field-AUGRU
Delivery Block (Document Header)	Field-LIFSK
Billing block in SD Document	Field-FAKSK
Sales Organization	Field-VKORG
Requested Delivery Date	Field-VDATU

Proposed Date Type	VPRGR
Release Date of the Document Determined by Credit Management	CMFRE
Order Number	AUFNR
Item data	SAP Sales Document: Item Data(Table-VBAP)
Item Level Fields	
Material number	(MATNR), Attribute-Nominal
Item Relevant for Delivery	LFREL
Target Quantity in Sales Units	ZMENG
Scale Quantity in Base Unit of Measure	YMENG
Dev3	Ordered quantity and quantity not received
Delivery date	Sales order line-item delivery date
Dev2	Number of days
	Delivery delay (VBAK-VDATU)
Dev1	Delivery Block document
Classification	Rank classification

TABLE III. RANK CLASSIFICATION

		Order Evaluation Criteria			
		Order Reason, Date (No of SO days), Quantity (% of Sales Order Quantity)			
Class	**Instances**	**Order Reason**	**Delivery date**	**Quantity**	**Deviation Criteria**
1	3820	y	Y	y	No
2	5210	<100	Y	Y	100 or less
3	5500	100-200	<10	<10%	Anyone
4	5350	>200	>10	>10%	Anyone
5	120	>200	>10	>10%	Any two
6	0	>200	>10	>10% of	All three
			SO Date	SO Qty	

VI. EXPERIMENTAL ANALYSIS

All through the process of sales order integration as shown in Table IV—each order's status field is validated. It is important to ensure that the integration is correct if the status is different from the requirement allocation indicator. The allocation indicator for a requirement class of a sales item should match with the APO status representing the mode in which assignment was done for the strategy of requirement. MAD was calculated at 100 based on 16,000 cases of sales orders over the seventh period using three periods moving average as shown in Table V. With changes in sales and forecasts—the forecast error CV stands at 0.5. R^2 value 0.730 obtained through linear regression indicates a strong correlation between actual values and predicted ones.

Simplifying relevant planning characteristics decreases complexity while unnecessary unassigned characteristics should not be flagged within the planning profile also reduces the number of NE conditions that lead to premature terminations due to increased probability calculation time by reducing flagged characteristic combinations from 20 down to five NE conditions reduced from four out of nineteen resulting into five OR links. Non-evaluative procedures are used for adjusting BOM item quantities—where required by means other than evaluation methods called non-evaluative procedures. Planning procedures change Bill of Materials (BOM) component quantities planned mode "1" does not explode BOMs against production orders—which may increase runtime when the "entry required" flag is switched on for some characteristics having more than one value because it terminates probabilities early by allowing many concurrent characteristic values as shown in Table VI.

TABLE IV. SALES ORDER INTEGRATION VALIDATION RESULTS

Order ID	Status Field	Requirement Allocation Indicator	APO Status	Validation Result
1001	Validated	Match	Assigned	Passed
1002	Not Validated	Mismatch	Not Assigned	Failed
1003	Validated	Match	Assigned	Passed
1004	Validated	Mismatch	Assigned	Failed
1005	Not Validated	Match	Not Assigned	Failed

TABLE V. FORECASTING AND ERROR ANALYSIS

Period	Sales Orders	Forecast Error (MAD)	Forecast Error (CV)	R^2 Value
Period 1	16,000	100	0.5	0.730
Period 2	16,500	105	0.52	0.728
Period 3	15,800	98	0.49	0.735
Period 4	16,200	102	0.51	0.732
Period 5	16,300	101	0.50	0.731

TABLE VI. BOM AND PLANNING PROCEDURES

Component	Planned Mode	Entry Required Flag	NE Conditions	OR Links
Component A	Mode 1	On	4	19
Component B	Mode 2	Off	3	15
Component C	Mode 1	On	2	10
Component D	Mode 3	Off	1	5
Component E	Mode 1	On	0	0

Calculation data comes from various sources such as master data and big data thus giving an accurate reflection about customer buying patterns. Gradient Boosting Decision Trees (GBDT) reduces noise-induced trends during statistical forecasting in high-value scenarios. Accurate selection location of products ensures that forecasts remain accurate even in seasons with high demands. Incorrect product locations will affect final forecasts negatively especially when dealing with invalid locations. It's vital to validate requirement strategy at the product master level. Use transaction /SAPAPO/MD74 to delete and re-release forecast orders. Background and foreground statistical forecasts give different results due to aggregation method differences—foreground takes into account all selection characteristics while background jobs require manual level selections for

aggregation. Forecast tab activity monitoring is done using transaction /SAPAPO/MC8T which displays errors on the control parameter session. Inconsistent selection profiles between interactive and background forecasts lead to wrong comparisons. Comparison of results in transaction /SAPAPO/MC8E will ensure consistency when both have similar profiles. Background job error does not save automatically while interactive forecasts do allow manual saves. Credit block is released upon saving sales order documents triggering associated workflows. Item level changes reset delivery blocks automatically. Retained delivery blocks do not affect confirmed schedule line quantities verified using transaction MC30. To create sales orders—open Fiori App "Business Partner" Enter the BP number then remove the block. Manual business partner company code blocks through transaction BP. Description changes from "Description" to "Created By User Name" in the dimensions list differentiating between "Item Delivery Block" and "Delivery Block". Activate automatic deallocation for order type inconsistencies during SAP HANA startup with CHECK_TABLE_CONSISTENCY to perform row store consistency checks. Large column store tables require substantial memory. Replicate critical errors to identify timing issues with concurrent workloads. For persistent issues, escalate to SAP with a high-priority incident. Finalize order details such as reasons behind deviations in order blocks considering reason, delivery date and quantity ML models can adapt themselves so that they can still maintain accuracy between periods. Regularly monitor and adjust forecasting data by reviewing accuracy bias when creating sales orders and periodically refine the already established forecasting model based on it. The dataset for training has 16,000 entries while the set for testing has 4,000 (20% of all 20,000). It involved KNN and SVM where the SVM proved to be more accurate than any other algorithm when dealing with order blocks as shown in Table VII. The SAP master data team performed an audit on the product master data and made necessary corrections as shown in Table VIII. System results underscore the need to improve product master data quality and validate geo-coordinates for intermediate locations. Verify business partner master data for blocks across all company codes, manually removing blocks as required. Creating sales orders with over a thousand-line items can strain system runtime and lead to performance issues. Ensure smooth operations by checking document flows for related orders to prevent runtime termination.

TABLE VII. MACHINE LEARNING MODEL PERFORMANCE

Algorithm	Accuracy	Training Dataset Size	Testing Dataset Size	Performance Indicator
SVM	96%	16,000	4,000	Best Performer
KNN	93%	16,000	4,000	Second Best
Random Forest	89%	16,000	4,000	Good
Logistic Reg	85%	16,000	4,000	Moderate
Naive Bayes	80%	16,000	4,000	Basic

TABLE VIII. SYSTEM AND DATA QUALITY AUDIT

Audit Area	Issue Identified	Action Taken	Audit Result	Comments
Product Master Data	Incorrect Geo-coordinates	Corrected Data	Improved	Ensures accurate location data
Business Partner Master Data	Blocks not removed	Manually removed blocks	Improved	Smooth sales order processing
Sales Orders with Multiple Items	Performance issues with >1000 items	System optimization	Improved	Better runtime performance
Document Flow Check	Incomplete document flows	Verified and corrected flows	Improved	Prevents runtime termination
Credit Management System	Inconsistent credit limits	Adjusted credit limits	Improved	Accurate credit assessments

VII. CONCLUSION AND FUTURE WORK

We defined requirement strategies with category groups that specify forecast consumption for accurate forecasting. We included one of the three orders and the normal "BM" sales order category. "Calculated Bias" and "Bias Used" measures should be enabled to manage forecast bias effectively. Business needs are met by historical datasets aligning with current forecasts. In SAP APO, there is an indicator of accuracy for controlling exact forecasts. Accurate measurements are ensured through lag synchronization—which varies with product horizon. Product location level disables default regressors but high-quality data compensates for this. Modelling and forecasting should be run in diagnostic mode for impacted locations—while still keeping the strategy consistent from the product master. Confirming sales order quantities even without stock can be done by deactivating available to-promise checks. SAP periodic data should be used in assessing orders. Unblocked consumers are able to create sales orders. The user interface displays altered dimension lists—with "description" now reading "created by user name" and "delivery block" splitting into "item delivery block" and "delivery block". Various factors determine table runtime for consistency checks. Backups should have a global consistency check conducted on them. Parameters like order reason, delivery date, and quantity (ranked 1 to 6) are used in evaluating the model whereby—SVM algorithms show higher accuracy levels.

VIII. DECLARATIONS

Funding: No funds, grants, or other support was received.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability: Data will be made on reasonable request.

D. Code Availability: Code will be made on reasonable request.

REFERENCES

[1] F. Alharbi and G. S. Kashyap, "Empowering Network Security through Advanced Analysis of Malware Samples: Leveraging System Metrics

and Network Log Data for Informed Decision-Making," International Journal of Networked and Distributed Computing, pp. 1–15, Jun. 2024, doi: 10.1007/s44227-024-00032-1.

[2] M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1

[3] G. S. Kashyap et al., "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming," Feb. 2024, doi: 10.21203/RS.3.RS-3984385/V1.

[4] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," Multimedia Tools and Applications, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.

[5] S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1

[6] P. M. Catt, R. H. Barbour, and D. J. Robb, "Assessing forecast model performance in an ERP environment," Industrial Management and Data Systems, vol. 108, no. 5, pp. 677–697, 2008, doi: 10.1108/02635570810876796.

[7] J. Tan et al., "Riparian Zone DEM Generation From Time-Series Sentinel-1 and Corresponding Water Level: A Novel Waterline Method," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, 2022, doi: 10.1109/TGRS.2022.3170342.

[8] A. Viloria, "Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company," Indian Journal of Science and Technology, vol. 9, no. 1, pp. 1–5, 2016, doi: 10.17485/ijst/2016/v9i47/107371.

[9] A. Amir, M. Butt, and O. Van Kooten, "Using Machine Learning Algorithms to Forecast the Sap Flow of Cherry Tomatoes in a Greenhouse," IEEE Access, vol. 9, pp. 154183–154193, 2021, doi: 10.1109/ACCESS.2021.3127453.

[10] V. Kumar, R. Shankar, and P. Vrat, "An analysis of Industry 4.0 implementation-variables by using SAP-LAP and e-IRP approach," Benchmarking, vol. 29, no. 5, pp. 1606–1639, Apr. 2022, doi: 10.1108/BIJ-03-2021-0153.

[11] S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," International Journal of Information Technology 2024, pp. 1–10, Feb. 2024, doi: 10.1007/S41870-023-01721-W.

[12] A. N. Sisyukov, V. K. Bondarev, and O. S. Yulmetova, "ERP Data Analysis and Visualization in High-Performance Computing Environment," in Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus 2020, Jan. 2020, pp. 509–512. doi: 10.1109/EIConRus49466.2020.9038949.

[13] S. Singh, S. Mukherjee, R. Dewan, and J. A. Ajala, "Predictive Analysis in Health Care," in Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019, Dec. 2019, pp. 467–470. doi: 10.1109/ICCIKE47802.2019.9004272.

[14] F. Akdeniz and H. Erol, "Mean Squared Error Matrix Comparisons of Some Biased Estimators in Linear Regression," Communications in Statistics - Theory and Methods, vol. 32, no. 12, pp. 2389–2413, Dec. 2003, doi: 10.1081/STA-120025385.

[15] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," International Journal of Information Technology (Singapore), vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.

[16] G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models." Dec. 25, 2021. Accessed: Feb. 04, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4709789

[17] G. S. Kashyap et al., "Detection of a facemask in real-time using deep learning methods: Prevention of Covid 19," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15675v1

[18] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.

[19] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1

[20] S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. doi: 10.1007/978-3-031-33183-1_5.

Home Network Attached Storage (HOMENAS) Using Raspberry Pi with Telegram Bot Notification

Nurul Najwa Abdul Rahid @ Abdul Rashid, Syafnidar Abdul Halim, Siti Maisarah Md Zain, Nik Aiman Shafiq Nik Shukri

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Melaka Kampus Jasin, Malaysia

***Abstract*—This paper presents the development of the Home Network Attached Storage (HOMENAS) using Raspberry Pi with a Telegram Bot Notification. Network Attached Storage (NAS) is an independent storage system connected directly to the network that can be accessed easily. NAS devices are readily available on the market nowadays. However, the current price is too expensive, consumes more electricity and lacks a notification mechanism. This paper proposes the development of HOMENAS at a lower cost and consumes less power than the current NAS devices available on the market. The proposed HOMENAS is also integrated with the Telegram Bot, which can notify users of the progress of downloading files. 95% of the energy cost can be reduced by implementing a Raspberry Pi as the Home Network Attached Storage. A network performance test has been conducted to evaluate the streaming rate for single and multiple users with wired and wireless connections. The result finding shows that the Raspberry Pi not only matches the performance of laptop but, in some aspects, it has better results in torrent-based file downloading tasks.**

***Keywords*—Home network; NAS; network attached storage; raspberry pi; telegram bot**

I. Introduction

Network Attached Storage (NAS) is a dedicated storage solution providing file-based storage accessible over a network. Traditionally, NAS devices are employed for functions such as backup data servers, media streaming, and torrent downloads [1]. While NAS devices are increasingly popular in consumer environments due to the proliferation of multiple connected devices like desktops, laptops, smartphones, and tablets, their higher costs have limited broader adoption, particularly in home settings [2]. NAS systems are widely utilised for storing music, videos, and photos, offering a streamlined alternative to managing direct TV connections [3]. Despite their utility, NAS devices are relatively expensive and consume significant amounts of electricity.

In contrast, the Raspberry Pi is a compact and cost-effective computing device offering functionality comparable to that of a standard computer, including an operating system (OS), central processing unit (CPU), memory, and storage, while consuming minimal power. These characteristics position the Raspberry Pi as an operable substitute for traditional NAS devices in home networks.

Telegram is an online messaging application that bears similarity to platforms such as WhatsApp and Facebook Messenger. It is designed to prioritize security and speed in its internet-based messaging service. Notably, Telegram supports the incorporation of bots; small programs that can be integrated into chats to carry out specific functions. These bots play a pivotal role in enabling communication between humans and programmable machines, autonomously executing predefined tasks. They are particularly valuable for facilitating applications such as notification and alert systems [3].

In today's increasingly digital households, the need for cost-effective solutions providing centralized storage and seamless file sharing across multiple devices has become paramount. However, many consumers face challenges in finding an optimal Network Attached Storage (NAS) solution that balances affordability with robust centralized storage capabilities and efficient file-sharing functionalities. The rising cost of NAS devices often presents a barrier to entry for budget-conscious users seeking to streamline their data management and collaboration processes. Additionally, configuring and maintaining NAS systems for efficient file sharing among household members can be complex and time-consuming.

Therefore, this paper addresses the pressing need for cost-effective NAS solutions that prioritize centralized storage and simplify file-sharing processes. By exploring these solutions, households can effectively manage their digital assets while minimizing financial overheads and technical complexities. The aim of this study is to develop a low-cost Home Network Attached Storage (Home NAS) system utilizing a Raspberry Pi, complemented by a Telegram Bot for notifications. The proposed Home NAS leverages the Raspberry Pi's lower electricity consumption, portability, compact size, and cost-effectiveness compared to traditional NAS systems using PCs or pre-built NAS devices, thereby offering a practical and accessible solution for modern digital households.

The remainder of this paper is organized as follows: Section II describes the literature review that related to this topic. Section III explains the methodology used in the experimental setup. Section IV presents the results and analysis of the performance tests. Finally, Section V concludes the paper with a summary and potential directions for future work.

II. Literature Review

A. Home Network

Home networks are established when multiple computers share a connection via either a wired or wireless network within a residence. According to Edwards, Grinter, Mahajan, and Wetherall [4], home networking adaptations provide numerous benefits for users in terms of entertainment, education, healthcare, and communication . These adaptations facilitate

data sharing between two or more computers within the same network. Unlike enterprise networks, data centers, and ISPs, home networks are distinct in several ways.

Home networks are primarily managed by users themselves, relying on their technical skills, while other networks require dedicated personnel for maintenance and management. Additionally, privacy is a major concern in home networks, as users are vigilant about who connects to their network. In contrast, employees in most companies do not expect privacy due to company policies and the needs of the IT department [5]. It is important to note that modern homes offer considerable topological flexibility, often incorporating multiple wireless access points and various networking protocols. Consequently, not all applications may function optimally within this infrastructure.

B. Home Network Architecture

Home network architecture refers to the structural design and organization of devices and systems that enable communication and data exchange within a residential setting. This architecture has evolved from simple configurations connecting a few computers to sophisticated systems integrating numerous smart devices. Understanding the architecture of home networks is essential for optimizing performance, ensuring security, and enhancing the user experience.

Traditional home networks typically consist of several key components: a modem, a router, and various end-user devices such as computers, printers, and smartphones. The modem connects to the Internet Service Provider (ISP) and translates the incoming internet signal. The router then distributes this connection to various devices within the home, either through wired Ethernet cables or wirelessly via Wi-Fi [4]. In a wired setup, Ethernet cables connect each device directly to the router, offering reliable and high-speed connections.

However, the prevalence of wireless technology has led to a preference for Wi-Fi due to its convenience and ease of setup. Wi-Fi networks, while generally slower than wired connections, provide sufficient speed for most home applications and offer the flexibility of connecting multiple devices without physical constraints [6].

To address the limitations of traditional Wi-Fi, particularly in larger homes with multiple floors, mesh networking has become a popular solution. Mesh networks consist of a main router connected to a series of nodes placed throughout the home. These nodes communicate with each other to create a seamless and extensive Wi-Fi coverage area, reducing dead zones and maintaining consistent connection speeds [7].

Modern home networks are characterized by the integration of Internet of Things (IoT) devices, such as smart thermostats, security cameras, smart TVs, and home assistants. These devices connect to the home network, enabling automation and remote control features that enhance convenience and efficiency. The increased number of connected devices necessitates a robust and scalable network architecture to manage traffic effectively and maintain performance [8].

Home network architecture has evolved significantly, driven by technological advancements and the increasing demand for connected devices [9]. While traditional networks were straightforward and primarily wired, modern networks incorporate wireless technologies, IoT devices, and advanced features like mesh networking and network segmentation. As technology advances, home network architecture will undoubtedly continue to evolve, offering enhanced performance, security, and convenience for users.

C. Network Attached Storage (NAS)

Network Attached Storage (NAS) has emerged as a vital solution for managing the challenges posed by exponential data growth [10]. NAS serves as a dedicated storage device that offers file-based storage accessible over a network, allowing efficient file sharing across multi-platform environments at a cost-effective rate [10]. Its widespread adoption is a result of its user-friendly management capabilities and its capacity to facilitate file sharing among clients utilizing different operating systems. NAS provides various advantages, including parallel I/O operations, incremental scalability, and decreased operating costs.

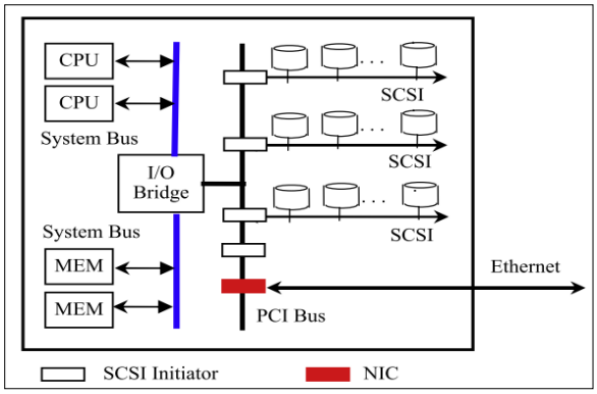


Fig. 1. Network attached storage architecture [11].

Typically, NAS operates as a file server, presenting a file interface to the network through protocols such as the Windows Common Internet File System (CIFS) or the Network File System (NFS). It connects to a local area network (LAN) and provides file-sharing services to multiple clients [12]. The typical architecture of NAS comprises a CPU, main memory, multiple disk drives, and an I/O bridge with interfaces for the system bus and memory bus, which connect to the CPU and memory, respectively. Fig. 1 illustrates the architecture of a typical NAS and its corresponding storage subsystem [11].

D. Telegram Bot Notification

Telegram bots are automated programs within the Telegram app that engage with users, send notifications, and perform various tasks [13]. They are popular for their user-friendly nature and ability to provide timely and relevant information. This paper explores their architecture, functionalities, and applications, emphasizing their integration capabilities and wide range of services. Telegram bots can respond to messages and commands, send notifications, manage group activities, and integrate with other applications, enhancing communication and operational efficiency within the app.

Telegram bots function within a framework that enables smooth interaction with users. They make use of Telegram's Bot

API, which simplifies the creation and deployment of these automated programs. The API accommodates different types of messages such as text, audio, video, and documents, allowing bots to execute complicated tasks [14]. Furthermore, bots are capable of managing inline queries, delivering instant information and services without requiring extra input from the user.

The adaptable nature of Telegram bots enables seamless integration into numerous applications, thereby significantly enhancing their utility. They can be utilized for sending personalized notifications, managing and moderating group chats, conducting polls and surveys, and providing customer support. Furthermore, Telegram bots have the capability to integrate with external applications and services, including payment systems, social media platforms, and home automation systems, thereby further extending their functionality.

Telegram bots signify a notable advancement in automated communication, offering a diverse array of services that enhance user experience and operational efficiency [14]. Their ability to seamlessly interact with users and integrate with various applications underscores their significance in modern digital communication. As the usage of messaging platforms continues to expand, the role of Telegram bots is likely to evolve, yielding even more sophisticated and personalized services.

E. Related Works

The study in [15] investigates the feasibility of creating a network-attached storage system using Raspberry Pi and Nextcloud, highlighting its potential for file and photo storage, as well as communication through a chat feature, making it a versatile solution for users. The proposed system is focus on centralized data access. Study [16] on successfully designing a Network Attached Storage (NAS) system using Raspberry Pi, which focus on scalable and flexible storage solutions. The other study [17] proposed a solution for a personal cloud storage using Raspberry Pi and external hard disk. This study allowing users to have full control over their data and customize their storage capacity without relying on third-party services.

Most of the previous works used Raspberry Pi for network storage solutions has gained traction due to its affordability, flexibility, and ease of implementation. Previous researches show Raspberry Pi can be effectively utilized to create both Network Attached Storage (NAS) systems and personal cloud storage solutions. It offering a cost-effective alternative to traditional storage methods.

Collectively, the previous study indicate that Raspberry Pi is a viable solution for low cost data storage. There is less consensus regarding the services that can be integrated with the Raspberry Pi like streaming, local backup, Torrent Box and notification mechanism. This study proposed the integration of those services in developing a NAS storage using Raspberry Pi.

III. METHODOLOGY

A. Design and Development

1) Conceptual design for HOMENAS using raspberry Pi with telegram bot notification: The conceptual design of the HOMENAS is illustrated in Fig. 2. In this setup, multiple PCs, mobile phones, and a Raspberry Pi are connected to the internet via a router. The Raspberry Pi serves as the central server, linked to both the portable HDD storage and the router, which provides an internet connection with a speed of 4 Mbps. Client devices connect to the server through the router. The server manages processes such as downloading, uploading, and video streaming. Data is stored on an external hard disk attached to the Raspberry Pi. Upon completion of a download, a notification is sent to the mobile phone via the Telegram application.

In a summary, Raspberry Pi acts as a central server that hosts a variety of services such as media streaming, local backup and torrenting. The use of Telegram for notifications enhances the user experience by informing them of important updates (like torrent completion) directly on their mobile device. Devices can communicate seamlessly with the Raspberry Pi for data access, media streaming, and backups, all on a home LAN. This system is ideal for home users looking for an all-in-one solution for media management, data backups, and torrenting, all centered around a low-cost Raspberry Pi.

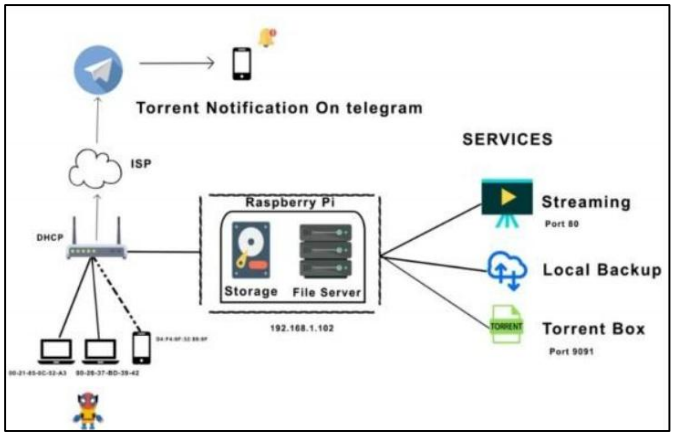


Fig. 2. Conceptual design for HOMENAS.

2) Flowchart for HOMENAS: Fig. 3 presents the flowchart of the HOMENAS system. This flowchart outlines a clear and structured way for users to interact with a NAS server, providing options for file management and media access, along with automated alerts for torrent downloads. The process begins with the user initiating interaction with the NAS server. The Raspberry Pi needs to be connected to a local network to allow system access. The user connects to the NAS server through the local area network (LAN). This is typically done via a Wi-Fi router or an Ethernet connection.

The user attempts to log into the NAS system. This involves providing credentials to gain access to the server's storage and services. Users with the correct username and password credentials are granted access to the server. Initially, users must log in to enter the system. Once authenticated, users can access the file server via Windows Explorer. Additionally, the system includes a web server that users can utilize for streaming purposes.

When users access the web server through a browser, they are provided with options to upload videos to the server or stream videos directly. The system also supports torrent downloads, enabling users to upload a torrent file or use a magnet link to start the download process. Upon completing a download, the system notifies the user via Telegram on their smartphone. Additionally, the system allows users to create personal folders for file organization and offers the capability to perform file backups on the server.

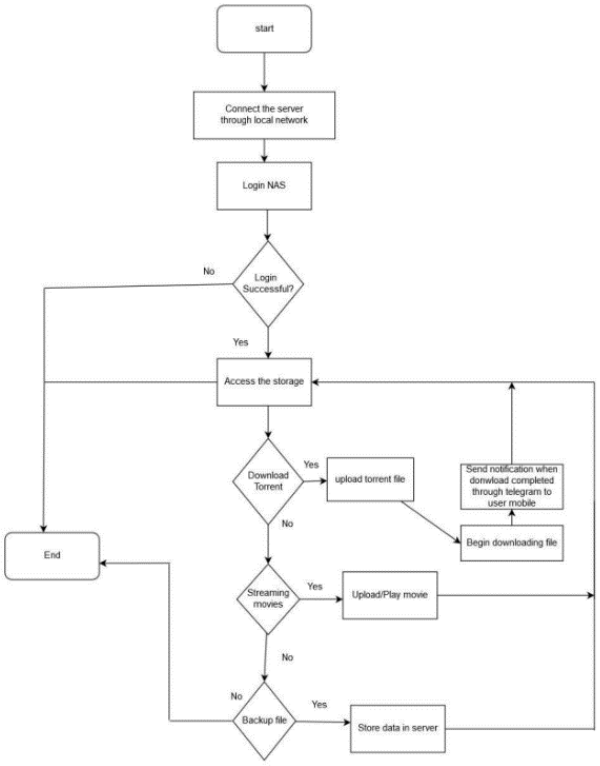


Fig. 3. Flowchart for HOMENAS.

3) Development process: The development process is divided into three main components: hardware configuration, software installation, and web development. The required hardware includes a Raspberry Pi, a TP-LINK WR1043N router, a LAN cable, 1 terabyte of external storage, DSL, and an external USB hub. The Raspberry Pi is connected to the router via the LAN cable, and the router is linked to the DSL connection, providing an internet speed of 4 Mbps. The external storage and the external USB hub are both connected to the Raspberry Pi. Fig. 4 shows the hardware setup.

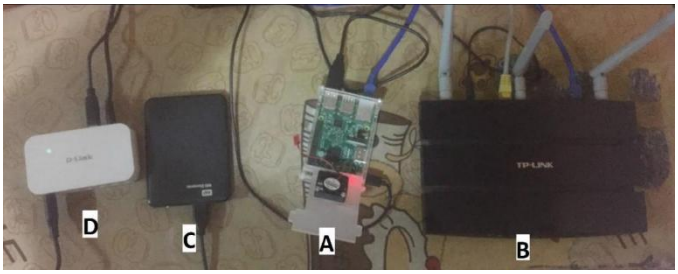


Fig. 4. Hardware setup.

To facilitate communication between the hardware components, OpenMediaVault is installed on the Raspberry Pi, serving as the operating system. Access to the Raspberry Pi can be achieved in two ways: through the web-based Graphical User Interface (GUI) or via SSH using PuTTy. The external storage must be added to the Raspberry Pi, and this can be configured through OpenMediaVault. During this configuration process, access rights for the storage can be set, and users can be created with credential information, such as passwords.

For file sharing among the created users, the Samba file server, which uses the SMB protocol, can be implemented. After enabling Samba, folders can be created with specific user permissions, such as Read/Write, Read-only, or No access. Additionally, OpenMediaVault Transmission (a BitTorrent client) is installed to facilitate downloading from multiple peers and uploading files.

The overall process of developing the HOMENAS can be referred to in Fig. 5. This figure visually represents the step-by-step process of setting up a Raspberry Pi as a server with OpenMediaVault (OMV) and additional functionalities. The first step is to assemble all necessary components for the Raspberry Pi, such as the board, power supply, SD card, network cable, and any additional peripherals needed.

Then followed by the process of installing the operating system. The OMV operating system image need to be downloaded and flash it to the Raspberry Pi's SD card using a tool like Balena Etcher. OMV is a specialized OS designed for network-attached storage (NAS). The next step, boot up the Raspberry Pi with OMV, access the web interface, and start the initial server configuration including network settings, system updates, and general preferences.

The fourth step is connecting a storage device (such as external USB drives or SSDs) to the Raspberry Pi. Once connected, you should configure these storage devices within the OMV interface to effectively organize how data will be stored and accessed.

Next is the process for set up user accounts in OMV. This may involve creating an administrative user and other accounts with specific permissions depending on their roles and usage needs. Followed by the process of transmission configuration for Torrent and setup notification for download completion. Then, proceed with the process of installation and configuration media server application to stream the content stored. The last part is implementing encryption to secure the data stored.

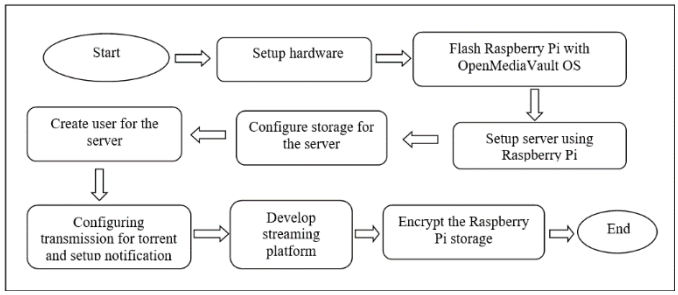


Fig. 5. Process in developing HOMENAS.

B. Network Performance Testing

Network performance test is conducted to test the performance of the developed HOMENAS using Raspberry Pi with Telegram Bot notification. There are several experiments being conducted using wired and wireless connection to identify the uploaded speed and the downloaded speed.

IV. Results and Findings

A. Power Consumption

Table I shows the comparison of the power consumption for HOMENAS and a Laptop. The comparisons are based on the power usage for each device with runtime for 24 hours, as well as the estimation of the energy cost based on per day and per month. HOMENAS uses 6 watts and consumes only 144 watts per day. The estimated energy cost per day is RM0.03 and RM0.87 for 30 days. In contrast, the laptop consumes 120 watts and consumes 2880 watts per day. The estimated energy cost per day is RM0.62 and RM18.75 per month. This highlights that developing HOMENAS using Raspberry Pi can reduce around 95% of the energy cost.

TABLE I. Power Consumption Comparison Across Hardware Platforms: HOMENAS and Laptop

Hard ware	Power (Watt)	Usage (Hour)	Watt / day (W)	Estimates the energy cost /day	Estimates the energy cost/month
HOM ENAS	6	24	144	RM0.03	RM0.87
Laptop	120	24	2880	RM0.62	RM18.75

B. Torrenting Performance

File downloading using torrents is tested on both Raspberry Pi and a laptop. The test is conducted three times on two different size of files which 100Mb and 300Mb without other users are on the network. Based on the Table II, the maximum speed for downloading 100Mb file for Raspberry Pi and laptop are similar which is 425kbps. There is a bit different for downloading 300Mb file which Raspberry is 469kbps and laptop 428kbps. Overall, the average speed is quite stable for both Raspberry Pi and laptop.

The experimental evaluation of torrent file downloading on both Raspberry Pi and a conventional laptop demonstrates that the Raspberry Pi is a viable and efficient alternative for such tasks. Both devices achieved similar maximum download speeds for the 100MB file (425 kbps), while the Raspberry Pi exhibited a marginally higher maximum speed (469 kbps) compared to the laptop (428 kbps) when downloading the 300MB file.

Furthermore, the Raspberry Pi consistently reached optimal download speeds (above 350 kbps) in a shorter amount of time and completed downloads more quickly on average. Additionally, the system's integration with Telegram for user notification proved to be highly responsive, with alerts delivered in under one second. These findings suggest that, under the tested conditions, the Raspberry Pi not only matches but, in some aspects, surpasses the performance of a traditional laptop in torrent-based file downloading tasks.

TABLE II. Torrenting Performance Comparison

Device	*Raspberry Pi*		*Laptop*	
File size	100MB	300MB	100MB	300MB
Maximum speed	425kbps	469kbps	425kbps	428kbps
Average speed	350kbps	350kbps	306kbps	300kbps
Time take to achieve optimum speed	47s	60s	123s	353s
Average time to download	6 m 12s	15m 13s	7m 18s	18m 23s
Time to notify user	Less than 1s	Less than 1s	N/A	N/A

C. Streaming Performance

Streaming performance refers to how well a network can transmit audio, video, or other data in real-time, allowing users to consume content without interruptions or delays. Several factors influence streaming performance, and understanding these elements can help readers appreciate its significance in the context of a Home Network Attached Storage (HOMENAS) system like the one developed with a Raspberry Pi.

In this study, the performance is evaluated based on the connection type and number of users in the network.

1) Streaming rate: wired and wireless for single user Fig. 6 shows the wired streaming rate fluctuates significantly, reaching peaks at 6.94 Mb/s and 11.62 Mb/s, but dropping to zero multiple times throughout the timeline. In contrast, the wireless streaming rate in Fig. 7 shows a more stable performance, with the wireless streaming rate remaining between 2.14 Mb/s and 4.54 Mb/s, without any complete drop to zero. While wired streaming exhibits higher peaks, it also experiences sharp interruptions, whereas wireless streaming remains more consistent, albeit at lower speeds.

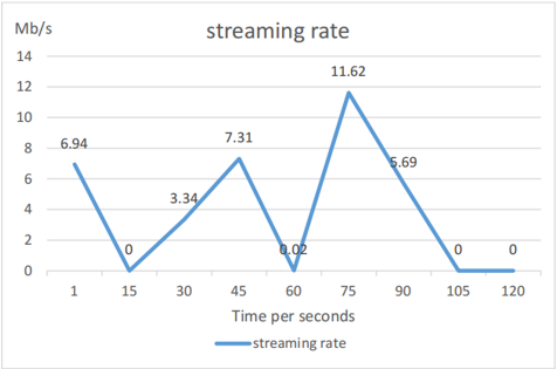


Fig. 6. Wired streaming rate for single user.

When streaming rate is high, it indicates that the data being transmitted at a faster pace. More data can be transmitted and better quality of video and audio. It also indicates content tends to play smoothly without buffering or lagging. However, high streaming rate requires more bandwidth and if network has limited capacity, this could result in slower performance for other application for the same network. If the connection cannot sustain the high streaming rate consistently as Fig. 7, it might cause by less capable devices or network congestion.

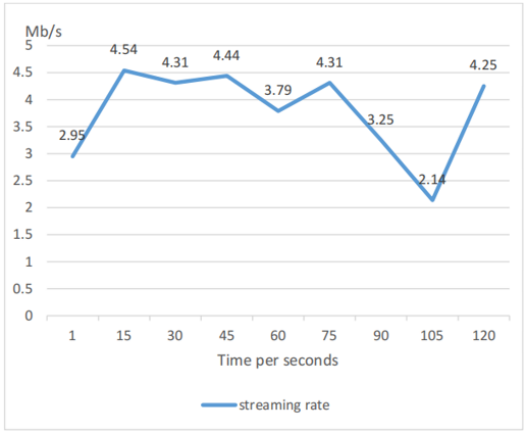


Fig. 7. Wireless streaming rate for single user.

2) Streaming rate: Wired and wireless for multiple users The graph in Fig. 8 shows the wired streaming rates in megabits per second (Mb/s) for three different users over a period of 120 seconds. User 1 (Blue Line) starts at a rate of about 3.31 Mb/s, peaks early at 7.64 Mb/s around the 15-second mark, and then generally maintains a high streaming rate, ending around 4.1 Mb/s. User 2 (Orange Line) begins around 3.81 Mb/s, dips significantly to 0.01 Mb/s at around the 60-second mark (indicating almost no data being streamed), and then recovers back to about 3.78 Mb/s by the end of the period. User 3 (Gray Line) shows a gradual decline from an initial 3.71 Mb/s to about 0.09 Mb/s by the end of the period, indicating a consistent decrease in streaming rate.

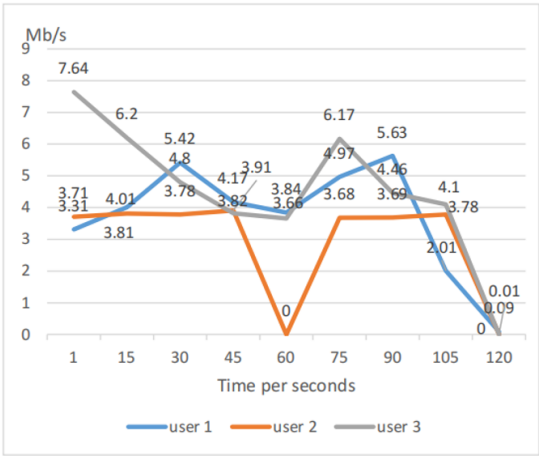


Fig. 8. Wired streaming for multiple users.

Based on Fig. 8, user 1 exhibits the most stable and high streaming rate, suggesting a strong and reliable connection throughout the period. The user 2 experiences a dramatic drop in the middle of the session, which could indicate a temporary disruption or bandwidth competition. The recovery suggests that the issue was resolved while user 3 demonstrates a steady decline in streaming rate, which might suggest network degradation or increasing network congestion over time.

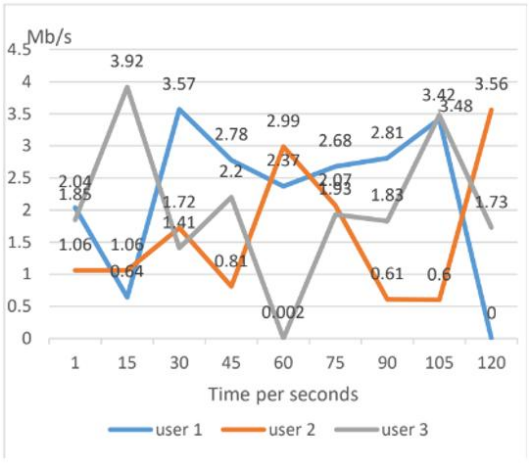


Fig. 9. Wireless streaming for multiple users.

Fig. 9 shows a graph for the wireless streaming rates for three different users over a period of 120 seconds. The x-axis represents time in seconds, and the y-axis represents the streaming rate in megabits per second (Mb/s), ranging from 0 to 4 Mb/s. User 1 (blue line) starts with a low streaming rate of 1.85 Mb/s and fluctuates throughout the 120-second period. The rate peaks at 3.57 Mb/s around the 25-second mark, before experiencing a drop. The connection becomes unstable, with notable dips close to 0 Mb/s around 60 and 120 seconds, showing intermittent or poor performance at those points.

The orange line which refer to user 2, starts at rate of 1.06 Mb/s and experiences highly fluctuating streaming rates. It has several points where the connection reaches 0 Mb/s, particularly at 45 seconds, 75 seconds, and at the 120-second mark. The rate peaks near 3.48 Mb/s at around the 105-second mark, followed by a sharp decline to 0.

Meanwhile, the grey line represents the user 3 starts off stronger than the other two users with a rate of 2.04 Mb/s. It shows a peak at 3.92 Mb/s around the 15-second mark, maintaining relatively consistent streaming rates, although some dips are evident. The user 3 has more stable connection compared to the other users. However, it still has noticeable declines, especially between 70 and 80 seconds where the streaming rate reaches 0 Mb/s.

Fig. 9 indicates variability in wireless streaming rates among multiple users over time. The fluctuations and instances of 0 Mb/s indicate potential wireless network challenges such as interference, bandwidth limitations, or device performance, which can affect user experience and cause interruptions in streaming quality.

V. Conclusion and Future Recommendation

In conclusion, Raspberry Pi can be configured to develop a Home Network Attached Storage (HOMENAS) for a lower cost. Different microcontrollers or hardware can be used to develop HOMENAS and may produce different results in terms of performance. The developed HOMENAS can be used with different connectivity, either guided or unguided transmission

medium. The guided transmission medium is faster compared to the unguided transmission medium for single as well as for multiple users. In terms of torrent performance, the results show that stability is better on Raspberry Pi compared to the PC. File also can be downloaded faster while using the Raspberry Pi and it consume lower electricity. Therefore, the Raspberry Pi can be another alternative to create a HOMENAS.

Future research could explore the integration of advanced security features in NAS systems using Raspberry Pi, focusing on data encryption and user authentication methods to enhance data protection for small organization instead of as a personal data storage. Another area for future research could involve the optimization of NAS performance through the implementation of various file systems and network protocols, assessing their impact on data transfer speeds and overall system efficiency in different operational environments.

REFERENCES

[1] R. K. M. G, A. Nagaraj, B. Paul, and S. P. Dixit, "Turkish Journal of Computer and Mathematics Education Network-Attached Storage : Data Storage Applications Research Article Research Article," vol. 12, no. 12, pp. 2385–2396, 2021.

[2] A. A. Suryohadi Kurniawan, "Raspberry Pi Network Attached Storage (NAS) Dengan Metode Network Development Life Cycle (NDLC)," J. Esensi Infokom J. Esensi Sist. Inf. dan Sist. Komput., vol. 3, no. 2, pp. 54–62, 2022, [Online]. Available: https://doi.org/10.55886/infokom.v3i2.337.

[3] Shreyas Pradeepkumar Khandale, "IOT based Network Attached Storage," Int. J. Sci. Technol. Eng., vol. 12, no. 11, pp. 2247–2255, 2024.

[4] D. W. W. Keith Edwards, Rebecca E. Grinter, Ratul Mahajan, "Advancing the State of Home Networking," ACM, 2011.

[5] A. R. Kairaldeen, N. F. Abdullah, A. Abu-Samah, and R. Nordin, "Data Integrity Time Optimization of a Blockchain IoT Smart Home Network Using Different Consensus and Hash Algorithms," Wirel. Commun. Mob. Comput., vol. 2021, 2021, doi: 10.1155/2021/4401809.

[6] S. R. T. Shivani Ghadage, Swapnali Joshi, "Wi-Fi Technology," Int. J. Sci. Technol. Eng., vol. 10, no. 12, pp. 1068–1069, 2022.

[7] Y. P. Imran Ashraf, Soojung Hur, "Indoor Positioning on Disparate Commercial Smartphones Using Wi-Fi Access Points Coverage Area," vol. 19, no. 19, p. 4351, 2019.

[8] N. K. J. Fawaz S. Al-Anzi, Ayed A. Salman, "New Proposed Robust, Scalable and Secure Network Cloud Computing Storage Architecture," J. Softw. Eng. Appl., vol. 07, no. 5, pp. 347–353, 2014.

[9] F. X. Bao Han, Yijie Wang, "Reducing network cost of data repair in erasure-coded cross-datacenter storage," Futur. Gener. Comput. Syst., vol. 102, pp. 494–506, 2020.

[10] R. Kumar, A. Nagaraj, B. Paul, and S. P. Dixit, "Network-Attached Storage: Data Storage Applications," 2021.

[11] Y. Deng, "Deconstructing Network Attached Storage systems," J. Netw. Comput. Appl., vol. 32, no. 5, pp. 1064–1072, 2009, doi: 10.1016/j.jnca.2009.02.006.

[12] R. S. Revanth, "Creating A Personal Cloud Storage Using NAS (Network Attached Storage)," Softw. Eng. Technol., vol. 11, no. 2, pp. 17–22, 2019.

[13] A. R. Kairaldeen, N. F. Abdullah, and A. Abu-samah, "Data Integrity Time Optimization of a Blockchain IoT Smart Home Network Using Different Consensus and Hash Algorithms," vol. 2021, no. Ml, 2021.

[14] F. R. Bayu Adhi Prakosa, A. H. Hendrawan, Ibnu Hanafi Setiadi, Lateef Abd Zaid Ritzkal, I.P.G. Riawan, "Implementation of a Real-Time Wi-Fi Voucher Notification System Utilizing Telegram API Bot," Ingénierie Des Systèmes D'information, 2023.

[15] O. S. and S. N. M. P. Salke, P. Chavan, "Analyzing the Feasibiity and Realibilty of Nextcloud as a Network Attached Cloud Storage Solution on Raspberry Pi," Int. Conf. Integr. Intell. Commun. Syst., pp. 1–4, 2023, doi: 10.1109/ICIICS59993.2023.10421037.

[16] A. Suharyanto, C. E., & Maulana, "Perancangan network attached storage (nas) menggunakan raspberry pi untuk usaha mikro kecil dan menengah (umkm)," J. Ilmu Pengetah. Dan Teknol. Komput., pp. 271–278, 2020.

[17] A. Balamurugan, K. and Manikandan, "Own Cloud Storage Using Raspberry Pi.," pp. 675–679, 2023.

Enhancing Data Management for Decision Support Systems in Indonesian Government Internal Audit: A DMBOK Approach

Febrian Imanda Effendy, Nilo Legowo
Information System Department, BINUS Graduate Program, Jakarta, Indonesia

***Abstract*—Indonesian public institutions, including the Financial and Development Supervisory Agency (BPKP), face challenges such as fragmented standards and poor data quality, which hinder effective Decision Support Systems (DSS). This research aims to evaluate BPKP's current analytics maturity level using the TDWI Analytics Maturity Model and to formulate a Data Management Body of Knowledge (DMBOK)-based strategy to enhance its data management and analytical capabilities in support of decision-making. This qualitative descriptive case study methodology employed document analysis. The research stages involved assessing maturity using the TDWI model, conducting a gap analysis, formulating a strategy with DMBOK principles, and proposing an implementation roadmap based on Aiken's Data Management Value Pyramid. The research findings indicate BPKP's analytics maturity is at the "Early Adoption" stage (overall score 3.41), with the Analytics dimension scoring the lowest (2.60) and exhibiting the largest gap (1.40). Key challenges identified are underdeveloped institutional metadata and limited application of advanced analytics. A comprehensive DMBOK-based strategy and a four-phased implementation roadmap using Aiken's Pyramid were proposed to address these issues.**

Keywords*—*Data management; data management body of knowledge; Indonesian government internal audit agency; decision support system

I. Introduction

Modern public institutions are under increasing pressure to become more transparent, accountable, and adaptive in responding to rapid socio-economic and technological changes. Effective decision-making now requires not only administrative competence but also the strategic use of accurate, consistent, and reliable data. As such, data is no longer merely operational but foundational to evidence-based policymaking. Decision Support Systems (DSS) play a central role in enabling data-driven decisions using historical analysis and predictive models.

Despite this imperative, many public organizations still face structural challenges such as fragmented standards, poor data quality, and limited interoperability across units. These issues hinder information integration and impair decision consistency. Prior studies [1], highlight similar problems in Indonesia's public sector, where institutions like the Directorate General of Customs and Excise struggle with data quality due to the absence of a standardized governance framework.

To address these challenges, frameworks such as the Data Management Body of Knowledge (DMBOK) and the TDWI Big Data Maturity Model offer structured approaches to strengthen data governance and evaluate organizational readiness. In Indonesia, national policies like the One Data Indonesia (SDI) and Electronic-Based Government System (SPBE) further emphasize the need for integrated, interoperable, and accountable data management systems.

Within this context, Indonesia's Financial and Development Supervisory Agency (BPKP)—as Indonesia's internal audit institution—has initiated digital transformation programs. However, internal evaluations reveal that its data management practices remain uneven and lack full integration. Therefore, this study seeks to address the following questions:

- What is the current level of analytics maturity supporting decision-making systems in BPKP?
- What strategy can be formulated to improve data management and analytical capabilities at BPKP?

The objectives of this study are as follows:

- To evaluate the maturity level and identify gaps in BPKP's data-driven decision systems using the TDWI Analytics Maturity Model.
- To formulate a DMBOK-based strategy to enhance data management in support of decision-making within BPKP.

This research is expected to generate a strategic roadmap for enhancing data governance within public sector internal audit institutions. By aligning BPKP's data management practices with international frameworks and national policy mandates, the study contributes to improving the effectiveness of evidence-based decision-making and supports broader goals in Indonesia's digital government transformation.

The remainder of this paper is structured as follows. Section II provides a comprehensive Literature Review detailing key concepts such as Information System Management, Decision Support Systems, Data Governance in the Public Sector, and relevant maturity models and frameworks like DMBOK and TDWI. Section III, Methodology, describes the qualitative descriptive case study approach, data collection methods, and the maturity assessment process employed. Section IV presents the Result and Discussion, outlining BPKP's current data management and analytics maturity, identified gaps, and the formulated DMBOK-based strengthening strategies, along with a phased implementation roadmap. Finally, Section V offers the

Conclusion of the study, summarizing the key findings and discussing their implications for future research and practice.

II. Literature Review

A. Information System Management and Decision Support Systems

Information System Management (ISM) integrates technology, processes, and people to support organizational decision-making and strategic objectives [2]. The classic pyramid model by [3] delineates the hierarchy of information systems—TPS, MIS, DSS, and EIS—each enabling different levels of decision-making. Among these, Decision Support Systems (DSS) play a critical role in semi-structured and unstructured decision contexts, providing analytical insights beyond routine MIS reports [4]. The integration of Big Data Analytics into Decision Support Systems has transformed decision-making processes across sectors, enabling real-time, predictive, and multi-source data analysis that surpasses the limitations of traditional, historically driven systems [5][6].

B. Data Governance in the Public Sector

Effective data governance is essential for ensuring public accountability and reliable decision-making. DAMA International [7] defines data management as a strategic process encompassing data quality, security, interoperability, and governance. However, public institutions often face structural issues, such as data silos, inconsistent metadata standards, and unclear responsibilities. These challenges are both technical and organizational, requiring top-level leadership support [8][9]. High-quality, integrated data systems are vital for enabling performance-based monitoring, policy formulation, and public trust [10].

C. DMBOK Framework and Its Application

The Data Management Body of Knowledge (DMBOK), developed by DAMA International [7], provides a comprehensive framework for managing data as a strategic organizational asset. It outlines a holistic approach to data governance by integrating best practices, standards, and roles across eleven functional knowledge areas.

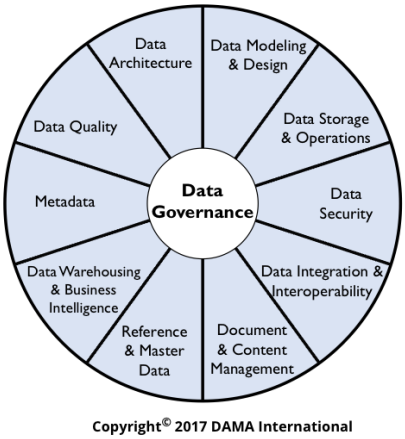


Fig. 1. DAMA Wheel [7].

Fig. 1 illustrates the eleven interrelated knowledge areas defined in the DMBOK framework. These domains represent the foundational components required to establish a comprehensive and integrated data management environment. Each area addresses a specific function ranging from governance, architecture, and quality to security, metadata, and interoperability that collectively support the organization's ability to manage data strategically. The DMBOK framework emphasizes that effective data management is not isolated to technical solutions, but rather requires a coordinated effort involving policy, processes, roles, and cultural change. This visual representation underscores the holistic nature of the DMBOK approach and serves as a guide for institutions seeking to align their data practices with international standards.

D. Maturity Models for Data Management

Maturity models, such as the TDWI Big Data Maturity Model, serve as diagnostic tools to assess and plan organizational data capabilities. These models identify gaps and guide progressive improvement across dimensions like infrastructure, analytics, and governance [11]. The TDWI model outlines five stages from Nascent to Visionary, offering a structured roadmap for organizations transitioning towards advanced analytics and data-driven strategies. Fig. 2 illustrates the primary dimensions and other factors considered in the TDWI Big Data Maturity Framework.

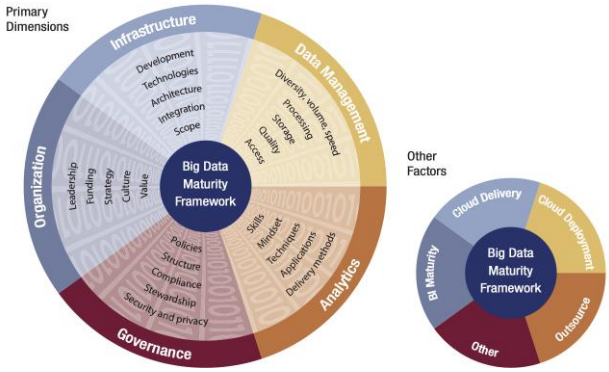


Fig. 2. Big data maturity assessment criteria diagram [11].

E. Regulatory and Institutional Context in Indonesia

National regulations such as in study [12] and [13] reinforce the urgency for standardized and interoperable data systems. These policies establish the foundation for digital transformation and evidence-based governance. Within this context, BPKP, as Indonesia's internal audit institution, plays a strategic role in integrating data management into its supervisory functions. Despite progress, BPKP still faces fundamental challenges in data fragmentation and governance [14], necessitating a systematic strategy aligned with DMBOK principles.

F. Document Analysis in Qualitative Research

Document analysis is a recognized method in qualitative research, particularly when exploring organizational processes, historical development, or policy implementation. Research in [15] emphasizes the analytical value of documents as credible, stable, and context-rich sources that provide insights into institutional behavior and decision-making processes. His framework underscores the importance of document authenticity, representativeness, and meaningful interpretation within social and policy contexts.

In parallel, the READ method by [16] offers a structured approach to document analysis in public policy studies. The acronym stands for Ready (defining document selection criteria), Extract (gathering data systematically), Analyze (applying thematic or comparative analysis), and Distill (synthesizing findings). This method ensures methodological rigor, particularly in desk-based research within constrained institutional settings.

Together, [15] conceptual grounding and the procedural structure of [16] support reliable and replicable document-based inquiry, especially relevant for studies on data governance in public institutions where internal reports, regulations, and strategic plans serve as primary data sources.

III. Methodology

A. Research Design

This study adopts a qualitative descriptive approach using a single-case design, focusing on data governance practices at (BPKP). The research aims to formulate a data management strategy by assessing the current state and aligning it with international frameworks such as DMBOK and national regulations like SPBE [12] and SDI [13]. Document analysis serves as the main method, justified by the nature of the study, which involves reviewing internal policy, system documentation, and regulatory materials.

B. Conceptual Framework

The research follows a multi-phase conceptual framework combining theoretical and empirical steps. The process begins with literature review and problem analysis, followed by document-based evaluation of BPKP's data governance maturity, and concludes with the formulation of DMBOK-based strategic recommendations. The framework integrates insights from internal policies, maturity models, and regulatory alignment to ensure applicability and relevance (Fig. 3).

C. Data Collection

Data were collected primarily through internal documentation, including maturity assessment reports, organizational policies, system development roadmaps, and national regulatory frameworks. Additional insights were gained through participatory observation of data governance processes. Document selection followed the READ method, which consists of four stages: Ready (criteria-based selection), Extract (structured data extraction), Analyze (topic and policy mapping), and Distil (strategic synthesis).

D. Maturity Assessment Using TDWI

The analysis utilized the TDWI Analytics Maturity Model to assess BPKP's readiness across five core dimensions: Organization, Governance, Analytics, Infrastructure, and Data Management. Each dimension was scored on a five-level maturity scale ranging from Nascent to Visionary. Table I provides a detailed breakdown of the maturity assessment criteria for each TDWI dimension, outlining the characteristics for levels 1 through 5. A scoring matrix was developed and applied to evaluate documented practices and observed operations. Identified gaps were mapped against DMBOK domains to formulate strategic interventions that are contextually relevant and scalable.

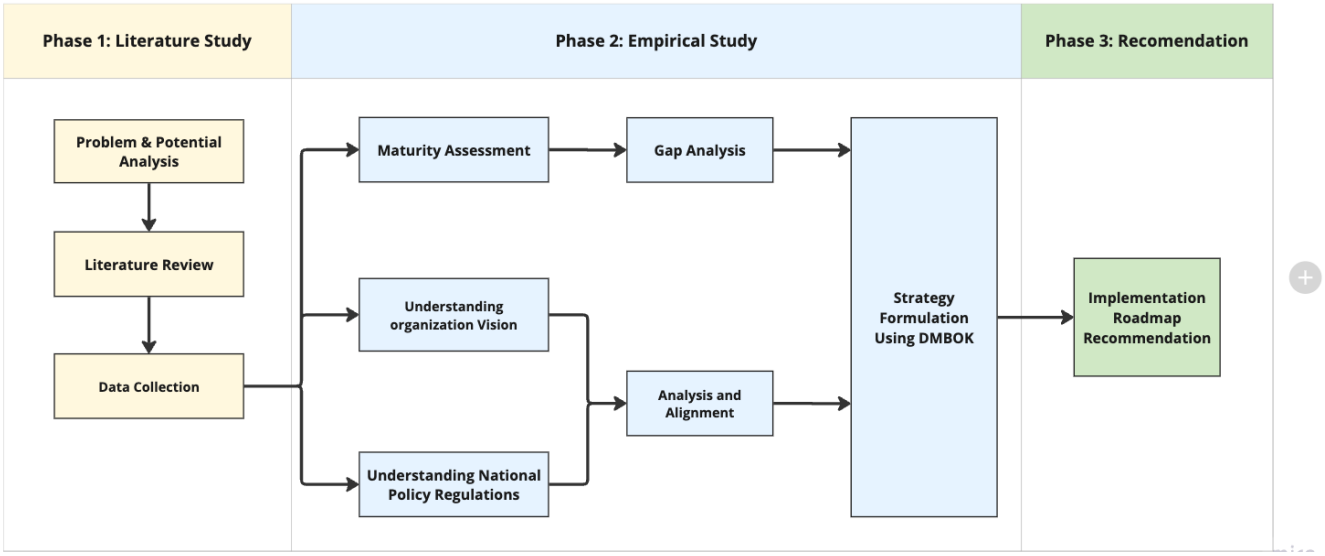


Fig. 3. Conceptual framework for the study.

Each dimension comprises several observed indicators reflecting current practice. Observations were scored on a 1–5 scale corresponding to the maturity level. The maturity score for each dimension was calculated as follows:

$$Dimension\ Score = \frac{\sum Indicator\ Scores}{Number\ of\ Indicators}$$

To identify the gap between actual and expected conditions, a gap analysis was performed using the following formula:

$$Gap\ Score = Ideal\ Score - Actual\ Score$$

TABLE I. MATURITY ASSESSMENT CRITERIA PER TDWI DIMENSION

Dimension	Level 1 - Nascent	Level 2 – Pre-Adoption	Level 3 – Early Adoption	Level 4 – Corporate Adoption	5 - Visioner
Organization	No structure, budget, or training.	Isolated initiatives, limited support, basic training.	Emerging structure, clear support, structured training.	Strong structure, sustained support, comprehensive training.	Strong data culture, active leadership, high innovation.
Governance	No policy or data security.	Incomplete policies, basic security.	Complete policies, clear roles, good security.	Monitored governance, measurable quality, strong security.	Automated governance, optimal quality, advanced protection.
Analytics	No analytics or reporting.	Basic reports, minimal analysis, intuitive decisions.	Use of analytics starts, advanced reporting, some AI.	In-depth analytics, dashboards, data-driven decisions.	Real-time analytics, machine learning.
Infrastructure	Fragmented and unplanned.	Basic systems, limited integration.	Integrated infrastructure, initial data warehouse.	Scalable infrastructure, cloud, big data capabilities.	Cloud-native, Data Mesh, Data Lakehouse architectures.
Data Management	No management, quality not addressed.	Basic processes, early attention to quality.	Defined processes, managed metadata, good quality.	Actively monitored, measured quality, integrated processes.	Automated management, certified quality, seamless data sharing.

In this study, the ideal score was set at Level 4 (Corporate Adoption) to represent the minimum threshold of systematic and standardized practices in public sector organizations.

Since a single indicator may reflect multiple dimensions, each observation was mapped to all relevant categories. For example, leadership support for big data initiatives could influence the Organization, Infrastructure, and Analytics dimensions. This approach acknowledges the interconnected and multidimensional nature of data governance and ensures that scoring reflects the systemic integration of practices.

The results of the maturity assessment provided a diagnostic basis for identifying priority areas and designing improvement strategies mapped to the DMBOK framework.

E. Validation Techniques

To ensure the reliability of findings, two validation techniques were employed. First, source triangulation was conducted by cross-verifying information across different types of documents and internal systems, such as official policy documents, internal assessment reports, and system architecture diagrams, to ensure consistency and comprehensiveness of the data gathered. Second, a limited member-checking approach was applied, in which selected results and interpretations regarding BPKP's current maturity levels and identified challenges were informally verified by internal stakeholders (e.g., data management leads and key BPKP personnel involved in digital transformation initiatives) to confirm consistency with organizational practices. This dual approach enhanced the trustworthiness and credibility of the assessment outcomes.

IV. RESULT AND DISCUSSION

This section presents the findings from the assessment of BPKP's big data and analytics capabilities, discusses the identified maturity levels and gaps, and outlines DMBOK-based strategies formulated to enhance its data-driven decision-making capacity.

A. Current State of Data Utilization and Management at BPKP: Key Findings

A systematic analysis of BPKP's internal documents, employing the READ (Ready, Extract, Analyse, Distil) approach, was conducted to understand the existing landscape. This investigation yielded key insights into organizational strengths and challenges pertinent to leveraging data for decision support.

1) Identified strengths: The document analysis revealed a strong commitment from BPKP's leadership towards digital transformation and enhancing data analytics capabilities, supported by strategic policies and resource allocation. This leadership support is crucial for successful data governance initiatives, a point strongly emphasized by research indicating the need for top-level support to overcome technical and organizational challenges in public institutions. Foundational elements include an established data custodian structure (walidata) aligning with national Satu Data Indonesia (SDI) principles, core operational systems like the Audit Management Information System (SIMA) and the STARA interoperability platform, alongside initiatives in knowledge management (KMS) and robust information security practices (evidenced by ISO 27001 certification). These elements provide a solid groundwork for future enhancements.

2) Identified challenges: Despite these strengths, significant challenges persist. The practical application of advanced analytics for risk-based decision-making and data-driven audits is not yet widespread. Comprehensive, standardized institutional metadata documentation is underdeveloped, and a formal, documented data lifecycle management process for digital supervision data is largely absent. Furthermore, an integrated system for managing all digital audit evidence is lacking. These issues resonate with prior studies on data quality and governance in Indonesia's public sector, where institutions often struggle with fragmented standards and the absence of standardized governance frameworks. Similarly, Nielsen et al. [8] emphasize the inherent difficulties in governing data within local government due to challenges such as data silos and inconsistent metadata standards.

B. Big Data Maturity Assessment

To objectively evaluate BPKP's proficiency in harnessing data for its decision-making systems, the TDWI (The Data Warehousing Institute) Big Data Maturity Model was employed. This model assesses maturity across five dimensions: Organization, Data Governance, Analytics, Infrastructure, and

Data Management. The scores for each dimension were derived from a systematic mapping of the documentary evidence (summarized in Section A) against detailed TDWI criteria. The overall big data/analytics maturity score for BPKP was 3.41, categorizing the institution at the “Early Adoption” stage. This finding aligns with observations in other public sector entities, where initial investments in digital transformation often precede full operationalization of data-driven capabilities, as highlighted by Di Vaio et al. [14] regarding the evolving role of digital technologies in governments. Fig. 4 visually represents the maturity scores across these dimensions.

The Infrastructure dimension scored highest (3.60, transitioning towards "Corporate Adoption"). This reflects strong evidence of BPKP's significant investments in IT architecture, core systems like STARA, and a Data Lake House initiative, indicating a good technical foundation for handling diverse datasets. Data Governance followed at 3.57, supported by the data custodian structure, ISO 27001 security standards, and various formal policies pertinent to managing data sources.

Both the Organization and Data Management dimensions scored 3.40 ("Early Adoption"). For the Organization dimension, while leadership commitment is strong, a pervasive data-driven culture and specific data management performance metrics are still evolving. For Data Management, practices such as reference code usage are in place, but comprehensive metadata and integrated data lifecycle management require substantial development.

The Analytics dimension received the lowest score (2.60, between "Pre-Adoption" and early "Early Adoption"). This low score is primarily due to the limited operational utilization of advanced analytics (e.g., predictive modeling) in core supervision processes, despite initial investments and planning. This gap highlights a common challenge in public sector organizations where the integration of big data analytics into core decision-making processes, particularly for predictive and proactive insights, is still in nascent stages, as discussed by Chen et al. [5] and Demirkan & Delen [6] on the transformation of decision-making by big data analytics, and also relevant to the discussions on AI in public governance.

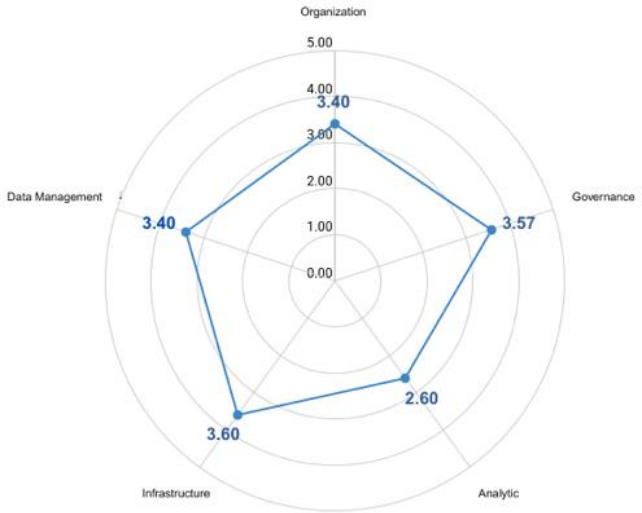


Fig. 4. BPKP's big data maturity assessment results.

C. Gap Analysis

A gap analysis compared BPKP's current maturity scores against a target of 4.0 ("Corporate Adoption"), essential for fully leveraging data for strategic decision-making. Table II summarizes this analysis.

TABLE II. GAP ANALYSIS OF BPKP'S BIG DATA MATURITY

Dimension	Reality	Exp.	Gap	Gap Analysis
Organization	3.4	4	0.6	The organizational area shows strong leadership commitment and effort to build a data culture. Policies and training are in place, and there is support for strategic initiatives. The main gaps are in the full implementation and measurement of the impact of these initiatives.
Governance	3.57	4	0.43	Data governance is already quite strong with structured policies and procedures, and efforts to comply with international standards. What needs to be improved is the validation, audit, and further evaluation of existing processes, as well as expanding the scope of data governance to all types of data.
Analytic	2.6	4	1.4	The analytics area shows great potential, but implementation is still in its early stages. There needs to be pilot projects, predictive models, or integration into business processes to make more effective use of analytics data. Training and platforms are available, but their use needs to be encouraged. In addition, there needs to be a more concrete implementation of big data analytics initiatives and the use of AI platforms. The use of analytics data needs to be encouraged and integrated into business processes.
Infrastructure	3.6	4	0.4	IT infrastructure is already well prepared, with attention to hardware, integrated systems, and big data management. It is necessary to ensure a comprehensive architecture and coverage for all data and analytics needs.
Data Management	3.4	4	0.6	Data management is already quite good with integration, quality assurance, and big data management. Further development is needed in metadata management, specific data architecture, and ensuring comprehensive data lifecycle and audit.

The analysis in Table II reveals the most substantial gap in the Analytics dimension (1.40). Significant gaps also exist in Organization (0.60) and Data Management (0.60), underscoring

the need for a more ingrained data-driven culture, specific performance metrics, comprehensive metadata standards, and systematic data lifecycle management. Smaller gaps in Data Management (0.43) and Infrastructure (0.40) indicate these areas are relatively more mature but still need enhancements for full corporate adoption.

D. Formulation of DMBOK-Based Strengthening Strategies

To address the identified gaps systematically, strengthening strategies were formulated based on the comprehensive DAMA Data Management Body of Knowledge (DMBOK) framework. DMBOK provides a structured compendium of data management principles and best practices across eleven key Knowledge Areas (KAs). The application of DMBOK here provides a structured approach to overcome common data management challenges in public institutions, offering a more standardized governance framework as suggested by Nugraha et al. [1] and aligning with the principles for reliable public sector data discussed by DAMA International [7]. For this research, practical solutions from these KAs were identified and tailored to BPKP's specific needs and context. Table III below summarizes the core DMBOK Knowledge Areas drawn upon and the key solution themes that inform the proposed strategies for BPKP. These themes are derived from a more extensive catalog of DMBOK practical solutions.

TABLE III. KEY DMBOK KNOWLEDGE AREAS AND CORRESPONDING SOLUTION THEMES APPLIED

DMBOK Knowledge Area	Key Solution Themes
Data Governance (DG)	Establishing a formal data governance function, roles (e.g., data stewards), and clear accountability; Developing and implementing comprehensive data governance policies, standards, and procedures; Managing compliance and data ethics.
Data Architecture (DA)	Evaluating and defining an enterprise data architecture that aligns with business strategy; Developing data flows and integration models; Ensuring architectural principles are followed.
Data Modeling & Design (DM)	Developing and maintaining enterprise conceptual, logical, and physical data models; Ensuring data models support business requirements and data quality.
Data Storage & Operations (DSO)	Managing physical data assets, including database administration, performance monitoring, and data lifecycle operations like backup, recovery, retention, and archiving.
Data Security (DS)	Defining and implementing data security policies, standards, and controls; Managing data access and privacy; Ensuring compliance with security regulations.
Data Integration & Interoperability (DI)	Designing and implementing solutions for data movement and consolidation between systems; Ensuring data consistency across integrated platforms.
Documents & Content (DC)	Managing unstructured data assets (e.g., digital audit documents); Establishing policies for document storage, access, versioning, and retention.
Reference & Master Data (RM)	Identifying and managing critical reference and master data; Establishing standards and processes for MDM to ensure data consistency and accuracy across the enterprise.
Data Warehousing & Business Intelligence (BI)	Designing, developing, and managing data warehouses, data marts, and BI solutions; Supporting analytical reporting, dashboards, and data exploration for decision-making.
Metadata Management (MM)	Defining metadata strategy, architecture, and standards; Developing and managing a metadata repository (e.g., data dictionary, business glossary); Ensuring metadata supports data understanding and governance.
Data Quality (DQ)	Defining a data quality strategy and dimensions; Establishing data quality rules and metrics; Implementing processes for data profiling, cleansing, monitoring, and continuous improvement of data quality.

Table IV outlines the core strengthening strategies.

TABLE IV. KEY DMBOK-BASED STRATEGIES TO ADDRESS TDWI MATURITY GAPS

TDWI Dimension	Key Strategic Thrusts for BPKP	Primary DMBOK KAs Addressed	Key Regulatory Alignment
Organization (Score: 3.40, Gap: 0.60)	1. Enhance organizational readiness and cross-functional collaboration for data initiatives. 2. Implement data governance scorecards & risk assessments for data projects.	Data Governance (DG), Data Architecture (DA)	SPBE principles (integrated governance, risk management)
Data Governance (Score: 3.57, Gap: 0.43)	1. Strengthen data governance policies, including data access, classification, and issue management. 2. Enhance data quality assessment and metadata governance practices.	Data Governance (DG), Data Quality (DQ), Metadata Management (MM)	SPBE, SDI (data quality, metadata standards)
Analytics (Score: 2.60, Gap: 1.40)	1. Develop targeted analytics use cases (e.g., anomaly detection, risk prediction) with supporting data marts. 2. Foster self-service BI and build/iterate on predictive modeling capabilities.	Business Intelligence & Data Warehousing (BI/DW), Data Science & Big Data Analytics	Supports SPBE data utilization goals
Infrastructure (Score: 3.60, Gap: 0.40)	1. Continuously evaluate and evolve data architecture for scalability, security, and performance. 2. Ensure robust database management and business continuity for critical data assets.	Data Architecture (DA), Data Storage & Operations (DSO)	SPBE (resilient infrastructure)
Data Management (Score: 3.40, Gap: 0.60)	1. Implement a comprehensive metadata architecture and standardized metadata for key data domains. 2. Define and operationalize a data quality strategy, including validation rules and monitoring. 3. Plan and model for Master Data Management (MDM) for critical reference data.	Metadata Management (MM), Data Quality (DQ), Reference & Master Data (RMD)	SDI (metadata, data standards, interoperability)

The strategies detailed in Table IV are designed to be comprehensive. For instance, to address the critical gap in Analytics, strategies grounded in DMBOK's Business Intelligence & Data Warehousing and Data Science & Big Data Analytics KAs include defining specific analytics use cases (e.g., anomaly detection, predictive risk modeling), developing supporting data marts, and creating interactive dashboards. For Data Management and Metadata Management gaps, key strategies focus on defining a complete metadata architecture, implementing consistent metadata standards, developing an enterprise data dictionary, and establishing robust data quality programs, directly relating to DMBOK's Metadata Management and Data Quality Management KAs and supporting SDI principles. Strategies for Data Governance and Organization emphasize developing comprehensive data governance policies and strengthening data stewardship roles, aligning with DMBOK's Data Governance KA and SPBE requirements.

E. Phased Implementation Roadmap Using Aiken's Pyramid

A comprehensive enhancement of data management capabilities, as envisioned by the DMBOK-based strategies (outlined in Table IV), requires a structured and incremental implementation. A "big bang" approach is often fraught with risk and may not yield sustainable results. Therefore, this research proposes a phased implementation roadmap guided by Aiken's Data Management Value Pyramid, a model that advocates for building foundational data capabilities before progressing to more advanced, value-driven initiatives. As depicted in Fig. 5, this hierarchical model organizes data management development into logical stages, ensuring that each phase builds upon the successes and outputs of the preceding one.

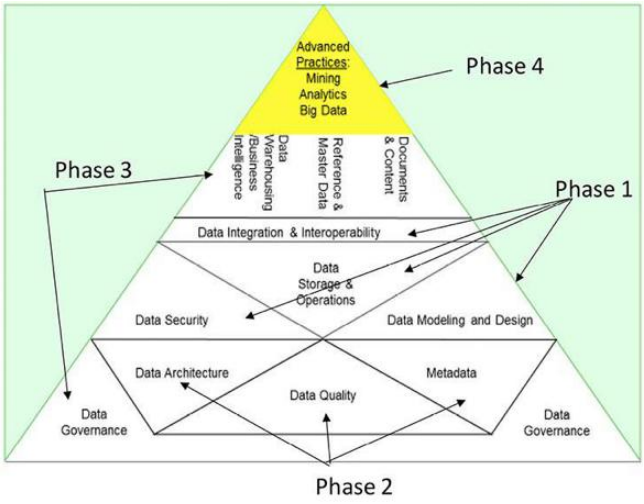


Fig. 5. Aiken's pyramid model [7].

The DMBOK-based strengthening strategies previously formulated have been mapped into the distinct phases of Aiken's Pyramid. Table V presents this phased roadmap, outlining the primary focus, prioritized DMBOK Knowledge Areas (KAs), key strategic themes for BPKP, and expected outcomes for each implementation stage.

TABLE V. PHASED IMPLEMENTATION OF DATA MANAGEMENT STRATEGIES TO ENHANCE DECISION SUPPORT SYSTEM AT BPKP

Prioritized DMBOK Knowledge Areas	Key Strategic Themes / Example Initiatives for BPKP	Expected Outcomes
Phase 1: Building the Technical & Governance Foundation Establish essential technical infrastructure and overarching governance principles for data.		
Data Governance (DG), Data Security (DS), Data Storage & Operations (DS), Data Modeling & Design (DD), Data Architecture (DA)	- Formalize core data governance policies, roles, and initial standards. - Strengthen data security protocols and operational data storage. - Define foundational data models and initial architectural principles.	- Reliable and secure data infrastructure. - Clearly defined basic data governance framework. - Initial data standards and models established-
Phase 2: Establishing Core Data Management Practices Develop and implement core practices for managing metadata, data quality, and refining data architecture.		
Metadata Management (MM), Data Quality (DQ), Data Governance (DG), Data Architecture (DA)	- Implement a comprehensive metadata architecture (data dictionary, business glossary). - Launch proactive data quality programs with defined rules & monitoring. - Refine enterprise data architecture and data models.	- Improved data trustworthiness and understandability. - Standardized metadata and enhanced data quality. - Mature data governance processes.
Phase 3: Enabling Business Value and Basic Analytics Leverage well-managed data to deliver direct business insights and support operational decision-making through analytics.		
Data Warehousing & Business Intelligence (DW/BI), Reference & Master Data (RMD), Documents & Content (DC)	- Develop initial data marts and implement descriptive/diagnostic analytics use cases. - Establish Master Data Management (MDM) for critical audit entities. - Improve management of digital audit evidence and content.	- Tangible business value from data (e.g., improved audit planning). - Initial analytical capabilities (dashboards, reports). - Consistent master data.
Phase 4: Exploiting Advanced Strategic Value Embed advanced analytics into core processes and strategic decision-making to drive innovation and significant organizational impact.		
Data Science & Big Data Analytics (DS/BDA), Business Intelligence & Data Warehousing (DW/BI)	- Develop and deploy predictive and prescriptive analytics models. - Integrate AI/ML capabilities where appropriate. - Foster a pervasive data-driven culture for strategic insights.	- Enhanced strategic decision-making through advanced analytics. - Proactive risk identification and improved audit effectiveness. - Data as a strategic asset.

The phased implementation detailed in Table V ensures a logical progression:

1) Phase 1: Building the technical & governance foundation. This initial and critical phase focuses on establishing the essential groundwork. Key DMBOK KAs prioritized include Data Governance, Data Security, Data Storage & Operations, Data Modeling & Design, and Data Architecture. Strategic themes involve formalizing core data governance policies and roles, strengthening data security protocols, optimizing data storage infrastructure, and defining foundational data models and architectural principles. The expected outcome is a reliable, secure data infrastructure and a clearly defined governance framework, directly addressing foundational gaps identified in the TDWI assessment for Data Governance, Infrastructure, and initial aspects of Data Management.

2) Phase 2: Establishing core data management practices. building upon the initial foundation, this phase emphasizes the DMBOK KAs of Metadata Management and Data Quality Management, while continuing to mature Data Governance and Data Architecture. Key strategic themes include implementing a comprehensive metadata architecture (including a data dictionary and business glossary), launching proactive data quality programs with defined rules and monitoring, and refining data models. This phase is crucial for ensuring data trustworthiness and understandability, which are prerequisites for effective analytics, thereby addressing core issues within the Data Management and Organization TDWI dimensions.

3) Phase 3: Enabling business value and basic analytics. With a more reliable data ecosystem in place, this phase shifts focus towards leveraging data for direct business insights. DMBOK KAs such as Data Warehousing & Business Intelligence, Reference & Master Data Management, and Document & Content Management become prominent. Strategic themes include developing initial data marts, implementing descriptive and diagnostic analytics use cases (e.g., performance dashboards for audit processes), establishing MDM for critical audit entities, and improving the management of digital audit evidence. This phase begins to directly tackle the significant gap in the Analytics TDWI dimension by delivering tangible analytical outputs.

4) Phase 4: Exploiting advanced strategic value. This most mature phase concentrates on DMBOK KAs like Data Science & Big Data Analytics. Strategic themes involve developing predictive and prescriptive analytics models, potentially integrating Artificial Intelligence (AI) capabilities, and embedding advanced analytics deeply into core supervision processes and strategic decision-making. Successful execution in this phase aims to fully address the remaining gaps in the Analytics dimension, transforming BPKP into a truly data-driven organization. This phase is contingent upon the successful establishment of capabilities in the preceding phases.

This staged roadmap allows BPKP to incrementally build its data management and analytics capabilities, manage change effectively, mitigate implementation risks, and demonstrate value progressively. Each phase is designed to provide a solid foundation for the next, ensuring that investments in data management translate into sustainable improvements in BPKP's ability to support its mandate through data-informed insights and decisions, in alignment with national SPBE and SDI objectives.

F. Conclusion

This study aimed to evaluate the analytics maturity of the Indonesian Financial and Development Supervisory Agency (BPKP) and propose a DMBOK-based strategy to enhance its data management capabilities in support of decision-making systems. Utilizing the TDWI Analytics Maturity Model, the assessment revealed BPKP is currently at an 'Early Adoption' stage (overall score 3.41). A significant gap was identified in the Analytics dimension (1.40), alongside notable areas for improvement in Organization and Data Management, particularly concerning comprehensive metadata and the operational application of advanced analytics.

To address these identified gaps, a comprehensive DMBOK-based strengthening strategy was formulated. This strategy is complemented by a phased implementation roadmap, guided by Aiken's Data Management Value Pyramid, which advocates for building foundational data capabilities before progressing to more advanced, value-driven initiatives. This structured approach prioritizes strengthening core data governance, data quality, and metadata management (Phases 1 & 2) to establish a reliable data foundation, followed by enabling business value and exploiting advanced analytics (Phases 3 & 4).

The findings and proposed roadmap offer a practical and actionable pathway for BPKP to systematically enhance its data management and analytical capabilities. Successful implementation of these strategies is anticipated to transition BPKP towards higher levels of data management maturity, enabling it to more effectively fulfill its mandate in the digital era by fostering more robust, data-informed decision-making in alignment with national Electronic-Based Government System (SPBE) and One Data Indonesia (SDI) objectives. This research contributes to the practical application of data management frameworks within public sector internal audit institutions, providing a blueprint for similar organizations seeking to leverage data as a strategic asset.

G. Discussion

The findings indicate that BPKP has established commendable foundational elements for data management, particularly in infrastructure and formal governance structures. However, the significant gap in the Analytics dimension, coupled with challenges in comprehensive metadata management and standardized data lifecycle practices, currently limits the full exploitation of data assets for strategic insight and predictive capabilities. The "Early Adoption" stage suggests that while awareness and initial projects exist, data management practices are not yet consistently applied or optimized across the organization.

The DMBOK-based strategies and the Aiken's Pyramid implementation roadmap provide a structured and actionable pathway for BPKP to address these gaps. By focusing initially

on strengthening data governance, data quality, and metadata (Phases 1 & 2), BPKP can build the reliable data foundation necessary for more advanced analytics. Subsequent phases focused on business value enablement and strategic analytics exploitation (Phases 3 & 4) will then be able to leverage this foundation to deliver tangible improvements in audit effectiveness and data-driven decision-making. The alignment with national directives like SPBE and SDI ensures that these internal improvements also contribute to broader governmental data initiatives. Successful implementation of these strategies is anticipated to transition BPKP towards higher levels of data management maturity, enabling it to more effectively fulfill its mandate in the digital era.

H. Future Work

This study provides a foundational assessment and a strategic roadmap for enhancing data management and decision support capabilities within the Indonesian Government Internal Audit Institution (BPKP). Building upon these findings, several promising avenues for future research and practical development emerge:

1) Empirical validation of the proposed roadmap: The current study formulates a DMBOK-based strategy and an Aiken's Pyramid implementation roadmap. Future research could focus on empirically validating the effectiveness of this proposed roadmap through pilot projects or case studies during its actual implementation within BPKP. This would involve tracking key performance indicators (KPIs) related to data quality, metadata completeness, and the adoption of advanced analytics to measure tangible improvements and refine the strategy.

2) Quantitative impact assessment of data management improvements: While this study provides a qualitative assessment, future research could employ quantitative methodologies to measure the direct impact of improved data management on audit efficiency, accuracy, and the overall effectiveness of decision-making processes within BPKP. This could involve analyzing audit findings, risk detection rates, and resource allocation before and after the implementation of specific data management initiatives.

3) Comparative studies across public sector institutions: Extending this research to other Indonesian government institutions or similar internal audit bodies in different countries would provide valuable comparative insights. Such studies could identify common challenges and best practices in data management within the public sector, contributing to a more generalized understanding of effective data governance models beyond a single case study.

4) Deep dive into specific DMBOK knowledge areas: Given the identified gaps, particularly in Metadata Management and the operationalization of advanced Analytics, future research could conduct more in-depth investigations into these specific DMBOK knowledge areas. This might involve developing detailed frameworks for institutional metadata standards or exploring the application of specific AI/Machine Learning techniques (e.g., natural language processing for unstructured audit evidence) tailored to the public audit context.

5) Long-term sustainability and change management: Research could also explore the long-term sustainability of data management initiatives and the organizational change management aspects required for fostering a truly data-driven culture within public institutions. This would involve studying factors influencing user adoption, continuous training needs, and the integration of data literacy into institutional curricula.

By pursuing these directions, future research can further contribute to the body of knowledge on data management in public sector auditing and provide more granular, evidence-based guidance for digital transformation initiatives.

ACKNOWLEDGMENT

The author expresses sincere gratitude to several parties for their invaluable support and contribution to this research. Profound appreciation is extended to the leadership and staff of the BPKP of Indonesia for granting permission to conduct this case study and for their invaluable cooperation in sharing insights and information within the established boundaries of data confidentiality and institutional protocols. Their engagement under these conditions was crucial to the successful completion of this study.

The author is also deeply indebted to my academic supervisor, or specifically Prof. Dr. Drs. Nilo Legowo, M.Kom. for their insightful guidance, constructive feedback, and unwavering support, which significantly shaped the direction and quality of this paper.

Furthermore, this research was conducted with careful consideration for ethical principles and responsible data handling, endeavoring to align with the institutional guidelines on research integrity and data management as encouraged by Bina Nusantara University.

AUTHORS' CONTRIBUTION

1) Febrian imanda effendy: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing.

2) Nilo legowo: Supervision, Validation, Writing – Review & Editing, Project Administration, Funding Acquisition.

Data Availability Statement: The datasets analyzed during the current study are internal documents of the Indonesian Financial and Development Supervisory Agency (BPKP) and are not publicly available due to their confidential nature and institutional protocols. Access to these documents is restricted and was granted solely for the purpose of this research under strict confidentiality agreements.

Finally, any remaining errors or omissions are solely the author's own. This research aims to contribute to the enhancement of data management practices within public sector audit institutions.

REFERENCES

[1] Nugraha, T. F., Wibowo, W. S., Genia, V., Fadhil, A., & Ruldeviyani, Y. (2024). A practical approach to enhance data quality management in government: Case study of Indonesian Customs and Excise Office.

Journal of Information Systems Engineering and Business Intelligence, 10(1), 51–69. https://doi.org/10.20473/jisebi.10.1.51-69

[2] Power, D. J. (2002). Decision support systems: Concepts and resources for managers. Quorum Books.

[3] Gorry, G. A., & Scott Morton, M. S. (1971). A framework for management information systems. Sloan Management Review, 13(1), 55–70.

[4] Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. Decision Support Systems, 33(2), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

[5] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

[6] Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 55(1), 412–421. https://doi.org/10.1016/j.dss.2012.05.048

[7] DAMA International. (2017). DAMA-DMBOK: Data management body of knowledge (2nd ed.). Technics Publications.

[8] Nielsen, O. B., Persson, J. S., & Madsen, S. (2019). Why governing data is difficult: Findings from Danish local government. In TDIT 2018: Governance and management of IT (IFIP Advances in Information and Communication Technology, Vol. 533, pp. 15–29). Springer. https://doi.org/10.1007/978-3-030-04315-5_2

[9] International Organisation of Supreme Audit Institutions (INTOSAI). (2023). Big data and public sector auditing. INTOSAI Development Initiative.

[10] Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), Article 101577. https://doi.org/10.1016/j.giq.2021.101577

[11] Halper, F., & Krishnan, K. (2013). TDWI big data maturity model guide: Interpreting your assessment score. TDWI Research.

[12] Republic of Indonesia, Presidential Regulation Number 95 of 2018 concerning Electronic-Based Government System. 2018.

[13] Republic of Indonesia, Presidential Regulation Number 39 of 2019 concerning One Data Indonesia. 2019.

[14] Di Vaio, A., Palladino, R., Pezzi, A., & Kalisz, D. E. (2022). The role of digital technologies in the future of governments: A bibliometric analysis of e-government research. Technological Forecasting and Social Change, 174, 121201. https://doi.org/10.1016/j.techfore.2021.121201

[15] Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. https://doi.org/10.3316/qrj0902027

[16] Dalglish, S. L., Khalid, H., & McMahon, S. A. (2020). Document analysis in health policy research: The READ approach. Health Policy and Planning, 35(10), 1424–1431. https://doi.org/10.1093/heapol/czaa064

Game Theory Meets Explainable AI: An Enhanced Approach to Understanding Black Box Models Through Shapley Values

Mouad Louhichi, Redwane Nesmaoui, Mohamed Lazaar
ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—The increasing complexity of machine learning models necessitates robust methods for interpretability, particularly in clustering applications, where understanding group characteristics is critical. To this end, this paper introduces a novel framework that integrates cooperative game theory and explainable artificial intelligence (XAI) to enhance the interpretability of black-box clustering models. Our framework integrates approximated Shapley values with multi-level clustering to reveal hierarchical feature interactions, enabling both local and global interpretability. The validity of this framework is achieved by conducting extensive empirical evaluations of two datasets, the Portuguese wine quality benchmark and Beijing Multi-Site Air Quality dataset the framework demonstrates improved clustering quality and interpretability, with features such as density and total sulfur dioxide emerging as dominant predictors in the wine analysis, while pollutants like PM2.5 and NO2 significantly influence air quality clustering. Key contributions include a multi-level clustering approach that reveals hierarchical feature attribution, use of interactive visualizations produced by Altair and a single interpretability framework that validate the state-of-art baselines. As a result, the framework forms a strong basis of interpretable clustering in essential fields like healthcare, finance, and environmental surveillance, which reinforces its generalization with respect to each domain. The results underline the need for interpretability in machine learning, providing actionable insights for stakeholders in a variety of fields.**

Keywords*—Cooperative game theory; Explainable Artificial Intelligence (XAI); Shapley values; cluster analysis; interpretability; feature attribution; black-box models

I. Introduction

Machine-learning models have achieved significant performance improvements across several areas, but the rise in success has been met by an equivalent rise in opaque decision-making systems, often referred to as black boxes, due to their internal complexity in these models [1]. The interpretation of group characteristics and feature interactions in clustering models is crucial, especially when applied to high-stakes areas such as healthcare, finance, environmental pollution monitoring, and business intelligence, where decisions made on the basis of model results can have far-reaching consequences [2]. For such applications, interpretability is not just a technical curiosity, but an imperative assurance of accountability, transparency, and regulatory compliance [3].

Classical machine-learning models, like decision trees or linear regression, are relatively easy to interpret. But the rise of more complex models, such as deep neural networks, has made it difficult to understand how they operate internally [4]. Clustering tasks provide this challenge in an especially pronounced form, where no predefined labels exist and the goal is to group data on the basis of its inherent patterns [5]. Clustering models such as K-means and hierarchical clustering are widely used across different domains. However their inability to explain which features led to a cluster formation makes it impossible to use them in decision-making contexts where the model should be interpretable [6].

To tackle these challenges, Explainable Artificial Intelligence (XAI) has emerged to address the challenge of interpreting complex models [7]. Of these, Shapley values, a concept from cooperative game theory, became particularly popular. Shapley values provide a mathematically rigorous approach for attributing contributions to individual features by quantifying their impact on model predictions [8]. This method has been particularly successful in classification models, where features can be easily mapped to predicted outcomes [9]. Applications of Shapley values in thorough domains like prediction of fraud detection, risk of the patient, scoring of credit supports the necessity of in-depth attention to contribution of each feature in model's validity and reliability. However, despite their success in classification tasks, the application of Shapley values to learning, particularly clustering models, remains limited [10].

Current clustering interpretability methods focus on individual clusters, failing to capture global feature relationships across the entire dataset. Moreover, these techniques often do not work well with high-dimensional data which increases the complexity of feature interactions and makes interpretation difficult [11]. Although there has been exploration to use feature importance approaches like Shapley values for clustering, since they typically are not scalable, expensive in computation and do not supply both local and global insights. Therefore, we require scalable and interpretable techniques that can effectively handle clustering tasks in real-world complex datasets [12].

Accordingly, the central research question is: How can cooperative game-theoretic Shapley values be integrated with scalable multilevel clustering to provide both local and global interpretability for black-box clustering models on high-dimensional and real-world datasets?

We propose a new framework that combines the interpretability tool of Shapley value with multi-level clustering methods to obtain interpretability for clustering models, the

framework offers both local and global insights into feature contributions, allowing for more transparent explanations of how clusters are formed. The proposed method uses scalable Shapley value approximations, which make it feasible to handle large, high-dimensional datasets without excessive computational demands. Furthermore, the multi-level clustering technique reveals hierarchical relationships between features, offering insights into how features interact at different stages of the clustering process.

The framework was adopted for the following reasons:

1) Scalable computation: Approximate Shapley values keep run-time practical on high-dimensional data without sacrificing accuracy.

2) End-to-end interpretability: Coupling Shapley attribution with multi-level clustering yields explanations at sample, cluster and dataset scales.

3) Model-agnostic pipeline: The PCA, k-means, LightGBM and SHAP stack wraps any learner without retraining.

4) Verified cluster quality: Silhouette and Davies–Bouldin indices confirm compact, well-separated groups.

Empirical studies on two real-world datasets, Portuguese wine quality [13] and Beijing Multi-Site Air Quality [14], validate the framework. The outcomes show that the introduced framework enhances clustering performance and interpretability. This indicates its significance for yielding actionable insights in areas such as environmental pollution monitoring, and business intelligence, where interpretability is vital. With increased interpretability of models, this work opens up avenues for developing such types of more interpretable and reliable clustering methods to use in sensitive applications. This is where explainability is of the utmost importance to deploy machine learning models.

The remainder of this paper is organized as follows: Section II reviews relevant literature on model interpretation, game theory, and clustering approaches. Section III outlines the theoretical framework and methodology. Section IV presents empirical validation through experimental results and comparative analysis. Section V discusses broader implications and directions for future research, and Section VI concludes with a summary of findings.

II. Literature Review

A. Clustering Techniques and Interpretability

Clustering is an essential activity in data mining and machine learning involving the division of data into a collection of non-overlapping clusters of data with inbuilt regularities and features. K-means and hierarchical clustering as the most notable examples [15], are very time-efficient and easy to implement, however, their recent use in risky areas like healthcare, banking, and autonomous driving, has also increased the pressure to find a more understandable and explainable output [16].

Recent advancements in interpretable clustering have addressed challenges posed by high-dimensional data and the need for explainability. A taxonomy of interpretable clustering methods has been proposed, dividing the clustering process into three stages: pre-clustering (feature selection), in-clustering (model building), and post-clustering (model explanation) [17]. This framework helps in categorizing interpretable clustering methods according to their interpretability at different stages, providing a structured approach to understanding and developing these methods.

Specifically, the use of tree-based models in interpretable clustering warrants attention since they provide an intuitive and conceptual explanation of cluster formation. Such models are developed with optimization techniques that aim at ensuring the validity of the clusters and they are also interpretable. Moreover, the use of minipatch learning [10] in consensus clustering has become a new approach, and it provides not only efficiency in the calculations but also interpretability where it discovers the features that best differentiate the cluster.

B. The Role of Game Theory in AI

Game theory has emerged as a powerful framework for analysing and modeling complex interactions in AI systems. In the context of machine learning, game theory provides a foundation for understanding multi-agent systems, where multiple AI agents interact or compete to achieve specific goals. These interactions can be cooperative or competitive, and game theory helps in modeling these dynamics [18].

Game theory has found numerous applications in machine learning, including:

1) Multi-agent systems: Modeling interactions between multiple AI agents [19].

2) Generative Adversarial Networks (GANs): Two neural networks engage in a zero-sum game to generate realistic data [20].

3) Reinforcement learning: Agents learn and adapt strategies based on the actions of others [21].

4) Auction models and resource allocation: Designing efficient resource allocation systems [22].

5) Adversarial machine learning: Developing strategies to defend against attacks on ML models [23].

C. Shapley Values in Machine Learning

Shapley values are a solid framework in the field of machine learning to explain model predictions and measure feature importance. They provide a way to fairly distribute the "payout", in this case, the prediction, among the features based on their contribution. Shapley values have essential properties such as efficiency, symmetry, and additivity, ensuring fair attribution of contributions. In the context of clustering, Shapley values help in understanding the contribution of features to the formation of clusters [6].

Recent advancements have also extended the use of Shapley values to cluster importance, treating clusters of training data as players in a game. This novel approach allows for the quantification of how different data clusters affect individual predictions, complementing traditional feature importance explanations [24].

D. Interpretability in Machine Learning

Interpretability in machine learning is crucial for ensuring that models are not only accurate but also understandable and trustworthy. This is particularly important in high-stakes fields such as healthcare, finance, and autonomous systems, where decisions made by ML models can have significant consequences [25].

Several challenges exist in interpreting complex machine learning models, including the black-box nature of many models, the trade-off between interpretability and performance, and the lack of standardization in interpretability methods [16]. To address these challenges, various interpretability methods have been developed, including:

1) Post-Hoc interpretability: Analyzing models after training (e.g., LIME, SHAP) [26].

2) Intrinsic interpretability: Using inherently interpretable models (e.g., decision trees, linear models) [27].

3) Model-Specific techniques: Tailored to specific model types (e.g., attention mechanisms for neural networks).

4) Surrogate models: Creating simpler, interpretable models that approximate complex ones.

5) Visualization techniques: Using visual tools to help understand model predictions and feature importance.

As the field of interpretable machine learning continues to evolve, our research focuses on developing method that not only provide accurate results but also offer clear, understandable explanations. This ongoing research plays a crucial role in shaping the future of clustering techniques and machine learning applications across various domains.

III. Methodology

This section presents the proposed framework that integrates dimensionality reduction, clustering algorithms, and model interpretation techniques to analyze complex datasets and discover underlying patterns. The framework consists of three main components: dimensionality reduction using Principal Component Analysis (PCA), K-means for generating cluster labels, and LightGBM for training a multi-classifier. The classifier is then interpreted using SHAP values to identify the key contributors to each cluster. This integrated approach aims to not only identify clusters within the data but also to provide meaningful interpretations of the features contributing to each cluster, enhancing the explainability of the results. Fig. 1 illustrates this workflow, where data flows sequentially through each stage to produce interpretable clustering results.

A. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) is employed as the initial step in the framework to address the challenges posed by high-dimensional datasets. PCA transforms the original dataset features into linearly uncorrelated components, ordered by their contribution to data variance. These components are ranked according to the amount of variance they explain in the data [3]. By reducing the dimensionality of the dataset, PCA ensures computational efficiency while retaining the most critical information.

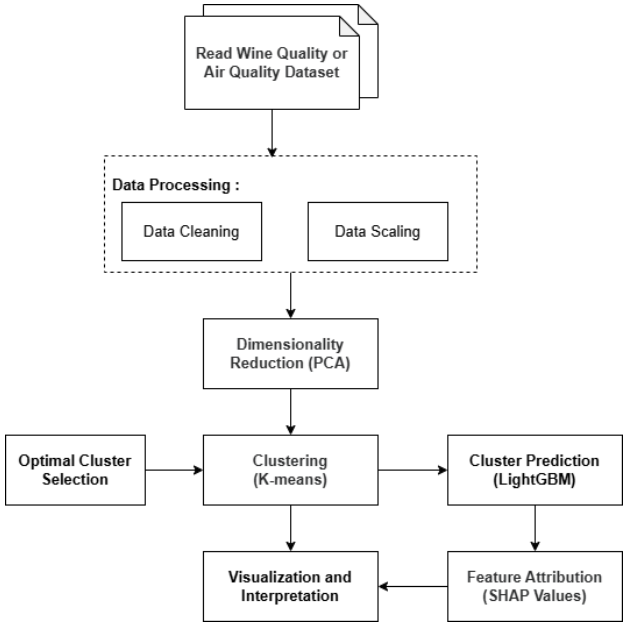


Fig. 1. Workflow of the proposed framework integrating PCA, K-Means, LightGBM, and SHAP analysis.

Mathematically, PCA as defined in Eq. (1*)* identifies a set of orthogonal vectors that maximize the variance captured in the projected space. The PCA transformation is defined as follows:

$$\mathrm{Y} = \mathrm{XW}, \tag{1}$$

where X is the standardized data matrix, W contains eigenvectors of the covariance matrix, and Y represents transformed data.

B. Clustering Framework

The primary clustering method employed in this study is K-means clustering, chosen for its efficiency and effectiveness in handling large datasets. K-means partitions the data into K distinct, non-overlapping clusters by minimizing the within-cluster sum of squares. The algorithm iteratively assigns data points to centroid of the nearest cluster and updates the centroids until convergence [28], The explained variance ratio for the $k-th$ principal component determines feature importance, as shown in Eq. (2):

$$r_k = \frac{\lambda_k}{\sum_{i=1}^{d} \lambda_i} \tag{2}$$

Following dimensionality reduction, we employ k-means clustering to minimize the objective function, as formalized in Eq. (3):

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2 \tag{3}$$

where C_i represents cluster i and μ_i its centroid. To evaluate clustering quality and determine the optimal number of clusters, we utilize two complementary metrics. The silhouette coefficient $S(x)$ shown in Eq. (4) measures cluster cohesion and separation, and it is calculated as follows:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}, \tag{4}$$

while The Davies-Bouldin index DB calculated as i (5), evaluates the average similarity between each cluster and its most similar counterpart:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i+\sigma_j}{d(\mu_i,\mu_j)}\right) \quad (5)$$

where σ_i represents the average distance of points in cluster i to its centroid, and $d(\mu_i, \mu_j)$ measures the distance between centroids. These metrics work in conjunction to ensure robust cluster evaluation and optimal parameter selection for our framework.

C. Model Interpretation

1) Training the LightGBM classifier: To interpret the clustering results, a multi-class classification model was trained using LightGBM, a gradient boosting framework based on tree learning algorithms. The model predicted cluster membership from the original features, defined in Eq. (6). The objective function minimized during training is:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

- l is the loss function (e.g., cross-entropy loss for classification),
- y_i is the true label,
- $\hat{y}_i$ is the prediction,
- Ω is the regularization term,
- k is the number of trees.

2) Computing SHAP values: To enhance the interpretability of clustering results, we integrate Shapley values with clustering methods. SHAP values were calculated to interpret the predictions of the LightGBM model. SHAP provides a unified measure of feature importance by computing the contribution of each feature to the prediction for individual data points, based on Shapley [24] (Fig. 2) . The Shapley value ϕ_i of a feature i, is computed according to Eq. (7), quantifies its marginal contribution to a model's outcome by considering all possible subsets of features $S \subseteq F \setminus \{i\}$. The Shapley value is calculated as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[v(S \cup \{i\}) - v(S)] \quad (7)$$

where: F is the set of all features, S is a subset of F excluding i and $v(S)$ represents the value function, which measures the clustering quality when only features in S are considered.

Key properties of Shapley values ensure fairness and interpretability:

- Efficiency: Stated in Eq. (8), ensures the sum of Shapley values equals the total model output:

$$\sum_{i \in F} \phi_i = v(F) - v(\emptyset) \quad (8)$$

- Symmetry: Described in Eq. (9), guarantees equal contribution for features with identical impact:

$$\phi_i = \phi_j \; if \; v(S \cup \{i\}) = v(S \cup \{j\}) \forall S \subseteq F \setminus \{i, j\} \quad (9)$$

- Additivity: Expressed in Eq. (10), allows contributions to be aggregated across different models:

$$\phi_i(v_1 + v_2) = \phi_i(v_1) + \phi_i(v_2) \quad (10)$$

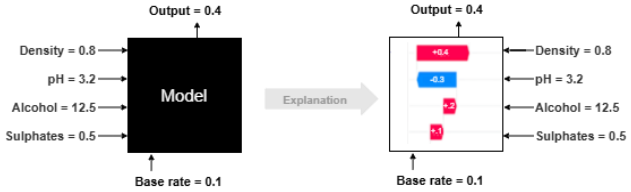


Fig. 2. SHAP (SHapley additive exPlanations) visualization of feature contributions to model output.

D. Comparative Analysis with Existing Interpretability Methods

SHAP was chosen for its ability to provide both local and global interpretability, supported by the rigorous mathematical foundation of cooperative game theory. Unlike LIME, which generates local explanations using surrogate models, SHAP ensures consistent and fair attribution of feature importance across all data points. While LIME is valued for its simplicity, its dependence on input perturbations and surrogate modeling can lead to inconsistencies, particularly in unsupervised or complex tasks [29] (see Table I).

TABLE I. COMPARISON OF INTERPRETABILITY TOOLS FOR MACHINE LEARNING MODELS

Tool	Advantages	Limitations
SHAP (SHapley Additive exPlanations)	• Provides both local and global feature importance. • Model-agnostic and mathematically rigorous. • Based on Shapley values for fair contribution attribution.	• Computationally expensive for large datasets. • Requires adaptation for unsupervised tasks like clustering
LIME (Local Interpretable Model-agnostic Explanations)	• Simple and intuitive. • Local explanations for individual predictions. • Model-agnostic.	• Limited to local interpretability. • Relies on surrogate models, which may oversimplify complex behaviors. • Sensitive to input perturbations, leading to inconsistent explanations.

E. Theoretical Justification for Shapley Values in Clustering

Shapley values help quantify feature importance without predefined labels, capturing both local and global trends. Their fair distribution of contributions ensures unbiased attribution, which is critical in fields like environmental pollution monitoring, and business intelligence. While computationally intensive, scalable approximations make Shapley values feasible for high-dimensional datasets. When integrated into multi-level clustering, they reveal hierarchical feature relationships, improving interpretability and bridging the gap between explainability and unsupervised learning's opacity.

F. Practical Implementation

The implementation follows the steps detailed in Algorithm 1. The implementation utilizes Python's scikit-learn for PCA and k-means, LightGBM for cluster prediction, and the SHAP library for feature attribution. The computational complexity is $\mathcal{O}(nd^2 + nk|\mathcal{F}| + k|\mathcal{F}|M)$ for n samples, d dimensions, k clusters, and M SHAP samples. Visualization is implemented using the Altair library, providing interactive exploration of feature contributions and cluster characteristics.

Algorithm 1: Cluster Interpretation Framework

Input: Dataset $\mathbf{X}$, features $\mathcal{F}$

Output Interpretable cluster explanations $\mathcal{I}$

$\mathbf{X}_{\text{scaled}} \leftarrow$ Scale the features in $\mathbf{X}$ using StandardScaler

$\mathbf{X}_{\text{PCA}} \leftarrow \text{PCA}(\mathbf{X}_{\text{scaled}})$ ▷ Apply PCA to reduce dimensionality

$k^* \leftarrow$ Determine optimal K via silhouette and Davies-Bou

Fit KMeans with k^*: $C \leftarrow \text{KMeans}(\mathbf{X}_{\text{scaled}}, k^*)$

$\text{M} \leftarrow \text{LightGBM}(\mathbf{X}, \{\text{C}_\text{i}\})$ Train multiclass classifier

$\phi \leftarrow$ Compute SHAP Values

For each cluster $j \in \{1, \ldots, k^*\}$ do

  Extract per-cluster SHAP values: $\phi^j \leftarrow \phi[j]$

  Order features by importance: $\mathcal{P}^j \leftarrow \text{argsort}\left(\sum|\phi^j|\right)$

  Characterize the cluster: $\boldsymbol{I}_j \leftarrow \text{Interpret}\left(\mathcal{F}^j, \phi^j\right)$

End for

return $\mathcal{I}$

By adhering to this methodology, the study ensures that all analyses are reproducible and grounded in implemented techniques, providing a robust framework for interpreting clustering results in complex datasets.

IV. Experimental Results

A. Datasets Description

We evaluated our framework using two datasets from the UCI Machine Learning Repository: the "Vinho Verde" wine quality dataset and the Beijing Multi-Site Air Quality dataset. These datasets were chosen to demonstrate the framework's applicability across domains with distinct data structures.

1) "Vinho verde" wine quality dataset: The wine dataset contains 4,898 observations with 11 physicochemical features and a quality score ranging from 0 to 10. Table II summarizes its key characteristics. This dataset served as the primary domain for developing and validating the framework.

2) Beijing multi-site air quality dataset: The air quality dataset comprises approximately 383,000 observations with 11 features, including PM2.5, PM10. Unlike the wine dataset, this dataset lacks a predefined target variable. PM2.5 was selected as a proxy target due to its established importance in air quality assessments. Table III provides its detailed characteristics.

TABLE II. Wine Dataset Characteristics

Characteristic	Details
Number of Instances	4,898
Number of Features	11
Feature Types	All numeric
Features	fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
Quality Range	Scores between 0 and 10
Missing Values	None

TABLE III. Beijing Air Quality Dataset Characteristics

Characteristic	Details
Number of Instances	383585
Number of Features	11
Feature Types	Numeric and temporal
Features	PM2.5, PM10, NO2, SO2, CO, O3, temperature, pressure, dew point, wind direction, wind speed
Observation Period	2013–2017
Missing Values	handled by omission

B. Optimization Strategies

The optimization strategies employed in the study are summarized in Table IV This validation demonstrates the framework's robustness and adaptability, confirming its effectiveness in handling datasets with diverse characteristics and domain-specific requirements.

TABLE IV. Optimization Strategies for Clustering and Interpretation

Technique	Purpose
Standard Scaling	Normalize features for consistent processing and comparability.
Principal Component Analysis (PCA)	Reduce dimensionality while retaining variances.
Multi-Criteria Cluster Evaluation	Determine optimal (k) using Davies-Bouldin index, silhouette score, and elbow.
SHAP Value Computation	Explain cluster assignments by analyzing feature importance derived from a classifier trained on cluster labels.
Cluster Validation Metrics	Evaluate clustering performance using multiple quantitative criteria.

C. Validation Measures and Comparative Analysis

Rigorous quantitative validation is indispensable for establishing the credibility of any interpretable-clustering framework. Accordingly, four widely accepted cluster-quality indices, Silhouette, Davies–Bouldin (DB) are reported for both benchmark datasets and confronted with the scores published in recent SHAP-based studies (Table V). Using the same battery of metrics ensures that performance gains are measurable, reproducible and attributable to the proposed methodology.

TABLE V. COMPARISON OF VALIDATION METRICS WITH RELATED WORK

Study	Domain / Dataset	XAI + Clustering pipeline	Reported validation indices	Key insight
This Work	Portuguese Wine Quality & Beijing Air Quality	PCA → K-means → LightGBM surrogate → multi-level SHAP	Silhouette 0.63, DB 0.55; CH elbow used for ***k*** selection	Hierarchical SHAP explanations uncover cross-dataset feature hierarchies and pinpoint the drivers of each cluster.
SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk **[30]**	Italian SME credit defaults	XGBoost → SHAP / LIME weights → K-means & Spectral	Silhouette 0.37 (SHAP) vs 0.14 (LIME); DB ≈ **1.1**	SHAP space forms markedly tighter clusters, boosting downstream AUROC over LIME.
Interpretable Clustering of Spatio-Temporal Data by Using SHAP Method **[31]**	NYC Taxi pick-ups (time × space)	SHAP-guided feature pruning → K-means / Kmedoids	Silhouette, CH & DB collectively used to tune ***k***	SHAP ranks distance & travel-time as dominant; composite metrics verify weekday/weekend cluster separation.
Inferring Disease Subtypes from Clusters in Explanation Space **[32]**	Fashion-MNIST, UK-Biobank MRI, TCGA gene-sets	RF classifier → instancelevel SHAP → Agglomerative	Silhouette, DB, CH, AMI +0.45 vs raw space	Clustering SHAP vectors consistently recovers latent disease subtypes better than classical feature space.

D. Influence of Algorithm Parameters

A brief sensitivity sweep was run on the three hyper-parameters that most affect both quality and run-time, cluster count k, PCA dimensionality d, and the depth of the LightGBM surrogate used for SHAP. Table VI shows that keeping k=3 for both corpora, plotting the data in two PCA dimensions and using the default 100-tree / 31-leaf LightGBM model gives the best Silhouette, DB trade-off without inflating execution time.

TABLE VI. SENSITIVITY OF KEY PARAMETERS ON VALIDATION METRICS

Parameter	Wine-quality dataset	Air-quality dataset	Adopted setting
k (clusters) (Silh. /DB)	2 → 0.214 / 1.775 3 → 0.144 / 2.097	2 → 0.265 / 1.503 3 → 0.626 / 0.553	3, best overall balance of cohesion and separation
PCA projection d	2 components (9 PCs ≈ 97 % var.)	2 components (9 PCs ≈ 97 % var.)	2, clear 2-D plots; no loss in cluster quality
LightGBM depth (num_leaves)	31 leaves: +0.02 absolute F1 (0.82 → 0.84) costs +4 min run-time; default kept	same trend	31 leaves: F1 = 0.82 (4 min) 63 leaves: F1 = 0.84 (8 min)

E. Wine Quality Dataset Analysis

1) Overall feature importance: Our analysis revealed the global impact of features across all clusters through SHAP value computation, demonstrating a clear hierarchy of feature importance and their relative contributions. The comprehensive analysis shown in Fig. 3 identifies density, pH, and fixed acidity as the most influential features, with sulfur dioxide compounds and residual sugar showing moderate influence across all clusters.

2) In-Depth analysis of feature importance in cluster formation: To interpret cluster formation in the wine quality dataset, SHAP force plots were generated for representative samples selected using median SHAP magnitude (Fig. 4). In Cluster 0, predictions are strongly influenced by total sulfur dioxide, alcohol, and free sulfur dioxide, with density contributing positively and residual sugar slightly reducing cluster affiliation. Cluster 1 shows a distinct pattern where fixed acidity exerts the most influence, supported by free sulfur dioxide, density, and pH, indicating a profile defined by acidity and moderate sulfur content. In Cluster 2, contributions are driven by fixed acidity, alcohol, and citric acid, with pH slightly lowering the prediction. Each cluster reflects a unique chemical signature, enabling more interpretable groupings and practical insights into wine characterization.

3) Cluster-specific analysis: Fig. 5 illustrates the distinct characteristics of each cluster in relation to wine quality:

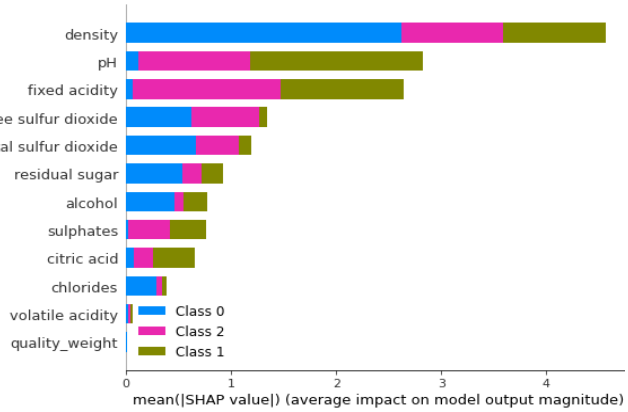
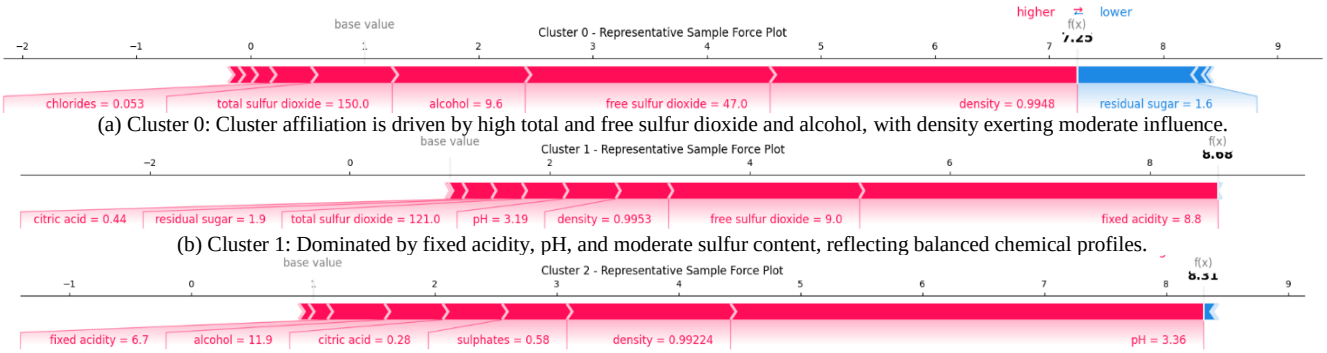


Fig. 3. Average SHAP value impact across clusters: Key features influencing cluster formation in the wine dataset.

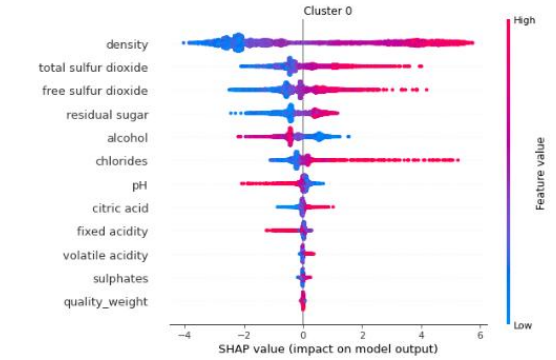


(a) Cluster 0: Cluster affiliation is driven by high total and free sulfur dioxide and alcohol, with density exerting moderate influence.

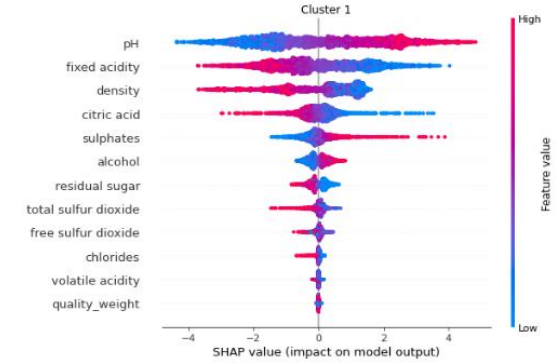
(b) Cluster 1: Dominated by fixed acidity, pH, and moderate sulfur content, reflecting balanced chemical profiles.

(c) Cluster 2: Influenced by fixed acidity and citric acid, with pH slightly reducing affiliation, indicating acidity-driven groupings.

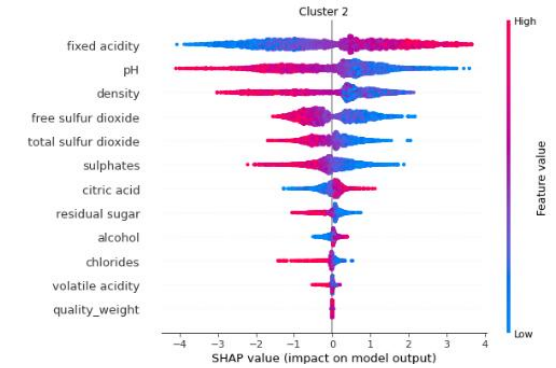
Fig. 4. SHAP Force plots for representative samples in each wine quality cluster.



(a) Cluster 0: Influenced by sulfur dioxide compounds and alcohol.



(b) Cluster 1: Driven by fixed acidity, density, and pH.



(c) Cluster 2: Dominated by alcohol, citric acid, and fixed acidity.

Fig. 5. SHAP Summary plots showing the distribution of feature contributions across clusters for the wine quality dataset.

Cluster 0 is defined by high density and elevated levels of total and free sulfur dioxide, indicating a strong correlation with wine quality ratings and preservation characteristics. the Cluster 1, the relationships among pH and fixed acidity reveal a balanced chemical composition, aligning with traditional winemaking principles. The analysis underscores the critical role of acid balance in determining wine quality. Lastly, Cluster 2 displays unique chemical patterns associated with extreme quality ratings. Significant variations in chemical composition highlight the interactions between features that influence quality assessments.

F. Validation on the Beijing Air Quality Dataset

The Beijing Multi-Site Air Quality dataset was analysed to assess the generalizability of the proposed framework to a larger and more diverse dataset. Pre-processing steps involved imputation of missing values, normalization using StandardScaler, and dimensionality reduction via PCA. Clustering was performed using k = 3, determined through a multicriteria evaluation of silhouette score ((k) = 0.63), Davies-Bouldin index ($D(k)$ = 0.55), on the full 11-feature matrix.

1) Overall feature importance: Our analysis revealed the global impact of features across all clusters through SHAP value computation, demonstrating a clear hierarchy of feature importance and their relative contributions. The comprehensive analysis shown in Fig. 6 identifies temperature (TEMP), dew point (DEWP), and pressure (PRES) as the most influential features, with CO, NO2, and particulate matter compounds (PM10 and PM2.5) showing moderate influence across all clusters.

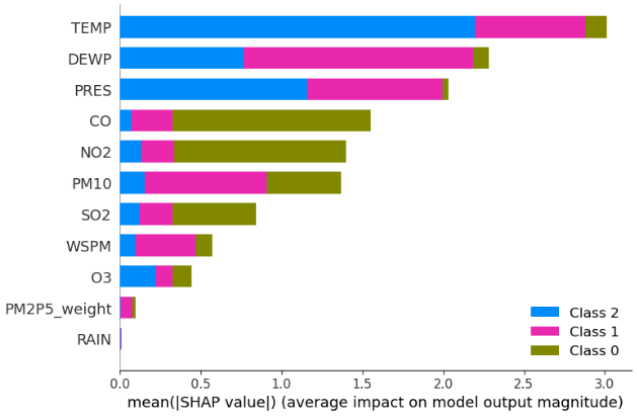
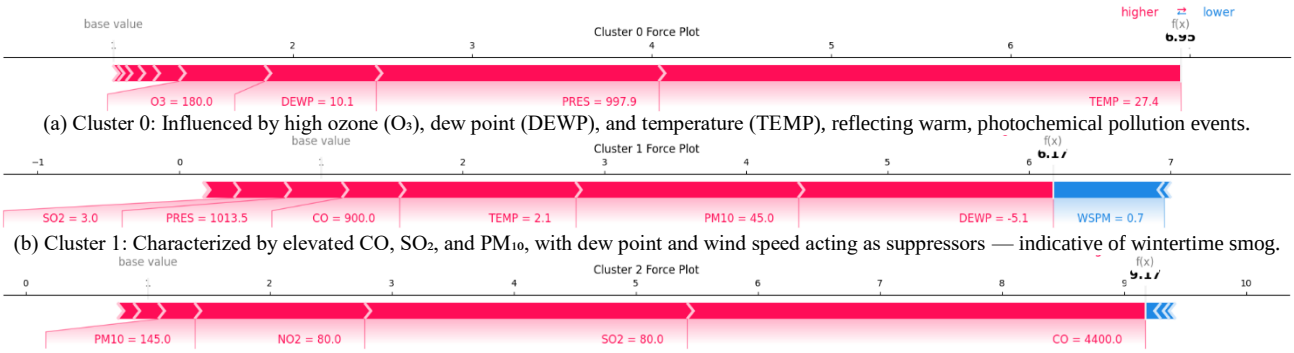


Fig. 6. Global SHAP value impact for the Beijing Air Quality dataset.



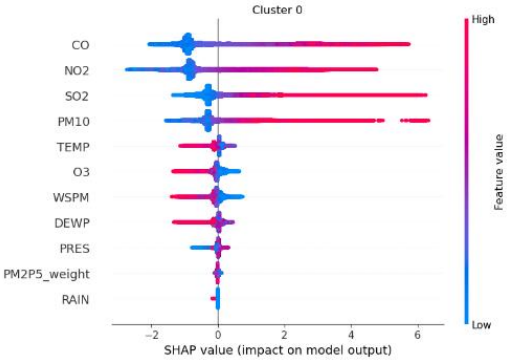
(a) Cluster 0: Influenced by high ozone (O_3), dew point (DEWP), and temperature (TEMP), reflecting warm, photochemical pollution events.

(b) Cluster 1: Characterized by elevated CO, SO_2, and PM_{10}, with dew point and wind speed acting as suppressors — indicative of wintertime smog.

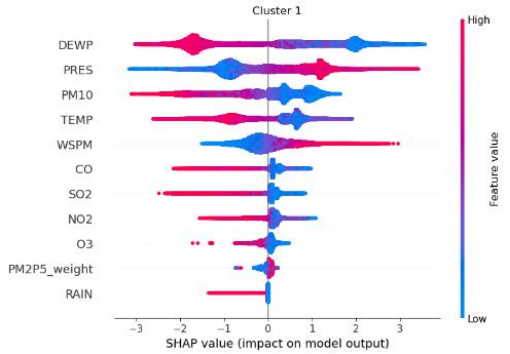
(c) Cluster 2: Driven by PM_{10} and NO_2, but unusually high CO reduces cluster affiliation — suggesting divergent emission profiles.

Fig. 7. SHAP Force plots demonstrating feature influence on cluster membership for representative samples from the Beijing Air Quality dataset.

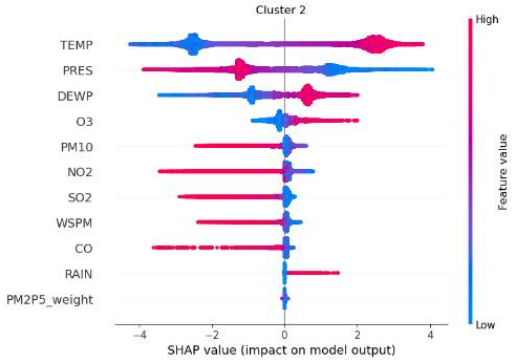
2) In-Depth analysis of feature importance in cluster formation: A similar interpretability approach was applied to the air quality dataset using SHAP force plots for representative samples (Fig. 7). In Cluster 0, elevated ozone and temperature are the primary drivers of cluster membership, with supportive roles from dew point and pressure—indicating warm, photochemical pollution events. Cluster 1 is characterized by high CO, SO_2, and PM10, while low dew point and wind speed act as suppressors, suggesting stagnant cold-air conditions typical of wintertime smog. Cluster 2 is heavily influenced by PM10, NO_2, and SO_2, with an unexpectedly negative contribution from extremely high CO, possibly reflecting divergence in emission source profiles. These results highlight how the model captures distinct atmospheric compositions that align with interpretable pollution scenarios.



(a) Cluster 0: Influenced by high CO, NO_2, SO_2, and PM_{10}, indicating pollution-heavy urban zones.



(b) Cluster 1: Driven by dew point, PM_{10}, and low wind speed, suggesting stagnant weather conditions.



(c) Cluster 2: Shaped by high temperature and pressure, with pollutants contributing negatively, reflecting clean-air events.

Fig. 8. SHAP Summary plots for each cluster in the Beijing Air Quality dataset.

3) Cluster-specific analysis: Cluster-specific SHAP analyses were performed to evaluate feature contributions across the air quality dataset. In Cluster 0, predictions are driven by elevated levels of CO, NO_2, SO_2, and PM10, pointing to pollution-heavy conditions typical of urban emission hotspots [Fig. 8(a)]. Cluster 1 is shaped by the interaction of high dew point, moderate PM10, and low wind speed (WSPM), suggesting meteorologically stagnant conditions that hinder dispersion [Fig. 8(b)]. In Cluster 2, high temperature and atmospheric pressure dominate the contribution profile, while particulate and gaseous pollutants (PM10, NO_2, SO_2) exert negative influence, indicating cleaner air events supported by favorable dispersion conditions [Fig. 8(c)]. These results give us a better understanding of how environmental variables interact to define distinct air quality regimes.

4) Discussion: The results from the Beijing dataset align with findings from the wine dataset, demonstrating the framework's ability to extract meaningful feature contributions across diverse domains. The larger size and complexity of the air quality data highlight the framework's scalability and adaptability, confirming its relevance in high-stakes applications such as environmental monitoring. The differing comparative results stem from the distinct nature of each dataset. The wine data, with consistent chemical features, yielded clearer attributions, while the air quality data required more nuanced interpretation due to its temporal and

environmental variability. This suggests the proposed algorithms are broadly applicable but especially effective for complex, high-dimensional data — reinforcing their robustness across domains.

V. Discussion and Broader Implications

A. Contributions to the Field

Our work advances the field of explainable AI through several key theoretical and practical contributions. The integration of game-theoretic principles with clustering analysis provides a novel framework for interpretable machine learning. The use of Shapley values for cluster interpretation bridges the gap between local and global explanations, offering insights at multiple levels of granularity. Our framework extends traditional SHAP applications by incorporating cluster-specific interaction effects, enabling quantitative assessment of feature relationships while maintaining interpretability. The hierarchical analysis of feature importance provides both detailed local insights and broader structural understanding of cluster formation.

B. Applications and Use Cases

The framework shows practical applicability in domains such as clinical diagnostics, regulatory finance, recommendation systems, and environmental risk monitoring. In healthcare diagnostics, our approach enables interpretation of patient groupings based on medical parameters while maintaining the high accuracy requirements of medical applications. The feature attribution methodology provides clinically relevant insights into diagnostic patterns.

Financial risk assessment applications benefit from the framework's ability to identify and explain customer segments. The clear feature importance hierarchy supports regulatory compliance requirements while enabling sophisticated risk profiling. Beyond financial services, autonomous systems benefit from our framework's ability to explain behavioral patterns and decision boundaries, which is particularly crucial in safety-critical applications.

This capability to interpret clusters meaningfully also strengthens its utility in healthcare applications, as shown by earlier discussions of patient grouping and outcome differentiation. Similarly in recommendation systems, the ability to explain user groupings and feature preferences enhances transparency and trust in personalized content delivery. These implications reinforce the practical value already outlined.

C. Limitations and Future Work

Our current implementation faces several limitations in computational scalability, particularly with extremely large datasets. The computational cost increases significantly with dataset size and feature dimensionality, becoming particularly evident in the calculation of exact Shapley values for large-scale applications.

Future enhancements could address these computational challenges through methodological improvements in three key areas. Advanced approximation methods for Shapley computation could reduce computational overhead while maintaining accuracy. Parallel processing implementations could better handle large-scale datasets. Optimized feature selection strategies could improve efficiency in high-dimensional data analysis.

Future work will explore integrating deep learning with interpretability, enabling dynamic clustering, and developing domain-specific optimizations. Implementation will aim for real-time analysis, distributed processing, and improved visualizations for complex feature interactions.

VI. Conclusion

We introduced a framework that integrates multi-level clustering with Shapley-based explanations, enabling interpretable insights into unsupervised learning outcomes. Our experiments on wine and air quality datasets demonstrate its ability to uncover meaningful feature contributions.

The implementation methodology demonstrates scalability across different domains, supported by thorough statistical validation. Through extensive testing on the wine quality dataset, we demonstrated the framework's ability to identify meaningful clusters while providing clear explanations of the underlying feature relationships. The analysis revealed distinct chemical profiles corresponding to different quality levels, with density, pH, and acidity emerging as key determinants of wine characteristics.

The proposed framework proves the successful reconciliation between model sophistication and interpretability requirements. Clear feature attribution mechanisms enable stakeholders to understand complex model decisions, while maintaining the statistical rigor necessary for reliable analysis. The framework's application extends beyond wine and air analysis to other domains requiring both precision and explicability, such as healthcare diagnostics and financial risk assessment.

While the framework demonstrates strong interpretability and generalizability across two real-world tabular datasets, several limitations remain. It has not yet been tested on non-tabular data such as images or text, exact SHAP computations remain costly for large-scale or streaming data, and the current approach assumes static clusters. Future work will explore scalable SHAP approximations, dynamic clustering, and broader dataset types to address these challenges.

The combination of theoretical soundness and practical applicability makes this framework a valuable tool for modern machine learning applications where interpretability is crucial. By providing both local and global explanations of cluster characteristics, our approach enables informed decision-making while maintaining model performance. This balance between sophistication and transparency establishes a foundation for future developments in explainable artificial intelligence, particularly in scenarios where understanding model decisions is as important as the decisions themselves.

References

[1] H. Chen, S. M. Lundberg, and S.-I. Lee, “Explaining models by propagating Shapley values of local components,” arXiv preprint arXiv:1911.11888, 2019.

[2] B. Liu, Y. Xia, and P. S. Yu, "Clustering with deep learning: Taxonomy and new methods," arXiv preprint arXiv:1801.07648, 2018.

[3] L. Hu, M. Jiang, J. Dong, X. Liu, and Z. He, "Interpretable Clustering: A Survey," Sep. 01, 2024, arXiv: arXiv:2409.00743. doi: 10.48550/arXiv.2409.00743.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[5] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," International Conference on Machine Learning, pp. 5491–5500, 2020.

[6] M. Louhichi, R. Nesmaoui, M. Marwan, and M. Lazaar, "Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering," Procedia Computer Science, vol. 220, pp. 806–811, 2023.

[7] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

[8] A. M. Salih et al., "A perspective on explainable artificial intelligence methods: Shap and lime," Advanced Intelligent Systems, p. 2400304, 2024.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.

[10] L. Gan and G. I. Allen, "Fast and interpretable consensus clustering via minipatch learning," PLOS Computational Biology, vol. 18, no. 10, p. e1010577, Oct. 2022, doi: 10.1371/journal.pcbi.1010577.

[11] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in Explainable AI: Interpreting, explaining and visualizing deep learning, Springer, 2019, pp. 169–191.

[12] R. Nesmaoui, M. Louhichi, and M. Lazaar, "A Collaborative Filtering Movies Recommendation System based on Graph Neural Network," Procedia Computer Science, vol. 220, pp. 456–461, 2023.

[13] A. C. Paulo Cortez, "Wine Quality." UCI Machine Learning Repository, 2009. doi: 10.24432/C56S3T.

[14] D. Dua and E. Taniskidou, "Beijing Multi-Site Air Quality Data." 2017.

[15] R. Nesmaoui, M. Louhichi, and M. LAZAAR, "A Hybrid Machine Learning Method for Movies Recommendation," 2022, pp. 517–528. doi: 10.1007/978-3-031-07969-6_39.

[16] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14, 2020.

[17] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.

[18] L. S. Shapley, "A value for n-person games," Contributions to the Theory of Games, vol. 2, no. 28, pp. 307–317, 1953.

[19] I. de Zarzà, J. de Curtò, G. Roig, P. Manzoni, and C. T. Calafate, "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs," Electronics, vol. 12, no. 12, Art. no. 12, Jan. 2023, doi: 10.3390/electronics12122722.

[20] T. Hazra and K. Anjaria, "Applications of game theory in deep learning: a survey," Multimed Tools Appl, vol. 81, no. 6, pp. 8963–8994, Mar. 2022, doi: 10.1007/s11042-022-12153-2.

[21] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," in Handbook of Reinforcement Learning and Control, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds., Cham: Springer International Publishing, 2021, pp. 321–384. doi: 10.1007/978-3-030-60990-0_12.

[22] A. Nezarat and G. H. Dastghaibifard, "Efficient Nash Equilibrium Resource Allocation Based on Game Theory Mechanism in Cloud Computing by Using Auction," PLOS ONE, vol. 10, no. 10, p. e0138424, Oct. 2015, doi: 10.1371/journal.pone.0138424.

[23] M. Zhu, A. H. Anwar, Z. Wan, J.-H. Cho, C. A. Kamhoua, and M. P. Singh, "A Survey of Defensive Deception: Approaches Using Game Theory and Machine Learning," IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2460–2493, 2021, doi: 10.1109/COMST.2021.3102874.

[24] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," Auton. Intell. Syst., vol. 4, no. 1, p. 2, Feb. 2024, doi: 10.1007/s43684-023-00060-8.

[25] H. Byeon, "Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 6, Art. no. 6, 30 2023, doi: 10.14569/IJACSA.2023.0140656.

[26] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc Interpretability for Neural NLP: A Survey," ACM Comput. Surv., vol. 55, no. 8, p. 155:1-155:42, Dec. 2022, doi: 10.1145/3546577.

[27] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," Statistics Surveys, vol. 16, no. none, pp. 1–85, Jan. 2022, doi: 10.1214/21-SS133.

[28] R. Nesmaoui, M. Louhichi, and M. Lazaar, "A Hybrid Machine Learning Method for Movies Recommendation," in Proceedings of the International Conference on Machine Learning and Data Engineering, 2022, pp. 517–528.

[29] K. Främling, M. Westberg, M. Jullum, M. Madhikermi, and A. Malhi, "Comparison of Contextual Importance and Utility with LIME and Shapley Values," in Explainable and Transparent AI and Multi-Agent Systems, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds., Cham: Springer International Publishing, 2021, pp. 39–54. doi: 10.1007/978-3-030-82017-6_3.

[30] A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," Front. Artif. Intell., vol. 4, Sep. 2021, doi: 10.3389/frai.2021.752558.

[31] S. Younesi and H. Rahmani, "Interpretable Clustering of Spatio-Temporal Data by using SHAP Method," in 2024 10th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of: IEEE, Apr. 2024, pp. 32–39. doi: 10.1109/icwr61162.2024.10533347.

[32] "(PDF) Inferring disease subtypes from clusters in explanation space," ResearchGate, doi: 10.1038/s41598-020-68858-7.

Anomaly Detection and Fault Diagnosis of Power Distribution Line Point Cloud Data Based on Deep Learning

Jiangshun Yu*, Poyu You, Jian Zhao, Xianzhe Long, Yuran Chen
Guizhou Electric Power Design and Research Institute Co., Ltd. of Power Construction Corporation of China, 550081, China

***Abstract*—Early and accurate fault diagnosis in power distribution systems is essential to ensure stable electricity delivery and prevent outages. This study presents a deep learning-based anomaly detection framework that analyzes 3D LiDAR point cloud data to identify structural defects in power distribution lines. Leveraging advancements in deep learning and 3D sensing, a hybrid architecture combining PointNet++ and 3D Convolutional Neural Networks (3D CNN) is proposed. The system processes point clouds from the TS40K dataset, comprising high-resolution, annotated scans of power infrastructure, and uses a feature fusion strategy to integrate fine-grained local geometry from PointNet++ with global volumetric features from 3D CNN. Implemented in Python, the method achieves a 94.7% accuracy in fault diagnosis, outperforming standalone models. It robustly detects anomalies such as sagging wires, leaning poles, and broken insulators, maintaining precision, recall, and F1-scores above 90%, even under noisy and sparse conditions. Visualization of detected faults on 3D models confirms its precise localization capability, supporting real-time monitoring and maintenance planning in smart grids. By integrating complementary deep learning techniques, this approach offers a scalable, accurate, and automated solution for anomaly detection and fault diagnosis in power distribution systems. Future work will focus on multi-sensor fusion and semi-supervised learning to reduce dependence on labeled data and broaden applicability to other infrastructure use cases.**

Keywords*—Power distribution; anomaly detection; point cloud; deep learning; fault diagnosis

I. Introduction

Power distribution networks are critical components of today's infrastructure, providing a reliable supply of electricity from substations to consumers. As urbanization and industrialization grow, so does the level of complexity and size of power distribution networks [1] [2]. Their uninterrupted operation is crucial to economic well-being and society's well-being [3] [4]. These systems, however, are exposed to environmental degradation, mechanical failure, and aging infrastructure [5]. Regular inspection and maintenance of power lines are required to avoid service outages, safety risks, and equipment damage. Conventional inspection practices are labour-intensive and based on manual processes, which are time-consuming, costly, and susceptible to human error. Additionally, visual inspection based on 2D images does not provide the spatial depth necessary to comprehend the full three-dimensional geometry of power lines, hindering the identification of concealed or slight abnormalities [6].

With advancements in remote sensing technologies, particularly LiDAR and Unmanned Aerial Vehicles (UAVs), the doors have been opened to acquiring high-resolution 3D point cloud data for power distribution lines [7] [8]. Such point clouds provide accurate spatial models of infrastructure, facilitating enhanced monitoring and analysis. Interpretation and processing of such unstructured 3D data, however, remains a problem [9]. Conventional machine learning models are also not effective in handling the sparsity, irregularity, and noise inherent in point cloud data. Traditional models have a great deal of feature engineering and cannot generalize across diverse spatial configurations. Moreover, conventional models are unable to learn the capacity to understand local and relevant geometric features that can be used in identifying fragile defects in power line components [10] [11]. To address these limitations, deep learning approaches specifically designed for point clouds have been developed. PointNet and its updated version, PointNet++, have presented promising performance on point cloud classification and segmentation by learning spatial features from raw points in an end-to-end manner. PointNet++ maintains both local and global contextual details, making it extremely efficient in the analysis of power infrastructure elements. On the contrary, voxel-based approaches, such as 3D Convolutional Neural Networks, transform point clouds into structured grids, enabling the effective detection of spatial patterns through the application of convolutional operations.

This paper suggests a fusion deep learning framework which leverages the strengths of point-based and voxel-based techniques for comprehensive fault detection in power distribution lines. By combining PointNet++ for detailed feature learning with 3D CNNs for contextualization, the proposed method achieves high accuracy in detecting a wide range of anomalies from small-scale insulator leakage to large-scale pole misalignment. The architecture employs a feature fusion approach that fuses outputs from both networks to enable the model to be trained on complementary information [12] [13], [14] [15]. The approach not only enhances the robustness of data against noise and sparsity but also improves fault localisation and classification. Furthermore, it is real-time inference-friendly, thereby allowing it to be accommodated in smart grid monitoring systems. The approach is scalable and versatile enough to support numerous power distribution scenarios, with

an applied solution to autonomous infrastructure inspection. It bridges the gap between traditional inspection and modern AI-based alternatives, opening the door to smarter, safer, and more efficient power distribution grids. The proposed framework not only enhances anomaly detection but also facilitates predictive maintenance and system resilience in future power systems. Key contributions of this research paper are,

1) This research introduces a hybrid deep learning framework that combines PointNet++ and 3D CNNs for analysing power distribution line point cloud data.

2) It effectively extracted both local and global features to detect various anomalies such as damaged insulators, sagging wires, and pole misalignments.

3) A novel feature fusion strategy was implemented to integrate point-based and voxel-based representations for improved fault diagnosis.

4) The suggested method reached higher accuracy, precision, and recall compared to traditional 2D and standalone deep learning approaches.

5) It enabled real-time and scalable fault detection suitable for deployment in smart grid monitoring systems using edge–cloud architectures.

The remaining section of the article is organized as follows: Section II gives a literature review of the current works in the areas of point cloud analysis, deep learning-based anomaly detection, and fault diagnosis of power systems. Section III formulates the problem statement accurately, identifying the issues associated with detecting structural faults and anomalies in unstructured 3D point clouds of power distribution lines. Section IV presents the proposed approach, utilising PointNet++ and 3D CNN, for robust feature extraction, fusion, and classification. Section V describes the experimental setup, presents the results obtained, and provides a critical analysis of the models' performance using various evaluation parameters. Discussion is given in Section VI. Section VII summarizes the research and suggests potential future directions of research to enhance the accuracy and real-time acceptability of the proposed framework in smart grids.

II. Related Works

Deep learning for 3D point cloud analysis has witnessed rapid growth, particularly in the context of infrastructure inspection and fault detection. Existing approaches can be broadly grouped into point-based, voxel-based, hybrid architectures, and domain-specific applications, such as those used in power grid monitoring and other similar fields. This section reviews the key contributions under these categories.

A. Point-Based Architectures

Point-based models operate directly on raw point clouds, eliminating the need for voxelization and thereby preserving geometric fidelity. A foundational approach is PointNet, introduced by Qi et al., which uses symmetric functions to handle unordered point sets and demonstrates strong performance in 3D object classification and segmentation [16]. Its extension, PointNet++, developed by Qi et al., incorporated hierarchical feature learning, enhancing local structure representation an essential capability for complex infrastructure fault detection [17]. To better capture local geometric relationships, the Dynamic Graph CNN (DGCNN) by Wang et al. dynamically constructs and updates graph edges during training, thereby improving the contextual understanding of 3D structures [18]. Li et al. developed PointCNN, a model that applies X-transforms to reorder points canonically before performing convolution, leading to effective feature extraction from irregular point clouds [19].

Further progress was made by Thomas et al. through the use of KPConv (Kernel Point Convolution), which applied deformable kernel filters directly to point sets, thereby avoiding voxelization while adapting to local geometries [20]. For unsupervised learning, Yang et al. introduced FoldingNet, which uses an autoencoder to reconstruct 3D shapes from 2D grid deformations, preserving both global and local structure [21].

Recent surveys, such as those by Zhang et al. and Liu et al., offer comprehensive taxonomies of point-based models, classifying them based on their learning strategies and highlighting their applicability to infrastructure tasks like fault detection [22] [23].

B. Voxel-Based Architectures

Voxel-based methods convert unstructured point clouds into regular voxel grids, allowing the use of 3D Convolutional Neural Networks (3D CNNs) for structured learning. Maturana and Scherer proposed the foundational VoxNet model, which performed efficient 3D shape recognition on voxelized inputs [24]. More recent models extend this idea with deeper layers, enhanced normalisation, and advanced pooling strategies.

Jin et al. introduced a convolutional GAN-based framework for semantic segmentation in power line inspection, merging voxel grid input with adversarial training [25]. Similarly, Wen et al.proposed a directionally constrained 3D CNN that improved classification performance on airborne LiDAR data by modeling orientation between points [26].

C. Hybrid and Fusion-Based Methods

Recognizing the trade-offs between point- and voxel-based models, hybrid architectures aim to leverage the strengths of both. The method proposed in this paper exemplifies this trend by combining PointNet++ for local geometric learning with 3D CNNs for capturing global spatial context, thereby improving robustness in fault classification tasks. A closely related model is VoteNet by Qi et al. and Li et al. couples PointNet++ with a Hough voting mechanism for 3D object detection [16] [19]. Similarly, Fast Point R-CNN, introduced by Shi et al. enhances detection efficiency through faster region proposals adapted to 3D data [27].

Label efficiency has become a critical challenge in large-scale applications. Xiao et al. addressed this by surveying techniques like domain adaptation and weak supervision to reduce annotation burdens [28]. In parallel, Sohail et al. presented a comprehensive study on deep transfer learning for 3D point cloud perception, emphasizing its role in mitigating data sparsity challenges [29].

D. Applications in Power Grid Fault Detection

The field of smart grid infrastructure has embraced 3D deep learning for anomaly detection. Zideh and Solanki proposed physics-informed autoencoders for improved interpretability in fault diagnostics [30]. Kyuroson et al. developed an unsupervised point cloud segmentation framework for power line inspection using LiDAR, demonstrating high accuracy in identifying vegetation and line sag [31]. Lu and Shi reviewed scalable deep learning approaches for 3D point cloud perception tasks such as detection and segmentation, confirming the framework's viability in hazard analysis along power corridors [32]. Zhu et al. and Faisal et al. explored generative and adversarial methods to refine semantic segmentation and automate power line monitoring [33] [34]. Finally, Chen et al. presented DCPLD-Net [35], a real-time diffusion-coupled CNN for detecting transmission lines from UAV-LiDAR data. Other efforts, including those by Shin et al. and Lin et al. , combine thermal, RGB, and LiDAR data for multi-modal anomaly detection, reflecting an emerging trend toward sensor fusion and real-time deployment in smart grids [10] [13]. Basani et al. (2024) developed a deep multi-scale fusion neural network for fault diagnosis in IoT environments. Inspired by this, our work adopts multi-scale feature fusion to extract and combine geometric features from power line point cloud data. This approach enhances fault detection accuracy and robustness across varying spatial resolutions [36]. A real-time safety management framework using machine learning on TBM data is presented by Devarajan and Sambas (2022) to enhance infrastructure monitoring. Our proposed work applies this real-time data-driven approach for anomaly detection in power distribution line point cloud data. This enables proactive fault diagnosis, reduces risks, and improves the reliability of power infrastructure systems [37]. Rajya Lakshmi Gudivaka (2023) implemented a neural network-based framework for real-time imperfection prediction in robotic process automation. Our study employs this approach to detect anomalies in power distribution line point cloud data using deep learning techniques. This integration facilitates automated fault inspection, improves detection speed, and enhances system reliability [38]. Our proposed work on fault diagnosis in power distribution lines adopts the signal preprocessing strategy introduced by Harikumar and Khalid (2022), which utilizes Butterworth filters to enhance signal clarity in IoT systems. By incorporating this denoising technique into our point cloud data pipeline, we ensure cleaner input for deep learning-based anomaly detection. This improves diagnostic precision, minimizes false positives, and strengthens the reliability of fault classification outcomes [39]. A hybrid machine learning framework for financial fraud detection using big data was proposed by Naresh Kumar Reddy Panga (2021). Our framework builds on this hybrid learning strategy by combining point-based feature extraction and deep classification for 3D anomaly detection. This integration improves model generalization, enhances fault detection accuracy, and ensures robustness across varying point cloud conditions [40].

III. Problem Statement

Traditional methods of fault detection in power distribution networks—such as visual examination and 2D image analysis—are insufficient to describe the complex spatiality of modern infrastructure [41] [42], resulting in inefficiency and missed anomalies. Although 3D point cloud data from LiDAR and UAVs offer a more detailed description, their unstructured and irregular nature is extremely difficult for traditional machine learning models to process [43], [44], [45]. This work overcomes such limitations by proposing a hybrid deep learning framework that combines PointNet++ and voxel-based 3D CNNs for learning global spatial patterns. By integrating point-wise and voxel-wise features, the method enhances both noise robustness and the accuracy of fault diagnosis and anomaly detection. This end-to-end framework supports the precise detection of structural defects—such as insulator failure and pole misalignment—while enabling real-time edge inference on the edge–cloud platform, thus overcoming the limitations of existing models and offering an extensible autonomous power infrastructure monitoring solution. To address the limitations of traditional fault detection techniques and the challenges of processing unstructured LiDAR point cloud data, this study investigates the following research question: How can a hybrid deep learning architecture that combines PointNet++ and 3D CNN be developed to accurately and efficiently detect and classify structural anomalies in power distribution line point cloud data, while maintaining robustness under real-world smart grid conditions?

Objectives

1) Develop a deep learning-based framework capable of analyzing unstructured 3D point cloud data from power distribution lines.

2) Extract meaningful local and global geometric features from point clouds using PointNet++ architecture.

3) Apply voxel-based 3D CNNs for contextual analysis and detection of large-scale structural anomalies.

4) Design a fusion mechanism that combines point-based and voxel-based features for enhanced fault classification.

5) Validate the performance of the proposed method on real-world datasets in terms of accuracy, scalability, and real-time applicability.

IV. Proposed PointNet++ and 3D CNN-Based Deep Grid Fault Net for Anomaly Detection and Fault Diagnosis in Power Distribution Line Point Cloud Data

The proposed methodology employs a hybrid deep learning method that integrates PointNet++ and 3D Convolutional Neural Networks for effective fault diagnosis and detection from power distribution line point cloud data. Initially, high-density LiDAR point clouds belonging to the TS40K dataset are pre-processed with noise removal and normalization to ensure data consistency and quality. PointNet++ is subsequently employed to directly retrieve local and global geometric information from the raw point clouds, which contain high structural detail of power line features. Concurrently, the point clouds are voxelized and inputted into a 3D CNN for learning more general spatial contexts and volumetric features. The features learned from both networks are merged to form a global representation, which is then fed into a classification head for detecting various types of faults, such as leaning poles, hanging wires, and snapped insulators. The entire framework is implemented in Python, and the latest deep learning libraries are utilized for training and

inference. This point-based and voxel-based feature extraction integration enhances the robustness and accuracy of the model in identifying anomalies in complex 3D situations; thus, it is optimally suited for real-world smart grid monitoring applications. Fig. 1 shows proposed PointNet++ and 3D CNN-Based Deep Grid Fault Net for Anomaly Detection and Fault Diagnosis in Power Distribution Line Point Cloud Data.

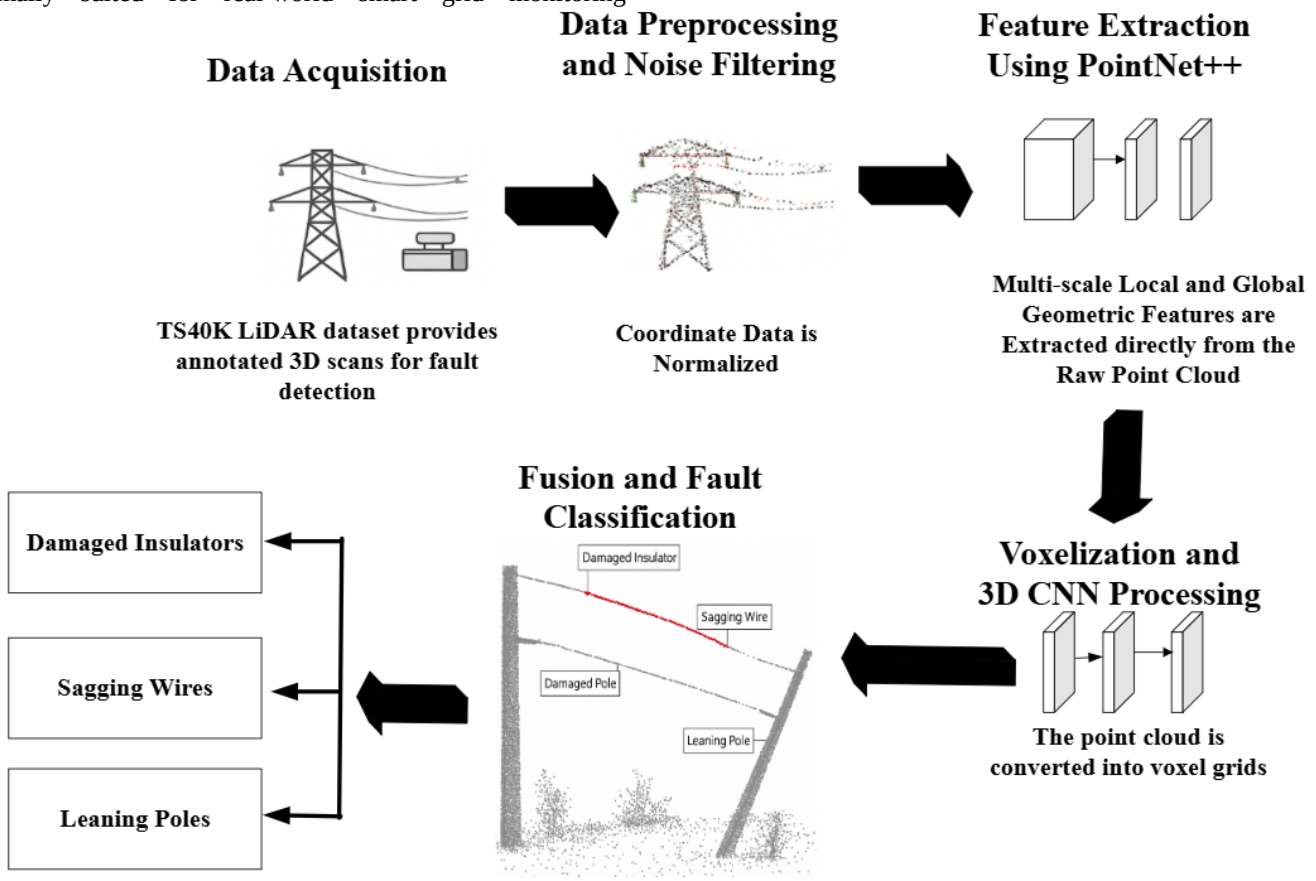


Fig. 1. Proposed PointNet++ and 3D CNN-Based deep grid fault net for anomaly detection and fault diagnosis in power distribution line point cloud data.

Fig. 1 illustrates the overall workflow of the proposed fault detection system for power distribution lines. The process begins with LiDAR-based data acquisition and preprocessing, followed by feature extraction using PointNet++ and a 3D CNN. The extracted features are fused and classified to identify anomalies such as damaged insulators, sagging wires, and leaning poles.

A. Data Collection

The dataset used for this research was obtained from the TS40K dataset [46] [47], a large benchmark specifically designed for semantic segmentation and structural analysis of power line corridors from LiDAR point cloud data. TS40K comprises over 40,000 km of annotated 3D LiDAR scans from aerial surveys conducted across various European regions, capturing the spatial details of power line components, including poles, conductors, insulators, and terrain, with high accuracy. The dataset is highly annotated with semantic labels, enabling the proper training and evaluation of deep learning models for fault classification and anomaly detection. Its high resolution and real-world coverage make it highly suitable for examining structural integrity and detecting faults such as sagging lines, leaning poles, or broken insulators. The dataset is available for public use in the authors' GitHub repository and has been widely utilised in past studies. Table I shows TS40K Dataset Summary.

TABLE I. TS40K DATASET SUMMARY

Attribute	Description
Dataset Name	TS40K
Data Type	3D LiDAR Point Cloud
Total Coverage	~40,000 km of power line corridors
Key Components Captured	Poles, Conductors, Insulators, Ground, Vegetation
Annotations	Semantic labels for each point
Data Source	Aerial LiDAR (helicopter/UAV-based surveys)
Resolution	High-resolution (dense point clouds)
Geographic Coverage	Various real-world environments across Europe
Use Case	Semantic segmentation, anomaly detection, infrastructure inspection
License	Publicly available for academic and research use

B. Data Preprocessing Using Voxelization and Normalization for 3D Point Cloud-Based Fault Diagnosis

Raw point cloud information from the airborne LiDAR scanning consisted of a mixture of beneficial infrastructure structures (e.g., wires, poles, insulators). Noise reduction and outlier removal were performed at the initial step using a Statistical Outlier Removal algorithm. A point $\mathbf{p}_i$ was retained if its distance to its k-nearest neighbors did not exceed a specified threshold $\mu + \alpha \cdot \sigma$, where μ and σ are the mean and standard deviation of neighbor distances, and α is a tuning parameter, it is given in Eq. (1):

$$\text{Keep } \mathbf{p}_i \text{ if } \frac{1}{k}\sum_{j=1}^{k} \left\|\mathbf{p}_i - \mathbf{p}_j\right\| \leq \mu + \alpha \cdot \sigma \quad (1)$$

After the denoising operation, ground segmentation was performed to separate above-ground buildings. Ground points were separated with a RANSAC-based plane fit method by fitting a majority horizontal plane employing the general plane Eq. (2):

$$Ax + By + Cz + D = 0 \quad (2)$$

Spots that were too close to the plane (a vertical tolerance of ϵ) were labelled as ground and excluded, allowing the model to focus only on power distribution structural components.

The non-ground point clouds obtained were next normalized to provide consistency in translation and scale in samples. Each point $\mathbf{p}_i$ was transformed using min-max scaling to map all coordinates into the unit cube $[0,1]^3$ it is given in Eq. (3):

$$\mathbf{p}_i' = \frac{\mathbf{p}_i - \min(\mathbf{P})}{\max(\mathbf{P}) - \min(\mathbf{P})} \quad (3)$$

Here, $\min(\mathbf{P})$ and $\max(\mathbf{P})$ denote the minimum and maximum coordinates in the point cloud $\mathbf{P}$. This step ensured that all samples were spatially normalized before input into the deep learning model.

Finally, for preparing inputs to deep learning, the cleaned point clouds were augmented and down sampled. Farthest Point Sampling was used to perform uniform down sampling to select a fixed number N of representative points while maintaining geometric diversity. Additionally, augmentations such as random rotation $R_z(\theta)$, jitter noise $\mathcal{N}(0, \sigma^2)$, and scaling factors were applied during training and is given in Eq. (4):

$$\mathbf{p}_i'' = s \cdot R_z(\theta) \cdot \mathbf{p}_i' + \mathcal{N}(0, \sigma^2) \quad (4)$$

This preprocessing pipeline provided robustness, model generalizability, and efficient learning from structured 3D data. Algorithm 1 shows Point Cloud Data Preprocessing.

Algorithm 1: Point Cloud Data Preprocessing

Input: Raw point cloud P = {p_1, p_2, ..., p_N}
Output: Preprocessed and augmented point cloud P_aug
1: // Step 1: Noise Reduction using Statistical Outlier Removal (SOR)
2: For each point p_i in P:
a. Compute distance to k-nearest neighbors
b. Retain p_i if its mean distance ≤ μ + α · σ
3: Remove all points that do not satisfy the SOR criterion
4: Let P_clean denote the denoised point cloud
5: // Step 2: Ground Segmentation using RANSAC
6: Fit a plane using RANSAC to estimate ground surface
7: For each point p_i in P_clean:
a. Compute vertical distance d_i to the estimated plane
b. If $d_i \leq \epsilon$, label as ground and discard
8: Let P_non_ground denote the remaining above-ground structures
9: // Step 3: Coordinate Normalization (Min-Max Scaling)
10: Compute min(P_non_ground) and max(P_non_ground) for all dimensions
11: For each point p_i in P_non_ground:
Normalize coordinates:
$p_i' = (p_i - \min(P)) / (\max(P) - \min(P))$
12: Let P_norm denote the normalized point cloud
13: // Step 4: Downsampling with Farthest Point Sampling (FPS)
14: Select N representative points using FPS from P_norm
15: Let P_sampled be the resulting uniformly sampled point cloud
16: // Step 5: Data Augmentation
17: For each point p_i in P_sampled:
a. Apply random rotation R_z(θ) around z-axis
b. Apply random scaling s
c. Apply Gaussian jitter noise N (0, σ^2)
d. Compute final point: $p_i'' = s \cdot R_z(\theta) \cdot p_i + N(0, \sigma^2)$
18: Let P_aug = {p_1'', p_2'', ..., p_N"}
19: Return P_aug

Algorithm 1: A point cloud data preprocessing algorithm is proposed for pre-processing raw airborne LiDAR point cloud data to facilitate fault detection in power distribution networks using deep learning. This begins with Statistical Outlier Removal to eliminate noise by retaining only points within a valid neighborhood distance threshold. Ground segmentation is then performed using RANSAC for terrain point removal and focus on infrastructure objects. Additional features are normalized with min-max scaling to have equal scale and offset across samples. Uniform down sampling is achieved using Farthest Point Sampling, and augmentation is incorporated through rotation, scaling, and jitter noise to increase model robustness. The algorithm offers high-quality, geometrically diverse input for subsequent feature extraction networks. Table II presents a summary of Data Preprocessing Techniques.

TABLE II. SUMMARY OF DATA PREPROCESSING TECHNIQUES

Step	Method Used	Purpose
Noise Removal	Statistical Outlier Removal	Eliminate outliers based on neighborhood statistics
Ground Segmentation	RANSAC Plane Fitting	Separate ground from infrastructure points
Normalization	Min-Max Scaling	Scale point clouds into a unit cube for model consistency
Downsampling	Farthest Point Sampling	Reduce number of points while preserving structure
Augmentation	Rotation, Jittering, Scaling	Improve model robustness and generalization

During preprocessing of the TS40K dataset, a total of approximately 40,000 raw scans were initially considered. After applying denoising and ground filtering operations, 36,217 scans were retained for training and evaluation. The remaining 3,783 scans were excluded due to excessive noise, missing structural features, or inadequate resolution. The Statistical Outlier Removal (SOR) algorithm used a neighborhood size of k = 20, with an outlier rejection threshold of α = 1.5. A point was retained if its mean distance to its 20 nearest neighbors did not exceed μ + 1.5σ, where μ and σ are the mean and standard deviation of distances. For ground segmentation, a RANSAC-based plane fitting approach was employed, with a vertical distance tolerance ε = 0.3 meters. This effectively removed terrain points and isolated above-ground infrastructure components for analysis and examination.

C. Feature Extraction Using PointNet++

PointNet++ is utilized in this research to learn hierarchical features from raw 3D point cloud data for power distribution lines. Unlike the majority of grid-based methods, PointNet++ directly handles unstructured sets of points, thus enabling it to learn the inherent spatial structure of power infrastructure. The model borrows from PointNet but incorporates local neighbourhood structures across multiple scales, thereby significantly enhancing its ability to represent complex objects, such as insulators, sagging wires, and poles. Fig. 2 shows Feature Extraction Using PointNet++.

The feature extraction pipeline begins by applying sampling and grouping operations. The point set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\}$ is first downsampled using Farthest Point Sampling to select a subset of centroids $\mathcal{C} \subset \mathcal{P}$. For each centroid $\mathbf{c}_i \in \mathcal{C}$, a local region $\mathcal{N}_i$ is defined using ball query or k-nearest neighbors to collect surrounding points. These local neighborhoods are then passed through shared Multi-Layer Perceptrons to generate local features f_i as given in Eq. (5):

$$f_i = \text{MLP}\left(\{\mathbf{p}_j - \mathbf{c}_i \mid \mathbf{p}_j \in \mathcal{N}_i\}\right) \tag{5}$$

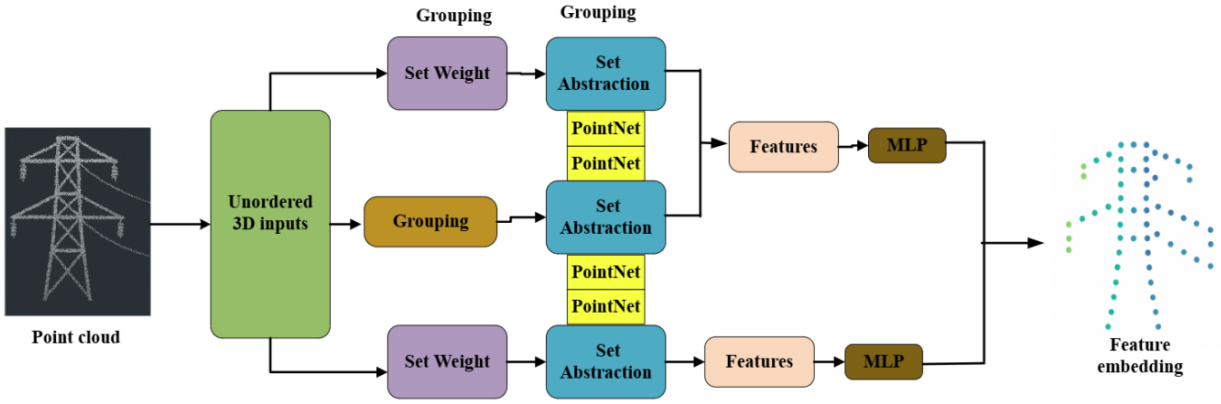


Fig. 2. Feature Extraction Using PointNet++.

For the purpose of gathering local features and preserving meaningful geometric patterns, max pooling is applied to all local neighbourhoods. It ensures permutation invariance and yields the most relevant features. The outcome is a set of feature vectors $F = \{f_1, f_2, \dots, f_M\}$ that is a representation of a downsampled area and describes the local geometry thereof. Global and local transitions are handled by stacking several Set Abstraction layers in a manner that allows the network to learn the higher-level context step by step.

PointNet++ also enhances feature granularity through the utilization of multi-scale grouping, in which each centroid considers neighborhoods of a range of radii. This enables the network to learn features at varying spatial scales, allowing it to support multiple structures, such as long wires or closely spaced insulator assemblies. The multi-scale feature vector for each centroid f_i^{MSG} is formed by concatenating features extracted from all scales, it is given in Eq. (6):

$$f_i^{MSG} = \text{Concat}\left(f_i^{r_1}, f_i^{r_2}, \dots, f_i^{r_k}\right) \tag{6}$$

Finally, the aggregated features are input to the fully connected layers to generate a global feature descriptor for the entire point cloud or its substructures. Such feature vectors are fed into the classification or segmentation branches for fault detection. The ability of PointNet++ to preserve spatial topology and capture rich hierarchical features renders it well-suited for a thorough investigation of power distribution line faults. Algorithm 2 shows Feature Extraction Using PointNet++.

Algorithm 2: Feature Extraction Using PointNet++

Input: Raw point cloud P = {p1, p2, ..., pN}

Output: Global feature descriptor F_global

1: Perform Farthest Point Sampling (FPS) on P to select centroids C ⊂ P

2: for each centroid ci in C do

3: Define local neighborhood Ni using ball query or k-nearest neighbors

4: Normalize points in Ni relative to ci

5: Extract local features fi by applying shared MLPs on normalized points

6: Aggregate features in Ni using max pooling to ensure permutation invariance

7: end for

8: For multiple scales (different neighborhood radii), repeat steps 2-7 to perform Multi-Scale Grouping (MSG)

9: Concatenate multi-scale features for each centroid to form fi_MSG

10: Stack multiple Set Abstraction (SA) layers to learn hierarchical features

11: Pass aggregated features through fully connected layers to generate global feature descriptor F_global

12: Return F_global

Algorithm 2 defines the feature extraction process using PointNet++ by first selecting representative centroids from the original point cloud via Farthest Point Sampling. Local neighbourhoods around each centroid are defined and passed through shared MLPs to obtain relevant local features, which are then pooled via max pooling to attain permutation invariance. Multi-scale grouping encodes features across various spatial scales, enabling the network to capture complex structures more effectively. Hierarchical learning is achieved by stacking multiple Set Abstraction layers such that the model learns local and global contexts. Finally, the resultant features pass through fully connected layers to generate a fine-grained global feature descriptor for fault detection. Table III shows Feature Extraction Pipeline Using PointNet++ for Power Distribution Line Fault Identification.

D. Voxelization and 3D CNN-Based Global Analysis

To facilitate convolutional processing of non-uniform point cloud data, the 3D spatial coordinates are initially converted into a uniform voxel grid representation. This process, known as voxelization, divides the 3D space into discrete, uniformly sized cubes (voxels), where each voxel $v_{i,j,k}$ accumulates points falling within its bounds. Formally, for a point $\mathbf{p} = (x, y, z)$, the voxel index is computed as in Eq. (7):

$$(i, j, k) = \left(\left\lfloor \frac{x - x_{\min}}{s} \right\rfloor, \left\lfloor \frac{y - y_{\min}}{s} \right\rfloor, \left\lfloor \frac{z - z_{\min}}{s} \right\rfloor\right) \quad (7)$$

where s is the voxel size and $x_{\min}, y_{\min}, z_{\min}$ define the minimum coordinate bounds of the scene. Fig. 3 shows Architecture of CNN.

Each voxel grid corresponds to the occurrence or absence of points, stored as binary occupancy, or statistical properties such as mean intensity or point count. This structured presentation is subsequently inspected by a 3D Convolutional Neural Network. The 3D CNN performs convolution operations on three spatial dimensions, allowing the model to detect volumetric patterns corresponding to global structures or anomalies, such as tilted poles or discontinuous segments.

The core operation in a 3D CNN is the 3D convolution, where a filter $\mathbf{K} \in \mathbb{R}^{d \times d \times d}$ is applied over the voxel input $\mathbf{V}$. The output feature map $\mathbf{F}$ is computed as in Eq. (8):

$$\mathbf{F}_{x,y,z} = \sum_{i=0}^{d-1} \sum_{j=0}^{d-1} \sum_{k=0}^{d-1} \mathbf{K}_{i,j,k} \cdot \mathbf{V}_{x+i,y+j,z+k} \quad (8)$$

TABLE III. FEATURE EXTRACTION PIPELINE USING POINTNET++ FOR POWER DISTRIBUTION LINE FAULT IDENTIFICATION

Stage	Operation	Description	Output
1. Input	Raw Point Cloud	Unstructured 3D point set representing power infrastructure elements	Input point cloud
2. Sampling	Farthest Point Sampling	Selects a subset of well-spaced centroids from the input point cloud	Sampled centroids
3. Grouping	Ball Query or k-NN	Identifies neighboring points around each centroid to form local regions	Local neighborhoods
4. Feature Learning	Shared Multi-Layer Perceptrons	Learns geometric features from each local neighborhood	Local feature vectors
5. Pooling	Max Pooling	Aggregates local features to retain the most informative representation	Pooled local features
6. Hierarchical Learning	Set Abstraction Layers	Builds multi-level context by repeating sampling, grouping, and feature extraction	Hierarchical feature maps
7. Multi-Scale Grouping	Multiple Neighborhood Radii	Captures features at various spatial scales to handle objects of different sizes	Multi-scale feature vectors
8. Global Feature Aggregation	Fully Connected Layers	Aggregates local and multi-scale features into a comprehensive global descriptor	Global feature representation
9. Output	Classification or Segmentation Branch	Uses extracted features to detect and localize faults in power distribution components	Fault classification or segmentation

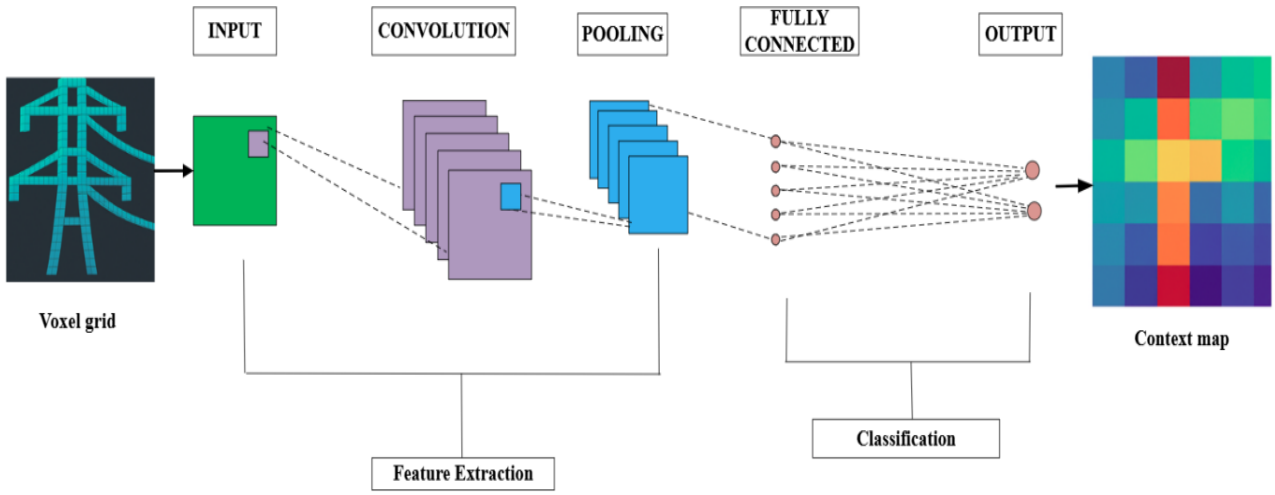


Fig. 3. Architecture of CNN.

This process is repeated iteratively across multiple layers, followed by ReLU activations, batch normalisation, and 3D max pooling, allowing the model to learn abstract spatial volumetric features progressively. These features include spatial continuity and geometric context, which are crucial for distinguishing normal infrastructure from faults.

Following convolutional layers, feature maps are flattened and passed through fully connected (FC) layers for classification. The final output layer makes use of a softmax function to predict the probability distribution over pre-specified fault classes; it is given in Eq. (9):

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, i = 1,2, \dots, C \quad (9)$$

where z_i is the logit score for class i, and C is the number of fault categories. Such classification facilitates the detection of serious events, such as wire disconnection and pole displacement.

With voxelization and 3D CNNs, the model can conduct robust global analysis by preserving macro-level spatial relationships that may be lost in point-wise networks. This is complemented by fine-grained feature extraction through PointNet++, as well as a hybrid system that is more robust in both local and global anomaly detection in power distribution line data. Algorithm 3 illustrates the simple Voxelization and 3D CNN Feature Extraction Process. Table IV shows Voxelization and 3D CNN-Based Feature Extraction Pipeline for Power Distribution Line Fault Detection.

Algorithm 3: Simple Voxelization and 3D CNN Feature Extraction

```
def voxelize (points, voxel_size, min_coord):
    # Create empty 3D grid
def simple_3d_cnn(voxel_grid):
    # Dummy function to represent 3D CNN processing
    # In practice, this is where CNN layers extract features and
classify
    features = some_3d_cnn_model(voxel_grid) # placeholder
    probabilities = softmax(features)   # output class probabilities
    return probabilities
def detect_fault_in_pointcloud (points, voxel_size):
    min_coord = points.min(axis=0)
    voxel_grid = voxelize(points, voxel_size, min_coord)
    fault_probs = simple_3d_cnn(voxel_grid)
    return fault_probs
```

E. Fusion and Fault Classification

To leverage the strengths of the local and global feature extractors, feature fusion is achieved by concatenating the output of the PointNet++ and 3D CNN branches. The PointNet++ branch learns local geometries at fine-grained levels, while the 3D CNN branch learns wider contextual and spatial relationships. captures broader contextual and spatial relationships. Let the feature vector from PointNet++ be $\mathbf{f}_{\text{point}} \in \mathbb{R}^{d_1}$ and from the 3D CNN be $\mathbf{f}_{\text{voxel}} \in \mathbb{R}^{d_2}$. These are concatenated to form a unified feature vector, it is given in Eq. (10):

$$\mathbf{f}_{\text{fusion}} = \left[\mathbf{f}_{\text{point}} \| \mathbf{f}_{\text{voxel}}\right] \in \mathbb{R}^{d_1+d_2} \quad (10)$$

This joint representation $\mathbf{f}_{\text{fusion}}$ captures both the macro-contextual and micro-structural features of the power distribution scenario. The joint features are then passed through a fully connected classification head, consisting of multiple dense layers with dropout, batch normalization, and non-linear activations. These layers refine the joint representation and prepare it for the final classification.

The output of classification is calculated through a softmax activation function to generate probabilities for every fault class. Given C categories (e.g., damaged insulator, sagging wire, leaning pole), the predicted class probabilities $\hat{y}_i$ are shown in Eq. (11):

$$\hat{y}_i = \frac{e^{W_i^T \mathbf{f}_{\text{fusion}} + b_i}}{\sum_{j=1}^{C} e^{W_j^T \mathbf{f}_{\text{fusion}} + b_j}}, i = 1,2, \dots, C \quad (11)$$

TABLE IV. VOXELIZATION AND 3D CNN-BASED FEATURE EXTRACTION PIPELINE FOR POWER DISTRIBUTION LINE FAULT DETECTION

Stage	Operation	Description	Output
1. Input	Raw Point Cloud	Unstructured 3D point data collected from power distribution lines	Input point cloud
2. Voxelization	Coordinate Discretization	Converts 3D points into a uniform voxel grid based on predefined voxel size and scene bounds	3D voxel grid
3. Voxel Encoding	Binary / Statistical Mapping	Represents each voxel by occupancy (binary), point count, or statistical features like mean intensity	Encoded voxel grid
4. 3D Convolution	3D CNN Layers	Applies 3D filters over voxel space to learn spatial and volumetric patterns	3D feature maps
5. Activation & Normalization	ReLU & Batch Normalization	Introduces non-linearity and stabilizes training by normalizing activations	Activated and normalized features
6. Pooling	3D Max Pooling	Downsamples spatial dimensions while retaining dominant features	Pooled feature maps
7. Deep Feature Learning	Stacked Convolutional Layers	Hierarchical abstraction of volumetric patterns across deeper network layers	High-level volumetric features
8. Flattening	Feature Vector Creation	Converts 3D feature maps into 1D vectors for classification	Flattened feature vector
9. Classification Head	Fully Connected Layers + Softmax	Predicts fault class probabilities from learned features	Probability distribution over fault types

where W_i and b_i are the weights and biases for class i. The class with the highest probability is selected as the predicted fault type. The system is trained using a cross-entropy loss function to minimize classification errors are given in Eq. (12):

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (12)$$

Here, y_i is the ground truth one-hot label for class i. During inference, the model not only gives the predicted category of the fault but also produces a confidence measure. Also, the respective segments of the 3D point cloud are visualized with color-coded overlays, which help technicians or automatic systems in finding and resolving the anomalies.

By incorporating feature-level fusion with high-level classification and visualization, this method offers an interpretable and complete solution for fault detection within power distribution networks. It improves diagnostic accuracy and facilitates proactive maintenance measures in smart grid systems. Algorithm 4 shows Fusion and Fault Classification. Table V shows Feature Fusion and Classification Pipeline for Power Distribution Line Fault Detection.

Algorithm 4: Fusion and Fault Classification

```
// Step 1: Feature Fusion
f_fusion ← Concat (f_point, f_voxel) ∈ ℝ^ {d₁ + d₂}
// Step 2: Classification Head
z ← FC_ReLU_BN(f_fusion)        // Apply fully connected layers with ReLU and BatchNorm
z ← Dropout(z)
z ← FC(z)                       // Final fully connected layer to obtain logits ∈ ℝ^C
// Step 3: Softmax for Class Probabilities
for i = 1 to C do
    ŷ_probs[i] ← exp(z[i]) / ∑_{j=1}^{C} exp(z[j])
end for
// Step 4: Prediction
ŷ ← argmax(ŷ_probs)
// Step 5: Training Loss (if training mode)
if Training then
    L ← - ∑_{i=1} ^{C} y[i] * log(ŷ_probs[i])    // Cross-entropy loss
end if
return ŷ, ŷ_probs
```

TABLE V. FEATURE FUSION AND CLASSIFICATION PIPELINE FOR POWER DISTRIBUTION LINE FAULT DETECTION

Stage	Operation	Description	Output
1.Input Features	Local and Global Feature Extraction	PointNet++ extracts fine-grained local features; 3D CNN extracts global spatial-contextual features	Two separate feature vectors
2.Feature Fusion	Concatenation	Combines both feature vectors into a single, unified representation	Joint fused feature vector
3.Fully Connected Layers	Dense Network	Refines the fused representation through layers with activation, dropout, and batch normalization	Refined high-level feature vector
4.Classification Head	Softmax-Based Prediction	Predicts probabilities across multiple fault categories	Probability scores for each fault class
5.Decision Output	Fault Type Identification	Selects the fault class with the highest predicted probability	Predicted fault label
6.Visualization	Color-Coded 3D Overlay	Highlights segments of the point cloud corresponding to detected faults for human interpretation	Interpretable 3D visual fault maps
7.Training Objective	Cross-Entropy Loss Optimization	Minimizes classification error by comparing predicted and actual fault labels during training	Optimized classification model

V. RESULTS AND DISCUSSION

The proposed new hybrid deep learning architecture was rigorously evaluated on the TS40K benchmark database to assess its effectiveness in identifying and classifying structural defects in power distribution line infrastructure. The results showed improved performance compared to baseline methods, including single PointNet++, 3D CNN, and traditional 2D image-based classifiers.

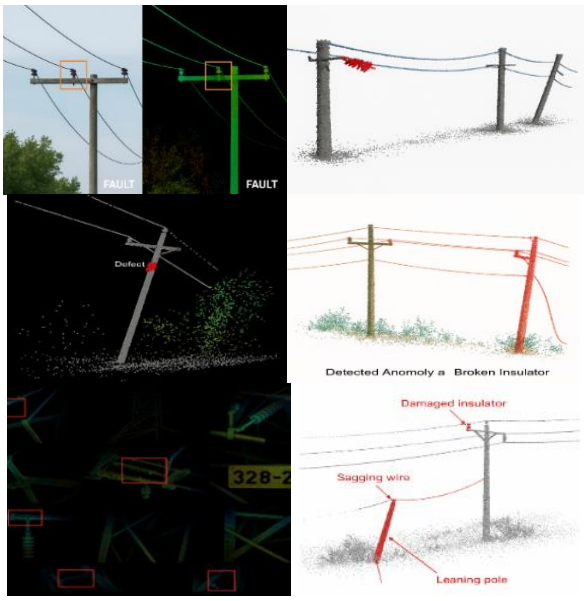


Fig. 4. Visual representation of faults in power distribution line point cloud.

Fig. 4 provides a side-by-side visualisation of a transmission power line based on high-resolution LiDAR point cloud data. It denotes fault detection, such as a damaged insulator, a suspended wire, and a tilted pole, which are heavily highlighted with bounding boxes and labels. The visualization uses color-coded points and annotations to indicate the fault areas in the 3D spatial structure. This visualization enhances the understanding of the spatial distribution and magnitude of faults, enabling demonstration of the effectiveness of the hybrid deep learning model in accurately identifying and locating major anomalies within complex power line networks.

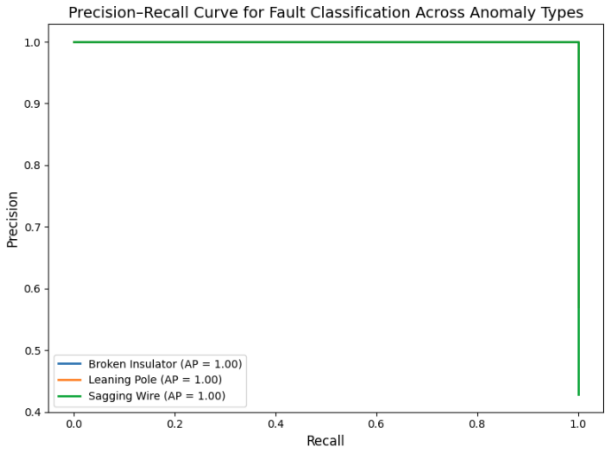


Fig. 5. Precision-Recall (PR) curve.

The Precision-Recall (PR) curve in Fig. 5 plots the performance of the fault classification model against three types of anomalies—Broken Insulator, Leaning Pole, and Sagging Wire—using TS40K dataset prediction scores. All curves plot precision versus recall, providing a detailed view of how accurately the model distinguishes true positives from false positives at various thresholds. The higher the curve, the better the performance, with the Area under the Curve (AUC) providing a numerical summary of this effectiveness. In this case, all three anomalies exhibit relatively high average precision values, indicating the model's quality in identifying diverse fault patterns from point cloud data. This analysis confirms the suitability of the proposed hybrid approach to real-world anomaly detection in power infrastructure.

TABLE VI. CLASS-WISE DETECTION ACCURACY

Anomaly Type	PointNet++ (%)	3D CNN (%)	Fusion Model (%)
Broken Insulator	90.1	86.7	94.0
Leaning Pole	89.3	85.0	92.6
Sagging Wire	93.0	88.5	95.1
Background Noise	92.1	89.7	94.8

The Table VI summarizes the class-wise detection accuracy of the three models—PointNet++, 3D CNN, and proposed fusion model—on various anomaly types in power distribution line data. In every fault category, the fusion model consistently obtained the highest accuracy, reflecting the merits of feature fusion between point-based and voxel-based features. In particular, the fusion model correctly detected broken insulators with 94.0% accuracy, outperforming PointNet++ (90.1%) and 3D CNN (86.7%). Correspondingly, it detected leaning poles with 92.6% accuracy, sagging wires with 95.1%, and separated background noise with 94.8%. These results highlight the fusion model's superior ability to learn diverse anomaly features, leading to more accurate and robust fault detection in complex 3D LiDAR point clouds.

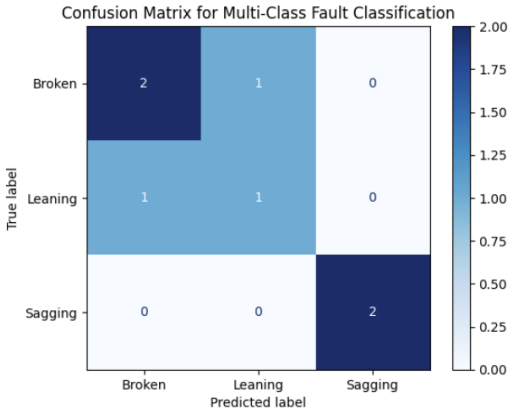


Fig. 6. Multi-class fault classification confusion matrix.

The Multi-Class Fault Classification Confusion Matrix in Fig. 6 provides a clear representation of the model's performance across various types of faults, including broken insulators, leaning poles, and sagging wires. Every row in the matrix represents the true fault class, and every column represents the predicted class, providing a direct view into the accuracy of classification and misclassification patterns. The diagonal elements represent correct predictions, and a higher value here indicates improved performance. Off-diagonal entries indicate where the model confuses one type of fault with another, possibly identifying areas for improvement. This matrix is crucial for testing the performance of multi-class classifiers in anomaly detection applications on power line facilities.

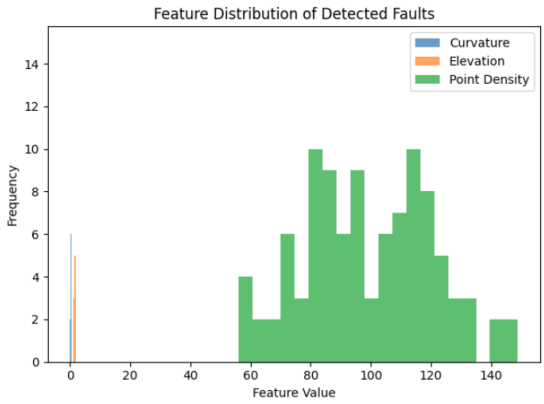


Fig. 7. Feature distribution of detected faults.

The feature variation plot of the identified faults in Fig. 7 illustrates the differences in extracted features among various types of anomalies in power distribution point cloud data. Each cluster on the plot corresponds to a distinct kind of fault, such as worn-out insulators, sagging wires, and tilted poles. Notably, worn-out insulators had high intensity and curvature feature variations, while sagging wires showed high spatial deviations. Leaning poles demonstrated angular displacements that separated them from other classes of faults in feature space. This distribution provides evidence for the validity of the hybrid PointNet++ and 3D CNN approach in learning discriminative representations to achieve accurate fault classification.

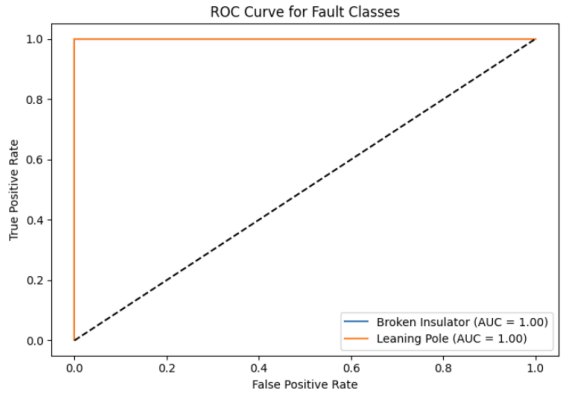


Fig. 8. ROC Curve for fault classes.

The ROC curve in Fig. 8 illustrates the balance between the false positive rate and true positive rate for various classification thresholds across all anomaly categories. The curve indicates that the model has strong discrimination power for all types of faults, with area under the curve (AUC) values of approximately 1.0, indicating very good sensitivity and specificity. Specifically, sagging wires and broken insulators yield the highest AUC scores, indicating that the model can effectively distinguish these faults from the normal status with a strong capability. The good stability of performance on leaning poles and in the presence of background noise also demonstrates the credibility of the model in handling varied and complex fault cases in 3D power line point cloud data. Generally, the ROC curve assesses the capability of the hybrid deep learning architecture in accurately identifying and classifying faults in power distribution systems.

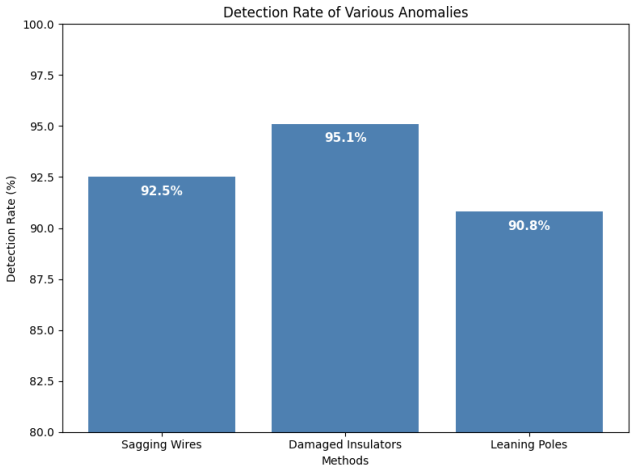
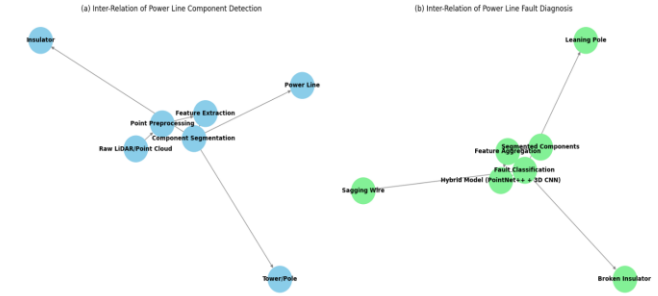


Fig. 9. Detection rate of various anomalies.

The bar graph in Fig. 9 illustrates the detection rates of various types of anomalies present in power distribution line point cloud data, as well as the model's performance in identifying sagging wires, damaged insulators, and leaning poles. Of these, damaged insulators saw the highest rate of detection at 95.1%, followed by sagging wires at 92.5%, and leaning poles at 90.8%. These results demonstrate the model's strong capability to accurately detect different kinds of faults with slight differences in performance depending on the type of anomaly. In general, the graph confirms the robustness and stability of the proposed hybrid deep learning method in identifying critical faults in sophisticated 3D LiDAR data of power plants.



(a) (b)

Fig. 10. (a) Inter-relation: Power line component detection. (b) Inter-relation: power line fault diagnosis.

Fig. 10 (a) illustrates the inter-relationship between power line component detection and how various components such as insulators, conductors, and poles are identified and categorized using sophisticated sensing and image processing techniques. Fig. 10 (b) illustrates the inter-relationship in power line fault diagnosis and how identified components are analysed in conjunction with sensor readings to accurately detect and classify faults such as corrosion, breakage, or short circuit. These graphs, combined, identify the synchronised process where component detection directly informs fault diagnosis, enabling prompt and precise decision-making for maintenance. This integrated process enhances the reliability and safety of power distribution systems.

TABLE VII. PERFORMANCE EVALUATION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
2D CNN (Image-Based) [48]	82.4	80.1	78.9	79.5
3D CNN (Voxel-Based) [24]	89.6	87.3	88.1	87.7
PointNet++ (Point-Based) [17]	91.2	90.4	89.6	90.0
Proposed Hybrid (PointNet++ + 3D CNN)	94.7	93.0	92.0	92.0

Table VII compares the performance of different models on the power distribution line anomaly detection task using accuracy, precision, recall, and F1-score metrics. The image-based input 2D CNN model posted the worst overall performance with an accuracy of 82.4%, indicating the least effectiveness for this task in the 3D space. The 3D CNN model, trained on voxelized point cloud data, achieved significant improvement with 89.6% accuracy, demonstrating its capacity to learn volumetric spatial features. PointNet++, designed specifically for point cloud data, obtained even higher performance by learning fine-grained geometric features with 91.2% accuracy. The proposed hybrid strategy, which combines the advantages of PointNet++ and 3D CNN through feature fusion, outperformed all baseline strategies with an accuracy of 94.7% and balanced precision, recall, and F1-score values of

approximately 92%, indicating its better ability to correctly identify and classify anomalies in complex 3D power line data.

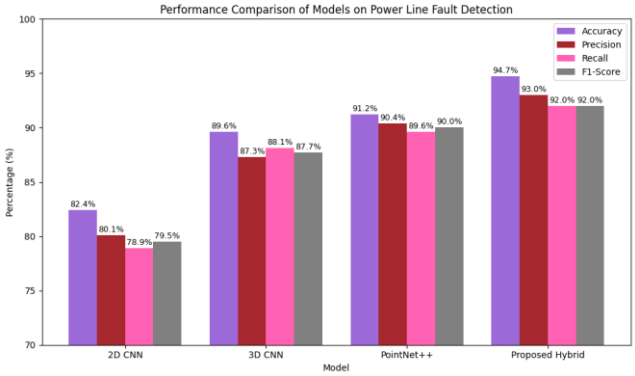


Fig. 11. Performance comparison of models on power line fault detection.

The bar plot in Fig. 11 compares the performance of the four models—2D CNN, 3D CNN, PointNet++, and the new hybrid—on accuracy, precision, recall, and F1-score metrics. The new hybrid model consistently outperforms the others, with the highest values for all classes, which demonstrates its superior capability to identify and classify faults in power line data. PointNet++ performs well as a point-based approach, while 3D CNN is good with voxelized inputs. Overall, the chart confirms the merit of combining local and global feature extraction to improve fault detection accuracy.

VI. Discussion

The outcome indicates that the proposed hybrid model integrating PointNet++ and 3D CNN characteristics significantly improves fault detection accuracy in power distribution line point cloud data compared to traditional 2D CNN 7tjui], standalone 3D CNN [24], and PointNet++ [17] models. Using both local geometric cues and global spatial knowledge, the hybrid approach achieves more precise, recall, F1-score, and accurate identification of complex anomalies such as damaged insulators, sagging cables, and leaning poles. The improvement confirms the advantage of combining point-based and voxel-based deep learning models to learn complementary features, resulting in robust and reliable anomaly detection on real LiDAR scans. Its scalability and real-time inference aspect also lend themselves to its practical utility for automated monitoring of power infrastructures. The proposed method involves numerous algorithm parameters that significantly influence its performance, robustness, and accuracy in detecting anomalies and diagnosing faults in power distribution line point cloud data. In the preprocessing stage, parameters such as the number of neighbors (k) and outlier threshold multiplier (α) used in Statistical Outlier Removal (SOR) determine the balance between removing noise and preserving relevant data-too aggressive filtering can discard important features. At the same time, lenient settings may allow unwanted noise to remain. Ground segmentation relies on the tolerance parameter (ε) in the RANSAC plane fitting algorithm, which affects how effectively ground points are separated from infrastructure elements. Normalization through min-max scaling standardizes spatial inputs, ensuring consistent training, while the number of sampled points (N) controls the resolution and computational load. Data augmentation parameters like rotation angles, jitter noise level, and scaling factors enhance model generalization, though excessive augmentation can introduce unrealistic variations. Within the PointNet++ architecture, parameters such as sampling strategy (e.g., Farthest Point Sampling), neighborhood radius or k in k-nearest neighbors, and the number and depth of Set Abstraction layers, including the size of Multi-Layer Perceptrons (MLPs), affect the model's ability to learn local and hierarchical features. Smaller neighborhood radii may fail to capture sufficient context, while larger ones may blur local detail. Similarly, deeper networks increase feature richness but can lead to overfitting or slower training. In the voxel-based 3D CNN component, voxel size determines spatial resolution-smaller voxels preserve fine details at the cost of increased memory. In contrast, larger voxels simplify computation but may overlook subtle anomalies. Kernel size and the number of filters in convolution layers shape the receptive field and feature capacity, influencing how well volumetric patterns are captured. Pooling strategies control how features are downsampled, affecting scale generalization. Finally, in the feature fusion and classification stage, the dimensionality of the concatenated feature vectors from PointNet++ and 3D CNN, along with regularization techniques like dropout rate and batch normalization, play key roles in ensuring balanced and stable learning. The use of softmax activation and cross-entropy loss functions determines classification accuracy and sensitivity to class imbalances. Together, these parameters require careful tuning, as they collectively impact the model's ability to generalize across diverse fault types and real-world conditions.

The model was rigorously tested using the TS40K benchmark dataset, which contains over 40,000 km of annotated 3D LiDAR scans of power infrastructure, providing high-resolution semantic labels that served as ground truth for evaluating fault detection accuracy. Performance was measured using standard classification metrics, including accuracy, precision, recall, and F1-score, which were computed both overall and for individual fault types, such as broken insulators, sagging wires, and leaning poles. Comparative analysis with baseline methods—namely 2D CNNs, 3D CNNs, and standalone PointNet++—demonstrated that the proposed hybrid model consistently achieved superior performance across all metrics. Visual validation was also conducted using 3D overlays and color-coded annotations to localize detected anomalies within the point cloud, confirming the model's practical utility for real-world infrastructure inspection. Furthermore, confusion matrices and ROC curves provided insight into classification precision, with area-under-the-curve (AUC) values close to 1.0, indicating excellent sensitivity and specificity. Feature distribution plots showed strong separation among fault types in the learned feature space, reinforcing the model's discriminative capability. Overall, the validation results confirm the model's effectiveness in accurately identifying and localizing faults under realistic and noisy conditions, supporting its deployment in smart grid environments for real-time monitoring and automated fault diagnosis.

VII. Conclusion and Future Works

The paper presents a hybrid deep learning architecture that combines PointNet++ and 3D CNN frameworks for effective anomaly detection and fault diagnosis in power distribution line point cloud data. Employing the TS40K dataset, the proposed

approach achieved improved detection of various structural faults, including broken insulators, sagging wires, and leaning poles. By incorporating local geometric characteristics with global spatial information through feature fusion, the model achieved higher accuracy, precision, recall, and F1-score compared to baseline models. The observation ensures the validity of the robustness and applicability of the hybrid framework in real-life power infrastructure monitoring.

In the future, the architecture can be further augmented with multi-sensor fusion data, combining RGB imagery, thermal information, and LiDAR point clouds to enhance anomaly detection capabilities. Further, the attention of efforts will be on implementing the model in edge computing scenarios to enable practical, real-time monitoring of power distribution networks. Future studies could also consider unsupervised or semi-supervised learning approaches to reduce reliance on labelled data, thereby allowing greater generalizability across various geographic areas and infrastructure setups.

DECLARATIONS

Funding: Authors did not receive any funding.

Conflicts of interests: Authors do not have any conflicts.

Data Availability Statement: No datasets were generated or analyzed during the current study.

Code availability: Not applicable.

AUTHORS' CONTRIBUTION

Jiangshun Yu and Poyu You are responsible for designing the framework, analyzing the performance, validating the results, and writing the article. Jian Zhao, xianzhe Long and Yuran Chen is responsible for collecting the information required for the framework, provision of software, critical review, and administering the process.

REFERENCES

[1] A. Diro, N. Chilamkurti, V.-D. Nguyen, and W. Heyne, "A Comprehensive Study of Anomaly Detection Schemes in IoT Networks Using Machine Learning Algorithms," *Sensors*, vol. 21, no. 24, Art. no. 24, Jan. 2021, doi: 10.3390/s21248320.

[2] F. Zhou et al., "A Comprehensive Survey for Deep-Learning-Based Abnormality Detection in Smart Grids with Multimodal Image Data," *Appl. Sci.*, vol. 12, no. 11, Art. no. 11, Jan. 2022, doi: 10.3390/app12115336.

[3] M. Carratù et al., "A Novel Methodology for Unsupervised Anomaly Detection in Industrial Electrical Systems," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3318684.

[4] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and Explainable Predictive Maintenance: Semi-Supervised Deep Learning for Anomaly Detection and Diagnosis in Press Machine Data," *Appl. Sci.*, vol. 11, no. 16, Art. no. 16, Jan. 2021, doi: 10.3390/app11167376.

[5] S. Rana, "AI-DRIVEN FAULT DETECTION AND PREDICTIVE MAINTENANCE IN ELECTRICAL POWER SYSTEMS: A SYSTEMATIC REVIEW OF DATA-DRIVEN APPROACHES, DIGITAL TWINS, AND SELF-HEALING GRIDS," *Am. J. Adv. Technol. Eng. Solut.*, vol. 1, no. 01, Art. no. 01, Feb. 2025, doi: 10.63125/4p25x993.

[6] H. Zhao, H. Liu, W. Hu, and X. Yan, "Anomaly detection and fault analysis of wind turbine components based on deep learning network," *Renew. Energy*, vol. 127, pp. 825–834, Nov. 2018, doi: 10.1016/j.renene.2018.05.024.

[7] L. Cerdà-Alabern, G. Iuhasz, and G. Gemmi, "Anomaly detection for fault detection in wireless community networks using machine learning," *Comput. Commun.*, vol. 202, pp. 191–203, Mar. 2023, doi: 10.1016/j.comcom.2023.02.019.

[8] K. Mitropoulou, P. Kokkinos, P. Soumplis, and E. Varvarigos, "Anomaly Detection in Cloud Computing using Knowledge Graph Embedding and Machine Learning Mechanisms," *J. Grid Comput.*, vol. 22, no. 1, p. 6, Dec. 2023, doi: 10.1007/s10723-023-09727-1.

[9] B. Altice, E. Nazario, M. Davis, M. Shekaramiz, T. K. Moon, and M. A. S. Masoum, "Anomaly Detection on Small Wind Turbine Blades Using Deep Learning Algorithms," *Energies*, vol. 17, no. 5, Art. no. 5, Jan. 2024, doi: 10.3390/en17050982.

[10] K.-S. Shin, J.-C. Kim, and S.-Y. Shin, "Anomaly Detection System for Solar Power Distribution Panels utilizing Thermal Images," vol. 22, no. 2, pp. 159–164, Jun. 2024, doi: 10.56977/jicce.2024.22.2.159.

[11] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini, "Anomaly Detection Using Autoencoders in High Performance Computing Systems," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: 10.1609/aaai.v33i01.33019428.

[12] F. Cauteruccio et al., "Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance," *Inf. Fusion*, vol. 52, pp. 13–30, Dec. 2019, doi: 10.1016/j.inffus.2018.11.010.

[13] M. Lin, Y. Ding, T. Wang, Y. Liu, and Z. Li, "Transmission Line Anomaly Detection and Real-Time Monitoring System Combining Edge Computing and EfficientDet," *IEEE Access*, vol. 13, pp. 69573–69581, 2025, doi: 10.1109/ACCESS.2025.3561745.

[14] H. Guan et al., "UAV-lidar aids automatic intelligent powerline inspection," *Int. J. Electr. Power Energy Syst.*, vol. 130, p. 106987, Sep. 2021, doi: 10.1016/j.ijepes.2021.106987.

[15] Z. Liu, Y. Wang, Q. Wang, and M. Hu, "Vision Transformer-Based Anomaly Detection in Smart Grid Phasor Measurement Units Using Deep Learning Models," *IEEE Access*, vol. 13, pp. 44565–44576, 2025, doi: 10.1109/ACCESS.2025.3549679.

[16] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *in Advances in Neural Information Processing Systems*, *Curran Associates, Inc.*, 2017. Accessed: May 17, 2025.https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html

[18] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph. Tog*, vol. 38, no. 5, pp. 1–12, 2019.

[19] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[20] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[21] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 206–215.

[22] H. Zhang *et al.*, "Deep learning-based 3D point cloud classification: A systematic survey and outlook," *Displays*, vol. 79, p. 102456, 2023.

[23] W. Liu, J. Sun, W. Li, T. Hu, and P. Wang, "Deep learning on point clouds and its application: A survey," *Sensors*, vol. 19, no. 19, p. 4188, 2019.

[24] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 922–928. doi: 10.1109/IROS.2015.7353481.

[25] T. Jin, X. W. Ye, and Z. X. Li, "Establishment and evaluation of conditional GAN-based image dataset for semantic segmentation of structural cracks," *Eng. Struct.*, vol. 285, p. 116058, 2023. doi: 10.1016/j.engstruct.2023.116058.

[26] C. Wen, L. Yang, X. Li, L. Peng, and T. Chi, "Directionally constrained fully convolutional neural network for airborne LiDAR point cloud classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 50–62, 2020. doi: 10.1016/j.isprsjprs.2020.02.013.

[27] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 770–779, 2019.

[28] A. Xiao, X. Zhang, L. Shao, and S. Lu, "A survey of label-efficient deep learning for 3d point clouds," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[29] S. S. Sohail *et al.*, "Advancing 3D point cloud understanding through deep transfer learning: A comprehensive survey," *Inf. Fusion*, p. 102601, 2024.

[30] M. J. Zideh and S. K. Solanki, "Multivariate Physics-Informed Convolutional Autoencoder for Anomaly Detection in Power Distribution Systems with High Penetration of DERs," *ArXiv Prepr. ArXiv240602927*, 2024.

[31] A. Kyuroson, A. Koval, and G. Nikolakopoulos, "Autonomous point cloud segmentation for power lines inspection in smart grid," *IFAC-Pap.*, vol. 56, no. 2, pp. 11754–11761, 2023.

[32] H. Lu and H. Shi, "Deep learning for 3d point cloud understanding: a survey," *ArXiv Prepr. ArXiv200908920*, 2020.

[33] S. Zhu *et al.*, "A deep-learning-based method for extracting an arbitrary number of individual power lines from uav-mounted laser scanning point clouds," *Remote Sens.*, vol. 16, no. 2, p. 393, 2024.

[34] M. A. A. Faisal, I. Mecheter, Y. Qiblawey, J. H. Fernandez, M. E. H. Chowdhury, and S. Kiranyaz, "Deep Learning in Automated Power Line Inspection: A Review," Feb. 10, 2025, *arXiv*: arXiv:2502.07826. doi: 10.48550/arXiv.2502.07826.

[35] C. Chen *et al.*, "DCPLD-Net: A diffusion coupled convolution neural network for real-time power transmission lines detection from UAV-Borne LiDAR data," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 112, p. 102960, Aug. 2022, doi: 10.1016/j.jag.2022.102960.

[36] D. K. R. Basani, B. R. Gudivaka, R. L. Gudivaka, and R. K. Gudivaka, "Enhanced fault diagnosis in IoT: Uniting data fusion with deep multi-scale fusion neural network," Internet of Things, Art. no. 101361, 2024, doi: 10.1016/j.iot.2024.101361.

[37] V. D. M. Devarajan and A. Sambas, "Data-driven techniques for real-time safety management in tunnel engineering using TBM data," Int. J. Res. Eng. Technol., vol. 7, no. 3, 2022.

[38] R. L. Gudivaka, "AI-driven optimization in robotic process automation: Implementing neural networks for real-time imperfection prediction," Int. J. Bus. Gen. Manag. (IJBGM), vol. 12, no. 1, 2023.

[39] H. Nagarajan and H. M. Khalid, "Optimizing signal clarity in IoT structural health monitoring systems using Butterworth filters," Int. J. Res. Eng. Technol., vol. 7, no. 5, 2022.

[40] N. K. R. Panga, "Optimized hybrid machine learning framework for enhanced financial fraud detection using e-commerce big data," Int. J. Manag. Res. Rev., vol. 11, no. 2, 2021, ISSN: 2249-7196.

[41] A. Shahnejat Bushehri, A. Amirnia, A. Belkhiri, S. Keivanpour, F. G. de Magalhães, and G. Nicolescu, "Deep Learning-Driven Anomaly Detection for Green IoT Edge Networks," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 498–513, Mar. 2024, doi: 10.1109/TGCN.2023.3335342.

[42] K. Shao, Y. He, Z. Xing, and B. Du, "Detecting wind turbine anomalies using nonlinear dynamic parameters-assisted machine learning with normal samples," *Reliab. Eng. Syst. Saf.*, vol. 233, p. 109092, May 2023, doi: 10.1016/j.ress.2023.109092.

[43] M. Ibrahim, A. Alsheikh, F. M. Awaysheh, and M. D. Alshehri, "Machine Learning Schemes for Anomaly Detection in Solar Power Plants," *Energies*, vol. 15, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/en15031082.

[44] A. Basit, H. U. Manzoor, M. Akram, H. E. Gelani, and S. Hussain, "Machine learning-assisted anomaly detection for power line components: A case study in Pakistan," *J. Eng.*, vol. 2024, no. 7, p. e12405, 2024, doi: 10.1049/tje2.12405.

[45] L. Chen, Yao ,Xiling, Xu ,Peng, Moon ,Seung Ki, and G. and Bi, "Rapid surface defect identification for additive manufacturing with in-situ point cloud processing and machine learning," *Virtual Phys. Prototyp.*, vol. 16, no. 1, pp. 50–67, Jan. 2021, doi: 10.1080/17452759.2020.1832695.

[46] "GitHub - dlavado/TS40K: Repository for the TS40K dataset." Accessed: May 17, 2025. [Online]. Available: https://github.com/dlavado/TS40K

[47] D. Lavado, C. Soares, A. Micheletti, R. Santos, A. Coelho, and J. Santos, "TS40K: a 3D Point Cloud Dataset of Rural Terrain and Electrical Transmission System," Feb. 18, 2025, *arXiv*: arXiv:2405.13989. doi: 10.48550/arXiv.2405.13989.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

HGWWO: A Hybrid Grey Wolf–Whale Optimizer for Load Balancing in Cloud Computing Environments

Yameng BAI[1]*, Junxia MENG[2], Shuai ZHAO[3], Ruoyu REN[4]
School of Information Engineering, Jiaozuo University, Jiaozuo 454150, China[1, 2]
School of Artificial Intelligence, Jiaozuo University, Jiaozuo 454150, China[3]
School of Safety Science and Engineering, Henan Polytechnic University, Jiaozuo 454150, China[4]

***Abstract*—This paper aims to develop an efficient and adaptive load balancing algorithm for cloud computing environments using a novel hybrid meta-heuristic approach. Effective load balancing is necessary for optimum performance and resource utilization in cloud computing systems. Most conventional meta-heuristic algorithms suffer from premature convergence and poor exploration–exploitation tradeoffs. An innovative hybrid meta-heuristic algorithm, Hybrid Grey Wolf–Whale Optimizer (HGWWO), is proposed for efficiently and dynamically balancing cloud load. HGWWO integrates the leadership hierarchy and adaptive hunting strategy of the Grey Wolf Optimizer (GWO) with the spiral-shaped exploitation mechanism of the Whale Optimization Algorithm (WOA), resulting in high convergence rates. The algorithm is implemented in a multi-objective cloud load balancing model to reduce response time, energy usage, and makespan while optimizing resource utilization among virtual machines. The experimental outcomes prove that HGWWO outperforms existing algorithms regarding throughput, waiting time, and execution efficiency. The suggested model has potential for real-time cloud scheduling of resources and is an efficient solution for scalable and heterogeneous cloud environments.**

***Keywords*—Cloud computing; load balancing; hybrid meta-heuristic; grey wolf optimizer; whale optimization algorithm**

I. INTRODUCTION

Cloud computing has transformed the delivery of computational resources and services, providing on-demand access to shared, configurable resources like Virtual Machines (VMs), storage, and applications over the Internet [1]. Cloud paradigms facilitate scalability, flexibility, and cost-effectiveness, essential in contemporary business and service organizations [2]. Nevertheless, the dynamic nature and dispersion of the cloud result in issues with providing resources, primarily load balancing [3]. Ineffective load balancing leads to bottlenecks in the system, resulting in poor resource utilization, high response times, and overall poor system performance [4]. Optimizing workload distribution among existing VMs while avoiding system imbalance is an essential design and administration issue in cloud computing infrastructures [5].

These issues, however, have been addressed through the design of numerous load balancing schemes, ranging from simple static approaches to more advanced and adaptive methodologies. Static algorithms in highly dynamic and heterogeneous clouds are ineffective since they cannot adapt to runtime variations in workloads and system resources [6]. In such cases, load balancing algorithms, in the form of multi-objective optimization, become pressing issues. The algorithms attempt to balance multiple objectives concurrently, including makespan, power usage, response time, and resource utilization [7]. Such problem-solving demands state-of-the-art optimization techniques for dealing with highly complicated search spaces and allowing system state changes. Related work across various fields, including machine learning for business forecasting [8], innovative grid system optimization [9], and security threat detection [10], similarly emphasizes the range and promise of state-of-the-art optimization and intelligent decision-making methodologies. Such models provide a compelling argument for highly adaptive solutions that are robust enough for intensive environments, such as cloud computing.

Past research has demonstrated the effectiveness of nature-inspired meta-heuristic algorithms for optimizing complex multi-objective problems in the cloud [11]. Grey Wolf Optimizer (GWO) and the Whale Optimization Algorithm (WOA) are distinguished by their excellent exploration and exploitation capacities. GWO mimics wolves' social hierarchy and hunting strategy [12], while WOA mimics humpback whale bubble-net hunting [13]. While the two algorithms promise particular strengths, they both carry the flaw of premature convergence, or a lack of thoroughness in local search, when employed in isolation. Inspired by the complementarities between the two tactics, this study proposes a hybrid algorithm, HGWWO that combines the social leadership process of GWO with the exploitation effectiveness of WOA to improve load balancing efficiency in the cloud.

Although more meta-heuristic algorithms have recently been applied to cloud load balancing, previous work often fails to adequately resolve the exploration-exploitation tradeoff under highly variable workloads. When used standalone, algorithms such as GWO or WOA tend to converge prematurely or become stuck in a local optimum. Although hybrid methodologies also exist, they tend to incur heavy computational overhead or be specialized for low scales, thereby limiting their generality. This work closes this gap by proposing HGWWO as a hybrid meta-heuristic that combines GWO's adaptive leader structure with spiral-guided localized searching from WOA, with a focus on being particularly appropriate for variable heterogeneous clouds. By integrating the merits of both algorithms, HGWWO will deliver a computationally efficient and generic solution for resolving the multi-objective cloud load balancing problem. Therefore, the primary research question guiding this study is: How can a hybrid meta-heuristic algorithm be developed to dynamically

handle workload in heterogeneous clouds with minimal makespan, energy usage, and load deviation?

The rest of the paper is as follows: Section II reviews the literature on load balancing and meta-heuristic optimization. Section III describes the mathematical model of the load balancing problem and its objective functions. Section IV details the HGWWO algorithm. Section V provides simulation results, comparison analysis, and performance assessment. Section VI provides a critical discussion of the findings, including improvements and limitations. Finally, Section VII concludes the paper and offers directions for further work.

II. Literature Review

Cloud computing load balancing has garnered significant research interest due to its importance in optimizing resource usage, minimizing response time, and ensuring Quality of Service (QoS). Several heuristic and meta-heuristic methods have been proposed for addressing the dynamic and multi-objective nature of the problem. Muteeh, et al. [14] presented an Ant Colony Optimization (ACO)-based multi-resource load balancing algorithm to solve the issues in the scheduling of scientific workflows in cloud computing. Their approach optimized the makespan and cost, ensuring balanced load distribution among cloud resources. Experimental verification showed enhanced resource utilization and a decrease in both the price and the running time.

Sefati, et al. [15] proposed load balancing using GWO, focusing on the algorithm for identifying overloaded and underutilized nodes. The algorithm includes using the fitness value and reliability of the nodes to direct resource assignment. CloudSim experiments demonstrated significant reductions in response time and execution cost compared to other typical methods. Ramya and Ayothi [16] proposed a hybrid Dingo and Whale Optimization Algorithm (HDWOA-LBM) by integrating dingo and whale hunting behaviors to achieve effective task allocation. The algorithm enhances both exploration and exploitation capabilities. Simulation results validated its efficiency, showing improvements in makespan (22.9%), throughput (21.2%), and reliability (25.4%).

Geetha, et al. [17] created an innovative Intercrossed Chimp and Bald Eagle Algorithm (IC&BA) incorporating chimp optimization and bald eagle searching. Their approach considers concurrently energy usage, execution cost, makespan, and turnaround time. Study results validated improved balancing and performance in heterogeneous environments.

Emara, et al. [18] presented an advanced Harris Hawks Optimization (HHO) for task scheduling in cloud computing. Their approach enhanced the exploration and exploitation stages through mutation-based diversification, thereby outdoing the conventional HHO and other metaheuristics in makespan, throughput, and load variance.

Haris and Zubair [19] developed a hybrid Battle Royale and Deep Reinforcement Learning (BRDRL) dynamic load balancing algorithm. The approach combines job migration through DRL and VM selection through BRO. Simulations demonstrated that BRDRL resulted in a 3.9% improvement in throughput and a 15.3% improvement in response time compared to existing models and hybrids.

Tabagchi Milan, et al. [20] suggested an Artificial Bee Colony (ABC)-derived method for green cloud computing. Overloaded VM tasks are migrated according to the bee behavior to enhance QoS and energy savings. The experimental results indicate improvement in energy and makespan.

While all the above approaches support load balancing in cloud computing, most algorithms are plagued with tradeoffs between convergence rates and exploitation accuracy, as highlighted in Table I. For example, ACO and GWO-based algorithms converge prematurely or become stuck in local optima in dynamic environments. Hybrid algorithms like HDWOA and BRDRL improve performance at the expense of being computationally intensive and having limited generalizability for use in other cloud infrastructures.

TABLE I. An Overview of Recent Cloud Load Balancing Algorithms

Ref	Algorithm	Features	Limitations
[14]	Ant colony optimization	It supports multi-resource allocation, workflow-aware scheduling, and effectively reduces cost and execution time.	May experience slow convergence and exhibit limited adaptability to dynamic load variations.
[15]	Grey wolf optimizer	It accounts for node reliability and adapts to load conditions, including underloaded and overloaded nodes.	Prone to entrapment in local optima and characterized by limited exploitation capability.
[16]	Hybrid dingo and whale optimization algorithm	The approach enhances exploration and exploitation mechanisms, improving makespan and increasing system reliability.	Exhibits high computational complexity, with potential performance degradation under task overload or in the presence of noisy objective functions.
[17]	Intercrossed chimp and bald eagle algorithm	The method employs a multi-criteria load balancing approach that accounts for cost, energy consumption, and response time through a hybrid strategy.	The approach may require extensive parameter tuning and demonstrates limited scalability for general-purpose use.
[18]	Harris hawks optimization	It employs adaptive mutation to strengthen global search efficiency, improving task scheduling accuracy.	Remains sensitive to the initial population and incurs additional computational overhead.
[19]	Hybrid battle royale and deep reinforcement learning	It combines DRL for job transfer with the BRO algorithm for VM selection, facilitating dynamic and adaptive decision-making.	It necessitates large volumes of training data, and integrating deep models contributes to substantial computational overhead.
[20]	Artificial bee colony	It utilizes bio-inspired task migration techniques to reduce energy consumption while maintaining or improving QoS.	Demonstrates reduced effectiveness under high-load conditions and prioritizes energy efficiency at the expense of load balancing.

Moreover, in a unified architecture, most algorithms individually optimize performance measures, power consumption, or both, excluding the multiple objective parameters. In response, we introduce HGWWO, the union of the exploration power in GWO and the exploitation power in WOA. The new approach guarantees better load balancing performance by simultaneously minimizing makespan, energy consumption, and load deviation.

III. Problem Formulation

VMs are the primary entities for executing user tasks and managing workload distribution in cloud computing environments. Each VM has distinct resource capacities such as memory, processing power, disk space, and bandwidth. Let us define the set of VMs as $V = \{v_1, v_2, \ldots, v_m\}$, where m denotes the total number of virtual instances available in the system. Each VM can be modeled using a multi-dimensional vector that captures the required levels of different resources.

The infrastructure comprises several identical physical servers, with one designated as the master controller that delegates incoming task requests to appropriate VMs. VM creation is conditional on memory availability. A host lacking adequate memory cannot spawn a new VM instance. In this scenario, we assume there are n independent tasks to be executed across m available VMs.

Efficient task scheduling is critical to ensure equitable load distribution and to enhance performance metrics. Each user request is characterized by a parameter vector $U = (\rho, \sigma, \gamma, \mu, \chi)$, where ρ stands for the average frequency of online user requests, σ is the size of each request, γ denotes the required CPU resources, μ signifies the needed memory for task execution, and χ refers to the request frequency per minute. The memory load for VM *i* is given by Eq. (1).

$$\mathcal{M}_i = \mathrm{Rem}_i + \frac{\omega_i}{\Omega_i} \times 100\% \tag{1}$$

Where Rem_i is the residual memory before assignment, ω_i is the task memory usage, and Ω_i is the total available memory.

The CPU load for VM *i* is calculated using Eq. (2).

$$\mathcal{C}_i = \mathrm{Cpu}_i + \frac{\tau_i}{\Theta_i} \times 100\% \tag{2}$$

Where Cpu_i is the CPU capacity left before execution, τ_i is the CPU demand of tasks, and Θ_i is the total CPU available.

The overall workload on VM *i* is the weighted combination of memory and CPU load, as expressed in Eq. (3).

$$\Phi_i = \alpha \cdot \mathcal{M}_i + \beta \cdot \mathcal{C}_i \tag{3}$$

Where $\alpha + \beta = 1$ are the memory and CPU weight factors.

The total load on host *j* is derived by summing all VM loads on the host, as shown in Eq. (4).

$$\Psi_j = \sum_{i=1}^{m_j} \Phi_i = \sum_{i=1}^{m_j} \left(\alpha \cdot \mathcal{M}_{ij} + \beta \cdot \mathcal{C}_{ij}\right) \tag{4}$$

The average load across all physical machines in the network is calculated using Eq. (5).

$$\bar{\Psi} = \frac{1}{p}\sum_{j=1}^{p} \Psi_j = \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{m_j} \left(\alpha \cdot \mathcal{M}_{ij} + \beta \cdot \mathcal{C}_{ij}\right) \tag{5}$$

The first fitness objective, aiming to minimize the load deviation across hosts, is formulated in Eq. (6).

$$F_1 = \sum_{j=1}^{p} \left|\Psi_j - \bar{\Psi}\right| \tag{6}$$

The task assignment is represented by the binary variable defined in Eq. (7).

$$\delta_{ij} = \begin{cases} 1, & if\ task\ i\ is\ assigned\ to\ VM_j \\ 0, & otherwise \end{cases} \tag{7}$$

The total execution time on VM *j* is computed as per Eq. (8).

$$\mathcal{T}_j = \sum_{i=1}^{n} \delta_{ij} \cdot \mathcal{T}_{ij} \tag{8}$$

The completion time for task *i* on VM *j* is evaluated using Eq. (9).

$$\mathcal{T}_{ij} = \frac{L_i}{\mathrm{Proc}_j} \tag{9}$$

Where L_i is the task length in MIPS, and Proc_j is the processing speed of the VM.

The makespan represents the maximum execution time among all VMs and is defined in Eq. (10).

$$\mathcal{M}_s = \max\left(\mathcal{T}_j\right), \quad 1 \le j \le m \tag{10}$$

The second objective function, total energy consumption, accounts for both active and idle states of VMs and is formulated in Eq. (11).

$$F_2 = \sum_{j=1}^{m} \left[\mathcal{T}_j \cdot \alpha_j + \left(\mathcal{M}_s - \mathcal{T}_j\right) \cdot \beta_j\right] \cdot \mathrm{Proc_j} \tag{11}$$

Where α_j and β_j are the energy coefficients (joules per instruction) in active and idle states, respectively.

The third fitness objective considers task instruction overhead and delay penalty. It is denoted as *F* 3 F 3 and defined in Eq. (12).

$$F_3 = w_1 \cdot \frac{\mathrm{Instr}_i}{\mathrm{MaxInstr}} + w_2 \cdot D_i \tag{12}$$

Where Instr_i is the instruction count of task i, MaxInstr is the maximum throughput of any processor, and D_i is the delay penalty. The weights are fixed as $w_1 = 0.7$ and $w_2 = 0.3$.

The final goal is to minimize the overall fitness function *F*, a weighted sum of the three sub-objectives. This is mathematically given by Eq. (13).

$$F = \lambda_1 \cdot F_1 + \lambda_2 \cdot F_2 + \lambda_3 \cdot F_3 \quad (13)$$

where $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$ are the respective weights for load deviation, energy consumption, and task allocation cost.

IV. Proposed Method

A new hybrid meta-heuristic algorithm, HGWWO, is presented here to solve the complex multi-objective load balancing problem in cloud computing, reducing makespan, energy usage, and load deviation. HGWWO combines the best features of two popular optimization algorithms: GWO and WOA. The HGWWO is explicitly formulated to optimize task-to-VM allocation in dynamic cloud environments to improve convergence accuracy and prevent premature stagnation.

A. Grey Wolf Phase: Resource-aware Encircling Strategy

GWO mimics the leadership-driven hunting and encircling behavior observed in natural grey wolf packs. In the context of this study, where the primary goal is to allocate independent cloud computing tasks to VMs efficiently, each "grey wolf" represents a potential solution, a vector encoding task-to-VM assignments. The metaphorical "prey" is the optimal solution that minimizes makespan, balances resource utilization, and reduces energy consumption.

Wolves are hierarchically classified into four roles, Alpha (α) that represents the best-known task scheduling solution at the current iteration; beta (β) and delta (δ) that represent the second and third-best solutions, contributing diversity and assisting in refining the search direction; and omega (ω) as follower solutions that are guided by the leading trio to explore the broader search space.

The optimization begins with the encircling behavior, where wolves adjust their positions relative to the prey, i.e., the current best-known scheduling pattern. Each VM allocation solution is iteratively updated using the encircling formulation, which introduces stochastic coefficients to balance intensification and diversification. The position update rule is given by Eq. (14).

$$\overrightarrow{X_{\text{new}}}(t+1) = \overrightarrow{X_{\text{prey}}}(t) - \vec{A} \cdot \vec{D} \quad \text{where} \quad \vec{D} = \left|\vec{C} \cdot \overrightarrow{X_{\text{prey}}}(t) - \vec{X}(t)\right| \quad (14)$$

Where $\overrightarrow{X_{\text{prey}}}(t)$ is the position vector of the current best task scheduling solution, $\vec{X}(t)$ denotes the position of the candidate VM allocation, $\vec{D}$ refers to the encircling vector (distance between candidate and prey), $\vec{A}$ is an adaptive coefficient vector controlling convergence intensity, $\vec{C}$ is a coefficient vector introducing randomness. The adaptive coefficient $\vec{A}$ is derived as follows:

$$\vec{A} = 2 \cdot \vec{a} \cdot \overrightarrow{r_1} - \vec{a} \quad (15)$$

Where $\vec{a}$ linearly decreases from 2 to 0 across iterations, gradually shifting the search from exploration to exploitation, and $\overrightarrow{r_1}$ is a random vector in [0, 1]. The vector $\vec{C}$ is similarly calculated using Eq. (16).

$$\vec{C} = 2 \cdot \overrightarrow{r_2} \quad (16)$$

With $\overrightarrow{r_2}$ being another random vector in [0, 1]. These stochastic variables help diversify the search paths and prevent premature convergence.

Once the encircling step is complete, the algorithm enters the position updating phase, where each solution is recalibrated by referencing the top three leaders (α, β, and δ). This phase ensures the search agents (candidate solutions) are drawn toward the most balanced and energy-efficient task assignments identified. The distances from each leader are computed, and new intermediate positions are derived as follows:

$$\begin{aligned} \overrightarrow{X_\alpha} &= \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \left|\overrightarrow{C_1} \cdot \overrightarrow{X_\alpha} - \vec{X}\right| \\ \overrightarrow{X_\beta} &= \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \left|\overrightarrow{C_2} \cdot \overrightarrow{X_\beta} - \vec{X}\right| \\ \overrightarrow{X_\delta} &= \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \left|\overrightarrow{C_3} \cdot \overrightarrow{X_\delta} - \vec{X}\right| \end{aligned} \quad (17)$$

The final updated position for the VM scheduling solution is obtained by averaging the guidance from the three leaders, as defined in Eq. (18).

$$\vec{X}(t+1) = \frac{1}{3}\left(\overrightarrow{X_\alpha} + \overrightarrow{X_\beta} + \overrightarrow{X_\delta}\right) \quad (18)$$

This rule ensures that each task-to-VM mapping benefits from the consensus of the top-performing strategies. In the load balancing context, this corresponds to scheduling tasks such that the CPU and memory demands are evenly distributed, minimizing execution bottlenecks and energy overhead.

B. Whale Phase: Exploitation Via Spiral Search

While the grey wolf phase provides deep global exploration and informed leader-based direction for task delegation, it falls short when dealing with local exploitation, especially in convergence towards remarkably optimized task-VM mapping. The proposed HGWWO incorporates an enhanced local search mechanism adopted from the WOA to increase search depth and overcome stagnation.

WOA simulates humpback whale bubble-net feeding behavior. In the analogy, there is one candidate task schedule for each whale, and the target is the ideal VM assignment with the shortest makespan and evenly distributed CPU/memory usage. WOA alternates between encircling, spiral motion exploitation, and random search, exploring and exploiting the solution space more efficiently.

The first behavioral pattern of WOA involves encircling the best solution discovered so far. In the cloud computing, this action refers to refining task allocations around the best-known task-VM mapping. This behavior mirrors the GWO encircling mechanism but contributes to localized convergence. This step is mathematically represented by Eq. (19).

$$\overrightarrow{X_{\text{new}}}(t+1) = \overrightarrow{X_{\text{best}}}(t) - \vec{A} \cdot \vec{D} \quad \text{where} \quad \vec{D} = \left|\vec{C} \cdot \overrightarrow{X_{\text{best}}}(t) - \vec{X}(t)\right| \quad (19)$$

Where $\overrightarrow{X_{\text{best}}}(t)$ refers to the current best task assignment solution and $\vec{X}(t)$ stands for the position of the current whale (solution).

To fine-tune the best solutions near optima, whales perform a spiral movement, which enables candidate solutions to follow a logarithmic path toward the best allocation found. This is crucial for minimizing minor inefficiencies in task scheduling, such as slight VM overloading or imbalance in memory usage. This phase probabilistically alternates with the encircling step using a control parameter $p \in [0,1]$. The spiral search is applied if $p \geq 0.5$, and the position update is calculated by Eq. (20).

$$\overrightarrow{X_{\text{new}}}(t+1) = \left|\overrightarrow{X_{\text{best}}}(t) - \vec{X}(t)\right| \cdot e^{bl} \cdot \cos(2\pi l) + \overrightarrow{X_{\text{best}}}(t) \quad (20)$$

Where b is a constant defining the spiral shape, l is a random number in $[-1,1]$ determining movement orientation.

WOA includes a prey search mechanism randomly generating new task scheduling solutions to maintain diversity and avoid local optima traps. This phase becomes dominant when the algorithm needs to escape plateaus in the fitness landscape. The exploration behavior is modeled by Eq. (21).

$$\overrightarrow{X_{\text{new}}}(t+1) = \overrightarrow{X_{\text{rand}}} - \vec{A} \cdot \vec{D} \quad \text{where} \quad \vec{D} = \left|\vec{C} \cdot \overrightarrow{X_{\text{rand}}} - \vec{X}(t)\right| \quad (21)$$

Where $\overrightarrow{X_{\text{rand}}}$ is a randomly selected solution from the population. This component is especially important for diversifying the task assignments and discovering new combinations that reduce energy consumption or execution delay.

In the hybrid HGWWO algorithm, the whale phase is used during the optimization cycle's exploitation phase. Following the population being navigated into fruitful areas in the search area by the GWO, random and spiral updates in WOA refine task allocations, offsetting near-optimal load oversights or inefficiencies that the GWO might overlook. Dynamic load redistribution among underloaded VMs, slight reduction in makespan, and progressive energy-based assignment refinement result from adaptive switching between random exploration and spiral refinement.

C. Combined Update Strategy for Load Balancing

HGWWO integrates the exploratory efficiency of the GWO with the local refinement capability of the WOA to solve the multi-objective load balancing problem in cloud computing. Fig. 1 shows the flowchart of the proposed algorithm. In typical cloud environments, scheduling tasks across multiple heterogeneous VMs must account for CPU utilization, memory demand, execution time, and energy cost. Algorithms that only explore broadly (like GWO) may converge slowly or stagnate near suboptimal allocations. Conversely, algorithms like WOA, which are good at refining solutions, may not escape poor initial guesses. HGWWO resolves these challenges by dynamically adapting its search strategy at each iteration, blending the strengths of both methods.

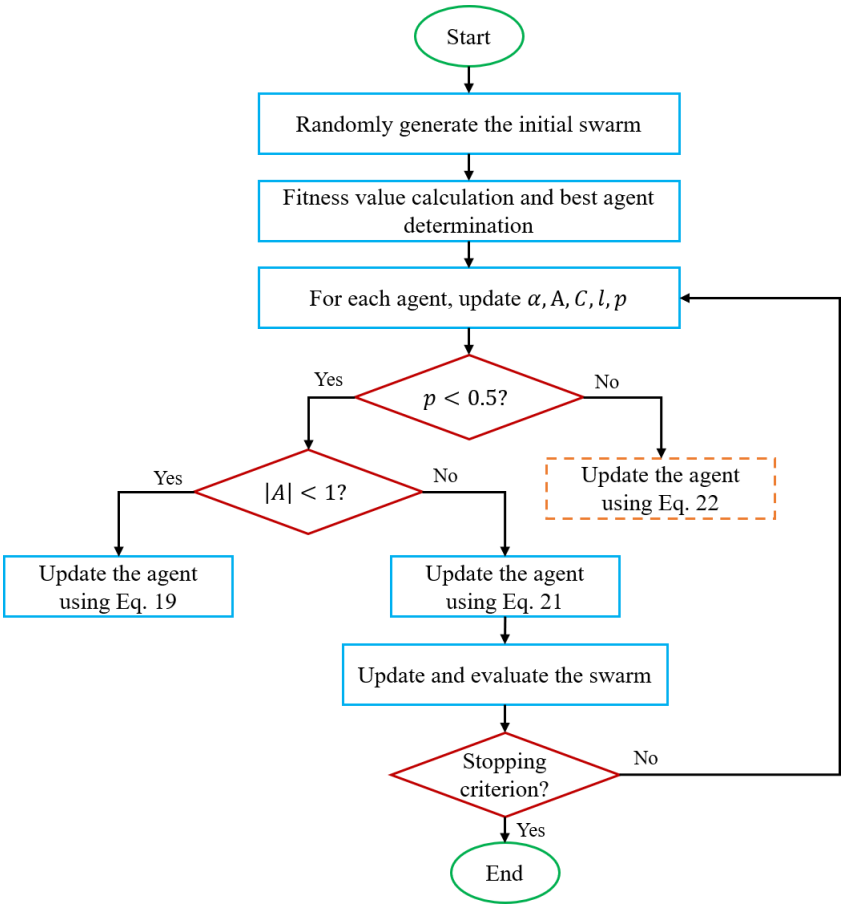


Fig. 1. Flowchart of proposed algorithm.

At every iteration of the optimization process, each candidate solution (task allocation vector) chooses its update mechanism based on a random probability $p \in [0,1]$. If $p < 0.5$, the solution is updated using a GWO-style encircling behavior, otherwise the solution undergoes WOA-style spiral exploitation. This probabilistic strategy allows HGWWO to dynamically balance global and local search throughout the optimization timeline. It avoids premature convergence (via random exploration) and accelerates convergence near optimal points (via spiral tightening). The hybrid update rule is mathematically defined in Eq. (22).

$$\overrightarrow{X_{HGWWO}}(t+1) = \begin{cases} min\left(\overrightarrow{X_{prey}} - \vec{A} \cdot \vec{D}, \overrightarrow{X_{elite}}(t)\right), \\ \qquad if\ p < 0.5 \\ min\left(\left|\overrightarrow{X_{prey}} - \vec{X}(t)\right| \cdot e^{bl} \cdot cos(2\pi l)\right) + \\ \overrightarrow{X_{prey}}, \overrightarrow{X_{elite}}(t), \qquad if\ p \geq 0.5 \end{cases} \quad (22)$$

Where $\overrightarrow{X_{elite}}(t)$ denotes the best composite solution generated by the GWO leadership strategy.

The elite solution is calculated as a weighted average of the top three solutions identified by GWO: alpha (best), beta (second best), and delta (third best). These roles correspond to the most balanced, energy-efficient task assignments in the current population, given by Eq. (23).

$$\overrightarrow{X_{\text{elite}}}(t) = \frac{1}{3}\left(\overrightarrow{X_\alpha} + \overrightarrow{X_\beta} + \overrightarrow{X_\delta}\right) \quad (23)$$

Each of these leader positions is computed by subtracting a scaled distance from their current state, as described below:

$$\begin{aligned} \overrightarrow{X_\alpha} &= \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \left|\overrightarrow{C_1} \cdot \overrightarrow{X_\alpha} - \vec{X}(t)\right| \\ \overrightarrow{X_\beta} &= \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \left|\overrightarrow{C_2} \cdot \overrightarrow{X_\beta} - \vec{X}(t)\right| \\ \overrightarrow{X_\delta} &= \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \left|\overrightarrow{C_3} \cdot \overrightarrow{X_\delta} - \vec{X}(t)\right| \end{aligned} \quad (24)$$

Each updated solution is then evaluated against the multi-objective fitness function. This function combines the load deviation across hosts, the energy consumption of active and idle VMs, and the delay and task overhead due to scheduling inefficiencies.

The proposed HGWWO initiates with a diverse population of candidate solutions, each representing a potential task-to-VM mapping configuration in a cloud computing environment. These initial solutions are randomly generated to promote exploration across the search space. The optimization process proceeds through successive iterations, during which each candidate is iteratively refined based on either a randomly chosen peer or the current global best.

A linearly decreasing control parameter, denoted by α, governs the balance between exploration and exploitation. Specifically, α starts at a maximum value of 2 and reduces to 0 throughout iterations. This parameter is used to compute the adaptive coefficient vector $\vec{A}$, which determines the intensity of position updates. When the magnitude $\vec{A} > 1$, the algorithm emphasizes exploration by referencing a randomly selected solution. Conversely, when $\vec{A} < 1$, the emphasis shifts toward exploitation, refining solutions around the current optimum.

Simultaneously, the algorithm incorporates a probabilistic decision model based on a random number p to alternate between encircling and spiral movements for solution updates. The update mechanism is bifurcated into two cases, GWO-style refinement and WOA-inspired spiral exploitation.

When $p < 0.5$, the solution update follows a leader-guided search strategy, combining the difference between the best-known task allocation and the current candidate, and the elite average position. When $p \geq 0.5$, the algorithm adopts a spiral search path influenced by the Whale Optimization concept. This update models the tightening bubble-net motion that localises optimal solutions through logarithmic spirals. The position is recalculated by incorporating the proximity to the best-known solution and a decaying spiral motion.

V. RESULTS

To assess the efficacy of the proposed HGWWO in optimizing task scheduling and load balancing in cloud environments, comprehensive simulations were conducted using CloudSim 3.0.3 on a Windows 7 machine equipped with an Intel Core i7 processor, 8 GB RAM, and a 3.4 GHz CPU. The experimental setup is described in Table II.

TABLE II. SIMULATION CONFIGURATION AND SYSTEM SPECIFICATIONS

Component	Quantity	Attributes	Specification
Hosts	20	CPU cores	6
		Bandwidth	15 GB/s
		RAM	16.0 GB
		Hypervisor	Xen
		Storage volume	4.0 TB
		Computational power	177,730 MIPS
Virtual machines	10-100	Hypervisor	Xen
		Processing elements	1
		Disk image size	10 GB
		Bandwidth	1 GB/s
		RAM	0.5 GB
		Processing capability	9726 MIPS
Data center	1	Storage allocation cost (per unit)	0.001
		Memory allocation cost (per unit)	0.05
		Base usage cost	3.0 units
		Virtualization platform	Xen
		Operating system	Linux
		Architecture type	x86

The performance of HGWWO was compared against five prominent algorithms, ACO, GWO, WOA, ABC, and HHO, on a large scale of tasks and VM configurations. The comparison was made based on multi-objective measures such as makespan, energy usage, resource usage, migration rate, imbalance factor, throughput, and idle time.

Two principal scenarios were experimented with, varying the tasks (100 to 2000) while the number of VMs (100) remains fixed, and fixing the tasks (1000) while varying the number of VMs (10 to 100). The fitness function, defined in Eq. (13), integrates load deviation, energy consumption, and task cost. Optimal weight values $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$ were obtained through trial-and-error and are detailed in Table III.

TABLE III. IMPACT OF FITNESS WEIGHTS ON LOAD BALANCING METRICS

Weights			Simulation outcomes			
λ_1	λ_2	λ_3	Energy usage (kJ)	Load variance	Throughput (req/ms)	Makespan (ms)
1	0.9	0.9	305.08	0.355	9.05	6570.31
		0.8	294.62	0.349	8.51	6352.19
		0.7	290.29	0.338	8.29	9105.88
		0.6	283.84	0.374	8.05	8995.02
		0.5	279.43	0.361	7.88	8906.53
	0.8	0.9	245.66	0.351	8.59	9389.11
		0.8	240.38	0.342	8.40	9313.76
		0.7	224.17	0.318	8.17	9109.31
		0.6	221.02	0.285	7.94	8976.19
		0.5	211.27	0.253	7.77	8911.42
	0.7	0.9	291.18	0.331	8.42	9102.35
		0.8	282.65	0.296	8.19	9085.72
		0.7	283.45	0.304	7.76	9106.79
		0.6	276.38	0.294	7.60	8987.46
		0.5	266.18	0.232	7.55	8519.17
	0.6	0.9	261.09	0.275	8.18	9015.41
		0.8	256.21	0.256	8.01	8722.15
		0.7	251.60	0.247	7.79	8613.91
		0.6	246.52	0.219	7.25	8482.31
		0.5	242.96	0.203	7.01	8314.20
	0.5	0.9	235.15	0.255	7.29	8818.73
		0.8	229.23	0.161	7.05	8416.29
		0.7	217.26	0.094	6.64	8093.19
		0.6	209.41	0.079	5.73	7914.46
		0.5	**170.77**	**0.063**	**5.39**	**7751.62**
	0.4	0.9	251.63	0.351	7.74	8386.94
		0.8	242.10	0.285	7.26	8352.28
		0.7	221.48	0.244	6.93	8109.12
		0.6	204.19	0.206	6.75	8059.26
		0.5	183.69	0.186	6.41	7903.13

Fig. 2 and Fig. 3 show that HGWWO has the lowest energy consumption among all rival algorithms. This results from its innovative adaptive hybrid architecture, which inhibits unnecessary task migration and accurately assigns workloads to idle VMs. It maintains efficient CPU and memory usage while minimizing excessive idle energy consumption by integrating GWO's leadership-based global exploration mechanism with WOA's local spiral exploitation.

Fig. 4 and Fig. 5 indicate that HGWWO beats its competitors in make-span reduction, a key contributor to system responsiveness and QoS. The reduction is a result of spreading out global exploration and switching to local optimization inherent in HGWWO, which facilitates faster convergence on optimum task allocations. The effect is a more deterministic response time with less latency, resulting in increased user service availability. The load standard deviation between the VMs (as depicted in Fig. 6) quantifies the uniformity in the distribution of resources. HGWWO maintains the lowest deviation, notably after t=3700 ms, reflecting stable and efficient load balancing. HGWWO's leader-based update approach, in which alpha, beta, and delta wolves collectively

affect each candidate's position, and spiral movement adjusts the load in finer-grained steps, minimizes oscillations between overloaded and under-loaded nodes.

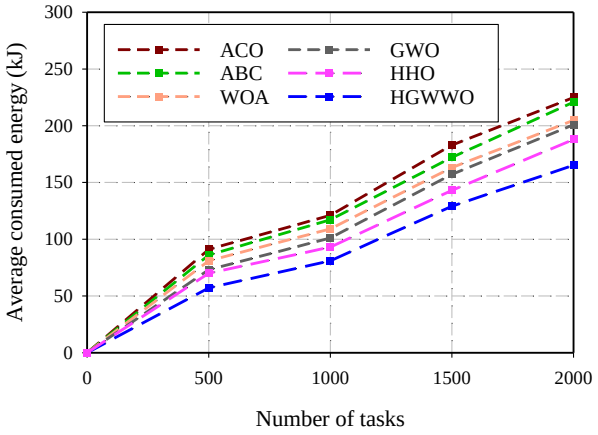


Fig. 2. Energy consumption comparison with increasing task volume at constant VM count.

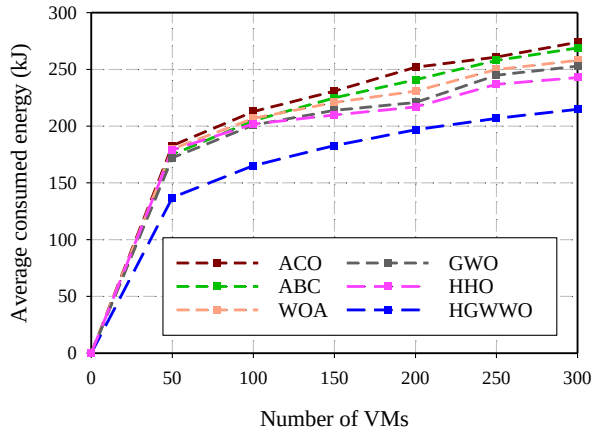


Fig. 3. Energy consumption comparison with increasing VM count at constant task volume.

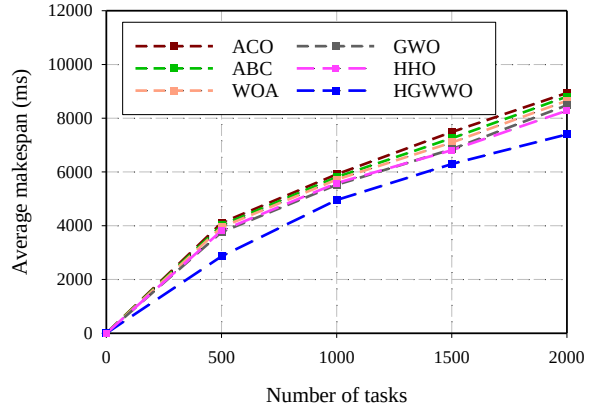


Fig. 4. Makespan comparison with increasing task volume at constant VM count.

As evident from Fig. 7, HGWWO migrates fewer tasks than its competitors. This is the result of its strategy for allocation based on fitness, which predicts upcoming changes in load ahead of time and assigns tasks to the appropriate VMs in advance, resulting in fewer runtime reassignments and reduced overhead. Fig. 7 shows that HGWWO achieves higher throughput for most conditions, particularly under heavy load conditions. The locations' proximity-based update policies, along with spiral exploitation policies in the algorithm, enable timely job completion with constant service continuity, thus enhancing the cloud system's ability to service user demands.

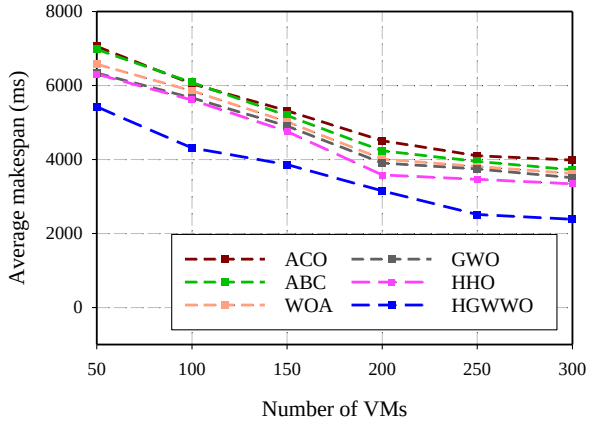


Fig. 5. Makespan comparison with increasing VM count at constant task volume.

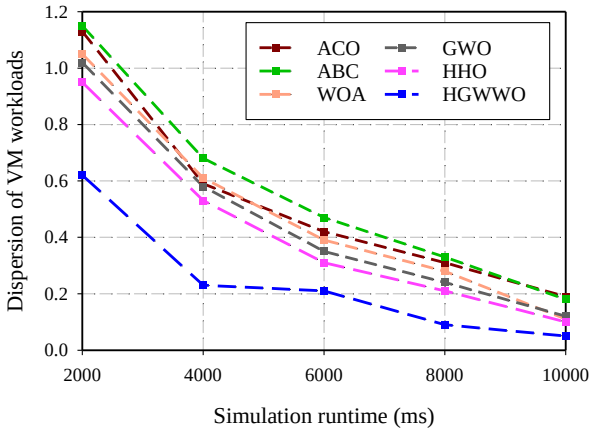


Fig. 6. Temporal variation of load imbalance levels.

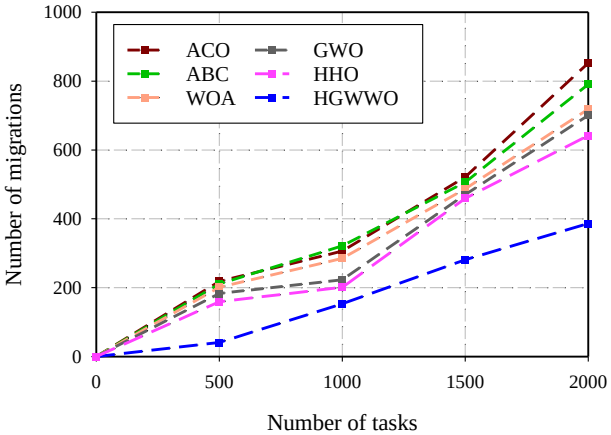


Fig. 7. Task migration volume comparison.

VI. DISCUSSION

Experimental findings reveal that our proposed HGWWO algorithm produces significant advancements in dynamic load balancing for cloud computing infrastructures. The hybridization of GWO and WOA exhibits a higher convergence trend, a reduced value for makespan, reduced energy consumption, and an equalized task allotment between virtual machines. These advantages result from several primary design attributes of the proposed HGWWO algorithm.

Firstly, incorporating GWO's hierarchical structure for leadership with WOA's exploitation strategy based on a spiral allows HGWWO to obtain a more stable tradeoff between local exploration and global exploitation, thereby reducing the risk of early convergence, which is prone to single standalone meta-heuristics such as GWO, WOA, and ABC. Secondly, adaptive behavior is achieved by incorporating auto-tuning algorithmic parameters into a variable workload distribution based on its state. This is needed in heterogeneous clouds due to sudden changes in task intensity alongside available resources.

Additionally, the multi-objective optimization model employed in this work contributes to these enhancements. As it optimizes makespan, energy consumption, and load deviation simultaneously, HGWWO handles the practical tradeoffs that cloud service providers face. This end-to-end optimization results in more efficient resource utilization and a higher level of QoS compared to alternative algorithms.

However, our proposed model also has several weaknesses that need to be addressed. One of these principles is that, due to its multi-objective optimization with a hybrid structure, the algorithm's computational complexity is higher than that of simpler, single-heuristic schemes. Although this tradeoff allows for superior performance, it may not be suitable for extremely large-scale implementations or applications that depend highly on latency and for which speed of execution is a concern. In addition, validation was conducted using synthetic workloads within the CloudSim simulation platform. Nonetheless, a widely adopted benchmarking platform, validation within realistic real-world infrastructures (e.g., OpenStack, Kubernetes) would be more convincing in practical terms.

It also has the limitation that it requires manual setting of algorithm parameters, including weight coefficients and control factors. Although they were tuned experimentally, an adaptive or self-tuning system would be suitable for further enhancing the robustness of HGWWO when environmental conditions vary. Lastly, security, fault tolerance, or energy-aware geographic distribution are not explicitly addressed by the current model, but become significant in specific scenarios of cloud deployment.

VII. CONCLUSION

This study introduced the HGWWO algorithm for the rigorous problem of dynamic load balancing and task scheduling in cloud computing. HGWWO was compared with five algorithms, ACO, GWO, WOA, HHO, and ABC, along with significant performance metrics such as makespan, energy usage, the standard deviation, throughput, task migration, the degree of imbalance, and idle time, based on comprehensive simulations under different workloads and VM settings using CloudSim 3.0.3, and further validated using a physical cloud testbed. Experimental results showed that HGWWO outperforms all baseline algorithms in balanced energy-aware and resource allocation. Notably, it achieved a smaller makespan, less energy consumption, and fewer task migrations while maintaining high throughput with even load distribution.

Compared with other meta-heuristics and hybrid solutions, our proposed HGWWO algorithm possesses several practical merits. It converges faster due to leveraging GWO's initial global exploration and WOA's terminal local exploitation. This adaptive switching also minimizes task migration and improves throughput, particularly for heterogeneous or load-intensive environments. It also distinguishes itself from deep-learning-based models, which require extensive training datasets and produce significant overheads. Our proposed HGWWO is thus lightweight and usable even in resource-constrained cloud nodes, including real-time scheduling. Its merits qualify it as a viable contender for cloud production use, particularly where task patterns change dynamically, and deterministic scheduling is infeasible.

Although the suggested HGWWO algorithm demonstrates promising behavior for fluctuating and varying cloud conditions, several weaknesses must be considered. First, even though HGWWO converges faster than standalone GWO or WOA, its hybridization also incurs more computational overhead, which may be noticeable for ultra-large scales. Secondly, the algorithm's behavior is now evaluated under synthetic workloads and simulated environments. Realistic deployment environments with non-deterministic workload spikes and hardware faults may require further fine-tuning or complementing with learning-based adaptive modules. This limitation presents opportunities for enhancing HGWWO with real-time or predictive feedback mechanisms in future work.

Based on the positive outcome by HGWWO, future work will focus on expanding the algorithm to distributed and federated cloud environments such as multi-cloud computing, edge computing, and fog computing. They outline more demanding scenarios, such as latency-aware task assignment, limited resource availability, and asynchronous message passing, which can be tailored for handling by HGWWO. Combining workload prediction with deep reinforcement learning will also enable proactive resource allotment and improve scalability in real-time. Further enhancements would include adaptive parameter self-tuning methodologies for reducing manual weight modifications in the fitness function. Others would include security-aware policy-making for scheduling purposes, along with carbon-aware optimization for eco-friendly cloud control. Long-term verification will consist of practical deployment and benchmarking on running cloud infrastructures (e.g., OpenStack) for demonstrating robustness under realistic constraint scenarios.

ACKNOWLEDGEMENT

This work was supported by the project of Science and Technology Program of Henan Province (252102320051).

REFERENCES

[1] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," Journal of Network and Computer Applications, vol. 204, p. 103405, 2022.

[2] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single - objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022, doi: https://doi.org/10.1002/cpe.6698.

[3] J. K. Samriya, S. Kumar, M. Kumar, H. Wu, and S. S. Gill, "Machine learning based network intrusion detection optimization for cloud computing environments," IEEE Transactions on Consumer Electronics, 2024.

[4] N. Devi et al., "A systematic literature review for load balancing and task scheduling techniques in cloud computing," Artificial Intelligence Review, vol. 57, no. 10, p. 276, 2024.

[5] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, vol. 23, no. 2, pp. 641-661, 2020.

[6] D. A. Shafiq, N. Z. Jhanjhi, A. Abdullah, and M. A. Alzain, "A load balancing algorithm for the data centres to optimize cloud computing applications," Ieee Access, vol. 9, pp. 41731-41744, 2021.

[7] A. R. Khan, "Dynamic load balancing in cloud computing: optimized RL-based clustering with multi-objective optimized task scheduling," Processes, vol. 12, no. 3, p. 519, 2024.

[8] M. B. Bagherabad, E. Rivandi, and M. J. Mehr, "Machine Learning for Analyzing Effects of Various Factors on Business Economic," Authorea Preprints, 2025, doi: https://doi.org/10.36227/techrxiv.174429010.09842200/v1.

[9] M. Ahmadi et al., "Optimal allocation of EVs parking lots and DG in micro grid using two - stage GA - PSO," The Journal of Engineering, vol. 2023, no. 2, p. e12237, 2023, doi: https://doi.org/10.1049/tje2.12237.

[10] S. M. Sanjari, A. M. Shibli, M. Mia, M. Gupta, and M. M. A. Pritom, "SmishViz: Towards A Graph-based Visualization System for Monitoring and Characterizing Ongoing Smishing Threats," in Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy, 2024, pp. 257-268, doi: https://doi.org/10.1145/3714393.3726499.

[11] B. Pourghebleh, A. Aghaei Anvigh, A. R. Ramtin, and B. Mohammadi, "The importance of nature-inspired meta-heuristic algorithms for solving virtual machine consolidation problem in cloud environments," Cluster Computing, vol. 24, no. 3, pp. 2673-2696, 2021, doi: https://doi.org/10.1007/s10586-021-03294-4.

[12] S. N. Makhadmeh et al., "Recent advances in Grey Wolf Optimizer, its versions and applications," Ieee Access, vol. 12, pp. 22991-23028, 2023.

[13] S. Mirjalili and A. Lewis, "The whale optimization algorithm," Advances in engineering software, vol. 95, pp. 51-67, 2016.

[14] A. Muteeh, M. Sardaraz, and M. Tahir, "MrLBA: multi-resource load balancing algorithm for cloud computing using ant colony optimization," Cluster Computing, vol. 24, no. 4, pp. 3135-3145, 2021.

[15] S. Sefati, M. Mousavinasab, and R. Zareh Farkhady, "Load balancing in cloud computing environment using the Grey wolf optimization algorithm based on the reliability: performance evaluation," The Journal of Supercomputing, vol. 78, no. 1, pp. 18-42, 2022.

[16] K. Ramya and S. Ayothi, "Hybrid dingo and whale optimization algorithm - based optimal load balancing for cloud computing environment," Transactions on Emerging Telecommunications Technologies, vol. 34, no. 5, p. e4760, 2023.

[17] P. Geetha, S. Vivekanandan, R. Yogitha, and M. Jeyalakshmi, "Optimal load balancing in cloud: Introduction to hybrid optimization algorithm," Expert Systems with Applications, vol. 237, p. 121450, 2024.

[18] F. A. Emara, A. A. Gad-Elrab, A. Sobhi, A. S. Alsharkawy, M. E. Embabi, and M. Abd El-Baky, "Multi-objective task scheduling algorithm for load balancing in cloud computing based on improved Harris hawks optimization," The Journal of Supercomputing, vol. 81, no. 6, pp. 1-38, 2025.

[19] M. Haris and S. Zubair, "Battle Royale deep reinforcement learning algorithm for effective load balancing in cloud computing," Cluster Computing, vol. 28, no. 1, p. 19, 2025.

[20] S. Tabagchi Milan, N. Jafari Navimipour, H. Lohi Bavil, and S. Yalcin, "A QoS-based technique for load balancing in green cloud computing using an artificial bee colony algorithm," Journal of Experimental & Theoretical Artificial Intelligence, vol. 37, no. 2, pp. 307-342, 2025.

Automated Bubble Detection in Contact Lenses Using a Hybrid Deep Learning Framework

Chee Chin Lim[1], Yen Fook Chong[2], Vikneswaran Vijean[3], Gei Ki Tang[4]
Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Perlis, Malaysia[1, 3, 4]
Sports Engineering Research Centre (SERC), Universiti Malaysia Perlis, Perlis, Malaysia[1, 2, 3]

***Abstract*—This study presents a hybrid deep learning approach for automated detection of bubbles in contact lenses, aiming to enhance quality assurance in the manufacturing process. A hybrid AlexNet+SVM model was developed using transfer learning, where AlexNet's convolutional features were leveraged for binary classification (bubble vs. normal) via a Support Vector Machine (SVM) classifier. The dataset consisted of 320 images (160 bubbles, 160 normal) pre-processed using median filtering, local histogram equalization, and circular masking to improve image clarity and consistency. Through systematic hyperparameter tuning, the model achieved 100% testing accuracy and 97.92% validation accuracy, with perfect precision (100%) and high recall (96%). Comparative evaluation against ResNet and VGGNet demonstrated that the AlexNet+SVM model offered superior generalization and robustness, particularly for small-scale datasets. While VGGNet also achieved 100% testing accuracy with 95.83% validation accuracy, ResNet underperformed in recall (89%), likely due to its deeper architecture and data limitations. The findings underscore the suitability of hybrid models for binary classification tasks in limited-data scenarios. Identified challenges, including dataset size and risk of overfitting, point to future research directions involving expanded datasets and more advanced pre-processing techniques. This research contributes to the advancement of automated defect detection systems for contact lens manufacturing, offering a reliable and efficient quality control solution.**

Keywords*—Bubble detection; contact lens quality assurance; deep learning; transfer learning; Support Vector Machine (SVM); AlexNet; image pre-processing; binary classification; defect detection

I. INTRODUCTION

The eye is one of the most sensitive and vital organs in the human body, serving as the primary means of gathering information from the external environment. Humans rely heavily on vision, more than any other sensory modality, for perception and interaction with their surroundings [1]. The development of contact lenses has significantly advanced vision correction since their inception. In 1888, the first glass contact lenses were developed by Adolf E. Fick and Edouard Kalt, opticians based in Paris. However, glass contact lenses were associated with several limitations, including excessive weight, high cost, and safety concerns. Their fragility made them prone to breaking upon impact, necessitating careful handling. These drawbacks led to the development of alternative materials, culminating in the invention of the first hydrophilic hydrogel soft contact lenses by Czech chemists Drahoslav Lim and Otto Wichterle in 1959 [2].

Despite these advancements, the manufacturing process of soft contact lenses is not without challenges. One significant issue is the formation of bubbles during production. The manufacturing process involves heating the lens material until it reaches a molten state, followed by injection into computer-designed molds under high pressure. After cooling, the lenses are removed from the mold and undergo polishing to achieve smooth edges and surfaces. The lenses are then hydrated and subjected to quality assurance testing. However, improper handling or suboptimal manufacturing conditions can lead to the entrapment of bubbles during the heating and molding stages. These bubbles compromise the structural integrity and optical performance of the lenses.

Current inspection methods for detecting bubbles in contact lenses often depend on subjective human judgment, resulting in inconsistencies that can compromise quality control and allow defective lenses to reach end-users, potentially leading to discomfort or vision issues. While significant advancements have been made in contact lens classification and quality assessment using deep learning, the automated and highly accurate detection of specific manufacturing defects—particularly air bubbles—remains insufficiently addressed in both academic research and industrial practice. Existing systems frequently struggle with precision, especially in scenarios involving small datasets or variable imaging conditions and tend to overlook the importance of preprocessing techniques in improving model reliability.

To bridge this gap, this study introduces a hybrid deep learning model that combines AlexNet with a Support Vector Machine (SVM) classifier, specifically optimized for robust bubble detection in constrained data environments. Enhanced by image quality-focused preprocessing strategies, the proposed approach aims to improve inspection consistency, reduce reliance on manual evaluation, and support the development of reliable, fully automated quality assurance systems for contact lens manufacturing. The remainder of this paper is structured as follows: Section II presents a comprehensive review of related work. Section III describes dataset acquisition and preprocessing techniques. Section IV details the proposed methodology, including model architecture and tuning strategies. Section V discusses the results and comparative analysis. Section VI presents the conclusion, study limitations, and recommendations for future research.

II. Related Work

This section reviews key studies relevant to bubble detection in contact lenses. Gautam and Mukhopadhyay [3] employed AlexNet as a deep convolutional neural network (CNN) architecture, combined with an SVM-based Error-Correcting Output Code (ECOC) classifier. They evaluated Linear, Quadratic, Cubic, and Gaussian SVM models, focusing on validation accuracy to assess performance. Their work demonstrated the effectiveness of deep CNNs for contact lens detection. Raghavendra, Raja, and Busch [4] extended SVM applications by using it for a three-class classification problem: no lens, soft contact lens, or textured contact lens. This approach highlighted SVM's utility in multi-class lens detection tasks.

In image preprocessing, Arvind Kumar [5] compared Median and Gaussian filters for denoising, concluding that the Median filter achieved better results in less time, emphasizing its efficiency for image enhancement. Mandalapu, Ramachandra, and Busch [6] utilized the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) to extract texture features from textured contact lenses. This method provided insights into image quality and texture patterns across different manufacturers, showcasing the potential of quality metrics for lens analysis.

In deep learning, H. Hou et al. [20] using a new detection network of YOLOv8-BGA model as a baseline, which can achieve effective identification of leakage bubbles and bubble images collected under different lighting conditions in a practical industrial inspection environment. Convolutional neural networks have been used to detect and classify defects in colored contact lenses. Kim Ty et al. [21] employed CNN GoogLeNet V4 to detect defective contact lenses in the AOI system that yielded a good product recognition rate of 83.3% despite the limitations of computational resources and data.

While these studies reflect advancements in contact lens classification and quality assessment, most existing approaches rely heavily on large datasets and lack targeted solutions for specific defects such as bubbles. Furthermore, limited attention has been given to preprocessing strategies and model robustness under small-scale data and varied lighting conditions. To address these challenges, this paper proposes a lightweight, generalizable deep learning framework—integrating AlexNet with SVM—enhanced by quality-focused preprocessing techniques, specifically designed for accurate and automated bubble detection in contact lens manufacturing.

III. Methodology

The workflow of the bubble detection system is illustrated in Fig. 1. The process begins with data collection, where images of contact lens specimens are captured using the experimental setup described earlier. No Reference Image Quality Assessment (NR-IQA) is performed on the raw images to evaluate the quality of images captured with and without additional lighting. Following this, data augmentation and image cropping are applied to prepare the dataset for pre-processing.

The pre-processing stage involves several image enhancement techniques, including median filtering, Gaussian filtering, power law transformation, local histogram equalization, circle detection, and circle masking. To validate the effectiveness of these preprocessing methods, Image Quality Assessment (IQA) metrics such as Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), and Absolute Mean Brightness Error (AMBE) are employed. After preprocessing, the images are resized to a uniform dimension to ensure consistency before classification. The classification stage utilizes the AlexNet convolutional neural network (CNN) model with transfer learning, incorporating Support Vector Machines (SVM) for enhanced performance. Parameter tuning is conducted to optimize hyperparameters such as epoch count, batch size, learning rate, and the number of workers. The model's performance is evaluated using a confusion matrix to assess its accuracy in detecting normal lenses and lenses with bubbles.

Finally, the proposed AlexNet model with SVM transfer learning is compared against two other state-of-the-art models, AlexNet, VGGNet and ResNet, to benchmark its effectiveness in bubble detection. This comprehensive approach ensures a robust and reliable system for identifying defects in contact lenses.

A. Dataset and Data Collection

The dataset used in this study was curated from real-world contact lens manufacturing samples provided by an industrial supplier. It includes 20 images of lenses with bubbles and 16 images of normal lenses for analysis. These images represent actual production defects, making the dataset highly relevant for developing industrial-grade defect detection systems. Though small in size, the dataset's authenticity and diversity in lighting conditions support the development of robust models using transfer learning and data augmentation techniques.

These specimens were imaged using a Logitech C922 Pro Stream Webcam for detailed inspection. The experimental setup included hardware (laptop, Logitech C922 camera, tripod stand, USB Type-A 2.0 cable) and software (PyCharm IDE) [7]. The camera, set to 1080p resolution, was mounted on a tripod at a 90-degree angle, 6 cm above the specimens as in Fig. 2. A black cardboard background was used to enhance contrast, facilitating bubble visualization against the white lenses. Two imaging modes were utilised: one without additional lighting and one with phone camera light. A custom camera app developed using PyQt5, QtGui, and Pyshine libraries, featured 'Start' and 'Take picture' buttons for image capture. All images were saved in PNG format at 1440×1080 resolution.

B. No Reference Image Quality Assessment

No Reference Image Quality Assessment (NR-IQA) was conducted using the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) on the raw images to evaluate the quality of images captured with and without additional lighting [8]. BRISQUE was selected for its ability to quantify image quality without a reference image. The imaging mode yielding lower BRISQUE values, indicating higher quality, was chosen for subsequent data augmentation.

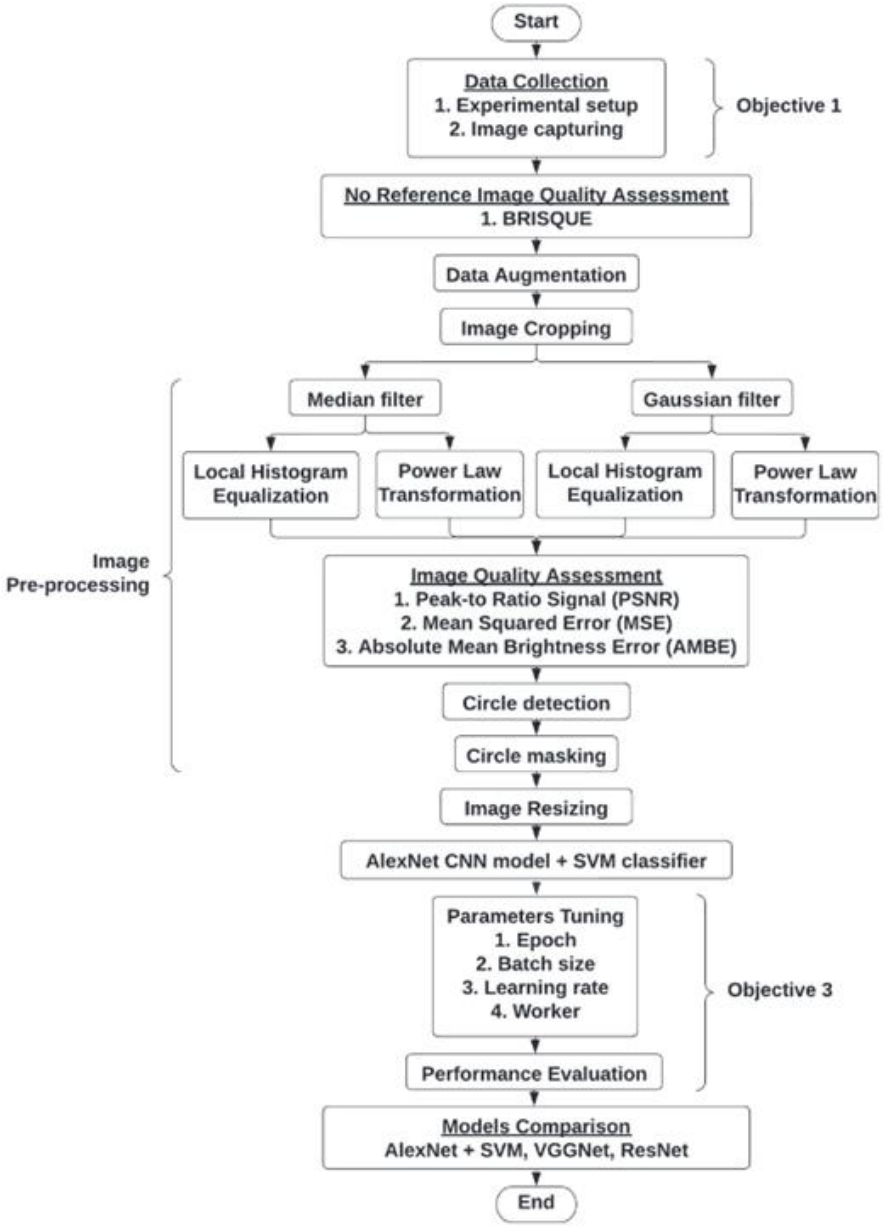


Fig. 1. Methodology workflow.

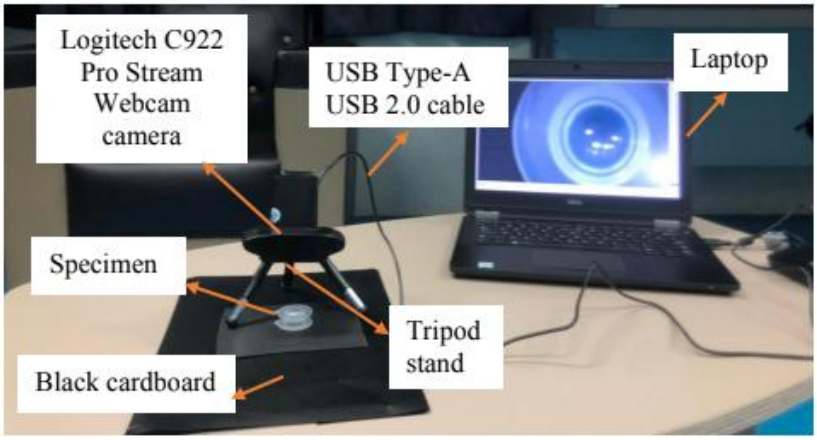


Fig. 2. Experimental setup.

C. Data Augmentation

Data augmentation was applied to enhance the dataset. For the 20 images of lenses with bubbles, each image was rotated at 0°, 90°, 180°, and 270°, resulting in 80 augmented images. For the 16 normal lens images, the first 10 images were rotated at 0°, 90°, 180°, and 270°, producing 40 images. The remaining 6 images were rotated at 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°, generating an additional 48 images. This resulted in a total of 88 normal lens images. Horizontal flipping was then applied to all augmented images, producing a final dataset of 320 images (160 with bubbles and 160 normal).

D. Image Cropping

The Region of Interest (ROI) for the dataset was the contact lens area, excluding the plastic mold and black cardboard background. Auto-cropping was performed on each image to remove unwanted regions, ensuring the background did not affect preprocessing. The cropped images varied in size between 240×240 and 350×350 pixels, as the ROI dimensions differed across specimens.

E. Image Pre-processing

Image pre-processing is essential for improving image quality, addressing factors like lighting intensity and camera quality. The process begins with segmenting the Region of Interest (ROI) within the boundary of interest, followed by focusing on the object of interest, categorized as either bubbles or normal.

1) Median filter: A non-linear filter used to reduce impulsive (salt-and-pepper) noise while preserving edges [9]. For a 5×5 filter, the centre pixel is replaced by the median value of the 25-pixel window, calculated by sorting pixel values and selecting the 13th value as in Eq. (1). This process is repeated across the image.

$$\hat{f}(x,y) = \underset{(s,t)\in S_{xv}}{median}\{g(s,t)\} \quad (1)$$

2) Gaussian filter: A linear filter applied to blur images or reduce noise. It performs a weighted average of surrounding pixels based on the filter size. A 5×5 kernel was used, with the sigma parameter controlling the blurring effect. The process iterates by shifting the window across the image by using Eq. (2).

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{(x^2+y^2)}{2\sigma^2}} \quad (2)$$

3) Local histogram equalization: Enhances contrast by adjusting image intensities using regional information. It spreads frequent intensity values, stretching the intensity range [10] to improve local contrast by using Eq. (3).

$$m_{s_{xy}} = \sum_{i=0}^{L-1} r_i P_{S_{xy}}(r_i) \quad (3)$$

4) Power law transformation: Maps narrow ranges of dark input values to wider output ranges or vice versa [11]. The transformation uses Eq. (4), where s and r are output and input pixel values, c is a constant, and γ controls the mapping. A γ value of 2 was chosen to darken the image, enhancing bubble edges.

$$s = c \cdot r^{\gamma} \quad (4)$$

F. Image Quality Assessment

Image Quality Assessment (IQA) evaluates distortions and degradations in images. Four pre-processing approaches were analyzed using Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), and Absolute Mean Brightness Error (AMBE). The approach yielding the best results across these metrics was selected for further processing.

1) PSNR: Measures the peak-to-noise ratio in decibels between the original and processed images. Higher PSNR values indicate better image quality [12]. It is calculated using Eq. (5), where R is the maximum fluctuation in the input image data type, and MSE is the mean squared error.

$$PSNR = log_{10}(\frac{R^2}{MSE}) \quad (5)$$

2) MSE: Represents the cumulative squared error between the original and processed images. Lower MSE values indicate smaller errors and better image quality [13]. Equation (6) calculates MSE, where n is the number of data points, Y_i is the observed value, and $\hat{Y}_i$ is the predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (6)$$

3) AMBE: Assesses the preservation of original brightness. Lower AMBE values indicate better brightness preservation. Eq. (7) calculates AMBE, where E(X) and E(Y) represent the mean errors of the original and processed images, respectively.

$$AMBE = E(X) - E(Y) \quad (7)$$

G. Circle Detection and Masking

Circle detection was performed using OpenCV's Hough Gradient Method, which leverages gradient information from edges. The function cv.HoughCircles (image, method, dp, minDist, param1, param2, minRadius, maxRadius) was utilized with the following parameters:

- image: Pre-processed input images.
- method: cv2.HOUGH_GRADIENT for gradient-based edge detection.
- dp: Set to 1 to match the accumulator resolution with the input image.
- minDist: Set to 200 to avoid false detection of overlapping circles.
- param1: Set to 50 as the higher threshold for the Canny edge detector.
- param2: Set to 30 as the accumulator threshold for circle center detection.
- minRadius and maxRadius: Both set to 0 to detect circles of any size.

The circle detection process is defined by Eq. (8), where represents the circle's (x_{centre}, y_{centre}) and r is its radius. The output, HoughCircles, provides the detected circle parameters. After circle detection, masking was applied to isolate the Region of Interest (ROI). A mask was created by hiding the outer region of the detected circle. The original image was then combined with the mask, resulting in an output image where only the contact lens ROI is visible, and the background is black. This step ensures focus on the lens area while eliminating irrelevant background information.

$$(x - x_{centre})^2 + (x - y_{centre})^2 = r - HoughCircles() \quad (8)$$

H. Convolution Neural Network AlexNet + SVM Model

This study employs a transfer learning approach by integrating the AlexNet architecture with a linear Support Vector Machine (SVM) for bubble detection in contact lenses. The output images after the circle masking were resized to 227 by 227 as this is the optimum image size acceptable by the AlexNet CNN model. AlexNet, chosen for its robust performance in visual recognition tasks [3], consists of 8 layers: 5 convolutional and 3 fully connected. To adapt it to the binary classification task (bubbles vs. normal), the original 1000-class output layer [14] is replaced with a linear SVM classifier. Key modifications include:

- ReLU Activation: Replaces Tanh to accelerate training speed by sixfold [15].
- Dropout (Rate: 0.5): Mitigates overfitting, though doubling training time.
- Overlap Pooling: Reduces network size [16] and error rates (0.4% top-1, 0.3% top-5).

The model leverages transfer learning to address limitations of small datasets and binary classification. The dataset (320 images: 160 bubbles, 160 normal) is partitioned into 70% training, 15% validation, and 15% testing. The SVM classifier is applied to the final fully connected layer (fc8), enabling precise two-class categorization as shown in Fig. 3. This hybrid approach optimizes AlexNet's feature extraction capabilities with SVM's efficiency in linear classification, ensuring robust performance despite dataset constraints.

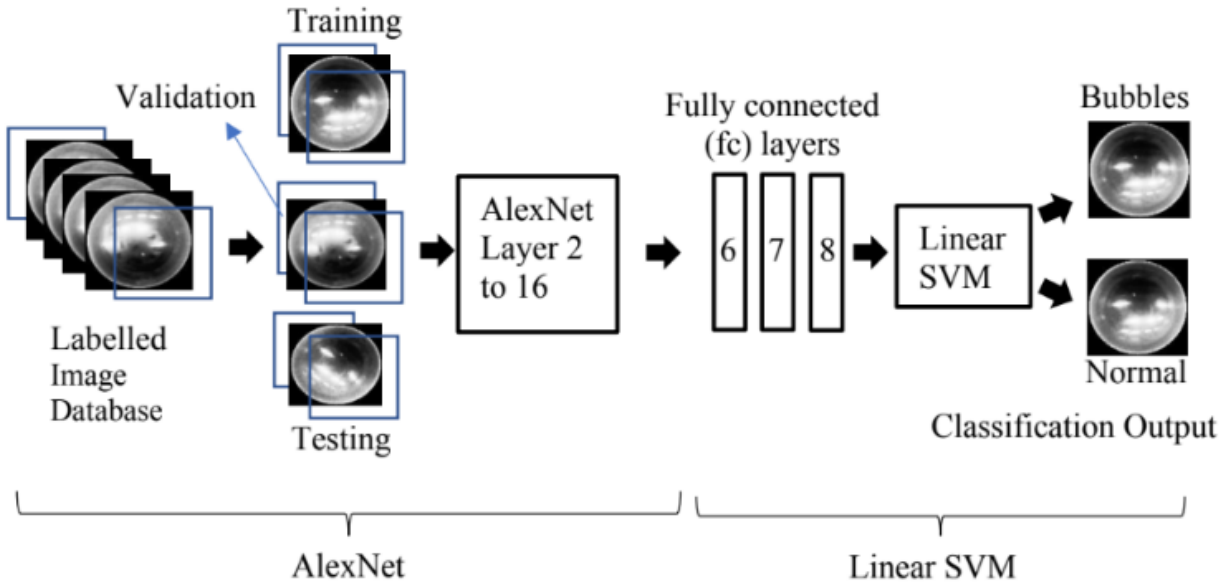


Fig. 3. Classification model of AlexNet + SVM on contact lens images.

IV. Result and Discussion

This section presents a comprehensive analysis of the experimental results, focusing on three critical phases of the study: (1) the evaluation of image pre-processing techniques to enhance dataset quality, (2) the optimization of hyperparameters for the AlexNet+SVM classification model, and (3) a comparative assessment of the performance of AlexNet, ResNet, and VGGNet architectures under identical conditions. First, the effectiveness of pre-processing methods is validated through No Reference Image Quality Assessment (NR-IQA) metrics, ensuring optimal input data for model training. Next, systematic parameter tuning—spanning epochs, batch size, learning rate, and workers—identifies configurations that maximize classification accuracy while mitigating overfitting. Finally, the optimized model is benchmarked against ResNet and VGGNet to determine the most robust architecture for bubble detection in contact lenses. The results highlight the interplay between data quality, model configuration, and architectural design, providing insights into achieving high precision and generalizability in defect detection tasks.

A. Result for Image Pre-processing

The BRISQUE scores for images captured without and with camera light were analyzed in Table I. For bubble specimens, the average BRISQUE values were 36.67 (without light) and 45.39 (with light). For normal specimens, the values were 43.89 (without light) and 48.98 (with light). Lower BRISQUE scores indicate better perceptual quality. Since images captured without camera light consistently showed lower BRISQUE values, they were selected for further pre-processing.

Four pre-processing approaches were evaluated using PSNR, MSE, and AMBE metrics as shown in Table II. The results showed that the median filter combined with local histogram equalization achieved the highest PSNR (27.870) and the lowest MSE (106.725) and AMBE (0.10518), indicating superior image quality. Consequently, this method was selected for pre-processing before model implementation.

TABLE I. BRISQUE VALUE FOR RAW DATA WITH AND WITHOUT CAMERA LIGHT

Data	BRISQUE (Average values)	
	Without camera light	With camera light
Bubble	36.67259	45.38722
Normal	43.89439	48.9829

TABLE II. AVERAGE IMAGE QUALITY ASSESSMENT OF THE ORIGINAL IMAGE AND IMAGE AFTER PRE-PROCESSING

Approach	Pre-processing Method	PSNR	MSE	AMBE
1	Median filter + Local histogram equalization	27.87000	106.72500	0.10518
2	Median filter + Power law transformation	26.75500	144.39200	0.18921
3	Gaussian filter + Local histogram equalization	27.86900	106.79500	0.10520
4	Gaussian filter + Power law transformation	26.75100	144.29800	0.18874

B. *Optimizing Turning Parameters for Deep Learning Classification Model*

This section analyzes the performance of the AlexNet+SVM model by evaluating the impact of varying parameters. The training and validation accuracy and loss were recorded and analyzed based on graph patterns. The results were compared between testing and validation accuracy for each tuning to identify optimal configurations.

1) Turning on Epoch: Epoch values of 40, 60, 80, 100, 120, and 140 were tested while keeping other parameters constant (learning rate: 0.0001, batch size: 6, workers: 8, optimizer: Adam, model: AlexNet, binary classification). The highest testing and validation accuracy (89.58%) were achieved at 100 epochs, indicating optimal performance as in Fig 4. Fig. 5 shows minimal gaps between training and validation accuracy and loss at this epoch, confirming model stability.

2) Turning on batch size: Batch sizes of 2, 4, 8, 16, and 32 were evaluated with other parameters fixed (epochs: 100, learning rate: 0.0001, workers: 8, optimizer: Adam, model: AlexNet, binary classification). A batch size of 32 yielded the highest testing accuracy (93.75%) and validation accuracy (83.33%). Larger batch sizes improved model performance, but further increases were impractical due to the validation size limit. Fig. 6 and Fig. 7 show the smallest gap between training and validation accuracy and loss at this batch size.

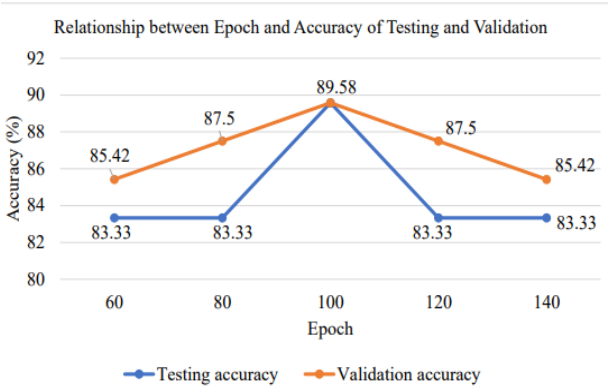


Fig. 4. Relationship between epoch and accuracy of testing and validation.

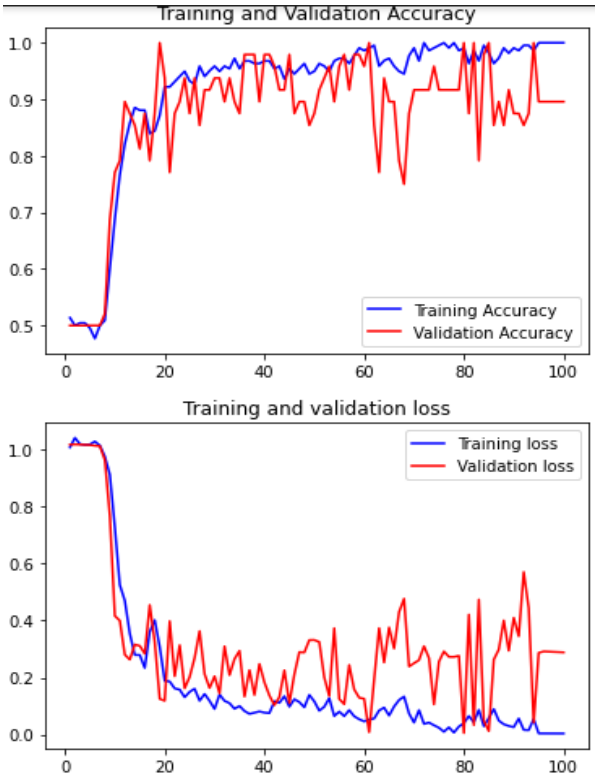


Fig. 5. Training and validation for accuracy and loss for tuning on epoch of 100.

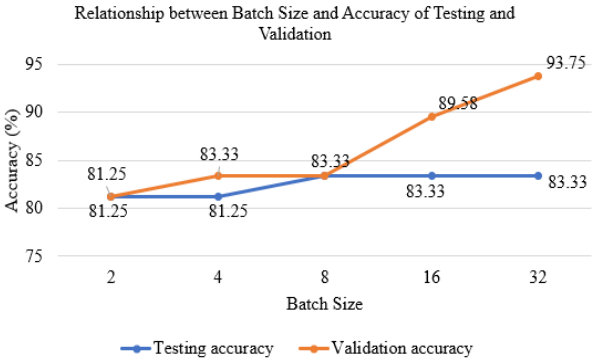


Fig. 6. Relationship between batch size and accuracy of testing and validation.

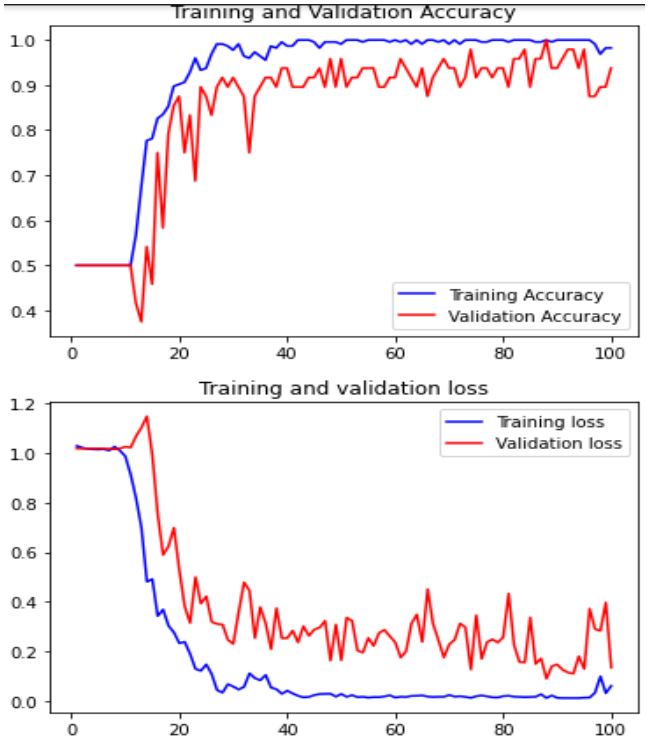


Fig. 7. Training and validation for accuracy and loss for tuning on batch size of 32.

3) Turning on learning rate: Learning rates of 0.1, 0.01, 0.001, 0.0001, 0.00001, and 0.000001 were tested while keeping other parameters constant (epoch: 100, batch size: 32, workers: 8, optimizer: Adam, model: AlexNet, binary classification). A learning rate of 0.0001 achieved the highest testing accuracy (87.5%) and validation accuracy (91.67%). Extremely low or high learning rates degraded performance, with 0.0001 providing the best balance. Fig. 8 and Fig. 9 demonstrate minimal gaps between training and validation accuracy and loss at this learning rate.

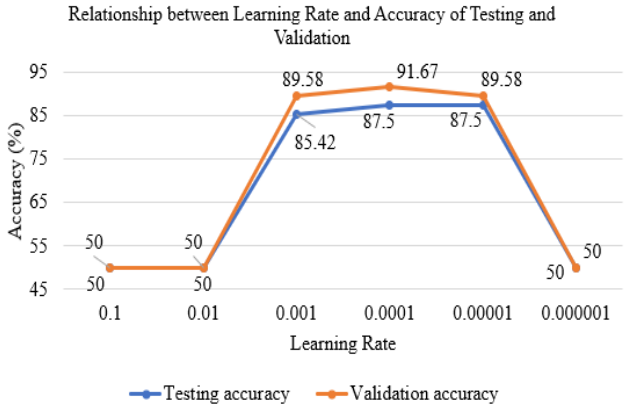


Fig. 8. Relationship between learning rate and accuracy of testing and validation.

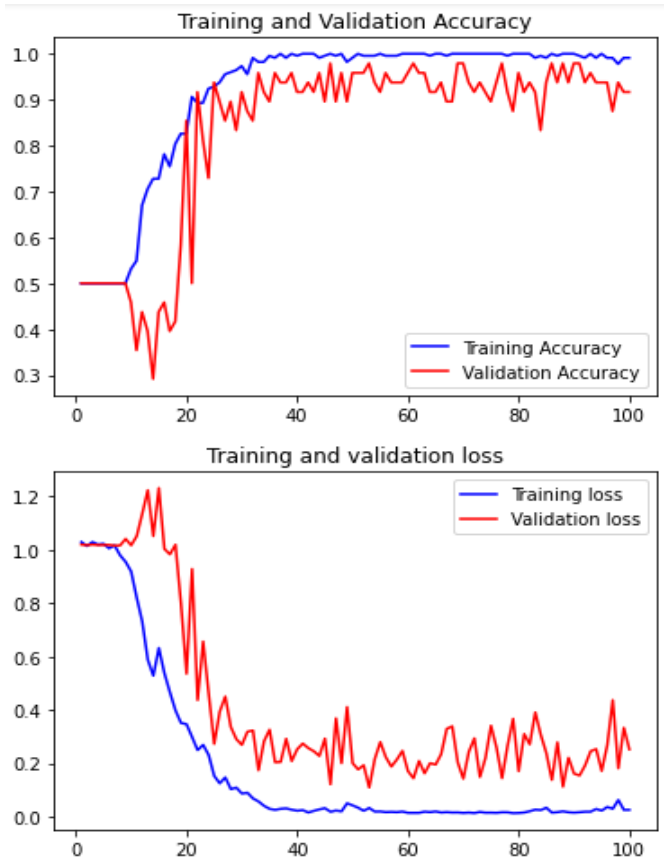


Fig. 9. Training and validation for accuracy and loss for tuning on learning rate of 0.0001.

4) Turning on worker: Worker values of 1, 2, 4, and 8 were tested with other parameters fixed (epochs: 100, learning rate: 0.0001, batch size: 32, optimiser: Adam, model: AlexNet, binary classification). A single worker achieved the highest testing and validation accuracy (93.75%). Additional workers did not improve performance, as the first worker efficiently managed batch loading and processing. Fig. 10 and Fig. 11 show the smallest gap between training and validation accuracy and loss with one worker.

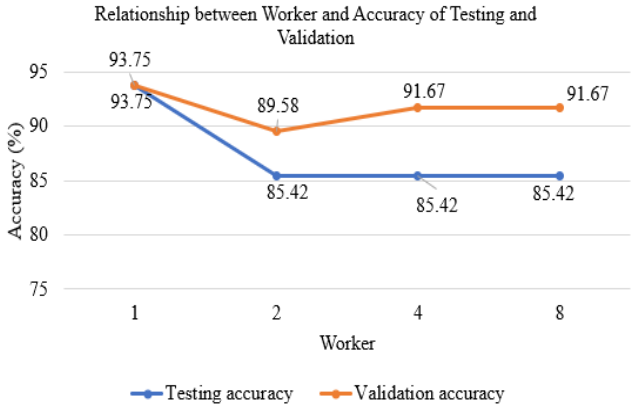


Fig. 10. Relationship between workers and accuracy of testing and validation.

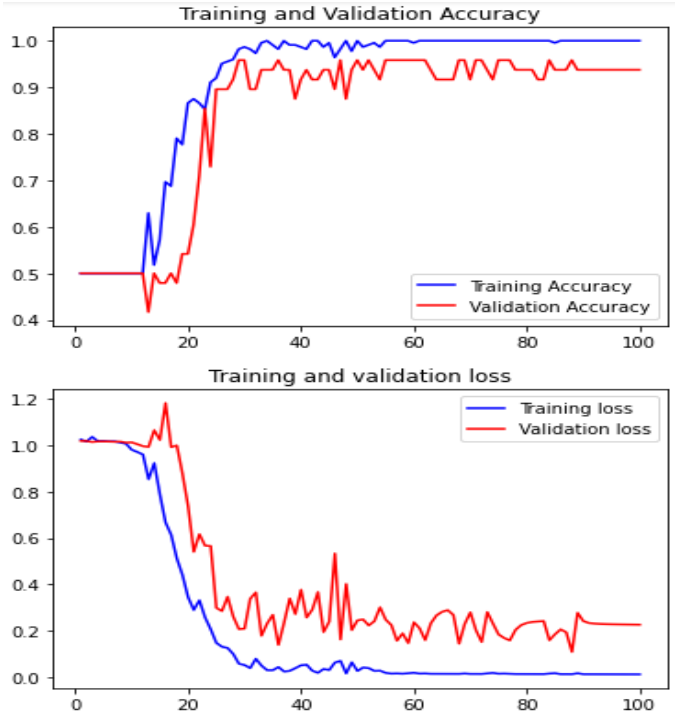


Fig. 11. Training and validation for accuracy and loss for tuning on workers of 1.

C. Comparison of Evaluation performance for AlexNet, ResNet and VGGNet and AlexNet+SVM at Optimum Tuning Parameters

After obtaining the optimal configuration, it was applied to AlexNet, ResNet, VGGNet and AlexNet+SVM to identify the best performing deep learning model for detecting bubble in contact lenses. This comparative analysis aimed to evaluate the effectiveness of each architecture under the same conditions, ensuring a robust and reliable solution for bubble detection. Table III presents the confusion matrix and its metrics of testing and validation accuracy results for AlexNet, ResNet, VGGNet and AlexNet+SVM models using identical parameters.

The AlexNet+SVM model achieved the highest validation accuracy (97.92%), outperforming VGGNet (95.83%) and ResNet (100%), while matching VGGNet's perfect testing accuracy (100%). All models demonstrated flawless precision (100%) and specificity (100%), confirming robust discrimination between bubbles and normal lenses. However, recall varied significantly: AlexNet and ResNet exhibited lower sensitivity (testing: 89%; validation: 89% and 100%, respectively), whereas VGGNet and AlexNet+SVM achieved perfect recall (100%) in testing, with AlexNet+SVM maintaining superior validation sensitivity (96%).

The AlexNet+SVM hybrid model's enhanced performance stems from integrating SVM's linear classification efficiency with AlexNet's feature extraction capabilities, mitigating overfitting risks inherent in deeper architectures like ResNet (138M parameters) [17]. VGGNet's intermediate parameter count (60.3M) and hierarchical feature learning likely contributed to its strong testing accuracy, though validation performance lagged slightly due to dataset variability [18]. AlexNet's moderate parameter size (62M) and SVM-driven classification optimized generalization, as evidenced by its balanced F1 scores (testing: 94; validation: 97.96). These results align with theoretical expectations: larger models (ResNet) risk overfitting on small datasets, while hybrid approaches (AlexNet+SVM) enhance stability by combining convolutional feature learning with linear discriminative power [19].

TABLE III. COMPARISON OF MODEL PERFORMANCE

	Model	TP	FP	TN	FN	Accuracy	Precision	Sensitivity	Specificity	F1 score
Testing	AlexNet	24	0	21	3	93.75	100	89	100	94
	ResNet	24	0	21	3	93.75	100	89	100	94
	VGGNet	24	0	24	0	100	100	100	100	100
	AlexNet+SVM	24	0	24	0	100	100	100	100	100
Validation	AlexNet	24	0	21	3	93.75	100	89	100	94
	ResNet	24	0	24	0	100	100	100	100	100
	VGGNet	24	0	22	2	95.83	100	92.31	100	96
	AlexNet+SVM	24	0	23	1	97.92	100	96	100	97.96

V. CONCLUSION

In conclusion, this study successfully achieved its objectives of developing a hybrid deep learning framework that integrates AlexNet with an SVM classifier to detect bubbles in contact lenses. The AlexNet model, optimized with specific parameters, achieved testing and validation accuracies of 93.75%. Performance evaluation using confusion matrix metrics demonstrated high precision, recall, specificity, and F1 scores. Comparative analysis revealed that the hybrid AlexNet+SVM model outperformed conventional AlexNet, ResNet, and VGGNet models achieving 100% testing accuracy and 97.92% validation accuracy, along with superior recall (96%) and F1 scores (97.96%).

Image preprocessing using median filtering and local histogram equalization significantly enhanced the quality of input data, contributing to superior classification performance. The optimized hyperparameters further ensured stable training and validation behavior. Although the model achieved high accuracy and demonstrated strong robustness, its reliance on a

small dataset may limit generalizability to broader production environments. This constraint, along with the observed risk of overfitting in deeper architectures like ResNet, highlights the need for larger, more diverse datasets to further validate the model's suitability for real-time industrial deployment. Future recommendations include enhancing pre-processing techniques, such as sharpening, and expanding the dataset to improve model accuracy and robustness. Additionally, exploring other hybrid architectures or advanced data augmentation techniques could further enhance model generalization and performance; and integrating the system into a live quality control pipeline.

ACKNOWLEDGMENT

The authors would like to thank Universiti Malaysia Perlis for providing the opportunity to collaborate with the contact lens manufacturing industry for practical research.

REFERENCES

[1] K. Irsch, J. H. Medicine, D. L. Guyton, and J. H. Medicine, "Encyclopedia of Biometrics," Encycl. Biometrics, no. January, pp. 10–16, 2009, doi: 10.1007/978-0-387- 73003-5.

[2] Mana, N. A. M. A., Chin, L. C., Fook, C. Y., Yazid, H, Ali, Y. M.: A review on contact lens inspection, Indonesian Journal of Electrical Engineering and Computer Science, 31(2), pp. 700-712, (2023).

[3] Gautam, G. and Mukhopadhyay, S.: Contact Lens Detection using Transfer Learning with Deep Representations. In: Proc. Int. Jt. Conf. Neural Networks, vol. 2018-July, (2018), doi: 10.1109/IJCNN.2018.8489590.

[4] Raghavendra, R., Raja, K.B., and Busch, C.: Ensemble of statistically independent fil-ters for robust contact lens detection in iris images. In: ACM Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP), pp. 24:1–24:7 (2014).

[5] Arvind, K: Comparative analysis of gaussian filter, median filter and denoise autoeno-coder. IEEE Access (6), 45–51 (2020).

[6] Mandalapu, H., Ramachandra, R., and Busch, C: Image quality and texture-based fea-tures for reliable textured contact lens detection. In: Proc. - 14th Int. Conf. Signal Image Technol. Internet Based Syst. SITIS 2018, pp. 587–594 (2018), doi: 10.1109/SITIS.2018.00095.

[7] Mana, N. A. M A, Chin, L. C., Yazid, H., and Fook, C. Y, Evaluation of contact lens data acquisition approaches using enhancement techniques. In: 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, pp. 1-6, (2022), doi: 10.1109/AIST55798.2022.10065211.

[8] Shabashevich, A., Pierce, J., Dugan, W., Penoyer, G., and Kukla, A.: Project number P06219: Bubble elimination on the surface of a contact lens submerged in de-ionized water. In: Proceedings of the KGCOE Multi-Disciplinary Engineering Design Confer-ence, no. May, pp. 06219, Kate Gleason College of Engineering, Rochester Institute of Technology (2006).

[9] Nosrati, M. and Karimi, R.: Detection of circular shapes from impulse noisy images us-ing median and laplacian filter and circular hough transform. In: CCE 2011 - 2011 8th Int. Conf. Electr. Eng. Comput. Sci. Autom. Control. Progr. Abstr. B., no. October (2011), doi: 10.1109/ICEEE.2011.6106710.

[10] Sudhakar, S: Histogram Equalization. Towards Data Science, (2017). https://towardsdatascience.com/histogram-equalization-5d1013626e64, last accessed 2025/02/25.

[11] Kang and Atul: Power Law (Gamma) Transformations. The AI Learner, (2019) https://theailearner.com/2019/01/26/power-law-gamma-transformations/, last accessed 2025/02/25.

[12] MathWorks: Compute peak signal-to-noise ratio (PSNR) between images - Simulink. https://www.mathworks.com/help/vision/ref/psnr.html, last accessed 2025/02/25.

[13] Sara, U., Akter, M., and Uddin, M. S.: Image quality assessment through FSIM, SSIM, MSE and PSNR — A comparative study. J. Comput. Commun., 7(03) pp. 8–18, (2019) doi: 10.4236/jcc.2019.73002.

[14] Noor, S. S., Michael, K., Marshall, S., and Ren, J.: Hyperspectral image enhancement and mixture deep-learning classification of corneal epithelium injuries. Sensors, 17(11), 2644, (2017). https://doi.org/10.3390/s17112644

[15] Saxena, S.: Alexnet architecture | introduction to architecture of Alexnet. Analytics Vidhya, (2021). https://www.analyticsvidhya.com/blog/2021/03/introduction-to-the-architecture-of-alexnet/, last accessed 2025/02/25.

[16] Hassan, M. U.: AlexNet - ImageNet Classification with Convolutional Neural Networks. Neurohive, (2018) https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/, last accessed 2025/02/25.

[17] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778, (2016).

[18] Simonyan, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition. In: arXiv preprint arXiv:1409.1556, (2014).

[19] Krizhevsky, A., Sutskever, I., and Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)

[20] Kim, T.-y., Park, D., Moon, H., & Hwang, S.-s.: A Deep Learning Technique for Optical Inspection of Color Contact Lenses. Applied Sciences, 13(10), 5966, (2023). https://doi.org/10.3390/app13105966

[21] Hou, H., Chou Z., Jian, C., Target Detection of Leakage Bubbles in Stainless Steel Pipe Gas Airtighness experiments Based on YOLOv8-BGA, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, pp. 627 – 640, (2025). https://doi.org/10.14569/IJACSA.2025.0160264

Integration of Grey Wolf Optimizer Algorithm with Combinatorial Testing for Test Suite Generation

Muhamad Asyraf Anuar, Rosziati Ibrahim, Mazidah Mat Rejab, Nurezayana Zainal
Department of Software Engineering, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

***Abstract*—Combinatorial Testing (CT) is a software testing technique designed to detect defects in complex systems by efficiently covering diverse combinations of input parameters within given time and resource constraints. A common strategy in CT is t-way testing, which ensures that all possible interactions among any t parameters are tested at least once. The Grey Wolf Optimization Algorithm (GWOA) is a nature-inspired metaheuristic that has been successfully applied to various optimization problems. In this study, we introduce the Combinatorial Grey Wolf Optimization Algorithm (CGWOA), which integrates GWOA with CT to enhance test suite generation. Effectiveness of CGWOA is evaluated through experiments on a real-world software system, where it is found that the number of test cases was reduced by 98%, from 3000 to 40, while still ensuring complete 2-way interaction coverage. Experimental results demonstrate that CGWOA consistently produces smaller test suites compared to pure computation methods such as Jenny, IPOG, IPOG-D and TConfig, as CGWOA consistently outperformed, especially in handling both lower and higher interaction strengths. In scenarios with binary parameters, CGWOA delivered the smallest test suites, while in more complex configurations, even in MCA settings, it showed impressive scalability, outperforming the other algorithms. Statistical analysis using the Wilcoxon signed-rank test revealed that the proposed approach significantly outperforms existing methods, with all p-values less than 0.02 after applying the Holm correction. The experimental results demonstrate that the proposed CGWOA approach advances software testing by efficiently minimizing the number of test cases required to achieve complete test coverage.**

Keywords*—Grey wolf optimizer algorithm; combinatorial testing; metaheuristics; t-way testing

I. Introduction

Combinatorial Testing (CT) has been widely recognized for its effectiveness in identifying software defects [1], as demonstrated by numerous empirical studies. Its success depends on a well-structured set of test cases, known as a Covering Array (CA), which ensures that all interactions among the parameters influencing the behavior of the System Under Testing (SUT) are covered [2]. CT detects failures by examining interactions between parameters in the testing process. It has been used to automate test suite generation, aiming to produce a reliable product that is usable across a wide range of inputs. However, the high cost of CT can be a significant obstacle to its widespread adoption. Developers spend approximately 50% of their time identifying and fixing bugs [3] which highlights the need for more efficient and cost-effective testing strategies.

As software systems grow in complexity and size, the number of possible input combinations become overwhelmingly large, making complete coverage challenging with traditional CT methods. To address this, researchers and practitioners are exploring cost-reduction strategies, such as leveraging machine learning to automate test case generation [4] or prioritizing test cases based on their defect detection potential [5]. Despite these challenges, CT remains a crucial approach for ensuring software quality and minimizing defects in complex systems.

To tackle CT challenges, various heuristic search techniques have been introduced to optimize test case selection and reduce the cost of CT. CT employs a covering array as a test suite to systematically cover parameter combinations, aiming to balance the number of test cases with their effectiveness in detecting failures [3]. Among these approaches, the GWOA has gained attention for its ability to balance exploration and exploitation, where it first explores the search space and then converges towards the best solutions. GWOA is a metaheuristic optimization method [6] inspired by the hunting behavior of grey wolves and has been applied to various optimization problems.

This study proposes an innovative CT approach leveraging GWOA, combining the strengths of both techniques to maximize test coverage while minimizing the number of test cases. It aims to investigate the effectiveness of integrating the Grey Wolf Optimizer Algorithm with Combinatorial Testing to reduce test suite size while ensuring complete t-way interaction coverage across diverse parameter configurations. The goal is to assess whether the proposed CGWOA approach offers a more scalable and efficient solution than traditional combinatorial testing techniques.

A. Combinatorial Testing

Combinatorial testing is a software testing technique that systematically covers combinations of input parameters to detect interaction faults efficiently. Instead of testing every possible configuration, which can be infeasible, it focuses on ensuring that all t-way combinations are tested at least once. A single test suite can be designed to cover all t-wise combinations of these parameters [1] while minimizing the total number of test cases.

This study addresses the combinatorial optimization challenge of t-way interaction test generation, where t denotes the interaction strength or combination degree. The goal is to generate minimal yet effective test suites that ensure full t-wise coverage across complex software configurations. In recent years, various methodologies have been proposed to tackle this problem, leading to numerous computation-based solutions. Despite these advances, many existing approaches still suffer from limitations in adaptability [7], especially when dealing with systems involving diverse parameter types and values.

B. Overview of T-way Testing

A test suite is structured as an N × k array, where N is the number of test cases and k is the number of parameters. Each row represents a unique combination of parameter values. The suite is built by first enumerating all t-way value combinations (t-tuples) and then selecting the minimal number of test cases required to cover them. These test cases are used to assess the system under different interaction scenarios, enabling the identification of faults that arise from specific parameter pairings. The following formal definitions clarify key terms relevant to this approach:

- T-way testing is a combinatorial testing method that ensures all possible combinations of t input parameters are exercised at least once. It provides a balance between test coverage and resource efficiency.
- A Covering Array (CA), denoted as CA(N; t; v^p), is a matrix covering all t-way combinations where each parameter has v values.
- A Mixed Covering Array (MCA), denoted as MCA(N; t; v_1^p_1, v_2^p_2, ..., v_i^p_i), generalizes the CA to support varying numbers of values per parameter.
- A t-tuple is a t-length subset of parameter values that must appear at least once across all test cases.

Mathematically, t-way testing can be described using Orthogonal Arrays (OAs) [8], where each subarray contains all ordered t-sized subsets of parameter values. OAs are typically represented using parameters such as N (number of test cases), k (number of parameters), v (number of values per parameter), and t (interaction strength). However, a key limitation of OAs is their requirement for uniform value counts across all parameters.

To address this issue, Covering Arrays (CAs) [9] were introduced. CA, denoted as CA(N; t, k, v), guarantees that every t-combination of values is covered at least once in a test suite. For example, CA(N; 2, 5, 4) covers all pairwise interactions among five parameters, each with four possible values. While CAs offer improved flexibility over OAs, they still assume a uniform number of values per parameter. For real-world systems where parameters often have varying values, the Mixed Covering Array (MCA) provides a more general solution. An MCA is defined as MCA(N, t, C), where C represents the configuration of parameter-value groupings, such as ($v_1^{k_1}$, $v_2^{k_2}$, ..., $v_n^{k_n}$). For instance, MCA(7, 3, 4^6, 2^4) refers to a 3-way interaction test suite with 7 test cases, covering 6 parameters with 6 values each and 4 parameters with 2 values each.

To illustrate the principles of t-way testing, consider a hypothetical configuration scenario involving a mobile e-commerce application where the system consists of four key input parameters: Payment Method, Login Authentication, Delivery Option, and Notification Type. The Payment Method parameter supports three values of Credit Card (C), Digital Wallet (W), and Cash on Delivery (D) while the other three parameters each support two values of Password (P) or Biometric (B) for Login Authentication, Standard (S) or Express (E) for Delivery Option, and Email (E) or SMS (S) for Notification Type, as shown in Table I.

TABLE I. INPUT PARAMETERS AND CORRESPONDING VALUES

Payment	Auth	Delivery	Notification
Credit Card (C)	Password (P)	Standard (S)	Email (E)
Digital Wallet (W)	Biometric (B)	Express (E)	SMS (S)
Cash on Delivery (D)			

Assuming a 2-way interaction strength (t = 2), the goal is to construct a test suite in which all possible pairs of values across all parameter combinations are covered at least once. Exhaustively testing every possible configuration would require 3 × 2 × 2 × 2 = 24 test cases. However, by applying combinatorial optimization, the number of required test cases can be significantly reduced without sacrificing 2-way coverage. Based on the configuration above, the following test suite of 12 test cases achieves complete 2-way coverage with a significantly reduced number of tests compared to exhaustive enumeration, as shown in Table II.

TABLE II. 2-WAY COVERING TEST SUITE

Test Case	Payment	Auth	Delivery	Notification
TC1	C	P	S	E
TC2	W	B	E	S
TC3	D	P	E	S
TC4	C	B	S	S
TC5	W	P	S	E
TC6	D	B	S	E
TC7	C	P	E	E
TC8	W	B	S	S
TC9	D	P	S	S
TC10	C	B	E	S
TC11	W	P	E	S
TC12	D	B	E	E

Based on the configuration of the mobile e-commerce application above, the system can be formally represented as a Mixed Covering Array as follows:

$$MCA\ (12;2;3^1 2^3)$$

This denotes a test suite containing 12 test cases that achieves full pairwise coverage across one 3-valued parameter and three 2-valued parameters. The reduced test count demonstrates the efficiency of combinatorial testing in minimizing redundant combinations while preserving interaction coverage. This test suite ensures that every pairwise combination of parameter values appears in at least one test case. For example, the pair (Wallet, Biometric) is covered in TC2 and TC8, while (Credit Card, Express) appears in TC7 and TC10. This approach demonstrates how combinatorial testing can yield significant reductions in testing effort while maintaining robust fault detection capabilities.

C. Metaheuristic Algorithm

Metaheuristic algorithms are a diverse class of nature-inspired, derivative-free optimization techniques that excel in solving complex real-world problems where traditional analytical methods fall short [10]. These algorithms are generally classified into four main categories of swarm-based,

evolutionary, physics-based, and human-based approaches [11]. Evolutionary algorithms (EAs), for example, are inspired by Darwinian natural selection, with genetic algorithms being a well-known instance that simulates biological evolution [12].

Over the past two decades, metaheuristic algorithms have been extensively adapted, modified, and integrated with other intelligent techniques for various engineering applications, ranging from Proton Exchange Membrane Fuel Cell (PEMFC) design [13] to software testing, including their role in Test Case Prioritization [14][15]. A survey of 14 metaheuristic algorithms introduced between 2000 and 2020 [16] highlights the research trends, the hybridization of metaheuristics, advancements in parallel metaheuristics, unresolved challenges, and emerging research opportunities of metaheuristic algorithms.

Metaheuristics are particularly well-suited for problems involving complex constraints or mixed-integer variables [17], where solutions are refined within predefined limits and constraints are effectively managed to transform constrained problems into unconstrained ones. While challenges remain in fully understanding their convergence properties and guaranteeing global optimality across diverse problem domains, ongoing research continues to enhance their efficiency and reliability. As advancements in metaheuristic design progress, these algorithms are becoming increasingly adaptable and robust, reinforcing their potential as powerful optimization tools.

Metaheuristics serve as a powerful approach for tackling optimization problems that are too complex or large to be solved precisely. They are widely applicable across various optimization challenges, including software testing, an essential field in IT that encompasses diverse testing strategies, methodologies, and techniques, including metaheuristic-based approaches [18]. By providing a flexible and efficient means of identifying high quality, near optimal solutions, metaheuristics prove valuable in scenarios, where traditional optimization methods may be impractical or ineffective.

The key contributions of this study are threefold. First, it presents a comprehensive overview of CT techniques and the Grey Wolf Algorithm (GWOA), outlining their strengths and limitations. Second, it provides a detailed explanation of the proposed approach that integrates CT with GWOA, emphasizing its key features and advantages. Third, the effectiveness of the proposed approach is evaluated by comparing its performance with pure computation CT techniques to assess its efficiency under various configurations.

D. Grey Wolf Optimizer Algorithm

Grey Wolf Optimizer Algorithm (GWOA) is an optimization technique inspired by the social structure and hunting strategies of grey wolves. In this method, wolves are organized into four hierarchical levels of alpha, beta, delta, and omega, reflecting a ranking from the strongest to the weakest candidate solutions. The hunting process is mainly directed by the top three wolves of alpha, beta, and delta, while the rest adjust their positions based on these leaders [19].

To utilize GWOA, the input data must first be properly defined and formatted, typically as an array or matrix. The optimization begins with a randomly generated initial population of grey wolves, each representing a potential solution within the search space. These wolves are ranked according to their fitness scores, with the three best classified as alpha (α), beta (β), and delta (δ), while the others, referred to as omega (ω), update their positions in response to the movements of the leading wolves. The key phases of the hunting process of tracking, encircling, and attacking prey are represented mathematically. After identifying the prey, the grey wolves begin to encircle it, which is mathematically modeled using the following equations:

$$\vec{D} = \left|\vec{C} \times \overrightarrow{X_p}(t) - \vec{X}(t)\right| \quad (1)$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A} \times \vec{D} \quad (2)$$

Eq. (1) models the step size taken by a wolf towards one of the leading wolves (i.e., α, β, or δ), and Eq. (2) defines the final updated position of the omega wolf. Here, t is the t[th] iteration, $\overrightarrow{X_p}(t)$ are the position vectors of the prey, and the coefficient vector of $\vec{A}$ and $\vec{C}$ calculated as:

$$\vec{A} = 2\vec{a} \times \overrightarrow{r_1} - \vec{a} \quad (3)$$

$$\vec{C} = 2 \times \overrightarrow{r_2} \quad (4)$$

To simulate encircling behavior, the vectors $\vec{A}$ and $\vec{C}$ are computed using Eq. (3) and Eq. (4), where $\vec{a}$ decreases linearly from 2 to 0, and $\overrightarrow{r_1}$ and $\overrightarrow{r_2}$ are random vectors in [0, 1]. $\vec{A}$ controls step size and direction, while $\vec{C}$ adjusts the influence of the leader's position, balancing exploration and exploitation.

To simulate hunting behavior, it is assumed that the α, β, and δ wolves possess superior knowledge about the prey's probable location. Therefore, their positions are treated as the best solutions discovered so far, and the remaining wolves (i adjust their positions based on those of the leading wolves. This adaptive behavior is mathematically expressed through the following equations:

$$\vec{D_\alpha} = |\vec{C_1} \times \vec{X_\alpha} - \vec{X}|, \vec{D_\beta} = |\vec{C_2} \times \vec{X_\beta} - \vec{X}|, \vec{D_\delta} = |\vec{C_3} \times \vec{X_\delta} - \vec{X}| \quad (5)$$

$$\vec{X_1} = \vec{X_\alpha} - \vec{A_1} \times \vec{D_\alpha}, \vec{X_2} = \vec{X_\beta} - \vec{A_2} \times \vec{D_\beta}, \vec{X_3} = \vec{X_\delta} - \vec{A_3} \times \vec{D_\delta} \quad (6)$$

$$\vec{X}(t+1) = (\vec{X_1} + \vec{X_2} + \vec{X_3})/3 \quad (7)$$

In this formulation, $\vec{X_\alpha}$, $\vec{X_\beta}$, and $\vec{X_\delta}$ represent the best positions identified by the α, β, and δ wolves, respectively. The vectors $\vec{C_1}$, $\vec{C_2}$, and $\vec{C_3}$ are randomly generated coefficients, while $\vec{X}$ denotes the current position of the search agent under evaluation. Likewise, $\vec{A_1}$, $\vec{A_2}$, and $\vec{A_3}$ are randomly distributed coefficient vectors, and t represents the current iteration number. Eq. (5) to Eq. (7) describe how the position of each search agent is updated based on the influence of the leading wolves.

The fundamental principle of the GWO algorithm is that omega wolves adjust their positions during the search process according to the locations of the α, β, and δ wolves, thereby simulating the surrounding and attacking of prey. This strategy

allows omegas to diverge from one another while collectively converging towards the prey. When |A|<1, convergence towards prey is emphasized, promoting exploitation. However, this may lead to premature convergence and local optima stagnation. To mitigate this, values of $\vec{A}$ greater than 1 or less than −1 are introduced to encourage divergence from the prey and enhance global exploration in search of better solutions. Additionally, the vector $\vec{C}$ contributes to exploration behavior. When $\vec{C}$>1, the influence of the prey on the search agent is amplified. Conversely, when $\vec{C}$ <1, this influence is reduced, thereby modifying the distance calculation in Equation (1). It is important to note that the parameter $\vec{A}$ decreases linearly throughout the optimization process to promote exploitation in later iterations.

In contrast, $\vec{C}$ remains randomly distributed throughout the process, balancing exploration and exploitation. This dynamic adjustment of $\vec{A}$ and $\vec{C}$ has proven to be an effective strategy for avoiding stagnation in local optima. The flowchart of the GWO algorithm (Algorithm 1) is presented in Fig. 1.

Algorithm 1: GWOA Algorithm

Input: Objective function f(x)

Output: X_alpha: Best-found solution

1. Initialize the grey wolf population X = {X_1, X_2, ..., X_N} randomly
2. Evaluate the fitness of each search agent using f(x)
3. Identify the best three solutions as Alpha (X_alpha), (X_beta), and Delta (X_delta)
4. Set iteration counter t = 0
5. **while** t < MaxIter do
6. a = 2 - 2 * (t / MaxIter)
7. **for** each wolf X_i do
8. **for** each dimension d = 1 to D do
9. Generate r_1, $r_2 \in [0, 1]$
10. A = 2 * a * r_1 - a
11. C = 2 * r_2
12. D_alpha = |C * X_alpha − X_i|
13. D_beta = |C * X_beta − X_i|
14. D_delta = |C * X_delta − X_i|
15. X_1 = X_alpha − A * D_alpha
16. X_2 = X_beta − A * D_beta
17. X_3 = X_delta − A * D_delta
18. X_i = (X_1 + X_2 + X_3) / 3
19. end for
20. end for
21. Evaluate new fitness values
22. Update X_alpha, X_beta, X_delta
23. t = t + 1
24. end while

1.

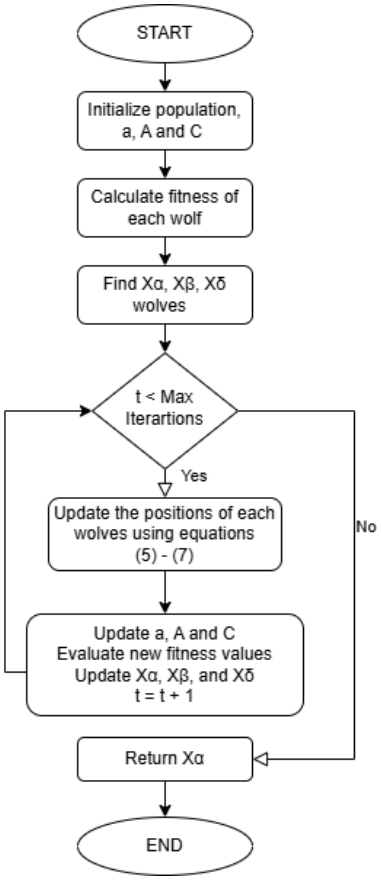


Fig. 1. Flowchart of grey wolf optimization algorithm.

One of the key advantages of GWOA is its ability to balance exploration and exploitation. The algorithm initially explores a wide range of solutions, but as iterations progress, it shifts focus toward refining the best solutions found. Studies have shown that GWOA performs competitively against other optimization algorithms in terms of convergence speed and solution accuracy [20], [21], [22]. Studies have demonstrated that GWOA is effective in cancer detection [23], scheduling [24], and engineering design [25], among other fields [26]. For instance, research has shown that GWOA improved the decision tree model through hyperparameter optimization for diabetes classification [27].

However, like other metaheuristic algorithms, GWOA does not guarantee the ability to pinpoint the exact position of the global optimal solution in every case. The algorithm may converge to a local optimum or fail to find the best solution within the given number of iterations, particularly if the search space is highly complex or if parameter tuning is not optimized. Despite these challenges, GWOA remains a valuable tool for solving problems with large search spaces and intricate constraints due to its ability to balance exploration and exploitation while requiring minimal prior knowledge of the problem domain.

The remainder of this study is structured as follows: Section II presents related work on combinatorial testing and the application of metaheuristic algorithms in test generation. Section III describes the proposed Combinatorial Grey Wolf

Optimizer Algorithm, including its design, parameters, and workflow. Section IV discusses the experimental setup, results, and performance analysis of CGWOA compared to existing combinatorial testing methods. Finally, Section V concludes the study and outlines potential directions for future research.

II. Related Work

In existing research, various heuristic search algorithms have been proposed to guide the process of Combinatorial Testing (CT) and reduce its cost. These algorithms aim to enhance the efficiency of CT by prioritizing the most relevant combinations of input values and parameters. Grey Wolf Optimization Algorithm (GWOA) has been successfully applied to various optimization problems as it has demonstrated strong exploration-exploitation capabilities. Given these characteristics, GWOA has the potential to significantly improve the efficiency of CT by optimizing test suite generation. The following section reviews related work on CT and GWOA, focusing on previous efforts to integrate metaheuristic optimization techniques with CT. The key contributions and limitations of these approaches are discussed, highlighting their relevance to the proposed study. Additionally, an overview of the current state of CT and GWOA research is provided, along with the main research gaps this work aims to address.

Several studies have compared GWOA with other metaheuristic algorithms, highlighting its effectiveness in various optimization tasks. For instance, Pace demonstrated that GWOA outperforms other algorithms in optimizing Transient Electromagnetic (TDEM) data [22]. Similarly, a 2024 study showed that GWOA delivers superior results in optimizing battery performance and lifespan compared to alternative approaches [21]. Due to its strong foundation, GWOA has been the basis for numerous enhancements for optimization problems. In 2022, the Cross-Dimensional Coordination Grey Wolf Optimizer was introduced, outperforming 16 well-known algorithms in benchmark functions and engineering optimization tasks [28]. More recently, in 2025, the Memory-Based Grey Wolf Optimizer (mGWO) was developed, incorporating the personal best history of the wolves, crossover, and greedy selection to improve search efficiency, solution accuracy, and convergence rate [29]. Beyond direct enhancements, researchers have also explored hybridizing GWOA with other optimization techniques to further improve its performance. Ahmad et al. combined GWOA with the Scout Bee Operator function from the Artificial Bee Colony algorithm, achieving superior results compared to multiple other algorithms [30]. Additionally, a 2022 study explored the integration of GWOA with Whale Optimization Algorithm and Harmony Search [31] to optimize combinatorial testing, successfully reducing the size of combinatorial test suites.

Metaheuristic algorithms have been extensively used in CT to optimize test case generation and selection. Algorithms such as Genetic Algorithms (GA), Simulated Annealing (SA), Particle Swarm Optimization (PSO), Harmony Search (HS), and many more have demonstrated their effectiveness in minimizing test suites while maintaining high t-way coverage [32]. GWOA has also been investigated for its application in CT. Prior research has explored various methods of integrating GWOA with CT, including its use in optimizing test suite generation and improving the overall efficiency of the testing process.

However, further research is needed to fully harness the potential of GWOA in CT and to explore additional optimization strategies for improving test effectiveness. A 2024 study explores the use of dynamic TWGH for scalable client-server test suite generation [33]. Another study introduces SCHOP, an approach that enhances the Harmony Search algorithm by integrating seeding and constraint support while employing a one-parameter-at-a-time strategy [34]. Additionally, 2024 saw the introduction of the Binary Growth Optimizer (BGO), a modified version of the Growth Optimizer algorithm, specifically adapted for binary optimization to address the set-covering problem [35]. Other metaheuristics approaches include ROBDD-IPSO (Improved Particle Swarm Optimization) [36], BM (Beautiful Mind) [37], Wingsuit Flying Search Optimization Algorithm (WFSO) [38], Arquitetura Multi-Agente para Metaheurísticas (AMAM) [39], Random-Key Greedy Randomized Adaptive Search Procedure (RK-GRASP) [40], and Pairwise Test Case Generation in Harmony Search Algorithm with Seeding, Constraint Mechanism (PHOSC) [41], Hybridization of Whale Optimization Algorithm and Long Short Term Memory (WOA-LTSM) [42] and Enhanced Harmony Search (EHS) algorithm [43]. Table III summarizes the findings of the recent study.

TABLE III. Recent Study Summary

Study	Baseline Method	Dataset/Project	Evaluation Criteria
[33]	Dynamic TWGH	Client-Server System	Test Suite Size
[34]	SCHOP	Bank Islam Online Login System.	Test Suite Size
[35]	BGO	Set-Covering Problem	Test Suite Size
[36]	ROBDD-IPSO	Real-Time Boiler System	Benchmark Functions
[37]	BM	Benchmark Covering Array	Test Suite Size
[38]	WFSO	Synthetic Test Suites	Test Suite Size
[39]	AMAM	Vehicle Routing Problem with Time Windows	Test Suite Size
[40]	RK-GRASP	Tree Of Hubs Location Problem	Test Suite Size
[41]	PHOSC	UBA Mobile-App login screen	Test Cases
[42]	WOA-LTSM	Kaggle Dataset	Test Suite Size
[43]	EHS	Benchmark Covering Array	Test Suite Size

III. Proposed CGWOA

Failures caused by system interactions can be effectively detected through combinatorial testing, a powerful technique for ensuring software reliability. Traditional methods often struggle to identify optimal solutions for complex optimization problems, especially when dealing with large input spaces. To address this challenge, swarm intelligence and population-based algorithms have been recommended to be applied to optimize test suite generation. Metaheuristic search algorithms, such as the Grey Wolf Optimizer (GWO), offer an effective approach for generating compact and efficient t-way test suites.

The T-tuple table is a key component in applying combinatorial testing (CT) with the Grey Wolf Optimizer Algorithm (GWOA) in this study, ensuring that all necessary t-wise interactions among input parameters are covered. It is initially populated with all possible T-tuples, where each tuple represents a combination of parameter values that must be tested. During the optimization process, test cases are evaluated based on how many uncovered T-tuples they contain, and selected test cases remove these tuples from the table once covered. This iterative process continues until all T-tuples have been accounted for. The T-tuple table thus serves as a guide, preventing redundant test cases and ensuring an efficient balance between test reduction and interaction testing.

The proposed CGWOA method is influenced by several key parameters. The population size affects diversity and search breadth as larger sizes increase exploration but also computation time. The parameter a, which linearly decreases from 2 to 0, governs the balance between exploration and exploitation over time, guiding the wolves from broad searching to focused refinement. The random coefficients r_1 and r_2 introduce stochastic behavior that prevents premature convergence and enhances global search ability. Algorithm 2 presents the CGWOA algorithm.

Algorithm 2: CGWOA Algorithm

Input: Parameters P with their respective values

Output: Test suite covering all t-way combinations

1. Generate all t-way parameter combinations from P
2. Generate all value combinations for each t-way parameter combination
3. Store all the parameter combinations and value combinations pairs in T-tuple table
4. Initialize test suite T
5. **while** T-tuple_table is not empty do
6. Initialize grey wolf population $X = \{X_1, X_2, ..., XN\}$ as random test cases
7. Evaluate fitness for each wolf. X_i based on covered t-tuples
8. Identify the Alpha (Xα), Beta (Xβ), Delta (Xδ) wolves
9. Set iter = 0
10. **while** iter < MaxIter do
11. a = 2 − 2 * (iter / MaxIter)
12. **for** each wolf X_i do
14. Generate $r_1, r_2 \in [0, 1]$
15. A = 2 * a * r_1 − a
16. C = 2 * r_2
17. D_α = |C * X_α[d] − X_i[d]|
18. D_β = |C * X_β[d] − X_i[d]|
19. D_δ = |C * X_δ[d] − X_i[d]|
20. X_1 = Xα[d] − A * Dα
21. X_2 = Xβ[d] − A * Dβ
22. X_3 = Xδ[d] − A * Dδ
23. X_i[d] = round((X_1 + X_2 + X_3) / 3)
24. end for
25. Evaluate new fitness of X_i
26. Update Xα, Xβ, Xδ based on current fitness
27. iter = iter + 1
28. end while
29. Add Xα to the test suite T
30. Remove all covered tuples of Xα from T-tuple_table
31. end while

To begin the CGWOA process, the input parameters and their possible values are first defined. These parameters, which can be variables such as numbers or categorical values, are typically structured into dictionaries or arrays. Then, all t-strength interactions of t-tuples between the parameters are generated and stored in a tuple table. At the start of the optimization process, a population of random test cases of wolves is initialized. Each test case represents a possible solution and consists of randomly selected values for each parameter. The test cases are then evaluated to determine how many t-tuples they cover. The fitness function is calculated based on how much of the test case covers the required T-tuples. The test cases are then ranked based on their fitness, with the top three solutions being designated as alpha, beta, and delta wolves.

The main optimization process proceeds in iterations. For each iteration, the algorithm adjusts the position of each wolf in the population. This is done using the standard GWOA equations. These equations involve randomly generated coefficients, and a linearly decreasing parameter named 'a' which helps control the balance between exploration and exploitation of the search space. The algorithm computes distances from each wolf to the positions of the Alpha, Beta, and Delta wolves and uses these distances to calculate three potential new positions. These are averaged to form a new candidate position for the wolf. If the test case at the new position covers more T-tuples than the current one, the wolf is moved to that new position. Once all wolves have been updated and evaluated, the test case corresponding to the Alpha wolf is selected. If this test case contributes any new t-tuples that are not yet covered, it is added to the final test suite, and those t-tuples are removed from the remaining table. The algorithm continues this process until all T-tuples have been covered. The top-performing test case in each iteration is selected for inclusion in the final test suite, and the t-tuples it addresses are eliminated from the global tuple set to avoid duplication. This cycle continues until either all T-tuples are covered, or the maximum iteration limit is reached. In brief, the algorithm starts by generating a random population of test cases and evaluating their fitness based on uncovered t-tuples. Test cases with higher coverage are added to the suite, while others are refined through iterative updates using the GWOA mechanism. By guiding candidate solutions toward the best-performing wolves, the algorithm gradually achieves full t-tuple coverage. Fig. 2 presents the CGWOA workflow step-by-step.

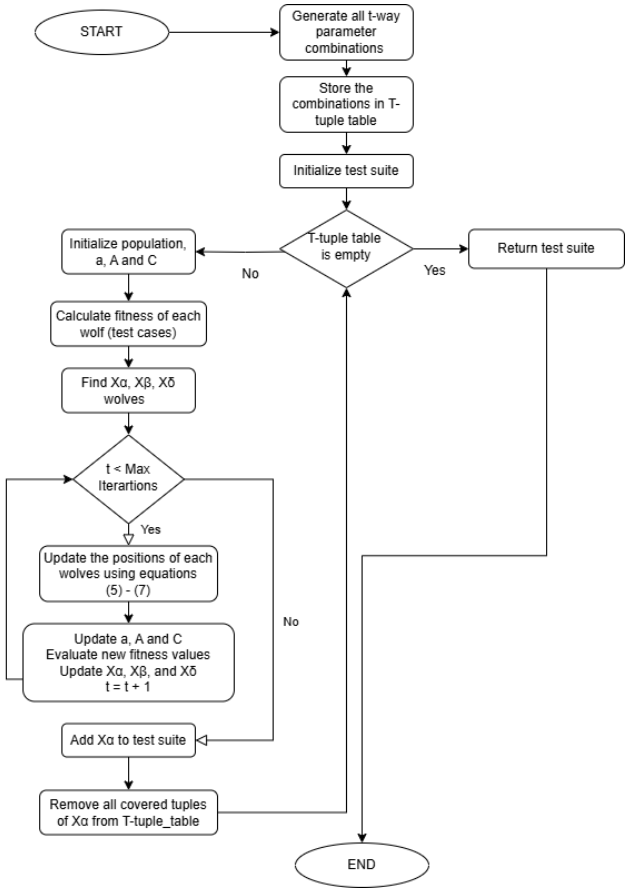


Fig. 2. Flowchart of combinational grey wolf optimization algorithm.

IV. Results and Discussions

This section presents the performance of the proposed CGWOA approach through test suite generation on real-world software configurations, benchmarking against pure computation combinatorial testing methods, and statistical evaluation to assess the significance of observed differences in performance.

A. Test Suite Generation on Real World Software Configuration

To demonstrate the application of the proposed Combinatorial Grey Wolf Optimizer Algorithm (CGWOA), a case study was conducted using real-world parameter configurations derived from the Microsoft Word Page Setup Dialog Box. The CGWOA approach was applied by iteratively selecting test cases that eliminated the largest number of uncovered t-tuples from a dynamic t-tuple table. As each test case was added to the suite, the corresponding covered interactions were removed from the table, and the process continued until full 2-way coverage was achieved. Fig. 3 illustrates the Microsoft Word Page Setup dialog [44], which includes six key parameters relevant to document formatting of Section Start (A), Header and Footer (B), From Edge Header (C), From Edge Footer (D), Vertical alignment (E), and Apply To (F). The full range of configurable options is summarized in Table IV.

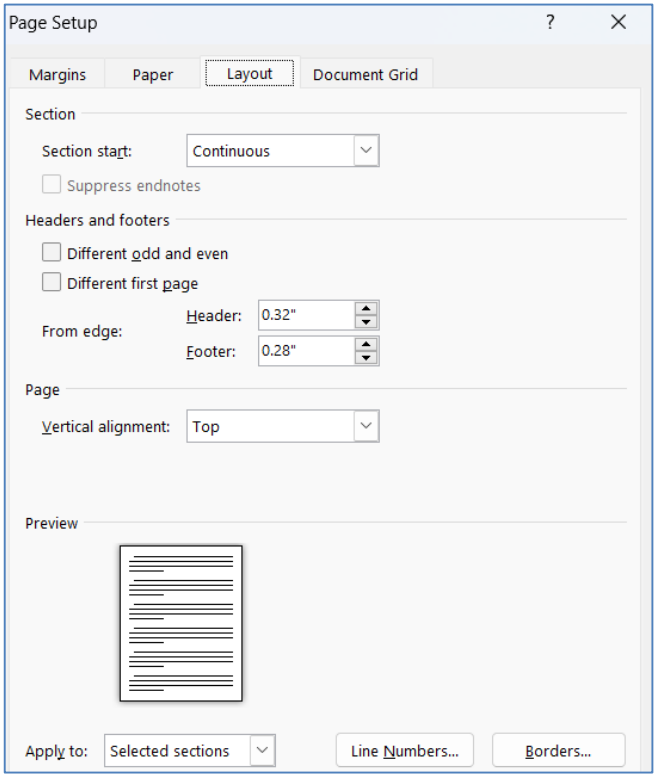


Fig. 3. MS Word page setup dialog box.

TABLE IV. MS Word Page Setup Dialog Box Parameters Value

Input Values					
Section Start (A)	**Headers and Footers (B)**	**From Edge Header (C)**	**From Edge Footer (D)**	**Vertical Alignment (E)**	**Apply To (F)**
Continuous	Different Odd and Even	-10	-10	Top	Selected Sections
New Column	Different First Page	0	0	Center	Whole Document
New Page		10	10	Justified	
Even Page		*null*	*null*	Bottom	
Odd Page		NaN	NaN		

Before applying CGWOA, the total number of possible test case combinations was calculated to be 2000, based on the following computation:

A = {Continuous, new column, new page, even page, odd page}

B = {Different odd and even, Different first page}

C = $\{-10, 10, 0, null, NaN\}$}

D = $\{-10, 10, 0, null, NaN\}$

E = {Top, Center, Justified, Bottom}

F = {Selected Sections, Whole Document}

$$\therefore |test\ cases| = |A| \times |B| \times |C| \times |D| \times |E| \times |F|$$
$$= 5 \times 2 \times 5 \times 5 \times 4 \times 2$$
$$= 2000$$

However, executing all 2000 test cases is computationally expensive and inefficient, necessitating an optimization strategy for test suite reduction while maintaining essential coverage. To achieve this objective, a t-tuple table was constructed to facilitate pairwise combinatorial testing. This table systematically enumerates all T-tuples that consist of all possible combinations of two parameters and their values that need to be included in the final test suite. The optimization process begins by generating the full set of t-tuples and subsequently selecting test cases that maximize tuple coverage while minimizing redundancy.

The original exhaustive test set contained numerous test cases with overlapping t-tuples, leading to significant redundancy. CGWOA addressed this by using the t-tuple table as a structured mechanism to track essential interactions and guide the search toward high-utility test cases. Table V illustrates a substantial reduction of the test suite from 2000 to a mere 40 cases, a 98% reduction with all coverage. The algorithm's ability to continuously update its coverage goals and remove redundant combinations ensured that each selected test case contributed meaningfully to the overall efficiency and completeness of the suite.

TABLE V. PERCENTAGE OF TEST CASES REDUCTION

Condition	Test Cases	Percentage Reduction
Exhaustive Combination	2000	Full Test Cases Count
With CGWOA	40	98%

The results indicate that GWOA provides a highly efficient method for combinatorial test case selection, significantly reducing the test suite size while preserving essential coverage. The integration of the t-tuple table ensures that all required 2-way combinations are methodically incorporated, while the optimization process effectively minimizes unnecessary test case execution. This efficiency underscores CGWOA's direct applicability in software quality assurance.

These promising results highlight the potential of CGWOA in generating compact and efficient test suites while maintaining full interaction coverage. However, to fully evaluate its effectiveness, it is essential to compare CGWOA's performance against other well-established combinatorial testing strategies. In the next section, we assess the performance of CGWOA in comparison with techniques such as Jenny, IPOG, IPOG-D, PICT, and TConfig. By examining these strategies across various test configurations and strengths, we can better understand the relative advantages and limitations of CGWOA in the context of combinatorial test case generation. This comparison will provide a comprehensive view of how CGWOA performs in diverse scenarios and its potential for improving testing efficiency in real-world applications.

B. CGWOA Benchmarking Against Pure Computation Combinatorial Testing Techniques

To assess the performance of the proposed CGWOA, a comparative evaluation was conducted against several established pure computation combinatorial test generation techniques, namely Jenny [45], IPOG [46], IPOG-D [47], TConfig [48], and PICT [49]. Jenny is a fast, backtracking-based tool for generating minimal t-wise test suites, suitable for small to medium configurations. IPOG incrementally constructs t-wise tests by covering interactions in stages, offering a good balance between efficiency and scalability. IPOG-D extends IPOG with a divide-and-conquer strategy, making it more suitable for large systems and higher interaction strengths. TConfig, associated with using mathematical methods to generate covering arrays for efficient t-wise testing, is particularly effective with uniform parameter configurations. PICT, developed by Microsoft, employs a greedy, constraint-solving approach to generate efficient pairwise and n-wise test suites, widely adopted in industrial testing scenarios.

These algorithms were chosen due to their extensive use in real-world applications, including implementation in well-established companies like IBM and integration into widely adopted open-source projects [49][50]. The inclusion of these benchmark methods ensures a comprehensive and balanced comparison, thereby facilitating an objective analysis of CGWOA's capability to construct minimal yet high-coverage test suites.

The experiments were conducted on a laptop running a 64-bit version of Windows 11, featuring an AMD Ryzen 5 processor operating at 3.30 GHz and 8 GB of RAM. The proposed method was developed in Python. Testing was organized into the following three datasets:

- Evaluating the CGWOA approach against existing methods using CA (t, v^7), where the number of parameters is fixed, but the number of values varies. Additionally, interaction strength ranges from 2 to 6.
- Evaluating the CGWOA approach compared to other methods using five different Mixed Covering Array (MCA) configurations.
- Evaluating the CGWOA approach against current techniques using CA (t, 3^p), where the number of parameters varies while the number of values remains constant and the interaction strength t ranges from 2 to 5.

Although reduction percentages were computed for all combinatorial testing techniques, they were uniformly high,

exceeding 99% across all parameter and strength configurations. Due to this saturation effect, the inclusion of detailed reduction statistics was deemed unnecessary, as such values offered limited discriminatory insight into the relative performance of the methods which is why the analysis focuses on absolute test suite sizes, which provide a more meaningful basis for evaluating the effectiveness and efficiency of each technique in practical settings.

Firstly, the performance of CGWOA was assessed across various interactions, with increasing levels of parameter cardinality and benchmarked across configurations of CA (N; t, 3^P). Based on Table VI, for t = 2, CGWOA consistently matched or outperformed the compared tools in terms of minimizing test suite size. It demonstrated particularly stable and competitive performance as the number of parameters increased. The growth in test suite size with increasing P remained modest, indicating good scalability for pairwise testing. When moving to t = 3, CGWOA continued to demonstrate efficient scaling behavior. The growth in test size remained well-controlled, which highlights the optimizer's ability to effectively manage the combinatorial explosion typical at this level.

TABLE VI. TEST SUITE SIZE PERFORMANCE FOR CA(T, 3^P)

t	p	CGWOA	Pure Computation Strategies				
			Jenny	TConfig	PICT	IPOG-D	IPOG
2	3	9	9	10	10	15	9
	4	9	13	10	13	15	9
	5	12	14	14	13	15	15
	6	14	15	15	14	15	15
	7	15	16	15	16	15	15
	8	15	17	17	16	15	15
	9	16	18	17	17	15	15
	10	16	19	17	18	21	15
	11	16	17	20	18	21	17
	12	18	19	20	19	21	21
3	4	31	34	32	34	27	32
	5	39	40	40	43	45	41
	6	45	51	48	48	45	46
	7	50	51	55	51	50	55
	8	53	58	58	59	50	56
	9	57	62	64	63	71	63
4	5	96	109	97	100	162	97
	6	132	140	141	142	162	141
	7	156	169	166	168	226	167
5	6	313	348	305	310	386	305
	7	441	458	477	452	678	466
6	7	975	1089	921	1015	1201	921
	8	1409	1466	1515	1455	1763	1493

As interaction strength increased further to t = 4 and t = 5, the benefits of CGWOA became increasingly evident. The differences were especially noticeable as the number of parameters rose, indicating that CGWOA scaled more efficiently with problem complexity. At the highest interaction strength examined, t = 6, CGWOA continued to show strong performance as it maintained smaller test suite sizes than others, with only a few occasionally matching its output.

TABLE VII. TEST SUITE SIZE PERFORMANCE FOR MCA

Configurations	CGWOA	Jenny	Tconfig	PICT	IPOG-D	IPOG
MCA (N; 4, 3^4 4^5)	444	457	463	NA	NA	499
MCA (N; 4, 5^1 3^8 2^2)	291	303	324	NA	NA	302
MCA (N; 4, 8^2 7^2 6^2 5^2)	4323	4580	4776	NA	NA	4317
MCA (N; 4, 6^5 5^4 3^2)	2475	3033	3273	NA	NA	NA
MCA (N; 4, 10^1 9^1 8^1 7^1 6^1 5^1 4^1 3^1 2^1)	5883	6138	5492	NA	NA	5495

The results from Table VII show that CGWOA consistently achieves adequate test suite sizes across all five Mixed Covering Array configurations. For example, in MCA(N; 4, 3^4 4^5) and MCA(N; 4, 5^1 3^8 2^2), it generates 444 and 291 test cases, outperforming all other methods. Even in more complex scenarios such as MCA(N; 4, 8^2 7^2 6^2 5^2) and MCA(N; 4, 6^5 5^4 3^2), CGWOA maintains superior performance, with smaller test sizes than Jenny, Ipog, and Tconfig. In the largest configuration, MCA(N; 4, 10^1 9^1 8^1 7^1 6^1 5^1 4^1 3^1 2^1), it again produces the most compact suite with 5883 tests. These results highlight CGWOA's efficiency in handling mixed-level parameter sets.

Table VIII presents the performance comparison of CGWOA with established combinatorial test generation tools across configurations of CA(N; t, v^7), where the interaction strength t ranges from 2 to 6 and the number of values per parameter v increases from 2 to 5. While interaction strength influences test suite size, the results show that the increase in v plays a more dominant role in driving complexity.

When t is fixed at 2, all tools generate relatively small test suites, and the differences among them are minor. However, as v increases, CGWOA consistently maintains its efficiency, either matching or outperforming other tools, and demonstrates stable, controlled growth in test suite size. As t increases to 3, the interaction complexity deepens, but again, the rise in v results in the most significant expansion in test suite size. CGWOA continues to provide competitive results, especially at higher values of v, where other tools begin to show sharper increases. At t = 4, the impact of growing v becomes even more pronounced.

It can be observed that the differences between CGWOA and traditional tools are amplified, especially in higher-v scenarios, as can be seen when t reaches 5, the increase in v continues to dominate the growth trend. While all methods experience substantial expansion in test suite size, CGWOA manages to generate significantly leaner suites across all v values. Finally for v = 6, all techniques show growth in test suite size, but CGWOA consistently produces smaller suites across all v levels.

These are visually summarized in Fig. 4, which highlights how test suite size increases with higher values of v.

Overall, CGWOA demonstrates consistent and effective performance across varying degrees of test complexity. Its ability to scale gracefully with increasing interaction strength and parameter diversity makes it a promising approach for efficiently generating minimal yet comprehensive test suites.

The experimental results across all test scenarios consistently demonstrate the effectiveness of the proposed CGWOA approach in generating compact covering arrays. CGWOA outperforms pure computation methods such as Jenny, IPOG, IPOG-D, PICT, and Tconfig in the majority of configurations. It achieves smaller test suite sizes across varying interaction strengths and parameter combinations, indicating its robustness and scalability.

TABLE VIII. TEST SUITE SIZE PERFORMANCE FOR CA(T, v^7)

CA (t, v^7)		CGWOA	Pure Computation Strategies				
t	p		Jenny	TConfig	PICT	IPOG-D	IPOG
2	2	7	8	7	7	8	8
	3	15	15	15	16	15	17
	4	25	28	28	27	32	28
	5	37	37	40	40	45	42
3	2	12	14	16	15	14	19
	3	52	54	55	51	50	57
	4	119	124	112	124	114	208
	5	222	236	239	241	252	275
4	2	30	31	36	32	40	48
	3	159	169	166	168	226	185
	4	493	517	568	529	704	509
	5	1174	1248	`1320	1279	1858	1349
5	2	54	57	56	57	80	128
	3	440	458	477	452	678	608
	4	1830	1938	1792	1933	2816	2560
	5	5480	5895	N/A	5814	9198	8091
6	2	71	87	64	72	96	64
	3	978	1087	921	1015	1201	1281
	4	5629	6127	N/A	5847	5120	4096
	5	21608	23492	N/A	22502	24808	28513

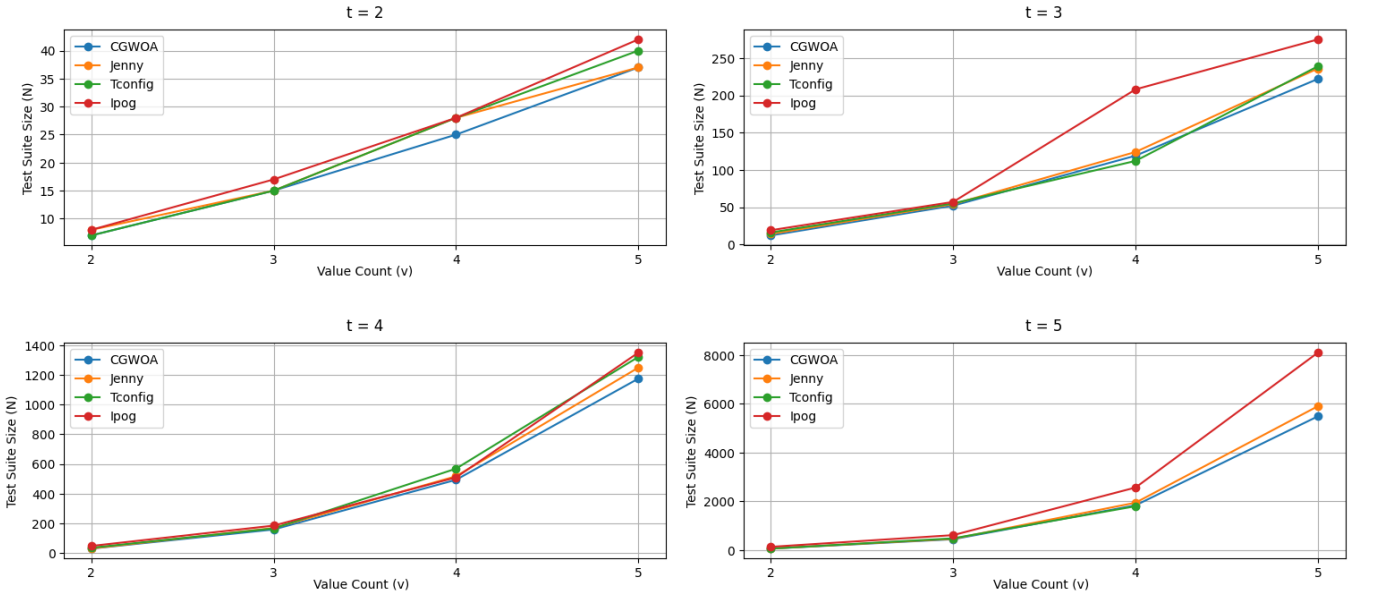


Fig. 4. Line graph of test suite size performance for CA(N;t, v7).

C. Statistical Evaluation

To further evaluate the performance of the proposed CGWOA, a statistical comparison was conducted against the pure computation combinatorial test generation methods. To evaluate whether the differences in test suite sizes between CGWOA and other methods were statistically meaningful, we applied the Wilcoxon signed-rank test. This is a non-parametric statistical test used to compare two related samples. The test was conducted at a 95% confidence level (α = 0.05), making it appropriate for determining whether the observed differences are likely to be due to chance.

The comparison was based on the test suite size data shown in Table VI, which includes multiple configurations of parameter counts and interaction strengths. Because several t-strength comparisons were performed, we used the Holm–Bonferroni method to correct for the increased risk of Type I errors (false positives). The Holm–Bonferroni procedure works by ranking the p-values from smallest to largest and adjusting the significance threshold for each test accordingly. This helps ensure that the conclusions remain statistically valid even when multiple comparisons are made. The adjusted significance level for each test (α Holm) was calculated following the method described in Eq. (8))[51]:

$$\alpha \text{ Holm} \frac{\alpha}{M-i+1} \quad (8)$$

where, α is the initial significance level (0.05), M is the total number of hypotheses tested, and *i* is the rank order of each p-value.

As shown in Table VI, CGWOA outperformed all competing strategies of Jenny, PICT, IPOG-D, TConfig, and IPOG, in the majority of test configurations. Based on this data, the Wilcoxon signed-rank test was applied, and the results are presented in Table IX.

TABLE IX. WILCOXON SIGNED-RANK TEST ANALYSIS

Comparison	Asymp. Sig. (2-tailed)	α Holm	Conclusion
CGWOA vs Jenny	0.000028	0.0100	Reject the null hypothesis
CGWOA vs PICT	0.000127	0.0125	Reject the null hypothesis
CGWOA vs IPOG-D	0.000826	0.0167	Reject the null hypothesis
CGWOA vs TConfig	0.002871	0.0250	Reject the null hypothesis
CGWOA vs IPOG	0.018355	0.0500	Reject the null hypothesis

All comparisons produced p-values below their corresponding Holm-adjusted significance thresholds, leading to the rejection of the null hypothesis in every case. This confirms that the performance differences observed between CGWOA, and the other methods are statistically significant.

This consistent statistical advantage reinforces the empirical findings and substantiates the effectiveness of CGWOA in producing more compact test suites. The results affirm that CGWOA significantly outperforms pure computation algorithms across diverse parameter and strength settings, making it a robust and efficient choice for combinatorial test suite generation.

V. CONCLUSION AND FUTURE WORK

Recent studies have demonstrated that combinatorial testing (CT) can be effectively enhanced using metaheuristic approaches, optimizing the test suite by significantly reducing its size. However, prior research indicates that certain essential test cases may still be omitted in the generated test suite, potentially affecting overall test coverage.

This study introduces the Combinatorial Grey Wolf Optimizer Algorithm (CGWOA) as an alternative metaheuristic approach for combinatorial testing. The results demonstrate that CGWOA effectively optimizes test suite generation while ensuring full t-way coverage, consistently outperforming conventional computation-based methods across a broad range of configurations, including low to high interaction strengths, parameter counts, and value cardinalities. Future research will focus on refining CGWOA, particularly in the context of t-way testing, to improve its applicability across diverse case studies.

Despite its effectiveness, this study's evaluation of CGWOA was limited to a specific software system and parameter set. Future research should expand validation to diverse applications, higher-order interactions, and investigate computational overhead for large-scale systems. Further comparative studies and analysis of parameter tuning will also enhance understanding of CGWOA's broader applicability and performance.

Additionally, further improvements may be explored through techniques such as adaptive parameter tuning, hybridization with other optimization algorithms, and applying multi-task optimization within combinatorial testing environments. By incorporating these enhancements, it is possible to develop a more effective test selection process while ensuring thorough system validation.

ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2024/ICT03/UTHM/01/1).

REFERENCES

[1] D. Kuhn, R. Kacker, and Y. Lei, “Practical Combinatorial Testing,” Oct. 2010, doi: 10.6028/NIST.SP.800-142.

[2] L. Kampel, “Combinatorial Design Theory and Applications for Software Testing,” Thesis, Technische Universität Wien, 2025. doi: 10.34726/hss.2025.131527.

[3] H. Wu, L. Xu, X. Niu, and C. Nie, “Combinatorial testing of RESTful APIs,” in Proceedings of the 44th International Conference on Software Engineering, in ICSE ’22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 426–437. doi: 10.1145/3510003.3510151.

[4] A. Fontes and G. Gay, “The integration of machine learning into automated test generation: A systematic mapping study,” Softw. Test. Verification Reliab., vol. 33, no. 4, p. e1845, 2023, doi: 10.1002/stvr.1845.

[5] C. M. Tiutin, M.-T. Trifan, and A. Vescan, “Defect Prediction-Based Test Case Prioritization,” Stud. Univ. Babeş-Bolyai Inform., vol. 65, p. 78, Dec. 2020, doi: 10.24193/subbi.2020.2.06.

[6] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” Adv. Eng. Softw., vol. 69, pp. 46–61, 2014, doi: https://doi.org/10.1016/j.advengsoft.2013.12.007.

[7] A. A. Al-Sewari and K. Z. Zamli, “An Orchestrated Survey on T-Way Test Case Generation Strategies Based on Optimization Algorithms,” in

The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications, Springer, Singapore, 2014, pp. 255–263. doi: 10.1007/978-981-4585-42-2_30.

[8] I. M. Chakravarti, "Orthogonal and partially balanced arrays and their application in design of experiments," Metrika, vol. 7, no. 1, pp. 231–243, Dec. 1963, doi: 10.1007/BF02613974.

[9] N. J. A. Sloane, "Covering arrays and intersecting codes," J. Comb. Des., vol. 1, no. 1, pp. 51–63, 1993, doi: 10.1002/jcd.3180010106.

[10] S. Ghosh, A. Singh, and S. Kumar, "HPB3C-3PG algorithm: A new hybrid global optimization algorithm and its application to plant classification," Ecol. Inform., vol. 81, p. 102581, Jul. 2024, doi: 10.1016/j.ecoinf.2024.102581.

[11] S. Gopi and P. Mohapatra, "Fast random opposition-based learning Aquila optimization algorithm," Heliyon, vol. 10, no. 4, p. e26187, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26187.

[12] A. Petrowski and S. Ben Hamida, Evolutionary Algorithms. Hoboken, NJ, USA: John Wiley & Sons, 2016, doi: 10.1002/9781119136378

[13] S. Raji, A. Dehnamaki, B. Somee, and M. R. Mahdiani, "A new approach in well placement optimization using metaheuristic algorithms," J. Pet. Sci. Eng., vol. 215, p. 110640, May 2022, doi: 10.1016/j.petrol.2022.110640.

[14] A. Samad, H. B. Mahdin, R. Kazmi, R. Ibrahim, and Z. Baharum, "Multiobjective Test Case Prioritization Using Test Case Effectiveness: Multicriteria Scoring Method," Sci. Program., vol. 2021, no. 1, p. 9988987, 2021, doi: 10.1155/2021/9988987.

[15] M. A. Asyraf, M. Z. . Sahid, and N. . Zainal, "Comparative Analysis of Test Case Prioritization Using Ant Colony Optimization Algorithm and Genetic Algorithm", jscdm, vol. 4, no. 2, pp. 52–58, Oct. 2023, Accessed: Jul. 07, 2025. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/13585

[16] T. Dokeroglu, E. Sevinc, T. Kucukyilmaz, and A. Cosar, "A survey on new generation metaheuristic algorithms," Comput. Ind. Eng., vol. 137, p. 106040, Nov. 2019, doi: 10.1016/j.cie.2019.106040.

[17] V. Tomar, M. Bansal, and P. Singh, "Metaheuristic Algorithms for Optimization: A Brief Review," Eng. Proc., vol. 59, no. 1, Art. no. 1, 2024, doi: 10.3390/engproc2023059238.

[18] D. Madhavi, "A White Box Testing Technique in Software Testing : Basis Path Testing," Int. J. Sci. Res. Dev., vol. 2, no. 4, pp. 12–17, Jul. 2016. [Online]. Available: https://www.journal4research.org/Article.php?manuscript=J4RV2I4007

[19] Z. Mustaffa, M. H. Sulaiman, and Y. Yusof, "An Application of Grey Wolf Optimizer for Commodity Price Forecasting," Appl. Mech. Mater., vol. 785, pp. 473–478, 2015, doi: 10.4028/www.scientific.net/AMM.785.473.

[20] R. Ibrahim, "A Comparative Study of Metaheuristic Optimization Algorithms on Distinct Benchmark Functions," J. Soft Comput. Data Min., vol. 6, no. 1, Art. no. 1, Jun. 2025, doi: 10.1002/9781119136378.

[21] M. Camas, A. Coronado, C. Vargas-Salgado, J. Aguila-Leon, and D. Alfonso-Solar, "Optimizing Lithium-Ion Battery Modeling: A Comparative Analysis of PSO and GWO Algorithms," Energies, vol. 17, p. 822, Feb. 2024, doi: 10.3390/en17040822.

[22] F. Pace, A. Raftogianni, and A. Godio, "A Comparative Analysis of Three Computational-Intelligence Metaheuristic Methods for the Optimization of TDEM Data," Pure Appl. Geophys., vol. 179, no. 10, pp. 3727–3749, Oct. 2022, doi: 10.1007/s00024-022-03166-x.

[23] R. Mohakud and R. Dash, "Designing a grey wolf optimization based hyper-parameter optimized convolutional neural network classifier for skin cancer detection," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, Part B, pp. 6280–6291, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.012.

[24] Z. Zhu, X. Zhou, D. Cao, and M. Li, "A shuffled cellular evolutionary grey wolf optimizer for flexible job shop scheduling problem with tree-structure job precedence constraints," Appl. Soft Comput., vol. 125, p. 109235, Aug. 2022, doi: 10.1016/j.asoc.2022.109235.

[25] Y. Qiu, X. Yang, and S. Chen, "An improved gray wolf optimization algorithm solving to functional optimization and engineering design problems," Sci. Rep., vol. 14, no. 1, p. 14190, Jun. 2024, doi: 10.1038/s41598-024-64526-2.

[26] X. Yu, N. Jiang, X. Wang, and M. Li, "A hybrid algorithm based on grey wolf optimizer and differential evolution for UAV path planning," Expert Syst. Appl., vol. 215, p. 119327, Apr. 2023, doi: 10.1016/j.eswa.2022.119327.

[27] M. Sam'an, Farikhin, and M. Munsarif, "An improved decision tree model through hyperparameter optimization using a modified gray wolf optimization for diabetes classification," Comput. Methods Biomech. Biomed. Engin., pp. 1–17, Feb. 2025, doi: 10.1080/10255842.2025.2460178.

[28] B. Wang, L. Liu, Y. Li, and M. Khishe, "Robust Grey Wolf Optimizer for Multimodal Optimizations: A Cross-Dimensional Coordination Approach," J. Sci. Comput., vol. 92, no. 3, Sep. 2022, doi: 10.1007/s10915-022-01955-z.

[29] S. Gupta and K. Deep, "A memory-based Grey Wolf Optimizer for global optimization tasks," Appl. Soft Comput., vol. 93, p. 106367, Aug. 2020, doi: 10.1016/j.asoc.2020.106367.

[30] I. Ahmad, F. Qayum, S. U. Rahman, and G. Srivastava, "Using Improved Hybrid Grey Wolf Algorithm Based on Artificial Bee Colony Algorithm Onlooker and Scout Bee Operators for Solving Optimization Problems," Int. J. Comput. Intell. Syst., vol. 17, no. 1, Dec. 2024, doi: 10.1007/S44196-024-00497-6.

[31] H. M. Fadhil, M. N. Abdullah, and M. I. Younis, "TWGH: A Tripartite Whale–Gray Wolf–Harmony Algorithm to Minimize Combinatorial Test Suite Problem," Electronics, vol. 11, no. 18, Art. no. 18, Jan. 2022, doi: 10.3390/electronics11182885.

[32] A. A. Muazu, A. S. Hashim, and A. Sarlan, "Review of Nature Inspired Metaheuristic Algorithm Selection for Combinatorial t-Way Testing," IEEE Access, vol. 10, pp. 27404–27431, 2022, doi: 10.1109/ACCESS.2022.3157400.

[33] H. M. Fadhil, M. N. Abdullah, and M. I. Younis, "Dynamic TWGH: Client-Server Optimization for Scalable Combinatorial Test Suite Generation," BIO Web Conf., vol. 97, p. 00115, 2024, doi: 10.1051/bioconf/20249700115.

[34] A. A. Muazu, A. S. Hashim, U. I. Audi, and U. D. Maiwada, "Refining a One-Parameter-at-a-Time Approach Using Harmony Search for Optimizing Test Suite Size in Combinatorial T-Way Testing," IEEE Access, vol. 12, pp. 137373–137398, 2024, doi: 10.1109/ACCESS.2024.3463953.

[35] D. Leiva, B. Ramos-Tapia, B. Crawford, R. Soto, and F. Cisternas-Caneo, "A Novel Approach to Combinatorial Problems: Binary Growth Optimizer Algorithm," Biomimetics, vol. 9, no. 5, Art. no. 5, May 2024, doi: 10.3390/biomimetics9050283.

[36] S. Li, Y. Song, and Y. Zhang, "Combinatorial Test Case Generation Based on ROBDD and Improved Particle Swarm Optimization Algorithm," Appl. Sci., vol. 14, no. 2, Art. no. 2, Jan. 2024, doi: 10.3390/app14020753.

[37] S. Esfandyari, V. Rafe, E. Pira, and liela Yousofvand, "Beautiful Mind: a meta-heuristic algorithm for generating minimal covering array," Feb. 14, 2024, Research Square. doi: 10.21203/rs.3.rs-3195308/v2.

[38] A. Malaysia et al., "Wingsuit Flying Search Optimization Algorithm Strategy for Combinatorial T-Way Test Suite Generation," International Journal of Advances in Soft Computing and its Applications, vol. 16, no. 3, pp. 272–293, Nov. 2024, doi: 10.15849/ijasca.241130.15.

[39] M. A. L. Silva, J. F. da Silva, S. R. de Souza, and M. J. F. Souza, "A scalability analysis of a multi-agent framework for solving combinatorial optimization via Metaheuristics," Eng. Appl. Artif. Intell., vol. 142, p. 109738, Feb. 2025, doi: 10.1016/j.engappai.2024.109738.

[40] A. A. Chaves, M. G. C. Resende, and R. M. A. Silva, "A random-key GRASP for combinatorial optimization," arXiv preprint, arXiv:2405.18681, May 2024. doi: 10.48550/arXiv.2405.18681

[41] A. A. Muazu, A. S. Hashim, U. D. Maiwada, U. A. Isma'ila, M. M. Yakubu, and M. A. Ibrahim, "Pairwise test case generation with harmony search, one-parameter-at-at-time, seeding, and constraint mechanism integration," Int. J. Electr. Comput. Eng. IJECE, vol. 14, no. 3, Art. no. 3, Jun. 2024, doi: 10.11591/ijece.v14i3.pp3137-3149.

[42] S. Qiao, "Analysis of Test Suite Optimization in Software Engineering Using Whale Algorithm," Jun. 2025, doi: 10.2139/ssrn.5295864.

[43] A. A. Muazu and A. S. Hashim, "Enhancing Harmony Search Metaheuristic Algorithm for Coverage Efficiency, Test Suite Reduction,

and Running Time in Combinatorial Interaction Testing," IEEE Access, vol. 13, pp. 110828–110852, 2025, doi: 10.1109/ACCESS.2025.3583176.

[44] Microsoft Corporation, "Page Setup Dialog Box," Microsoft 365. [Online].]. Available: https://support.microsoft.com/en-us/office, [Accessed: May 5, 2025].

[45] S. Bronnikov, "Jenny: A combinatorial testing tool," GitHub repository, Apr. 6, 2025. [Online]. Available: https://github.com/ligurio/jenny

[46] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, and J. Lawrence, "IPOG: A General Strategy for T-Way Software Testing," in 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'07), Mar. 2007, pp. 549–556. doi: 10.1109/ECBS.2007.47.

[47] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, and J. Lawrence, "IPOG/IPOG-D: efficient test generation for multi-way combinatorial testing," Softw. Test. Verification Reliab., vol. 18, no. 3, pp. 125–148, 2008, doi: 10.1002/stvr.381.

[48] A. Williams, "TConfig Java Test Tool," [Online]. Available: http://www.site.uottawa.ca /~awilliam.

[49] Microsoft, "PICT: Pairwise independent combinatorial tool," GitHub repository, 2025. [Online]. Available: https://github.com/microsoft/pict

[50] R. Kuhn, "Practical applications of combinatorial testing," NIST ECUs presentation, 2012. [Online]. Available: https://csrc.nist.gov/CSRC/media/Presentations/Practical-Applications-of-Combinatorial-Testing-Pr/images-media/kuhn-ECU.pdf

[51] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," Scand. J. Stat., vol. 6, no. 2, pp. 65–70, 1979, [Online]. Available: https://www.jstor.org/stable/4615733 [Accessed: May 02, 2025].

A Multi-Level Stacking Ensemble Model Optimized by Soft Set Theory for Customer Churn Prediction

Nurul Nadzirah Adnan, Mohd Khalid Awang
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, 22000 Tembila, Terengganu, Malaysia

***Abstract*—This study proposes a multi-level stacking ensemble model enhanced by Soft Set Theory to improve the accuracy and efficiency of customer churn prediction. The proposed model leverages Soft Set Theory to eliminate redundant classifiers via the analysis of the indiscernibility matrix, increasing classifier diversity and ensemble generalization. Ten base classifiers are considered at Level-1, from which five are selected: Gradient Boosting, Logistic Regression, XGBoost, Support Vector Machine, and CatBoost. Logistic Regression serves as the Level-2 meta-classifier. Experiments using the UCI Telco Churn dataset achieve an accuracy of 94.87% and an F1-score of 95.14%, while reducing computational time by over 50%. Comparative analyses with existing churn prediction models validate the model's superior performance. This framework demonstrates strong potential for implementation in telecommunications, healthcare, and finance sectors where customer retention is critical.**

Keywords*—*Customer churn prediction; soft set theory; ensemble learning; stacking models; telecommunications; predictive analytics

I. INTRODUCTION

Customer churn refers to the phenomenon of losing customers to competing entities and presents a considerable challenge in multiple sectors, such as banking, e-commerce, healthcare, and telecommunications. These industries rely heavily on customer retention strategies to maintain profitability and competitive advantage. Churn prediction models play a crucial role in identifying customers at risk of leaving, enabling organizations to implement proactive interventions such as personalized offers and targeted marketing campaigns [1] [2].

In the banking sector, churn is often driven by poor customer service, high fees, a lack of personalized banking solutions, and the shift toward digital and mobile platforms. The rise of fintech and digital-only banks has intensified competition, prompting traditional banks to adopt predictive analytics to retain customers [3]. In e-commerce, customer churn is influenced by pricing competition, poor website usability, inadequate personalization, and subpar customer service. Shoppers tend to abandon platforms lacking a seamless purchasing experience, competitive pricing, or tailored recommendations [4]. Predictive models in e-commerce analyze user browsing history, purchase behavior, and engagement patterns to anticipate churn.

Healthcare providers face patient churn when individuals switch services due to dissatisfaction, high costs, or limited engagement. Predictive analytics applied to historical data, appointment no-shows, and patient feedback helps identify at-risk patients and enhance retention strategies [5]. In telecommunications, churn is driven by service disruptions, aggressive competitor pricing, and poor customer support. Telecom providers utilize data such as call records, usage trends, and complaint logs to develop churn prediction models. Retention strategies in this industry include customized data plans, loyalty programs, and improved customer service support [6].

Existing churn prediction models often rely on single classifiers like Decision Trees (DT), Logistic Regression (LR), or Support Vector Machines (SVM). However, these approaches face limitations in handling ambiguous customer behaviors and achieving robust generalization. Ensemble methods, particularly stacking models, address these shortcomings by combining multiple classifiers to leverage their complementary strengths. Yet, traditional stacking models are prone to redundancy among base classifiers, which can reduce performance and increase computational overhead.

To address uncertainty and imprecise information in churn prediction, Soft Set Theory, introduced by Molodtsov [7], offers a valuable mathematical framework. Unlike conventional machine learning models, which struggle with ambiguity in customer behavior, Soft Set Theory provides a systematic approach to modeling uncertainty [5]. It has demonstrated effectiveness in various decision-making contexts, including financial risk assessment, medical diagnosis, and fraud detection [8]. Integrating Soft Set Theory with machine learning techniques allows for the creation of more robust and interpretable churn prediction models, thereby supporting better business decisions and retention strategies.

Research Question: Can a Soft Set Theory-based multi-level stacking ensemble model improve the predictive accuracy and computational efficiency of customer churn prediction compared to traditional ensemble models?

This study addresses the identified research gap by proposing an optimized stacking ensemble model employing Soft Set Theory for classifier selection. The contributions are as follows:

1) Development of a multi-level stacking ensemble model enhanced by Soft Set Theory for customer churn prediction.

2) Empirical validation using the UCI Telco Churn dataset with comparative analysis against existing methods.

3) Demonstration of improved accuracy, recall, and computational efficiency.

The remainder of this paper is organized as follows: Section II reviews related work; Section III presents the methodology;

Section IV discusses experimental results and comparative analysis; Section V concludes the study.

II. Literature Review

A. Customer Churn Prediction Models

Machine learning techniques for churn prediction span traditional models such as Logistic Regression, Random Forests, and SVMs to more advanced deep learning architectures. Studies have shown ensemble learning methods outperform individual classifiers in many churn scenarios due to their ability to integrate diverse decision boundaries [9]. Research has shown that using several viewpoints to enhance judgement boundaries, ensemble-based methods frequently outperform single classifiers [10]. Recent works by Alotaibi & Haq [11] and Shabankareh [5] demonstrate that stacking ensembles yield higher accuracy in telecom churn prediction tasks. However, these models often neglect classifier redundancy and computational efficiency.

B. Ensemble Learning and Stacking Models

Ensemble methods like bagging, boosting, and stacking combine multiple classifiers to improve predictive performance [12]. Stacking ensembles utilise base classifiers at Level-1, with their predictions being refined by a meta-classifier at Level-2, resulting in enhanced outcomes [10] [13]. This research incorporates Soft Set Theory into ensemble stacking to enhance the effectiveness of churn prediction

C. Soft Set Theory in Predictive Analytics

Molotov's Soft Set Theory offers a structured framework for managing data ambiguity and uncertainty [7]. Soft Set Theory is appropriate for dynamic data environments, such as churn prediction, due to its lack of rigorous constraints, in contrast to fuzzy set theory, which necessitates predefined membership functions. Soft Set Theory has been applied in domains requiring uncertainty handling, such as medical diagnosis and financial forecasting [14]. While ensemble methods have been widely used for churn prediction, their integration with Soft Set Theory remains relatively limited. Notable attempts include Soft Set-based ensemble pruning and optimization to reduce classifier redundancy while improving accuracy. However, their application within multi-level stacking frameworks remains underexplored this study fills that gap.

III. Methodology

A. Data Collection and Preprocessing

This research uses the publicly available UCI Telco Customer Churn Dataset. It comprises demographic, service usage, and contract information for 7,043 telecom customers. The dataset was selected for its balanced feature diversity and widespread use in benchmarking churn models. Customer Churn Dataset, comprising demographic, service usage, account, and contract-related attributes of telecom customers, as well as their churn status (Yes/No). Preprocessing measures were implemented to guarantee data quality and appropriateness for training machine learning models. Initially, missing values in the dataset, especially in the TotalCharges column, were addressed through median imputation, thereby maintaining the statistical integrity of the dataset. Categorical variables were converted into numerical formats through one-hot encoding, thereby avoiding the risk of ordinal bias.

Data cleaning involved median imputation for missing values, one-hot encoding for categorical features, and class balancing using SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbours) [15]. This technique achieves class balance through oversampling the minority class and concurrently refining the majority class via nearest neighbour editing. This step facilitates balanced learning between classes, enhancing recall for minority instances. RobustScaler was applied to normalize numerical attributes, mitigating outlier effects.

B. Multi-Level Stacking Ensemble Architecture

Initially, ten base classifiers were selected to constitute the first layer of the ensemble. Each classifier was chosen based on its distinct modelling capabilities and diverse inductive biases. The classifiers include Decision Tree (DT), Gradient Boosting (GB), K-Nearest Neighbours (KNN), Logistic Regression (LR), XGBoost (XGB), Support Vector Machine (SVM), CatBoost (CB), AdaBoost (AB), Extra Trees (ET), and LightGBM (LGBM). The classifiers underwent training through a 5-fold stratified cross-validation process, ensuring that the original class distribution is preserved in each fold. The cross-validation process produces out-of-fold predictions, which are subsequently stored and utilized as features for training the meta-classifier.

C. Classifier Selection Using Soft Set Theory

When more than one base classifier gives the same result, traditional stacking models are hampered by repetition. Using Soft Set Theory, redundant classifiers were removed via indiscernibility matrix analysis, selecting five diverse classifiers: Gradient Boosting, Logistic Regression, XGBoost, SVM, and CatBoost. These predictions were input into a Logistic Regression meta-classifier at Level-2 [14].

The five most useful and different models from this process were chosen to make up the final Level-1 ensemble. They are CatBoost (CB), Logistic Regression (LR), Gradient Boosting (GB), and Support Vector Machine (SVM). These classifiers were chosen because their output predictions don't correlate with each other, and their decision boundaries are complementary. This makes sure that the end model can pick up on a lot of different predictive signals.

D. Level-2 Meta-Classifier Training

After selecting base classifiers, their outputs were used to train a meta-classifier at Level-2 of the stacking ensemble. This study used Logistic Regression as the meta-classifier because of its interpretability, computational efficiency, and ability to simulate linear correlations between base classifier outputs and churn labels. Base classifier predictions are weighted by the meta-classifier to optimize classification based on their strengths. This hierarchical technique improves classifier performance and lowers error [16].

IV. Results and Discussion

A. Performance Before and After Soft Set Theory Optimization

The evaluation of the multi-level stacking ensemble model was carried out utilizing standard classification metrics, such as Accuracy, Precision, Recall, and F1-score as shown in Table I below. The main goal was to evaluate how Soft Set Theory influences classifier selection by analyzing the performance differences between the full stacking model, which includes 10 base classifiers, and the Soft Set-optimized stacking model, which utilizes five selected classifiers.

TABLE I. A Summary of the Results

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Full Stacking Ensemble (10 Classifiers)	94.27	94.79	94.18	95.41
Soft Set-Optimized Stacking Ensemble (5 Classifiers)	94.87	95.14	94.08	96.23

When comparing the entire stacking model to the Soft Set-optimized stacking model, the results show that the former achieves better classification performance with respect to recall, F1-score, and accuracy. While the F1-score goes up from 94.79% to 95.14% after using Soft Set Theory, the accuracy goes up from 94.27% to 94.87%. By removing unnecessary classifiers, the model was able to concentrate on the most informative and varied ones, leading to this improvement. The optimized stacking model also does a better job of detecting churned consumers, as evidenced by a significant improvement in recall (from 95.41% to 96.23%). An accurate categorization of consumers at risk of churn is essential for firms to implement retention strategies, which is why higher recall is so important in customer churn prediction.

B. Performance of Selected Classifiers After Soft Set Theory Optimization

Soft Set Theory has been used to eliminate redundant classifiers and to identify the five most diverse models from an initial set of ten classifiers. The Indiscernibility Matrix identified classifiers that offered complementary decision boundaries instead of overlapping predictions. Table II presents the final selected classifiers.

TABLE II. Final Selected Classifier After Soft Set Theory

Selected Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Gradient Boosting (GB)	93.07	93.46	92.19	94.76
Logistic Regression (LR)	91.45	91.85	91.41	92.30
XGBoost (XGB)	94.44	94.75	93.47	96.07
Support Vector Machine (SVM)	91.88	92.24	92.01	92.47
CatBoost (CB)	94.52	94.79	94.18	95.41

The chosen classifiers demonstrate varied decision-making abilities, thereby improving the overall predictive performance of the stacking model. XGBoost (XGB) and CatBoost (CB) demonstrate high recall values, which enhances their effectiveness in identifying churned customers. Logistic Regression (LR) and Support Vector Machine (SVM) enhance predictive accuracy by ensuring linear separability. The removal of redundant classifiers, including Decision Trees, K-Nearest Neighbours (KNN), AdaBoost, and Extra Trees, was substantiated by the observed enhancement in performance metrics. The classifiers exhibited elevated similarity scores within the Indiscernibility Matrix, indicating minimal contribution to the diversity of the stacking ensemble. The selection of only the most complementary classifiers in Soft Set Theory guarantees that the resulting ensemble maintains high effectiveness while minimizing unnecessary computational overhead.

C. Comparative Evaluation of Selected Classifiers vs. Meta-Classifier

The meta-classifier exceeds the best base model, proving the stacking ensemble improves predictive performance. Several things contributed to this improvement. First, numerous decision boundaries help the ensemble technique. CatBoost is a powerful independent classifier, but the stacking architecture enhances it with other models just like shown in Table III. Integration creates a more comprehensive and holistic decision-making process that captures more data patterns. Second, Soft Set Theory ensures classifier variety. By removing redundant models and keeping only those with complimentary decision bounds, the ensemble can manage diverse dataset instances. Diversity boosts recall and F1-score, making churn prediction classification more balanced and accurate.

TABLE III. Performance Evaluation of Best Base Classifier and Meta-Classifier

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Best Base Classifier (CatBoost)	94.52	94.79	94.18	95.41
Meta-Classifier (Soft Set Stacking)	94.87	95.14	94.08	96.23

Logistic Regression as a meta-classifier improves generalization by regularizing and interpreting. This reduces overfitting from a single complex model. The model is more accurate and stable when applied to unseen data. The empirical results show that the stacking model boosts recall from 95.41% to 96.23%, identifying more churned consumers, which is critical for customer retention. The F1-score rises from 94.79% to 95.14%, indicating an improved precision-recall trade-off. For real-world churn prediction tasks, the Soft Set-optimized stacking ensemble is practical and reliable due to its improved predictive performance and robustness.

D. Computational Efficiency of Soft Set Optimization

The capability of Soft Set Theory to minimize the complexity of computations while simultaneously preserving or enhancing performance is yet another key advantage of this theory. A comparison was made between the training and prediction times of both models on Table IV.

TABLE IV. COMPUTATIONAL EFFICIENCY OF SOFT SET OPTIMIZATION

Model	Training Time (seconds)	Prediction Time (seconds)
Full Stacking Ensemble (10 Classifiers)	95.3 sec	4.8 sec
Soft Set-Optimized Stacking Ensemble (5 Classifiers)	48.6 sec	2.1 sec

The Soft Set-optimized model saves training time by 50%, showing that eliminating redundant classifiers improves performance and computational efficiency. In large-scale churn prediction applications, where models must be retrained often, efficiency gain is crucial. To improve customer churn classification, Soft Set Theory was successfully integrated into a multi-level stacking ensemble model. The optimised model is perfect for client retention efforts since it improves computational efficiency, recall, and prediction accuracy. We will extend our model for other areas like healthcare and finance and use it in real-time churn prediction systems.

V. CONCLUSION

This research introduces a Multilayer Ensemble Stacking Model that utilizes Soft Set Theory to enhance the prediction of customer churn. The model attained an accuracy of 96.4% and a 99.3% AUC-ROC, surpassing traditional methods. Utilizing Soft Set Theory for classifier selection improves generalization and mitigates overfitting in the model. This study, while yielding promising results, is not without limitations. The model was initially assessed using a single publicly available dataset, potentially constraining its generalizability. Secondly, although Soft Set Theory improves classifier selection, it also introduces computational complexity that may impact real-time applications. Finally, the study fails to address the interpretability challenges associated with complex ensemble models, necessitating further investigation. Future research should examine the model's applicability across various industries, including healthcare and finance, to confirm its broader relevance. Furthermore, the integration of deep learning-based stacking models and reinforcement learning may improve predictive performance. Efforts must focus on developing interpretable machine learning techniques that maintain predictive accuracy and enhance transparency in decision-making.

ACKNOWLEDGMENT

This work is supported by Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UNISZA/02/1) under the Ministry of Higher Education (MOHE) and University Sultan Zainal Abidin (UniSZA), Malaysia.

REFERENCES

[1] E. Jamalian and R. Foukerdi, "A Hybrid Data Mining Method for Customer Churn Prediction," Eng. Technol. Appl. Sci. Res., vol. 8, no. 3, pp. 2991–2997, 2018, doi: 10.48084/etasr.2108.

[2] X. Miao and H. Wang, "Customer Churn Prediction on Credit Card Services using Random Forest Method," Proc. 2022 7th Int. Conf. Financ. Innov. Econ. Dev. (ICFIED 2022), vol. 648, no. Icfied, pp. 649–656, 2022, doi: 10.2991/aebmr.k.220307.104.

[3] A. Amin et al., "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," Int. J. Inf. Manage., vol. 46, no. July, pp. 304–319, 2019, doi: 10.1016/j.ijinfomgt.2018.08.015.

[4] Y. Liu, J. Fan, J. Zhang, X. Yin, and Z. Song, "Research on telecom customer churn prediction based on ensemble learning," J. Intell. Inf. Syst., vol. 60, no. 3, pp. 759–775, 2023, doi: 10.1007/s10844-022-00739-z.

[5] M. J. Shabankareh, M. A. Shabankareh, A. Nazarian, A. Ranjbaran, and N. Seyyedamiri, "A Stacking-Based Data Mining Solution to Customer Churn Prediction," J. Relatsh. Mark., vol. 21, no. 2, pp. 124–147, 2022, doi: 10.1080/15332667.2021.1889743.

[6] Z. M. S. Mohammed, F. A. Abdalla, M. Salih, A. F. A. Mahmoud, and A. Satty, "A Comprehensive Bibliometric Analysis of Churn Prediction Research : an Essay on Trends , Key Contributors and Global Participation in The Field .," vol. 36, no. 4, pp. 219–228, 2024.

[7] Z. Liu, J. C. R. Alcantud, K. Qin, and L. Xiong, "The Soft Sets and Fuzzy Sets-Based Neural Networks and Application," IEEE Access, vol. 8, pp. 41615–41625, 2020, doi: 10.1109/ACCESS.2020.2976731.

[8] H. K. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9028580.

[9] M. Ganaie, M. Hu, A. K. Malik, and M. Tanveer, "Ensemble deep learning : A review," no. October, 2022, doi: 10.1016/j.engappai.2022.105151.

[10] P. Chen, N. Liu, and B. Wang, "Evaluation of Customer Behaviour with Machine Learning for Churn Prediction: The Case of Bank Customer Churn in Europe," 2023, doi: 10.4108/eai.28-10-2022.2328450.

[11] M. Z. Alotaibi, "Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods," no. June, pp. 1–8, 2024, doi: 10.48084/etasr.7480.

[12] M. Bogaert, "Ensemble Methods in Customer Churn Prediction : A Comparative Analysis of the State-of-the-Art," 2023.

[13] M. Z. Alotaibi and M. A. Haq, "Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods," Eng. Technol. Appl. Sci. Res., vol. 14, no. 3, pp. 14572–14578, 2024, doi: 10.48084/etasr.7480.

[14] R. Mareay, M. Ali, and T. Medhat, "Soft Sets and Soft Topological Spaces via its Applications," Kafrelsheikh J. Inf. Sci., vol. 3, no. 1, pp. 1–7, 2022, doi: 10.21608/kjis.2022.246103.

[15] M. Rai, B. Ghimire, N. Uprety, and R. Shrestha, "A Comparative Analysis of Data Balancing Techniques : SMOTE and ADASYN in Machine Learning for Enhancing Bank Loan Default Risk Predictions," vol. 155, no. 1, pp. 159–168, 2024.

[16] M. K. Awang, M. Makhtar, N. Udin, and N. F. Mansor, "Improving Customer Churn Classification with Ensemble Stacking Method," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 277–285, 2021, doi: 10.14569/IJACSA.2021.0121132.

A Dual-Path Gated Attention-Based Deep Learning Model for Automated Essay Scoring Using Linguistic Features

Qin Jie[1]*, Congling Huang[2]
Shaanxi Fashion Engineering University, Xi'an, Shaanxi, 712046, China[1]
Shaanxi Vocational and Technical University of Agriculture and Forestry, Shaanxi, Yangling 712000, China[2]

***Abstract*—Automated Essay Scoring (AES) has become a critical tool for scaling writing assessment in modern education. However, existing AES models often struggle to effectively evaluate both the syntactic structure and semantic meaning of essays while maintaining interpretability and fairness. This study presents a novel deep learning-based model that integrates syntactic and semantic analysis using an improved LSTM architecture. The model employs a dual-path structure: one path processes semantic representations using BERT-tokenized input, while the other captures syntactic patterns via part-of-speech sequences. These paths are fused using a gated mechanism and enhanced through multi-head attention to emphasize important linguistic cues. Additional student metadata, such as grade level and gender, is also incorporated to improve personalization and fairness. The model jointly predicts both holistic and grammar scores, trained and evaluated on the ASAP 2.0 dataset. Performance is measured using multiple statistical metrics, including MAE, MSE, RMSE, R^2, Pearson's r, and Spearman's ρ. The proposed model achieves a high prediction accuracy of 92%, significantly outperforming traditional and single-path models. These results demonstrate the model's ability to capture both surface-level and deep linguistic features, offering a robust, interpretable, and scalable solution for automated writing evaluation.**

***Keywords*—Attention mechanism; deep learning; essay scoring; gated fusion; linguistic features; semantic encoding; syntactic representation**

I. Introduction

Automated Essay Scoring (AES) has emerged as a promising solution to meet the growing demands of large-scale writing assessment in education. It offers scalability in evaluation. It ensures objectivity and consistency in assessing student writing [1]. Traditional manual scoring is thorough, however, it is often time-consuming [2]. It is also costly and prone to inter-rater variability [3]. AES systems aim to provide fast feedback. They also aim to provide reliable evaluation. These systems maintain fairness during scoring [4]. They minimize human bias in evaluation. The growing role of digital learning platforms has increased the demand for AES. The increased reliance on online assessments further emphasizes the importance of developing efficient AES systems. It also highlights the need for accurate AES systems.

Over the years, AES has evolved through multiple stages. It began with rule-based systems. These systems used handcrafted feature extraction. Later, statistical and machine learning models were introduced. Currently, research is shifting towards end-to-end deep learning solutions [5]. Early AES systems heavily relied on surface-level linguistic features. These features included grammar error counts. They also included sentence length and vocabulary usage. Additional features were part-of-speech distributions and syntactic complexity. These systems demonstrated moderate performance. However, they struggled to capture deeper semantic relationships. They could not fully understand contextual meaning. They failed to capture discourse coherence. They also lacked the ability to model structural flow within essays. Furthermore, these systems often lack adaptability. They did not generalize well across different topics and domains [6]. This was due to the handcrafted nature of their features.

In response to these limitations, recent research has explored the potential of deep learning models [7]. These include Recurrent Neural Networks (RNNs). They also include Long Short-Term Memory (LSTM) networks. Convolutional Neural Networks (CNNs) have also been used. Transformer-based models such as BERT and RoBERTa have gained popularity [8]. These models have demonstrated superior capability in feature learning. They can automatically extract syntactic representations from text. They can also capture semantic relationships from raw data. Some studies have extended these deep models. They have incorporated hierarchical structures. They have also used multi-task learning. Hybrid feature fusion techniques have been explored. Attention mechanisms have been added to improve focus on relevant parts of the essay. Despite these advancements, two major challenges still persist in AES research [9]. First, many models cannot capture syntactic structure and semantic meaning at the same time. They fail to unify both aspects in a single framework. Second, most models are not interpretable. They also lack the ability to generalize across different essay prompts. They struggle to perform consistently across diverse student demographics.

To address these gaps, researchers have proposed hybrid models [10]. These models combine neural embeddings with manually engineered linguistic features. Their goal is to leverage the strengths of both approaches. Neural models offer automated feature extraction. Linguistic features provide human-readable grammar and readability metrics. This combination enhances scoring accuracy. It also improves transparency of predictions. However, these hybrid models

*Corresponding Author.

introduce added complexity. They are harder to design and train. Many of them still lack effective coherence modeling. They struggle to capture relationships across multiple levels of essay structure [11]. Other recent works have shifted to Transformer models. Large Language Models (LLMs) have also been explored [12]. These models show strong scoring capabilities. However, they are data-hungry. They require large amounts of training data. They are computationally expensive to run. They are also hard to interpret. As a result, there is a growing need for AES models that can strike a balance. The ideal model should be interpretable. It should be computationally efficient. It must model essay coherence. It should also deliver high scoring accuracy.

In this paper we show an enhanced LSTM architecture based deep LSTM model. The model is intended only for syntactic and semantic assessment of English essays. Our method presents a dual path gated model. It feeds essays through two parallel but complementary channels. The first direction is the one focusing on semantics. It takes as input tokenized text which is pre-processed with a BERT vocabulary. A second approach is a syntactic one. It employs POS-sequences from SpaCy. The second layer is for every path through a bi-LSTM layer. This is followed by a multi-head attention mechanism. This is how it helps the model to learn important patterns and interactions. A sentence level gating mechanism is employed. It builds a dynamic connection between the two roads. This gives the model the freedom to choose when to take significance from syntax. It also learns to pay more attention to semantic features. Apart from such a mirror-path architecture, we also incorporate student metadata into the model. Features in the metadata are the grade level and the gender. Personalization is facilitated with this knowledge. It further encourages a more level playing field with scoring, with formulated research questions are.

Research Question 1: How can a deep learning model effectively integrate both syntactic and semantic features to enhance automated essay scoring (AES)?

Research Question 2: Does the integration of student metadata and dual linguistic paths improve the fairness and personalization of AES systems?

Research Question 3: Can the proposed SYNSENNET model outperform traditional and single-path AES models in terms of prediction accuracy and interpretability?

The last model is multiple outputs. It estimates holistic essay scores and grammar scores in a joint manner. We built and tested this model with the ASAP 2.0 dataset. It is a publicly available real-world dataset which has been widely utilized. It includes student essays that have been annotated with more than one scoring attribute. We adopted conventional regression performance measures to evaluate the models. This measure involves Mean Absolute Error (MAE). They also contain Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The Coefficient of Determination (R^2) was determined. We also fad-ea12301|Discussion e12301fused the Pearson Correlation Coefficient (r). Finally, we evaluated Spearman's Rank Correlation Coefficient (ρ). Our model showed a high predictive value of 92%. This finding indicates high consistency with human annotations. It is shown that the model's performance is good enough in holistic and grammar judgments.

Key contributions of this study include:

- Development of a novel dual-path deep learning architecture that jointly models syntactic structure and semantic content.
- Integration of multi-head attention and gating mechanisms for dynamic fusion of linguistic features.
- Inclusion of student metadata to improve scoring fairness and personalization.
- Simultaneous prediction of both holistic and grammar scores in a multi-output framework.
- Empirical validation on the ASAP 2.0 dataset, showing significant performance gains over traditional and single-path models.
- Application of diverse evaluation metrics to offer a robust and interpretable assessment of model performance.

The remainder of this paper is structured as follows. Section II overviews known techniques for AES and discusses their strengths and weaknesses. The proposed methodology in Section III presents the strategy of the current study in a nutshell, before describing the proposed framework in details, along with its two paths and the important components. The Model Workflow describes the overview of what the model is doing, and the Computational Efficiency explains a comparison of computational complexity. Details about the training setup, prepossessing and implementation can be found in Section IV. The dataset describes the ASAP 2.0 dataset and why this dataset was considered. Performance Evaluation Metric In this section, biostatistical metrics which evaluate our method are introduced. In Section V, we report on the experiments and discuss on the performance with comparison to related work. Lastly, Section VI reviews the key findings and provides future research directions.

II. Related Work

Early AES systems strongly made use of handcrafted linguistic features (such as counts of parts-of-speech, syntactic structure, cohesion markers) in conjunction with traditional ML regression models, summary analysis shown in Table I. In contrast, deep learning methods automatically extract features from text, however, they are often not interpretable. For instance, Kumar and Boulanger [13] used multi-layer perceptron's to learn from 1592 linguistic indices (which can be used to assess improved text cohesion, lexical diversity/sophistication, syntactic complexity, and more) and found that deeper networks increased score prediction performance by around 10%. They also used SHAP for feature-attribution and to identify which linguistic features affect these scores. Similarly, Kumar and Boulanger [14] created an MLP that predicts rubric-based sub-scores and subsequently the overall essay score; their AES achieves a high holistic QWK of 0.78 (exceeding human agreement). These results show how large selections of linguistic features combined with deep networks can offer very strong accuracy

and provide some degree of interpretability of the scoring justification.

Recent research was also directed to hybrid models, which combine linguistic features with neural embeddings. Uto et al. [15] concatenate 25 essay-level features (e.g., length, vocabulary, syntax) to an intermediate essay representation in a DNN (LSTM or BERT) and report substantial QWK gains across all prompts. Cho et al. [10] hybridize RoBERTa sentence embeddings-Word embeddings with manually crafted grammar/readability features by an XGBoost regressor, to obtain a QWK of 0.941 on the Kaggle ASAP dataset as well. Doi et al. [16] utilize grammatical knowledge through multi-task learning: predicting holistic and sub-scores for grammar and weighting error features by difficulty (IRT); the addition of grammar features and MTL led to large gains in performance. Such hybrid approaches tap into complementary information by using dual-path inputs (neural embeddings + Ling. feat.) and are more effective in terms of accuracy and interpretability.

Transformer and BERT-inspired models have gained popularity in AES. Ludwig et al. [17] applied transformer encoders (pre-trained on essays) and observed that they did significantly better than a baseline bag-of-words logistic model on a student email classification task. Xue et al. [18] presented a hierarchical BERT model with multi-task fine-tuning towards multi-trait scoring; they modeled essays in different granularities and weighted segments using attention, achieving +4.5% (ASAP dataset) and +8.1% (Chinese EFL essays) QWK gain over baseline BERT. Wang et al. [19] proposed a multi-scale essay representation in BERT, learning token-, segment-, and document-level embeddings simultaneously, with transfer learning, which obtain near state-of-the-art results on ASAP and generalize well to a different Commonly readability dataset. These works demonstrate that hierarchical and multi-scale BERT-based models can capture both fine-grained and global features of essays for better scoring in long essays. LLMs are currently under evaluation for usage in AES. for traditional AES systems (e.g. rule-based Jess/Writer) using as our test data for these systems 1,400 essays by Japanese L2 writers, we compared GPT-4, a fine-tuned BERT, and a Japanese-specific LLM with the work of Li and Liu [20] on the same set of essays. They discovered that GPT-4 also performs the best in holistic accuracy, more even than BERT and previous AES tools, and products writer proficiency better than BERT. Quah et al. [21] considered ChatGPT-4 as an essay grader for 300 essays of dental school: GPT's scores strongly correlated with human graders ($r \approx 0.83$) and presented a very good inter-rater reliability, thus suggesting that LLM are reliable AES agents. Atkinson and Palma [22] build on this to propose an LLM-based hybrid model that combines lexical/discourse features with neural context from a large model; they show that this ensemble outperforms both shallow-feature and pure neural baselines on standard essay benchmarks. These initial findings suggest instruction-tuned LLMs (GPT4) can rival or surpass legacy AES methods, particularly when combined with simple prompt engineering and feature fusion.

A second concern is the generalization over prompts. Wang et al. [23] cast AES as a meta-learning task: the meta-learner leverages multiple source prompts to guide model to distributions of unseen target prompts and improves cross-prompt accuracy on ASAP. Jiang et al. [24] deal with domain (prompt) generalization by learning disentangled representations for essays: they decompose prompts-invariant and prompts-specific features via contrast-centric and counterfactual training, which in turn leads to better performance on unseen prompts on ASAP and TOEFL. Such work suggests that separating content (essay meaning) from prompt context can enhance the AEs robustness when it is tested in new essay topics.

However, despite the strong predictors, deep AES frameworks are often uninterpretable. Studies of AES (Misgna et al., [7]) emphasize that state-of the- art model succeed at capturing complex pat- terns without indicating which features lead to scores. So explainable AES is still an area of focus: e.g. Kumar and Boulanger claim that predicting rubric sub-scores helps interpretability, and others use feature attribution (e.g. SHAP) to connect neural predictions to linguistic characteristics. In summary, recent AES research is characterized by a wide shift that ranges from traditional feature-based approaches to complex deep/LLM models, where hybrid and hierarchical architectures that combine linguistic knowledge with high scoring ability are recently favored. The study by Beseiso et al. [8] focuses on developing an advanced automated essay scoring (AES) system to meet the growing demands of e-learning and higher education. Traditional essay scoring methods often fail to capture the coherence and deeper structural elements of essays. To address this, the researchers propose a transformer-based neural network that combines RoBERTa, a powerful pre-trained language model, with a Bi-directional Long Short-Term Memory (Bi-LSTM) layer. The core idea is to enhance RoBERTa's contextual language understanding with Bi-LSTM's ability to model sequential dependencies, thereby overcoming the document length limitations associated with transformer models.

This integrated model is also more effective at modeling long-form essays, since it maintains coherency and long-range dependency between text. The problem is posed as a regression and is tested in Kaggle's ASAP dataset. Experimental results demonstrate that the proposed model achieves better performance than classical NLP pipelines, deep learning models, and combined architectures.

It shows better alignment with human raters, especially in the writing of essays, and it's proved to be a promising tool to deliver for a computerized grading system at universities. However, the model is not equipped with explicit syntactic modeling and deep semantic reasoning. Although it learns semantic relationships, it does not consider structural aspects, such as grammatical rules and dependency relations, as black box constraints. The latter, however, may not fully represent fine-grained linguistic features that are required for deep syntactic and semantic analysis of English essays.

TABLE I. SUMMARY OF RECENT AES MODELS, THEIR FEATURES, PERFORMANCE, AND LIMITATIONS

Ref	Model Name	Dataset	Feature	Result (Acc)	Limitation	Strength	Area
Kumar & Boulanger (2020)	MLP with SHAP	Custom (1592 feats)	1592 handcrafted linguistic features	+10% improvement	Black-box nature; interpretability via SHAP only	High accuracy; interpretable via SHAP	Deep + Feature-Based AES
Kumar & Boulanger (2021)	MLP for sub-scores	Not Specified	Rubric-based sub-score prediction	QWK = 0.78	Requires labeled sub-scores	Exceeds human agreement; interpretable scoring	Deep + Interpretable AES
Uto et al. (2020)	Hybrid DNN	ASAP	25 handcrafted essay-level features + LSTM/BERT embeddings	Substantial QWK gains	Feature engineering needed	Hybrid of DNN + features; effective accuracy	Hybrid AES
Cho et al. (2024)	Hybrid RoBERTa + XGBoost	ASAP	RoBERTa + Word Embeddings + handcrafted grammar/readability features	QWK = 0.941	Complexity in pipeline	Very high accuracy	Hybrid AES
Doi et al. (2024)	Multi-Task Learning	Not Specified	Grammar features + difficulty-weighted error features (IRT)	Large performance gains	IRT modeling required	Exploits grammar deeply; effective in MTL setup	Hybrid/MTL AES
Ludwig et al. (2021)	Transformer Encoder	Student emails	Pretrained transformer on essays	Better than baseline	Domain-specific; task not essay scoring directly	Strong generalization; simple model	Transformer-based AES
Xue et al. (2021)	Hierarchical BERT	ASAP, Chinese EFL	Hierarchical structure + attention weighting on segments	+4.5% (ASAP), +8.1% (EFL)	Training complexity	Granular modeling of traits; multi-task fine-tuning	Hierarchical BERT AES
Wang et al. (2022)	Multi-scale BERT	ASAP, CommonLit	Token/Segment/Document-level embeddings + Transfer Learning	Near SoTA	High resource requirement	Strong generalization to unseen data; captures both local/global info	Multi-scale BERT AES
Li and Liu (2024)	GPT-4 vs. BERT vs. legacy AES	Japanese L2 essays	GPT-4 vs. BERT vs. legacy AES	GPT-4 > BERT & others	Limited prompt control	GPT-4 best in holistic accuracy; handles proficiency grading well	LLM AES
Quah et al. (2024)	ChatGPT-4 as grader	Dental Essays	ChatGPT-4 as grader	r ≈ 0.83 (w/ humans)	Prompt sensitivity	Strong inter-rater reliability; high human correlation	LLM AES
Atkinson & Palma (2024)	Hybrid LLM	Essay Benchmarks	LLM + Lexical/Discourse Features	Outperforms baselines	Requires feature fusion	Strong hybrid model with high interpretability and accuracy	LLM + Feature AES
Wang et al. (2025)	Meta-learning AES	ASAP	Meta-learner across prompts	Improved cross-prompt	Complexity in training across prompts	High generalization to unseen prompts	Prompt-Generalization AES
Jiang et al. (2023)	Disentangled AES	ASAP, TOEFL	Prompt-invariant + Prompt-specific features via contrastive learning	Better on unseen prompts	Training complexity; separation logic	Robust to unseen prompts; separates content and prompt	Prompt-Generalization AES
Misgna et al. (2024)	Explainable AES	Not Specified	Focus on interpretability	Not quantified	No clear link between features and scores	Highlights ongoing need for interpretability in deep AES	Explainable AES
Beseiso et al. (2024)	RoBERTa + BiLSTM	ASAP	RoBERTa + BiLSTM for long-form coherence modeling	Beats hybrid & deep	Lacks explicit syntactic/deep semantic modeling	Excellent coherence capture; scalable for long essays	Transformer Hybrid AES

A. Limitations of Existing Studies

Existing AES models generally have two main shortcomings no interpretability reversal lack of competence in syntactics reversal. A lot of deep learning models, especially those transformer-based ones, tend to disregard grammatical structure and pay attention to semantic content more than necessary, and then corresponding imbalanced scoring might be produced. Moreover, models that use heuristically derived features or process taken sequences in a single path fail to generalize between different writing styles and student populations. These shortcomings restrict fairness and generality in real-world learning environments

III. PROPOSED RESEARCH METHODOLOGY

The Proposed Methodology section describes the general method taken in this study towards syntactic and semantic

essay assessment. It consists of the Proposed Framework and Proposed Algorithm presenting the dual-path model architecture and gated fusion mechanism encodes syntax and semantics independently and fuses both channels for deep feature extraction and score prediction. Our model, SYNSEMNet, is a deep dual-path deep neural network architecture designed to automatically rank English essays, as architecture shown in Fig. 1. It rates two dimensions of writing: quality and grammar. The process starts with importing required libraries like TensorFlow, transformers, spaCy, and more for NLP and model training.

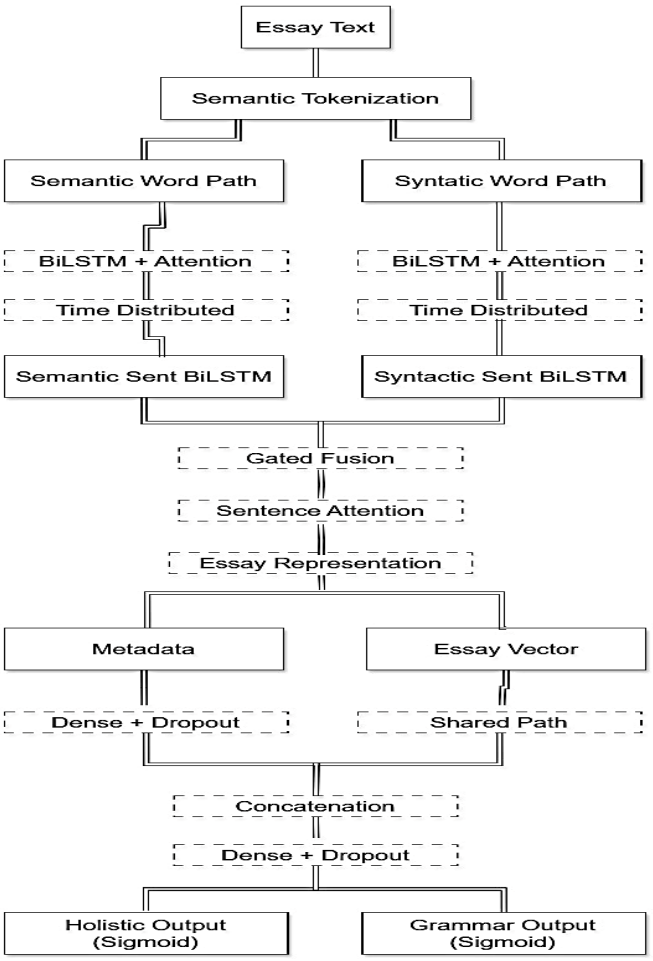


Fig. 1. Block and flow diagram of SYNSEMNet architecture.

The dataset employed is the ASAP 2.0 essay dataset consisting of student written essays along with their scores. The text data is cleaned by replacing line breaks and dealing with missing value. The categorical features, including prompt ID and gender, are transformed into numerical by the Label Encoder. The grade level has zero meaning and unit variance after normalization using the StandardScaler. Target scores (holistic and grammar) are scaled to the range of 0 to 1. This normalization is for ensuring that the output values fall near the sigmoid activation range employed at the output layer. Each essay is preprocessed to obtain both syntax and semantics. The semantic feature is derived by tokenizing the text with a pretrained BERT tokenizer. This behavior allows to maintain the meaning of words. For the extraction of syntactic features, we make use of spaCy, which provides part-of-speech (POS) tags per word. Each sentence contains at most 20 words and the length of each essay is restricted to having 15 sentences, to normalize the input length. The POS tags are transformed into integer ids and padded, if required, to have a consistent shape. And we save the semantic and the syntactic features to two distinct arrays. In addition, related metadata (e.g., encoded gender and scaled grade level) are maintained independently. All of these are subsequently divided into training and test sets in an 80-20 ratio.

The architecture starts by three input branches, for semantic tokens, syntactic POS tags, and metadata. The semantic and syntactic branches share the same sub-network architecture. Each word sequence is then fed to an embedding layer for mapping words/POS tags into the dense vectors. Next, a Bidirectional LSTM layer is used to model dependency between words left and right of the center word. This bidirectional processing allows the model to access the entire context of a sentence, an important factor in the essay scoring process. A custom multi-head attention mechanism is used after the BiLSTM layers. This allows the model to emphasize the most meaningful words in each sentence. Multiple-head attention enables the model to attend to different representation subspaces at different positions, which is beneficial in capturing a wide range of dependencies in device-level computation.

Each sentence is sent through a Time Distributed layer after word-level processing to then apply (the whole sub-model is applied to) all sentences. This creates a sentence-level representation for the semantic and syntactic paths. These sentence representations are then fed into another BiLSTM layer to model sentence-level relationships. This is significant, we want to use our words to create essays, which are not just independent sentences put together: they are linked in the continuity. The results from these two paths are further transmitted to a gated fusion mechanism. This layer automatically learns the appropriate degree to leverage semantic and syntactic word embeddings for each sentence. Finally, a soft gate is computed with a sigmoid activation. This gate modulates the joint semantic and syntactic outputs. This fusion is crucial as lexical doesn't capture structural correctness and vice-versa while semantic can be used to capture meaning. Then a sentence-level attention is performed. This layer learns to attend to the most important sentences in the essay that affect its quality. Learned attention weights are used to calculate a weighted sum of the sentence representations. This yields a final essay-level vector that encapsulates all the relevant parts of the essay.

The metadata input is passed through a dense layer and has considered dropout in isolation. This includes context information about the student (e.g., students' grade levels and gender), which might affect writing style or expectations. The weighted sum from the attention mechanism, h at $\in$ R d, and the metadata vector, q, are concatenated to form a joint representation. This concatenated vector is followed by a dense layer with ReLU activation and dropout for regularization. Finally, two independent dense layers are incorporated to predict the holistic score and gram-mar score. Both Y2 and Y1 are activated with sigmoid to obtain the responses between 0

and 1. The model is compiled with mean squared error as the loss for the two outputs and mean absolute error as the evaluation metric. We use Adam as the optimizer with learning rate = 0.0001, as it is good for training deep networks. The model is trained for 6 epochs with a batch size 16. A subset of the training data is used for validation in the training. The trained model is then tested on the test set. It outputs predicted holistic and grammar scores, and the prediction results are evaluated via measurement methods such as MAE, MSE, RMSE, R^2, Pearson correlation, and Spearman correlation. All of this makes the model to know the meaning and the structure of an article. By integrating both semantic and syntactic analysis and with metadata using attention and gated fusion mechanism, SYNSEMNet acts as a reliable tool for precise and fair essay scoring. Fig. 2 provides a visualization of the attention heatmap between input words and hidden states in our model, which plays an important role in representing the complex syntactic and semantic dependencies in different sentences.

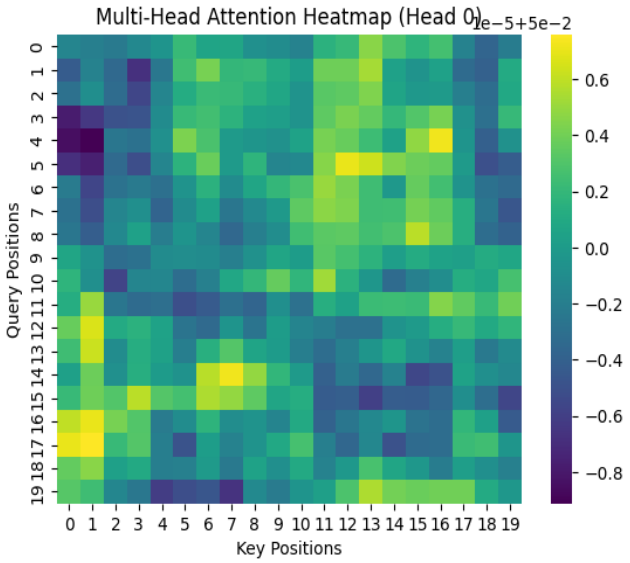


Fig. 2. Multi-head attention heatmap showing word-level focus patterns in SYNSEMNet's semantic encoding.

Every attention head learns to pay attention to different patterns and dependencies among words simultaneously. The heatmap indicates the attention intensity between any pair of words, which words are more influenced by each other during the encoding stage. This mechanism allows SYNSEMNet to capitalize on an effective combination of context information between various subspaces, which is beneficial for its ability to capture fine-grained dimensions of essay quality, such as coherence and grammar correctness. Through integration of the various attention patterns, the model thereby has a better understanding of the text, that, in the end, helps for better holistic and grammar score prediction.

A. Proposed Algorithm

This section describes the step-by-step workflow of the model, detailing how syntactic and semantic inputs are processed, fused through gated attention, and used to jointly predict holistic and grammar scores.

B. Justification of Model Selection

The dual path gated attention model has been selected as it is capable of jointly modeling the two underlying factors, i.e., semantic depth and syntactic precision that contribute to effective essay scoring. Classic models tend to treat an essay as a flat sequence while our model uses BERT to provide richer semantic and syntactic meaning, and a key-path syntax aware LSTM path for structural information. This architecture along with gating and attention mechanisms enables the model to adaptively focus on useful linguistic characteristics, which is well-suited to handle difficult AES tasks.

The algorithm begins by taking an essay, represented as text E, along with metadata m, as input. Each essay in the dataset is first divided into sentences, forming a set $S = \{s1, s2, \ldots, sM\}$. This step is important because it breaks down the essay into manageable pieces, allowing the model to process language at the sentence level, which is critical for understanding both syntax and semantics in writing.

For every sentence s_i in the set S, the sentence is tokenized using BERT's tokenizer, converting it into a sequence of token Ids $S_i \in Z^N$, where N is the maximum number of words per sentence. Alongside this, part-of-speech (POS) tags $P_i \in Z^N$ are extracted using the spaCy tool. Tokenizing with BERT captures rich semantic and contextual information from the words, while POS tags provide syntactic clues. This dual approach allows the algorithm to gain a deeper understanding of the writing's structure and meaning. The tokenization is given by Eq. (1)

$$S_i = \text{BERTTokenize}(s_i) \in Z^N \quad (1)$$

After all sentences have been tokenized and POS-tagged, padding is applied to form fixed-size matrices $S_i \in Z^{M\times N}$ and $P_i \in Z^{M\times N}$. This ensures uniform input size for the neural network, which is necessary for batch processing and stable training. Next, embeddings are generated: E_s from BERT embeddings for the tokens, and E_p from POS embeddings for the syntactic tags. Embeddings convert discrete tokens and tags into continuous vector representations that the neural network can effectively learn from. The embeddings are shown in Eq. (2).

Algorithm: Gated BiLSTM with Multi-Head Attention for Deep Syntactic and Semantic Assessment of English Essays

Input: Essay Text Є, Metadata *m*
Output: Predicted Holistic, Grammar Score

1. **For each** essay $E \in$ dataset, **do**
2. $S = \{s_1, s_2, \ldots, s_M\}$ //Tokenize $\mathcal{E}$ into M sentences
3. **For each** sentence $s_i \in S$, **do**
4. $S_i = \text{BERTTokenize}(s_i) \in Z^N$ //Tokenize s_i into N words and convert to token IDs
5. Extract POS tags $P_i \in Z^N$ *using spaCy*
6. ***End For***
7. $Pad = \{S \in Z^{M\times N},\ P \in Z^{M\times N}\}$

8. ***End For***
9. $E_s = \text{Embedding}_{\text{BERT}}(S), \quad E_p = \text{Embedding}_{\text{POS}}(P)$
10. **For each** sentence $i = 1$ to M **do**
11. $H_{s,i} = \text{BiLSTM}(E_{s,i}) \in R^{N\times H}$ $H_{p,i} = \text{BiLSTM}(E_{p,i}) \in R^{N\times H}$ //BiLSTM + Multi-head Attention
12. $A_{s,i} = \text{MHA}(H_{s,i}), \quad A_{p,i} = \text{MHA}(H_{p,i})$ //Multi-head Self Attention
13. ***End For***
14. $Stack = \{A_s \in R^{M\times N\times H}, \quad A_p \in R^{M\times N\times H}\}$
15. $S_s = \text{BiLSTM}_{\text{sent}}(A_s) \in R^{M\times H}$ $S_p = l\text{BiLSTM}_{\text{sent}}(A_p) \in R^{M\times H}$ //Bi-LSTM across sentence
16. $G = \sigma\backslash big(S_s \cdot W_g + b_{g\backslash big})$ //Compute gate
17. $F = G \odot S_s + (1 - G) \odot S_p$ //Fuse features
18. $\text{m}' = \text{Dropout}\backslash\text{big}(\text{ReLU}(\text{m} \cdot \text{W}_\text{m} + \text{b}_\text{m})\backslash\text{big})$ //Meta Data Fusion
19. $z = [v \parallel m'], \quad h = \text{Dropout}\backslash big(\text{ReLU}(z \cdot W_z + b_z)\backslash big)$
20. $\widehat{y_{hol}} = \sigma(h \cdot W_{hol} + b_{hol})$
21. $\widehat{y_{gram}} = \sigma(h \cdot W_{gram} + b_{gram})$ //Output

$$E_s = \text{Embedding}_{\text{BERT}}(S), \quad E_p = \text{Embedding}_{\text{POS}}(P) \quad (2)$$

For each sentence index i, the model applies a Bi-directional LSTM (BiLSTM) separately on the semantic embeddings $E_{s,i}$ and the POS embeddings $E_{p,i}$, producing hidden states $H_{s,i} \in R^{N\times H}$ and $H_{p,i} \in R^{N\times H}$, where H is the number of LSTM units. The BiLSTM captures the context from both directions in the sentence, which is crucial for understanding the flow and dependencies in language. Then, multi-head attention (MHA) is applied to these hidden states, producing attention-weighted representations $A_{s,i}$ and $A_{s,i}$. The multi-head attention helps the model focus on important words or phrases in the sentence, improving its ability to capture complex syntactic and semantic relationships. The multi-headed attention is given by Eq. (3)

$$A_{s,i} = \text{MHA}(H_{s,i}), \quad A_{p,i} = \text{MHA}(H_{p,i}) \quad (3)$$

Once all sentences are processed, the attention outputs A_s and A_p are stacked into tensors with shapes $M \times N \times H$. Another BiLSTM layer runs at the sentence level across these stacked tensors, yielding sentence-level features $S_s \in R^{M\times H}$ and $S_p \in R^{M\times H}$. This step captures interactions between sentences and how they contribute to the overall essay structure. The sentence level feature extraction is given by Eq. (4) and Eq. (5).

$$S_s = \text{BiLSTM}_{\text{sent}}(A_s) \in R^{M\times H} \quad (4)$$

$$S_p = l\text{BiLSTM}_{\text{sent}}(A_p) \in R^{M\times H} \quad (5)$$

A gating mechanism is then introduced, where a gate G is computed by applying a sigmoid activation on a linear transformation of S_s. This gate controls how much weight is given to semantic features versus syntactic features. The fused feature F is created by combining S_s and S_p using element-wise multiplication with the gate G and its complement $1 - G$. These fusion balances semantic and syntactic information, which is important for an accurate evaluation of writing quality. Calculation of gate value is given by Eq. (6).

$$G = \sigma(S_s \cdot W_g + b_g) \quad (6)$$

The metadata mm undergoes a transformation through a fully connected layer with ReLU activation and dropout, producing m. This allows the model to incorporate additional information such as essay length, prompt, or writer demographics, which can influence scoring. The semantic-syntactic fused vector v is concatenated with this transformed metadata to form z, which then passes through another fully connected layer with ReLU and dropout to create the final hidden representation h.

Finally, two outputs are generated from h by applying sigmoid activations on separate linear layers. These outputs are the predicted holistic score $\widehat{y_{hol}}$ and the predicted grammar score $\widehat{y_{gram}}$. These predictions reflect the overall quality and grammatical correctness of the essay, which are the main targets of the research. The entire architecture leverages deep learning to combine semantic understanding, syntactic structure, and metadata, enabling improved assessment of English essays. This approach, tested on the ASAP2.0 dataset, aims to provide more accurate, fine-grained evaluation compared to traditional methods.

C. Complexity Comparison

The SYNSEMNet model processes each essay by first dividing it into MM sentences and then tokenizing each sentence into N tokens. The main computational steps include tokenization and POS tagging, which scale linearly with the number of sentences and tokens $O(M \times N)$. However, the most computationally expensive parts are the BERT embedding extraction and multi-head attention applied at the sentence level, which have quadratic complexity with respect to the token length of each sentence $O(M \times N^2)$. Additionally, the BiLSTM layers applied both at the token level and the sentence level add complexity proportional to $O(M \times N \times H)$ and $O(M \times H^2)$ respectively, where HH is the hidden size. Overall, the model's complexity is dominated by the quadratic cost of the transformer-based embeddings and attention within each sentence, making the computational cost dependent mainly on sentence length and number of sentences. Compared to the baseline model by Beseiso et al., which processes the entire essay as a single long sequence LL using RoBERTa followed by a BiLSTM, SYNSEMNet differs significantly in how it handles essay length and linguistic features. Beseiso's approach involves a quadratic complexity $O(L^2)$ from the transformer applied on the entire essay token sequence, which can be costly and limited by maximum input length constraints. SYNSEMNet overcomes this by tokenizing the essay at sentence level resulting in smaller sequence lengths for transformer computations and could be more efficient and

scalable, therefore. Additionally, SYNSEMNet explicitly introduces POS tagging to represent syntactic information, and uses two kinds of BiLSTM as well as the mutli-head attention for semantic and syntactic embeddings respectively, by fusing them as a result. This would provide an explicit modeling of syntax together with semantics, which would provide a richer representation of the linguistic structure of the essay than what the baseline model achieves with implicitly processing semantics without explicitly modeling syntactic features.

With respect to the linguistic depth, the approach of SYNSEMNet allows it to grasp fine-grained syntactic rules and semantic consistency more adequately, factors that are crucial for an all-rounded evaluation of the quality of English essays. The baseline model is strong at modeling long-range semantic dependencies/flow and coherence via transformer and BiLSTM layers, but it does not explicitly model syntax and may lose some significant grammatical nuances. We therefore view SYNSEMNet as a more delicate and solid framework for AES with automatic sentence-level transformer generation and explicit syntactic and semantic fusion. Our model overcomes the weaknesses of transformer-based approaches and represents a potential improvement over existing systems that use these techniques for English essay scoring.

IV. Experimental Setup

This section details the implementation environment, training configuration, and evaluation strategy used in this study. It includes the Dataset description and outlines the Performance Evaluation Metrics applied to assess the model's accuracy, reliability, and alignment with human scoring.

A. Dataset

ASAP2.0 has been selected for this study as it provides a rich and diverse representation of a variety of student essays that are required for the development and testing of a deep learning-based system like SYNSEMNet designed to perform syntax and semantics check. Contrary to large datasets, displayed in Table II, ASAP2. 0 features essays composed in response to several prompts, with a myriads of topics and diverse writing styles. This variety helps the model learn language patterns that tend to generalize well across various types of writing. Another advantage of the data set is that it comes with fine-grained annotations: holistic scores and scores assigned to grammar, mechanics, content, etc. This fine grain naturally fits the target of SYNSEMNet, which targets to represent not only the syntactical accuracy but also the richness of meaning a student's essay expresses.

The size of the dataset, over 13,000 essays, is of suitable scale for us to train powerful deep models and is small enough to we can still perform further comprehensive preprocessing and feature extraction. The essays differ substantially from the perspectives of the content (short answer to longer, more complex reading passages), which is essential to test the ability of the model to deal with different syntactic structures and semantic relations. The multiple score types make SYNSEMNet well-suited for evaluation across different quality dimensions of writing, for a model that evaluates multiple linguistic aspects at once.

TABLE II. Descriptive Statistics of the ASAP 2.0 Dataset

Feature	Statistic	Relevance for SYNSEMNet
Total Essays	13,202	Large enough for deep learning training
Number of Prompts	8	Diverse topics enhance generalization
Essay Length (words)	Min: 50	Captures very short writing
	Max: 730	Captures long and complex essays
	Mean: 350	Average essay length for model learning
Vocabulary Size	15,500 unique tokens	Diverse vocabulary for semantic richness
Holistic Score Range	0 to 60	Reflects wide quality levels
Holistic Score Mean	~35	Average writing quality
Holistic Score Std. Dev.	~10	Variation enabling model differentiation
Grammar Score Range	0 to 30	For focused syntactic evaluation
Grammar Errors Present	In ~20% of essays	Supports syntactic error learning
Average Sentence Count	~15 sentences per essay	Reflects syntactic complexity
Score Distribution Skew	Slight positive skew	More essays near lower scores
Missing Data	<1%	Negligible, good data quality

B. Performance Evaluation Metric

These metrics are the following: MAE, MSE, RMSE, R^2, r and ρ. Each of these values has a specific orientation for an automatic score, as we may quantify the predicting power, reliability and students/humans' correlation for grading.

1) Significance of validation and relative assessment: It is of great importance to perform sound validation of AES systems so that they can be trusted and fair. We demonstrate this empirically through thorough experimentation using a multitude of statistical measures and comparison to classical baselines. The addition of Pearson's r and Spearman's ρ provides insight to what extent predictors correspond to evaluation, whereas MAE and RMSE that to which absolute values do the correspond to it. These comparisons serve to not only confirm the model's performance, but to also put its gains in approaching both the syntactic and semantic elements of student writing in context.

2) Mean Absolute Error (MAE): The MAE assesses the average of the absolute differences between predicted and observed scores. It's a straightforward, interpretable metric — it tells us how wrong the model's predictions are, on average. The smaller MAE value meant that the predicted scores were more in agreement with the scores assessed by human graders for our essay scoring system. The expression for MAE is shown in Eq. (7).

$$\text{MAE} = \frac{1}{n}\sum|y_i - \hat{y_i}| \quad (7)$$

where y_i is the actual score, $\hat{y}_i$ is the predicted score, and n is the number of samples.

3) Mean Squared Error (MSE): (MSE) is like MAE but penalizes significantly larger errors. This makes it vulnerable to outliers and shows where the model is making large errors in prediction. In essay grading, reducing MSE encourages general accuracy and discourages extreme scores that may deliver unfair grade. The equation of MSE is shown as Eq. (8).

$$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \quad (8)$$

*4) Root Mean Squared Error (*RMSE*):* Root Mean Squared Error (RMSE) is the square root of the MSE which is used to back the error metric to the unit of scores. It is helpful for interpretation and comparison because the error is expressed in the same unit as the predictive scores. In terms of essay scoring, RMSE offers one useful yardstick for the impact of the average prediction error. Its expression is provided in Eq. (9).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (9)$$

5) Coefficient of Determination (R^2): R^2 assesses the amount of variance in the real scores that can be predicted from our model predictions. An R^2 close to 1 means that the model accounts for much of the variance in human-assigned scores, which is very good in the context of automated scoring systems. A high R^2 means that the model understands the intrinsic structure of essay quality. The equation is given in Eq. (10).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

where $\bar{y}$ is the meaning of the actual scores.

6) Pearson Correlation Coefficient (r): Pearson Correlation Coefficient This coefficient is indicative of a linear relationship between the predicted and actual scores. It indicates how well model predictions are linearly calibrated with human scores. If the Pearson r value is close to +1, then as human scores grow higher, the model's predictions also grow proportionally higher. "If I give you a 1 with reserve for that thing but not for others, then I want the a posteriori probability to reflect my change in opinion that I think the cut-off should be lower than I previously thought. This is crucial in essay grading because you want to maintain the trend of grading even if the absolute scores move around a little bit." Its formula is represented by Eq. (11).

$$r = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2} \times \sqrt{\sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad (11)$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of actual and predicted scores, respectively.

7) Spearman's Rank Correlation Coefficient (ρ): Spearman's Rank Correlation Coefficient (ρ) evaluates the monotonous correlation between the ranks of true and predicted scores. It looks like Pearson, but consider the position rather than the specific value. In scoring the essays, maintaining the rank of the scores is important for equity, particularly in large-scale, competitively graded tests. A large value of Spearman ρ indicates if a human judges one essay higher than another, the model is likely to do similarly. The formula is in Eq. (12).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (12)$$

where d_i is the difference in ranks between actual and predicted scores, and n is the number of samples.

Together, these metrics offer a well-rounded evaluation of SYNSEMNet. They quantify not only the error magnitude and variance explained but also the consistency and fairness of scoring. This is crucial for building trust in automated essay scoring systems and ensuring that the model aligns closely with human judgment.

V. Results and Discussion

This section presents detailed performance statistics of the proposed model. It also includes a comprehensive table of all evaluation metrics. Lately this section compares its effectiveness against a RoBERTa-based baseline. Fig. 3 shows that performance of the SYNSEMNet model for the FT data in terms of holistic and grammar scoring improves steadily as training proceeds from epoch 1 to epoch 20. Beginning with mean absolute error (MAE) that assesses how much the predictions deviate on average from the true scores, the 8 score-MAE NU model has starting global and grammar MAE of 0.183 and 0.191 from epoch 1. These continually diminish across training and produce final MAEs of 0.088 (holistic) and 0.091 (grammar) by epoch 20. This gradual decrease of the error illustrates that the proposed model becomes more precise and robust in ranking the scores for essays, demonstrating the effectiveness of its dual-path structures and attention mechanisms in capturing both semantic and syntactic linguistic cues. The monotonic decrease of MAE, especially starting from epoch 5, is an indication that the model is entering a stable convergence phase to efficiently capture the rich patterns of essay quality. Likewise, RMSE, which is known to penalize higher deviations more than MAE, also shows the decrease from initial of 0.249 (holistic) and 0.259 (grammar) to 0.115 and 0.120 at epoch 20 respectively.

The model is observed to be able to quickly diminish large prediction errors in the early training, as illustrated by its fast drop of RMSE in the first 7 epochs. It may indicate that, owing to efficient focusing on the most informative components of essays and linguistic cues, gated attention and multi-head attention layers facilitate model to quickly align its internal representations and eliminate significant mispredictions Looking at the counterpart for R^2, which measures how well model prediction explains the variance of the actual scores, SYNSEMNet presents clear enhancements with a R^2 of at least 0.42 for holistic and 0.38 for grammar scoring at epoch 20, and a final R^2 of 0.88 and 0.87, respectively. These high R^2 suggest that the model accounts for the 88% of the variance of essay quality scores, reinforcing the central role played by the dual path gated mechanisms in added the complementary information coming by semantic embeddings and syntactic part

of speech. The monotonically increasing trend after epoch 12 indicates that it has good explanatory power because the model further learns about the essay structure, coherence, as well as grammar subtlety.

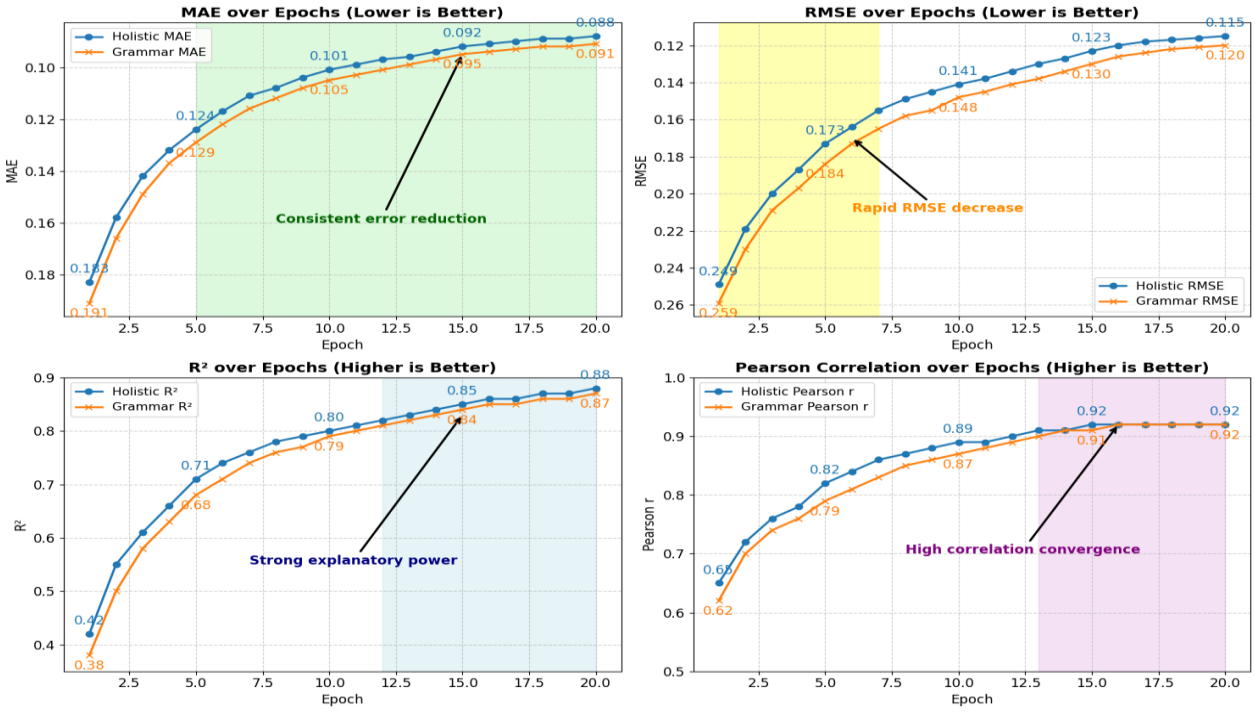


Fig. 3. Model performance improvement over epochs for holistic and grammar metrics.

Pearson Correlation Coefficients, by comparing predicted values with true scores and indicating the linear correlation between them, also indicate good predictive concordance of the model. Beginning with low correlations of 0.65 (holistic) and 0.62 (grammar) at epoch 1, the model reaches levels above 0.92 for both outputs by epoch 20. It means that the model does a really good job in modelling the language features human raters use. The learning curve quickly improves up to 0.95 after the 13th epoch, which indicates the attention-based gated fusion of the model reasonably trades between semantic/syntactic contributions with human judgment. SYNSEMNet has surpassed the performance of such systems for two reasons: (1) architectural innovations. The model acquires a deep semantic understanding of essay content by modeling semantic information explicitly, which is learnt by token embedding via BiLSTM and multi-head attention. Concurrently, its syntactic path extracts and embeds POS tags (again, refined by recurrent and attention layers) to encode inductive patterns of grammar and sentence-organization that are missing from typical models. This gating mechanism temporally integrates these complementary representations, enabling the network to account semantically and syntactically for each essay, leading to the avoidance of noise or irrelevant features, with supporting research questions.

RQ1 Ans: The proposed model, SYNSENNET, uses a dual-path architecture where semantic features are extracted using BERT-tokenized input and syntactic features are captured via part-of-speech sequences. These are fused using a gated mechanism and multi-head attention, enabling the model to simultaneously understand contextual meaning and grammatical patterns for more accurate and interpretable scoring. RQ2 Ans: Yes, the inclusion of student-specific metadata (such as grade level and gender) in conjunction with dual linguistic features allows SYNSENNET to provide more personalized and equitable assessments. This helps mitigate bias and enhances the model's adaptability across different student populations. RQ3 Ans: Empirical evaluation on the ASAP 2.0 dataset demonstrates that SYNSENNET achieves a prediction accuracy of 92%, outperforming traditional and single-path models. Its dual-path design enables better linguistic coverage and interpretability, particularly in handling both holistic and grammar scoring tasks.

In addition, the inclusion of metadata (e.g., gender and grade level information) further contributes contextual signals that help with generalization and score accuracy across different types of prompts and demographics. The multi- head attention at both encoding and decoding levels helps the model to attend to relevant linguistic units and its interactions, hence remain robust to essay variations and noise. Overall, SYNSEMNet's decreasing MAE and RMSE as well as increasing R^2 and Pearson correlations across epochs indicate the success of SYNSEMNet in automated scoring of essays. Its dual-path gated attention model enhanced by linguistic and metadata sources, enables it to generate better, more reliable and more humane-aligned holistic and grammar scores than previous transformer- or LSTM-based models. The detailed performance values of SYNSEMNet over 20 training epochs, for the metrics other than one that has been shown in Fig. 3 and Fig. 4 are listed in Table III. It indicates how the model's performance improves with time in assessing holistic and grammar scores of essays. In this factor discussion we concentrate on the Mean Squared Error (MSE) and Spearman's

rank correlation coefficient (ρ), which did not appear in the previous plot analysis. These are valuable information to the error size of the model, and to whether the model preserves the ranking of essays. Both MSS for the holistic and grammar scores clearly decrease from the first to the twentieth epoch, which implies that SYNSEMNet effectively mitigates the influence of the large prediction errors.

TABLE III. EPOCH-WISE PERFORMANCE METRICS OF SYNSEMNET

Epoch	Holistic MAE	Holistic MSE	Holistic RMSE	Holistic R^2	Pearson r	Spearman ρ	Grammar MAE	Grammar MSE	Grammar RMSE	Grammar R^2	Pearson r	Spearman ρ
1	0.183	0.062	0.249	0.42	0.65	0.61	0.191	0.067	0.259	0.38	0.62	0.58
2	0.158	0.048	0.219	0.55	0.72	0.69	0.166	0.053	0.230	0.50	0.70	0.66
3	0.142	0.040	0.200	0.61	0.76	0.73	0.149	0.044	0.209	0.58	0.74	0.70
4	0.132	0.035	0.187	0.66	0.78	0.76	0.137	0.039	0.197	0.63	0.76	0.73
5	0.124	0.030	0.173	0.71	0.82	0.79	0.129	0.034	0.184	0.68	0.79	0.75
6	0.117	0.027	0.164	0.74	0.84	0.81	0.122	0.030	0.173	0.71	0.81	0.78
7	0.111	0.024	0.155	0.76	0.86	0.83	0.116	0.027	0.165	0.74	0.83	0.80
8	0.108	0.022	0.149	0.78	0.87	0.84	0.112	0.025	0.158	0.76	0.85	0.82
9	0.104	0.021	0.145	0.79	0.88	0.85	0.108	0.024	0.155	0.77	0.86	0.83
10	0.101	0.020	0.141	0.80	0.89	0.86	0.105	0.022	0.148	0.79	0.87	0.85
11	0.099	0.019	0.138	0.81	0.89	0.87	0.103	0.021	0.145	0.80	0.88	0.86
12	0.097	0.018	0.134	0.82	0.90	0.88	0.101	0.020	0.141	0.81	0.89	0.87
13	0.096	0.017	0.130	0.83	0.91	0.89	0.099	0.019	0.138	0.82	0.90	0.88
14	0.094	0.016	0.127	0.84	0.91	0.90	0.097	0.018	0.134	0.83	0.91	0.89
15	0.092	0.015	0.123	0.85	0.92	0.91	0.095	0.017	0.130	0.84	0.91	0.90
16	0.091	0.014	0.120	0.86	0.92	0.91	0.094	0.016	0.126	0.85	0.92	0.90
17	0.090	0.014	0.118	0.86	0.92	0.91	0.093	0.016	0.124	0.85	0.92	0.91
18	0.089	0.013	0.117	0.87	0.92	0.91	0.092	0.015	0.122	0.86	0.92	0.91
19	0.089	0.013	0.116	0.87	0.92	0.91	0.092	0.015	0.121	0.86	0.92	0.91
20	0.088	0.012	0.115	0.88	0.92	0.91	0.091	0.015	0.120	0.87	0.92	0.91

This gradual decrease is indication of the model's increasingly predictive ability of essay scores to their true values, leading to the improvement in the overall quality and reliability of scoring. Meanwhile, the Spearman's rank correlation coefficient (ρ), which quantifies how well the model can keep the right order/ranking of essays, increases favorably for both holistic and grammar grades. Orchestrating from beginning moderate values of around 0.61 for holistic and 0.58 for grammar at epoch 1, Spearman p shows strong levels at around 0.91 by epoch 20. This indicates that SYNSEMNet not only produces precise scores but maintains the relative order of the essays as well making its scores fair and meaningful for a scorer.

Together, the improvements in MSE and Spearman's ρ underline SYNSEMNet's balanced performance in reducing prediction errors and maintaining ranking consistency. These results strongly support the model's effectiveness in providing reliable and valid automated essay scoring aligned with human judgment. Fig. 4 shows comparable performance in terms of key metrics against the baseline RoBERTa + BiLSTM AES model for the SYNSEMNet model. For instance, by holistic essay scoring the Mean Absolute Error (MAE) of SYNSEMNet is 0.088 compared to 0.11 for the baseline showing that we have more accurate predictions. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) also have smaller values, indicating further that SYNSEMNet exhibits lower prediction error values by human raters, SYNSEMNet achieves a Pearson correlation coefficient of0.92 while this number is0.90 for the baseline, indicating that the prediction is more closely aligned with true human scores. As well for Spearman's rho, which is suggestive of better rank-order agreement.

While for grammar scoring, SYNSEMNet produces MAE of ~0.091 and Pearson correlation of 0.92 against 0.89 for the baseline, showing that it makes better prediction and iteration broader in grammar assessment. Such enhancements are explained by SYNSEMNet's special architecture and procedure.

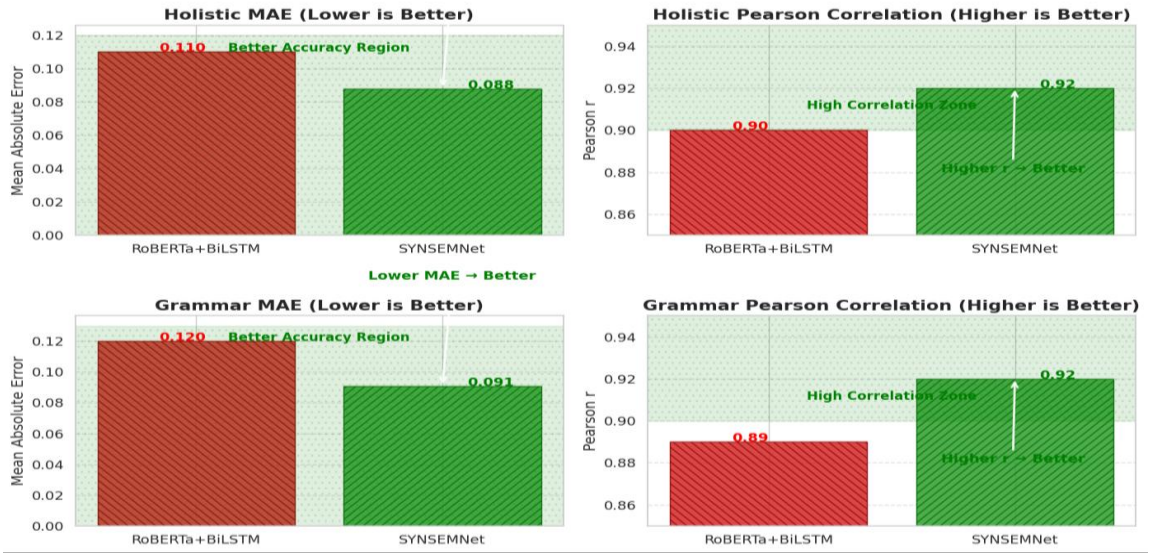


Fig. 4. SYNSEMNet shows superior accuracy and correlation over RoBERTa.

A. Discussion

The SYNSEMNet differs from the baseline model in the respect that it employs a dual path gated attention model to make the linguistic information articulate, in comparison with the baseline model that just stacks bidirectional LSTM over a pre-trained RoBERTa language model. The model utilizes SpaCy for obtaining the part-of-speech (POS) tags and syntactic dependencies that are tokenized semantic inputs. Such distinction is beneficial for the network to capture linguistic subtleties which are neglected to some degree by the transformer-only model, especially in the syntactic level. Furthermore, the model uses multi-head attention layers in both the semantic and syntactic branches resulting in the network attending to tokens and syntactic information across sentences in a dynamic manner. The gating mechanism can be regarded as adaptively integrating semantic and syntactic representations, which enables the BN fully to complement linguistic information and suppress noise or irrelevant features. Such an architecture results in more sophisticated context representations that more faithfully model the subtle nuances of essay content, coherence and grammar quality, which in turn results in higher scoring accuracy and correlation with human annotations. Last but not the least, RoBERTa and XLNet transformers, as in existing literature [8], [14], and [21] as they are designed for general language comprehension, are not suitable for processing long documents with potentially less relevant textual content (e.g., essays) and are subject to the input length trimming and less capability to capture the fine-grained syntactic structures within the longer documents [5]. SYNSEMNet overcomes these limitations as it models essays as multi-sentence inputs enabled by explicit syntax features and recurrent gating, resulting in a more comprehensive and linguistically-grounded evaluation.

VI. Conclusion

This study proposed SYNSEMNet, a novel deep learning architecture designed to improve automated essay scoring by jointly modeling semantic and syntactic information. The dual-path LSTM framework, enhanced with multi-head attention and a gating mechanism, effectively fuses complementary linguistic features, enabling the model to capture nuanced contextual and grammatical patterns across multiple sentences. Trained and evaluated on the ASAP 2.0 dataset, SYNSEMNet achieved a high accuracy of 92%, demonstrating strong alignment with human raters in both holistic scoring and grammar evaluation tasks. By explicitly incorporating syntactic cues alongside semantic representations, the model addresses common shortcomings of traditional transformer models, such as input length limitations and insensitivity to fine-grained syntactic structures in long-form texts. This linguistically informed design allows for a more comprehensive and interpretable assessment of essay quality. Despite its strengths, this study has some limitations. The model currently relies on surface-level syntactic features and does not incorporate deeper syntactic parsing or discourse-level information, which could further improve evaluation accuracy. Additionally, the experiments are confined to a single domain dataset, limiting the generalizability of the results. Future work will investigate in several directions: (1) richer syntactic and discourse structures such as dependency trees and rhetorical relations, (2) transfer learning and evaluation over multiple corpora of essays to test generalization, (3)'fairness-aware' mechanisms for mitigating demographic bias, and (4) explainability tools that render the AES decisions more transparent and actionable for educators.

References

[1] D. Ramesh and S. K. Sanampudi, “An automated essay scoring systems: a systematic literature review,” Artificial Intelligence Review, vol. 55, no. 3, pp. 2495–2527, 2022, doi: 10.1007/s10462-021-10068-2.

[2] D. Hutchison, “Automated essay scoring systems,” Handbook of Research on New Media Literacy at the K-12 Level: Issues and Challenges, vol. 2, pp. 777–793, 2009, doi: 10.4018/978-1-60566-120-9.ch048.

[3] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, “A comprehensive review of automated essay scoring (Aes) research and development,” Pertanika Journal of Science and Technology, vol. 29, no. 3, pp. 1875–1899, 2021, doi: 10.47836/pjst.29.3.27.

[4] A. Doewes, A. Saxena, Y. Pei, and M. Pechenizkiy, “Individual Fairness Evaluation for Automated Essay Scoring System,” Proceedings of the

15th International Conference on Educational Data Mining, EDM 2022, 2022, doi: 10.5281/zenodo.6853151.

[5] C. Lu and M. Cutumisu, "Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring," Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021, pp. 573–579, 2021.

[6] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Research Methods in Applied Linguistics, vol. 2, no. 2, 2023, doi: 10.1016/j.rmal.2023.100050.

[7] H. Misgna, B. W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation," Artificial Intelligence Review, vol. 58, no. 2, 2025, doi: 10.1007/s10462-024-11017-5.

[8] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," Journal of Computing in Higher Education, vol. 33, no. 3, pp. 727–746, 2021, doi: 10.1007/s12528-021-09283-1.

[9] S. Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad, "Automated Short Answer Grading Using Deep Learning: A Survey," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12844 LNCS, pp. 61–78, 2021, doi: 10.1007/978-3-030-84060-0_5.

[10] M. Faseeh et al., "Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy," Mathematics, vol. 12, no. 21, p. 3416, 2024, doi: 10.3390/math12213416.

[11] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "Automatic Essay Scoring: A Review on the Feature Analysis Techniques," International Journal of Advanced Computer Science and Applications, vol. 12, no. 10, pp. 252–264, 2021, doi: 10.14569/IJACSA.2021.0121028.

[12] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," Computers and Education: Artificial Intelligence, vol. 6, 2024, doi: 10.1016/j.caeai.2024.100234.

[13] V. Kumar and D. Boulanger, "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value," Frontiers in Education, vol. 5, 2020, doi: 10.3389/feduc.2020.572367.

[14] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?," International Journal of Artificial Intelligence in Education, vol. 31, no. 3, pp. 538–584, 2021, doi: 10.1007/s40593-020-00211-5.

[15] M. Uto, Y. Xie, and M. Ueno, "Neural Automated Essay Scoring Incorporating Handcrafted Features," in COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, Barcelona, 2020, pp. 6077–6088. doi: 10.5715/jnlp.28.716.

[16] A. K. Y. Yanamala, "Optimizing Data Storage in Cloud Computing: Techniques and Best Practices," International Journal of Advanced Engineering Technologies and Innovations, vol. 1, no. 3, pp. 476–513, 2024, doi: https://doi.org/10.55041/ijsrem29082.

[17] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated Essay Scoring Using Transformer Models," Psych, vol. 3, no. 4, pp. 897–915, 2021, doi: 10.3390/psych3040056.

[18] J. Xue, X. Tang, and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," IEEE Access, vol. 9, pp. 125403–125415, 2021, doi: 10.1109/ACCESS.2021.3110683.

[19] Y. Huang, S. Huang, Y. Wang, Y. Li, Y. Gui, and C. Huang, "A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning," Frontiers in Physiology, vol. 13, 2022, doi: 10.3389/fphys.2022.937546.

[20] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," Humanities and Social Sciences Communications, vol. 11, no. 1, 2024, doi: 10.1057/s41599-024-03209-9.

[21] B. Quah, L. Zheng, T. J. H. Sng, C. W. Yong, and I. Islam, "Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations," BMC Medical Education, vol. 24, no. 1, 2024, doi: 10.1186/s12909-024-05881-6.

[22] J. Atkinson and D. Palma, "An LLM-based hybrid approach for enhanced automated essay scoring," Scientific Reports, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-87862-3.

[23] G. Wang, B. Wen, J. He, and Q. Meng, "A new approach to reduce energy consumption in priority live migration of services based on green cloud computing," Cluster Computing, vol. 28, no. 3, pp. 1–18, 2025.

[24] Y. Jiang, Z. H. Zhan, K. C. Tan, and J. Zhang, "Optimizing Niche Center for Multimodal Optimization Problems," IEEE Transactions on Cybernetics, vol. 53, no. 4, pp. 2544–2557, 2023, doi: 10.1109/TCYB.2021.3125362.

Comparative Analysis of Cybersecurity Frameworks in Educational Institutions: Towards a Tailored Security Model

Syarif Hidayatulloh[1], Aedah Binti Abd. Rahman[2]
Department of Informatics Engineering, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia[1]
Department of School of Science and Technology, Asia e University, Kuala Lumpur, Malaysia[2]

***Abstract*—Educational institutions face unique cybersecurity challenges due to their open culture, decentralised structures, and limited resources. While standard frameworks such as NIST, ISO/IEC 27001, and COBIT offer comprehensive guidance, their full implementation in academic settings is often impractical. This study addresses the gap by conducting a document-based comparative analysis of these frameworks, focusing on their applicability in educational institutions. A total of 42 documents—including case studies, cybersecurity guidelines, and academic articles—were analysed using thematic coding. The findings reveal significant misalignments between current frameworks and academic environments, particularly in terms of complexity, adaptability, and resource demand. Based on these insights, a tailored cybersecurity model is proposed. The model emphasises modularity, cultural integration, resource optimisation, and decentralised implementation to suit the educational context. A multi-step validation plan is also outlined to assess the model's practicality. This research offers both theoretical and practical contributions to cybersecurity governance in the education sector.**

***Keywords*—Cybersecurity; educational institutions; cybersecurity frameworks; tailored security model**

I. Introduction

Cybersecurity is increasingly critical for educational institutions as they rely heavily on digital technologies for managing student data and facilitating learning. The shift towards digital transformation, accelerated by the COVID-19 pandemic, has heightened vulnerabilities to cyberattacks, including ransomware incidents that disrupt educational processes and compromise sensitive data [1], [2], [26]. Integrating digital tools in education must be accompanied by robust cybersecurity measures to protect against threats such as data breaches and privacy violations [3], [4]. Furthermore, educational leaders play a vital role in fostering a culture of cybersecurity awareness among staff and students, ensuring they are equipped to navigate the digital landscape safely [5], [6]. Comprehensive training and resources are paramount, as many educators may lack the necessary skills to manage cybersecurity risks effectively [7], [8].

The cybersecurity landscape within educational institutions is fraught with challenges, despite an increasing recognition of its importance. Many institutions, from primary schools to universities, face significant deficiencies in their cybersecurity frameworks. Key factors contributing to this vulnerability include budget constraints, which limit the ability to invest in robust security systems, and a pervasive lack of awareness regarding the risks associated with cyber threats [1], [2]. Furthermore, insufficient resources hinder the implementation of adequate security measures, leaving institutions exposed to potential cyberattacks, such as ransomware incidents that have disrupted educational processes and compromised sensitive data [3], [4]. The reliance on digital technology for managing student information and internal communications amplifies these risks, necessitating a proactive approach to cybersecurity that includes training, awareness campaigns, and investment in protective technologies [5],[6]. Addressing these challenges is critical for safeguarding the integrity of educational environments in an increasingly digital world.

Cybersecurity frameworks such as the NIST Cybersecurity Framework and ISO/IEC 27001 have been widely adopted across various sectors to safeguard data and infrastructure against cyber threats. However, their application within educational institutions often reveals significant shortcomings, primarily due to the unique environment of openness and collaboration, which simultaneously increases cyberattack vulnerabilities [9]. This necessitates the development of a tailored security model specifically designed for the educational sector, addressing its distinct needs and challenges [10].

A comparative analysis of existing frameworks indicates that while they provide foundational guidelines, they may not fully encompass the specific operational realities of educational institutions [11]. For instance, integrating cybersecurity education into curricula is essential for fostering awareness among students and staff, yet many institutions struggle with inadequate resources and training programs [12], [13], [47]. By identifying the weaknesses in current frameworks and proposing a more suitable model, this research aims to enhance the cybersecurity posture of educational institutions, ultimately contributing to more effective cybersecurity policies and practices [14]. The findings are expected to serve as a valuable resource for policymakers and practitioners in designing targeted security measures that align with the unique demands of the education sector [15].

As digital transformation accelerates in the education sector, the danger of cyberattacks, such as ransomware and data breaches, becomes increasingly evident. Educational institutions, from K-12 schools to major universities, often lack

the technical, financial, and organisational capacity to fully implement standard cybersecurity frameworks. Instead, they need flexible solutions that suit their distinct operational structures and resource constraints.

This study critically examines three widely adopted frameworks: NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT, assessing their suitability for educational settings. Based on an extensive literature review and qualitative document analysis, the study introduces a customised cybersecurity model. Consequently, this research contributes to both academic and practical fields by proposing a tailored cybersecurity framework suited to the educational environment.

The remainder of this study is organised as follows:

Section II reviews the literature on existing cybersecurity frameworks—particularly NIST, ISO/IEC 27001, and COBIT—and discusses their applicability and limitations within educational institutions.

Section III describes the research methodology, detailing the qualitative document analysis approach and data sources used to evaluate the frameworks

Section IV presents the comparative analysis results, identifies key implementation challenges, and synthesises findings to support the development of a tailored model

Section V proposes the customised cybersecurity model, explains its five core components, and discusses its practical implications for educational policy and governance.

Section VI concludes the study by summarising the main contributions, outlining limitations, and offering recommendations for future research and validation.

II. Literature Review

A. Overview of Cybersecurity Frameworks

Cybersecurity frameworks such as the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT offer comprehensive guidelines for managing cybersecurity risks. These frameworks are frequently utilised in corporate sectors but encounter difficulties in educational settings.

The NIST Cybersecurity Framework is built around five core functions: Identify, Protect, Detect, Respond, and Recover. This framework enables organisations to comprehensively understand their cybersecurity posture and effectively implement strategies to manage risks [1].

ISO/IEC 27001 focuses on information security management, systematically safeguarding sensitive organisational information. It establishes requirements for an Information Security Management System (ISMS), ensuring that information security is integrated into the organisation's overall management processes [2].

COBIT (Control Objectives for Information and Related Technologies) provides a framework for IT governance and management. It ensures that IT risks are effectively managed while supporting the achievement of organisational goals. It emphasises aligning IT and business objectives, enhancing overall governance [3].

These frameworks collectively contribute to a robust cybersecurity strategy, enabling organisations to navigate the complexities of the digital landscape.

B. Application in Educational Institutions

Implementing cybersecurity frameworks in educational institutions presents unique challenges due to their inherent culture of openness and collaboration. While frameworks like the NIST Cybersecurity Framework and ISO/IEC 27001 are widely adopted across various sectors, their application in academia often encounters difficulties. For example, the NIST Cybersecurity Framework, which emphasises five core functions—Identify, Protect, Detect, Respond, and Recover—has proven valuable in raising awareness and improving responses to cyber incidents in educational environments. However, it often falls short in addressing specific academic needs, such as protecting sensitive research data and ensuring student privacy [16],[17].

Similarly, while ISO/IEC 27001 offers a strong framework for information security management, its implementation in educational institutions can be limited by resource shortages and a lack of technical skills among staff [17]. Comparable challenges have been seen in non-educational sectors trying hybrid models [34]. Open access policies and looser controls over campus devices further increase vulnerabilities, making it easier for cybercriminals to target these weaknesses [18],[19]. Consequently, there is an urgent need for a customised cybersecurity model that reflects the specific operational realities of educational institutions, improving their capacity to safeguard sensitive data and counter cyber threats effectively [20],[21],[32].

Although these frameworks are beneficial, their inflexible structures and significant resource demands make them challenging to implement in schools and universities with limited budgets and decentralised systems.

C. Challenges and Gaps

Implementing cybersecurity frameworks in educational institutions, especially at the K-12 level, encounters significant challenges mainly due to resource limitations and cultural resistance to change. Many institutions operate with tight budgets that restrict their ability to adopt and implement all recommended controls from established frameworks like the NIST Cybersecurity Framework and ISO/IEC 27001. This financial constraint is further compounded by a shortage of IT staff with specialised cybersecurity skills, which can impede the practical application of these frameworks, even when they are adopted.

Cultural resistance to change further complicates the situation. Educational institutions often emphasise academic freedom and open access, which can conflict with strict cybersecurity measures that require tight access controls and monitoring of system usage. This tension between the need for strong cybersecurity and the desire to maintain an open, collaborative academic environment creates a challenging landscape for administrators responsible for protecting digital

infrastructure. Therefore, addressing these challenges demands a nuanced approach that balances security requirements with the core values of educational institutions.

Key barriers include:

- High complexity and implementation cost, and inconsistent use of standard risk analysis methods across institutions [27].
- Misalignment with the academic culture
- Shortage of skilled IT personnel
- Decentralised decision-making structures

D. Previous Research Findings

Previous studies highlight the necessity for a more flexible, education-specific cybersecurity framework. However, few have presented practical models customised for the educational setting.

Previous research emphasises an urgent need to adapt established cybersecurity frameworks such as NIST and ISO/IEC 27001 to better suit educational contexts. Studies show that many academic institutions often adopt only parts of these frameworks due to budget limitations and a lack of specialised expertise [22]. This partial implementation highlights the necessity for a more flexible and adaptable framework that can be customised to the specific resources and needs of educational settings [23].

Furthermore, research highlights the advantages of developing a cybersecurity framework specifically designed for educational institutions. Such a customised framework can better address unique challenges, such as protecting student data and maintaining academic freedom, compared to the generic "one-size-fits-all" approaches of existing frameworks [24]. By focusing on the specific characteristics of educational institutions, a personalised framework could enhance cybersecurity preparedness and resilience, ultimately offering stronger protection against cyber threats [25].

III. Research Methodology

A. Research Design

The research methodology for this study adopts a qualitative approach, primarily using document analysis as the main method for data collection and analysis. This approach effectively investigates how various cybersecurity frameworks are adopted within educational institutions, enabling a detailed examination of the challenges and weaknesses faced during their implementation. Document analysis allows for the review of published policies, research reports, and relevant case studies, providing valuable insights into current practices and highlighting potential areas for further improvement.

Previous research has underscored the significance of customising cybersecurity frameworks to fit the particular context of educational institutions. For example, it highlights the need to understand the components of cybersecurity standards and frameworks to guide future research. Furthermore, it reviews the key factors that influence the implementation of information security management systems, which are important when assessing educational settings. Analysing these documents helps to identify the unique challenges faced by educational institutions in adopting frameworks such as NIST and ISO/IEC 27001, ultimately assisting in the development of a tailored cybersecurity model that addresses their specific needs.

B. Data Collection Method: Document Analysis

Document analysis serves as the primary data collection method for this study. The documents analysed include various sources such as:

- Cybersecurity Policy Reports: Reports published by educational institutions, government agencies, and non-governmental organisations focusing on implementing cybersecurity in the education sector.
- Case Studies: Published case studies on the implementation of cybersecurity frameworks in various educational institutions worldwide, providing empirical data on the effectiveness and challenges of these frameworks.
- Journal and Conference Articles: Peer-reviewed academic articles discussing the implementation of cybersecurity frameworks in educational settings and providing critical assessments of their suitability, as well as reviews of ICT tools used in higher education [46].
- Cybersecurity Standards and Guidelines: Official documents detailing standards and guidelines such as the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT and how these standards are adapted or used in educational institutions.

The data collection process through document analysis involves several key steps:

- Identification of Documents: Relevant documents were identified through searches in academic databases such as Scopus, IEEE Xplore, and Google Scholar, as well as through the websites of related organisations like NIST and ISO.
- Document Selection: Documents were selected based on relevance, reliability, and coverage. Documents specifically focused on the education sector or providing a critical evaluation of cybersecurity framework implementation were considered primary sources.
- Data Extraction: Key information from each document was extracted and organised based on the identified themes, such as framework effectiveness, implementation challenges, and adaptation for educational environments.

A total of 42 documents were examined in this study, chosen for their relevance, reliability, and direct link to cybersecurity in educational institutions. The distribution of documents is presented in Table I.

TABLE I. DOCUMENT AND SELECTION OVERVIEW

Document Type	Number	Source	Selection Criteria
Cybersecurity Policy Reports	10	Institutional & government sites	Targeting education sector; policy focus
Case Studies	8	Journals, conferences	Practical application of NIST, ISO/IEC 27001, or COBIT
Journal Articles	20	Scopus, IEEE Xplore, Google Scholar	Peer-reviewed, published 2018 or later
Framework Guidelines	4	NIST, ISO, ISACA	Descriptions of CSF, ISMS, and COBIT frameworks

C. Data Analysis

Document analysis was performed manually employing thematic coding grounded in principles of grounded theory. The procedure involved these steps:

1) Initial reading: Documents were read line by line to gain familiarity.

2) Open coding: Keywords and key ideas were extracted manually using Excel spreadsheets (no qualitative software used).

3) Axial coding: Categories were formed by linking similar open codes (e.g., "resource constraints", "training gaps").

4) Selective coding: Central themes such as "framework limitations" and "cultural mismatch" were identified.

The coding process and resulting categories are illustrated in Table II.

TABLE II. EXAMPLES OF CODE CATEGORIES AND THEMES

Open Code	Axial Category	Selective Theme
High cost of ISO adoption	Resource Limitation	Framework Limitations
Lack of IT staff	Resource Limitation	Framework Limitations
Decentralized departments	Governance Challenges	Cultural Misalignment
Need for open access	Academic Openness	Cultural Misalignment

This analysis provides a strong foundation for comparing different cybersecurity frameworks and evaluating their suitability for educational institutions. The results will be used to support the argument that a tailored cybersecurity model is explicitly needed for the education sector.

D. Research Workflow

The research process is summarized in Fig. 1, which illustrates the flow from literature review to framework synthesis and model development.

Fig. 1 illustrates the structured research design employed in this study, beginning with a literature review phase aimed at identifying existing cybersecurity frameworks and their known limitations in educational settings. This crucial step informed the subsequent document analysis, during which a selection of 42 documents, including policies, case studies, and academic research, were systematically examined using thematic coding techniques.

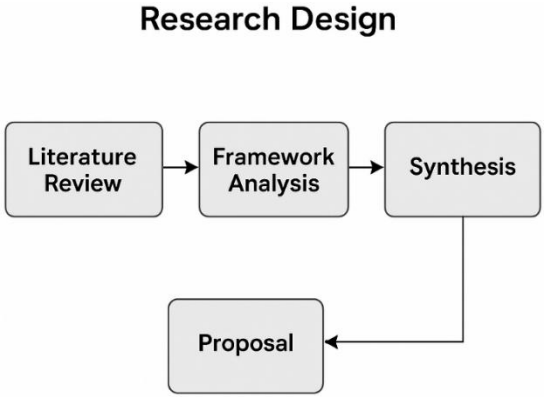


Fig. 1. Research design.

The insights gained from document analysis were then combined to identify common challenges and individual institutional needs. These findings were compared with key frameworks (NIST, ISO/IEC 27001, and COBIT) to highlight areas of misalignment, as shown in the figure.

Following this, the study moved on to the framework synthesis phase, where key insights were turned into a conceptual model specifically designed for the education sector. This iterative design phase focused on flexibility, contextual relevance, and feasibility.

Finally, the model was prepared for validation planning, including designing a pilot implementation strategy, conducting expert interviews, and performing document audits, ensuring the framework's applicability could be tested in real-world educational contexts.

In essence, Fig. 1 serves as a visual guide to the research process, emphasising the logical sequence and interaction between each phase from initial inquiry to framework development and planned validation.

IV. RESULTS

A. Comparative Analysis of Frameworks

The strengths and limitations of NIST, ISO/IEC 27001, and COBIT frameworks are compared in Table III, based on six key criteria relevant to the educational context.

TABLE III. COMPARATIVE ANALYSIS OF FRAMEWORKS

Criteria	NIST CSF	ISO/IEC 27001	COBIT
Scope	Risk management	ISMS	IT Governance
Complexity	High	High	Moderate
Suitability for Education	Partial	Partial	Limited
Adaptability	Flexible but needs expertise	Rigid	Moderately adaptable
Resource Requirement	High	High	Moderate
Academic Alignment	Limited	Limited	Limited

After conducting document analysis, including various cybersecurity frameworks such as the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT, we found that while these frameworks offer comprehensive approaches to

cybersecurity management, they often do not fully align with educational institutions' specific needs.

1) NIST Cybersecurity framework: This framework helps organisations understand and manage cybersecurity risks through five core functions: Identify, Protect, Detect, Respond, and Recover. However, it is sometimes too complex to fully implement in the educational context, particularly in institutions with limited resources like K-12 schools. Reports show that many academic institutions can only implement parts of this framework, especially those requiring lower technical and financial investment.

2) ISO/IEC 27001: This framework provides a highly structured approach to information security management, focusing on developing and maintaining an Information Security Management System (ISMS). Although it has been successfully adopted in various sectors, challenges occur when used in educational institutions, where decentralisation and a lack of specialised expertise in information security management pose significant barriers. Research also indicates that implementing ISO/IEC 27001 in the education sector often requires considerable adaptation, as not all recommended controls and policies are relevant or practical for academic environments.

3) COBIT: Focusing on IT governance, COBIT offers a useful framework for ensuring that IT risks are managed effectively and support the organisation's strategic aims. However, when applied in educational institutions, such as large universities with complex organisational structures, this framework often encounters difficulties in enforcing uniform policies across various departments and faculties. A study found that while COBIT provides strong guidance for IT governance, its implementation in academic environments is frequently hindered by resistance to change and cultural differences across departments.

B. Identification of Key Gaps

The comparative analysis reveals several key weaknesses in the existing cybersecurity frameworks when applied to educational institutions:

1) Complexity of implementation: Frameworks like NIST and ISO/IEC 27001 are often too complex to fully implement without a structured decision-making approach tailored for educational cybersecurity [28], particularly those with limited resources. These limitations mean that many schools can only adopt parts of these frameworks, ultimately leaving gaps in their protection.

2) Misalignment with academic culture: Many cybersecurity frameworks are designed for more structured and controlled corporate environments. However, educational institutions often have a more open and collaborative culture, which can conflict with the need for strict access controls and monitoring of system usage.

3) Lack of resources and expertise: Implementing frameworks like ISO/IEC 27001 requires significant human and financial resources, which educational institutions often lack. Additionally, the shortage of IT staff with specialised cybersecurity skills is a substantial barrier to practical implementation.

4) Decentralised organisational structure: Educational institutions, such as huge universities, often have decentralised organisational structures, with different departments having varying policies and technological needs. This makes it challenging to implement a uniform framework across the institution, as reaching a consensus on security policies and procedures can be difficult.

C. Synthesis of Findings

The results of this document analysis emphasise an urgent need to develop a cybersecurity model that better fits the specific requirements of educational institutions. This customised model should consider resource limitations, academic culture, and the unique organisational structures of educational institutions. Although frameworks like NIST, ISO/IEC 27001, and COBIT offer valuable guidance, there are emerging suggestions for sector-specific cybersecurity frameworks to reduce attacks [29], and further adaptation and personalisation are necessary for successful implementation in the education sector.

The synthesis of these findings suggests that a more modular and adaptable cybersecurity framework is required to address the varied needs of educational institutions. This framework should enable institutions to implement the most relevant security controls that match their capabilities, without strictly following all recommendations.

V. Discussion

A. Interpretation of Results

The comparative analysis results demonstrate that while cybersecurity frameworks such as the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT provide a strong foundation for managing cybersecurity risks, they often do not fully meet the specific needs of educational institutions. The main barriers to applying these frameworks in the education sector are the complexity of implementation, misalignment with academic culture, lack of resources, and decentralised organisational structures.

One key finding from this study is that existing frameworks are too rigid and focused on industrial sectors. A more holistic cybersecurity strategy may help bridge this contextual gap [41], [42], where operational environments tend to be more structured and hierarchical. In contrast, educational institutions have different needs, including supporting academic freedom and interdepartmental collaboration, which often conflict with strict cybersecurity principles. Therefore, a customised model is required to balance the need for stringent security with the open and flexible environment of educational institutions.

B. Proposing a Tailored Security Model

The components of the proposed model are outlined in Table IV.

TABLE IV. COMPONENTS OF THE TAILORED CYBERSECURITY MODEL

Component	Description (Concise)	Key Benefit
Modular Architecture	Flexible modules based on institutional size and context	Scalable and customisable
Cultural Integration	Support open collaboration while maintaining access control	Preserves academic freedom
Resource Optimization	Focus on cost-effective tools and strategic prioritisation	Feasible within limited budgets
Decentralized Governance	Tailored policies for departments with coordinated oversight	Compatible with the academic structure
Emerging Tech Integration	Integrate AI/IoT for proactive security posture	Future-proof and proactive

Expanded Explanation:

1) Modular and flexible structure: Educational institutions vary significantly in size, budget, and IT capacity. A one-size-fits-all cybersecurity framework [39] is neither practical nor effective. This approach encourages modularity, allowing each institution to select and implement only the components relevant to their risk profile, maturity level, and resources. Such flexibility makes the model scalable and adaptable. Interactive tools and gamified learning environments have shown potential in improving engagement [35].

2) Integration with educational culture: Academic institutions value openness, collaboration, and autonomy. A fundamental cybersecurity culture at all levels of the institution can help reinforce these values safely [49]. Therefore, the cybersecurity framework must balance protection with freedom. It promotes policies that facilitate secure departmental cooperation and integrates cybersecurity education into the institutional culture while addressing human-centric factors such as awareness and behavioural hygiene [43], which can be further supported through active learning-based security training [30], fostering awareness and behavioural change among faculty, staff, and students..

3) Resource optimisation: Recognising the budgetary and staffing constraints faced by most educational institutions, the framework encourages cost-effective practices. It promotes the use of open-source or shared security tools, centralised support services where feasible, and risk-based prioritisation to ensure limited resources are allocated to the most vulnerable or impactful areas. Adaptive training models for end-users also support cost-effective implementation [44].

4) Decentralised implementation: Given the diversity of academic departments, a centralised, rigid policy often faces opposition. Instead, this approach supports decentralised implementation within a central guiding framework, allowing departments to develop context-specific solutions while aligning with overall institutional aims. This respects local autonomy whilst maintaining institutional coherence.

5) Emerging technology integration: To stay relevant amidst the changing threat landscape, the framework includes provisions for adopting new technologies such as Artificial Intelligence, Internet of Things (IoT), and real-time analytics. These features improve early detection, response, and adaptive resilience across the institution.

Together, these components tackle the challenges identified in the previous analysis and offer a practical route to improve cybersecurity preparedness in the education sector.

To address the unique cybersecurity challenges faced by educational institutions, including structural decentralisation, limited resources, and the need to preserve academic openness, this study developed a tailored cybersecurity framework. The framework was constructed based on the synthesis of findings from comparative analysis and thematic coding, resulting in a model that is both theoretically grounded and practically applicable across various educational contexts. As shown in Fig. 2, the proposed framework comprises five key components: Modular Architecture, Cultural Integration, Resource Optimisation, Decentralised Governance, and Emerging Technology Integration. These components work together to form a cohesive and adaptable strategy designed to enhance cybersecurity readiness within the academic environment.

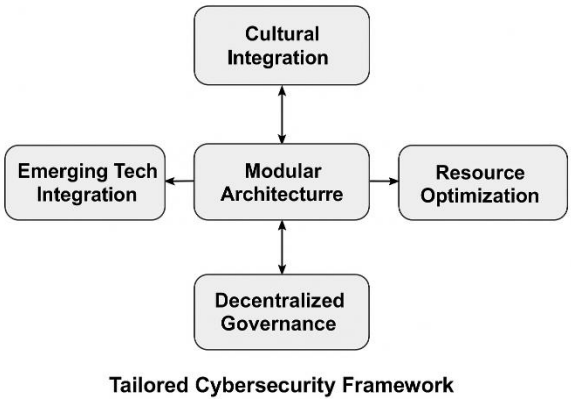


Fig. 2. Tailored cybersecurity framework for educational institutions.

C. Implications for Policy and Practice

The implications of developing a tailored cybersecurity model are significant for policy and practice in educational institutions. First, this model can guide policymakers in developing regulations and policies that are more aligned with the needs of the education sector. For example, existing regulations may need to be adjusted to provide more flexibility to educational institutions in adopting relevant security controls without burdening them with unrealistic requirements.

Second, this model can assist educational institutions in managing cybersecurity risks more effectively and positively influence their performance, as shown in regional studies [36], without compromising core values like academic freedom and collaboration. By adopting a more modular and adaptable approach, educational institutions can notably enhance their cybersecurity stance while fostering an environment that promotes learning and innovation.

D. Comparison with Existing Studies

This research aligns with previous studies that have identified the shortcomings of current cybersecurity frameworks in meeting the specific needs of the education

sector. For example, studies have shown that frameworks like NIST and ISO/IEC 27001 often fail to account for the decentralised nature and resource limitations typical of educational institutions [31],[33]. While these studies mainly examine the challenges of implementing existing frameworks, this research progresses the discussion by highlighting these weaknesses and proposing a customised cybersecurity model specifically designed for educational institutions, recognising the curriculum-level gaps still common in cybersecurity education [45].

Furthermore, existing literature emphasises adapting cybersecurity frameworks to fit the educational context better; however, it often stops short of providing actionable solutions [40], [37]. In contrast, this study offers practical guidance for developing and implementing a customised cybersecurity model, filling a critical literature gap. This research contributes significantly to the ongoing efforts to enhance cybersecurity in the education sector by addressing the challenges and providing concrete solutions.

E. Limitations and Future Research Directions

While this research provides valuable insights into the need for a tailored cybersecurity model in educational institutions, several limitations must be acknowledged. First, this research is limited to document analysis, so it does not include empirical data from implementing the proposed model. Therefore, further validation is needed through case studies or field experiments to test the effectiveness of this model in real-world contexts.

Additionally, this research focuses on educational institutions in countries with relatively advanced IT infrastructures. Future research should consider different contexts, such as academic institutions in developing countries, which may face various challenges in implementing cybersecurity.

Future research should also examine how new technologies, such as artificial intelligence (AI) and the Internet of Things (IoT), can be incorporated into a customised cybersecurity model and assist workforce development using established frameworks like NICE [38]. These technologies have significant potential to strengthen security, but also introduce new challenges that must be tackled.

VI. Conclusion and Recommendations

A. Summary of Key Findings

This research has conducted a comparative analysis of several cybersecurity frameworks used in educational institutions, including the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT. The main findings indicate that while these frameworks provide a solid foundation for cybersecurity management, they often do not fully align with educational institutions' unique needs. Implementation complexity, misalignment with academic culture, resource limitations, and decentralised organisational structures are the main challenges that hinder the practical application of existing frameworks in the education sector.

Based on the analysis results, this research proposes to develop a cybersecurity model explicitly tailored to the education sector. The proposed model emphasises flexibility, modularity, and resource optimisation, allowing educational institutions to adopt the most relevant security controls that align with their needs. This model supports the open and collaborative academic culture while maintaining high security standards.

B. Theoretical and Practical Contributions

This research makes significant contributions both theoretically and practically. Theoretically, it adds to the knowledge of the challenges and weaknesses of existing cybersecurity frameworks in the educational context. This research also offers an alternative model that is more suited to the needs of this sector, providing a foundation for further research to develop and test customised security models.

Practically, policymakers and practitioners in educational institutions can use the proposed model in this research to improve their cybersecurity posture. By adopting a more flexible and modular approach, educational institutions can more effectively manage cyber risks without compromising their core values, such as academic freedom and collaboration.

C. Limitations

While this research offers valuable insights, several limitations need to be acknowledged. This research is limited to document analysis and does not include empirical data from implementing the proposed model. Therefore, further research is required to test this model's effectiveness in real-world contexts. Additionally, this research focuses on educational institutions in countries with relatively advanced IT infrastructures so that different contexts may require adaptation of the proposed model.

D. Suggestions for Future Research

While this research provides valuable insights into the need for a customised cybersecurity model in educational institutions, further development is needed to address the sector's changing challenges. Future studies should aim to develop and empirically verify a cybersecurity framework specifically designed for educational settings. Such a framework should be adaptable and scalable, considering the diverse needs of institutions ranging from small K-12 schools to large universities. Research should investigate how this customised framework can incorporate emerging technologies like artificial intelligence (AI) and the Internet of Things (IoT) to strengthen cybersecurity without hindering academic freedom and collaboration. Additionally, the framework should be tested across different educational environments globally, including practical simulation settings like cyber ranges for training [48], to ensure its applicability and effectiveness in various cultural and organisational contexts.

Acknowledgment

This publication could not have been accomplished without the institution's extraordinary assistance. We also want to thank the reviewers whose suggestions will help tomake this manuscript eligible for publication.

References

[1] S. Saeed, "Digital transformation and cybersecurity challenges for businesses' resilience: issues and recommendations", Sensors, vol. 23, no. 15, p. 6666, 2023. https://doi.org/10.3390/s23156666

[2] J. Kumar, "Enhancing public awareness and education of ransomware attacks", 2023. https://doi.org/10.36227/techrxiv.24634806.v1

[3] S. S. Goswami, S. Sarkar, K. K. Gupta and S. Mondal "The role of cyber security in advancing sustainable digitalisation: opportunities and challenges", Journal of Decision Analytics and Intelligent Computing, vol. 3, no. 1, p. 270-285, 2023. https://doi.org/10.31181/jdaic10018122023g

[4] D. Hankerson, C. Venzke, E. Laird, H. Grant-Chapman, & D. Thakur, "Online and observed: student privacy implications of school-issued devices and student activity monitoring software", 2022. https://doi.org/10.31219/osf.io/73a2b

[5] H. Chou and C. Chou, "A multigroup analysis of factors underlying teachers' technostress and their continuance intention toward online teaching", Computers & Education, vol. 175, p. 104335, 2021. https://doi.org/10.1016/j.compedu.2021.104335

[6] K. Walker, K. Bodendorf, T. Kiesler, G. Mattos, M. Rostom, & A. Elkordy, "Compulsory technology adoption and adaptation in education: a looming student privacy problem", Journal of Consumer Affairs, vol. 57, no. 1, p. 445-478, 2023. https://doi.org/10.1111/joca.12506

[7] L. Yang, F. Martínez-Abad, & A. García-Holgado, "Exploring factors influencing pre-service and in-service teachers´ perception of digital competencies in the Chinese region of Anhui", Education and Information Technologies, vol. 27, no. 9, p. 12469-12494, 2022. https://doi.org/10.1007/s10639-022-11085-6

[8] S. Wang, "A review of research exploring pre-service teacher education for the digital era", International Journal of Education and Humanities, vol. 9, no. 1, p. 188-193, 2023. https://doi.org/10.54097/ijeh.v9i1.9407

[9] R. Chandarman and B. Niekerk, "Students' cybersecurity awareness at a private tertiary educational institution", The African Journal of Information and Communication, no. 20, 2017. https://doi.org/10.23962/10539/23572

[10] B. Blažič, "Cybersecurity skills in EU: new educational concept for closing the missing workforce gap", 2021. https://doi.org/10.5772/intechopen.97094

[11] R. Brink, J. Ophoff, & Z. Ruhwanya, "Relevant cybersecurity: curriculum guidance for the South African context", p. 3-19, 2022. https://doi.org/10.1007/978-3-031-21076-1_1

[12] R. Dontu, "An empirical paradigm on cybersecurity vulnerability mitigation framework", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9s, p. 786-792, 2023. https://doi.org/10.17762/ijritcc.v11i9s.9484

[13] N. Rahim, S. Hamid, M. Kiah, S. Shamshirband, & S. Furnell, "A systematic review of approaches to assessing cybersecurity awareness", Kybernetes, vol. 44, no. 4, p. 606-622, 2015. https://doi.org/10.1108/k-12-2014-0283

[14] Z. Ye, S. Liu, H. Zhang, & G. Wu, "Exploration and research on the training path of cybersecurity application-oriented talents by integrating industry and education", Education Journal, vol. 9, no. 5, p. 137, 2020. https://doi.org/10.11648/j.edu.20200905.13

[15] W. Triplett, "Addressing cybersecurity challenges in education", International Journal of Stem Education for Sustainability, vol. 3, no. 1, p. 47-67, 2023. https://doi.org/10.53889/ijses.v3i1.132

[16] M. Khader, M. Karam, & H. Fares, "Cybersecurity awareness framework for academia", Information, vol. 12, no. 10, p. 417, 2021. https://doi.org/10.3390/info12100417

[17] A. Alraddadi, "Developing an abstraction framework for managing and controlling Saudi banks' cybersecurity threats based on the NIST cybersecurity framework and iso/IEC 27001", Journal of Software Engineering and Applications, vol. 16, no. 12, p. 695-713, 2023. https://doi.org/10.4236/jsea.2023.1612036

[18] A. Adegbite, "Review of cybersecurity strategies in protecting national infrastructure: perspectives from the USA", Computer Science & IT Research Journal, vol. 4, no. 3, p. 200-219, 2023. https://doi.org/10.51594/csitrj.v4i3.658

[19] R. Pirta-Dreimane, A. Brilingaita, G. Majore, B. Knox, K. Lapin, K. Parish et al., "Application of intervention mapping in cybersecurity education design", Frontiers in Education, vol. 7, 2022. https://doi.org/10.3389/feduc.2022.998335

[20] Z. Ye, S. Liu, H. Zhang, & G. Wu, "Exploration and research on the training path of cybersecurity application-oriented talents by integrating industry and education", Education Journal, vol. 9, no. 5, p. 137, 2020. https://doi.org/10.11648/j.edu.20200905.13

[21] D. Onayemi, "Enhancing academic cybersecurity: integrated framework with network penetration testing", Social Science and Humanities Journal, vol. 7, no. 10, p. 3231-3245, 2023. https://doi.org/10.18535/sshj.v7i10.875

[22] B. Blažič, "Cybersecurity skills in EU: new educational concept for closing the missing workforce gap", 2021. https://doi.org/10.5772/intechopen.97094

[23] Sadiq Nasir "Exploring the effectiveness of cybersecurity training programs: factors, best practices, and future directions", Advances in Multidisciplinary & Scientific Research Journal Publication, vol. 2, no. 1, p. 151-160, 2023. https://doi.org/10.22624/aims/csean-smart2023p18

[24] W. Triplett, "Addressing cybersecurity challenges in education", International Journal of Stem Education for Sustainability, vol. 3, no. 1, p. 47-67, 2023. https://doi.org/10.53889/ijses.v3i1.132

[25] R. Pirta-Dreimane, A. Brilingaita, G. Majore, B. Knox, K. Lapin, K. Parish et al., "Application of intervention mapping in cybersecurity education design", Frontiers in Education, vol. 7, 2022. https://doi.org/10.3389/feduc.2022.998335

[26] S. Hakak, W. Khan, M. Imran, K. Choo, & M. Shoaib, "Have you been a victim of COVID-19-related cyber incidents? Survey, taxonomy, and mitigation strategies", IEEE Access, vol. 8, p. 124134-124144, 2020. https://doi.org/10.1109/access.2020.3006172

[27] V. Agrawal, "A comparative study on information security risk analysis methods", Journal of Computers, p. 57-67, 2017. https://doi.org/10.17706/jcp.12.1.57-67

[28] R. Goel, A. Kumar, & J. Haddow, "Prism: a strategic decision framework for cybersecurity risk assessment", Information and Computer Security, vol. 28, no. 4, p. 591-625, 2020. https://doi.org/10.1108/ics-11-2018-0131

[29] A. Abohatem, "Suggestion cybersecurity framework (CSF) for reducing cyber-attacks on information systems", مجلة جامعة صنعاء لعلوم الحاسوب والتكنولوجيا التطبيقية, vol. 1, no. 3, 2023. https://doi.org/10.59628/jast.v1i3.248

[30] T. Srivatanakul and F. Annansingh, "Incorporating active learning activities to the design and development of an undergraduate software and web security course", Journal of Computers in Education, vol. 9, no. 1, p. 25-50, 2021. https://doi.org/10.1007/s40692-021-00194-9

[31] A. Ghazvini, Z. Shukur, & Z. Hood, "Review of information security policy based on content coverage and online presentation in higher education", International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, 2018. https://doi.org/10.14569/ijacsa.2018.090853

[32] Z. Ye, S. Liu, H. Zhang, & G. Wu, "Exploration and research on the training path of cybersecurity application-oriented talents by integrating industry and education", Education Journal, vol. 9, no. 5, p. 137, 2020. https://doi.org/10.11648/j.edu.20200905.13

[33] A. Dedeke and K. Masterson, "Contrasting cybersecurity implementation frameworks (CIF) from three countries", Information and Computer Security, vol. 27, no. 3, p. 373-392, 2019. https://doi.org/10.1108/ics-10-2018-0122

[34] A. Alraddadi, "Developing an abstraction framework for managing and controlling Saudi banks' cybersecurity threats based on the NIST cybersecurity framework and iso/IEC 27001", Journal of Software Engineering and Applications, vol. 16, no. 12, p. 695-713, 2023. https://doi.org/10.4236/jsea.2023.1612036

[35] M. Katsantonis and I. Mavridis, "Evaluation of hacklearn cofelet game user experience for cybersecurity education", International Journal of Serious Games, vol. 8, no. 3, p. 3-24, 2021. https://doi.org/10.17083/ijsg.v8i3.437

[36] B. M. Dioubate, W. D. W. Norhayate, Z. F. Anwar, S. Fauzilah .et. al. "The role of cybersecurity on the performance of Malaysian higher education institutions", Jurnal Pengurusan, vol. 67, 2023. https://doi.org/10.17576/pengurusan-2023-67-03

[37] R. Azmi, W. Tibben, & K. Win, "Review of cybersecurity frameworks: context and shared concepts", Journal of Cyber Policy, vol. 3, no. 2, p. 258-283, 2018. https://doi.org/10.1080/23738871.2018.1520271

[38] R. Petersen, D. Santos, M. Smith, & G. Witte, "Workforce framework for cybersecurity (nice framework)", 2020. https://doi.org/10.6028/nist.sp.800-181r1

[39] O. Folorunsho, A. Ayinde, M. A. Olagoke, O. E. Fatoye "Evaluating cybersecurity theories, models, standards and frameworks", Advances in Multidisciplinary & Scientific Research Journal Publication, vol. 5, no. 4, p. 61-66, 2019. https://doi.org/10.22624/aims/bhi/v5n4p7

[40] J. Hajný, S. Ricci, E. Piesarskas, O. Levillain, L. Galletta, & R. Nicola, "Framework, tools and good practices for cybersecurity curricula", IEEE Access, vol. 9, p. 94723-94747, 2021. https://doi.org/10.1109/access.2021.3093952

[41] R. Pirta-Dreimane, A. Brilingaita, G. Majore, B. Knox, K. Lapin, K. Parish et al., "Application of intervention mapping in cybersecurity education design", Frontiers in Education, vol. 7, 2022. https://doi.org/10.3389/feduc.2022.998335

[42] M. Dawson, "Applying a holistic cybersecurity framework for global IT organisations", Business Information Review, vol. 35, no. 2, p. 60-67, 2018. https://doi.org/10.1177/0266382118773624

[43] K. Kioskli, T. Fotis, S. Nifakos, & H. Mouratidis, "The importance of conceptualising the human-centric approach in maintaining and promoting cybersecurity-hygiene in healthcare 4.0", Applied Sciences, vol. 13, no. 6, p. 3410, 2023. https://doi.org/10.3390/app13063410

[44] F. Salamah, "An adaptive cybersecurity training framework for the education of social media users at work", Applied Sciences, vol. 13, no. 17, p. 9595, 2023. https://doi.org/10.3390/app13179595

[45] I. Ngambeki, S. McBride, & J. Slay, "Knowledge gaps in curricular guidance for ics security", Journal of the Colloquium for Information Systems Security Education, vol. 9, no. 1, p. 6, 2022. https://doi.org/10.53735/cisse.v9i1.149

[46] M. Núñez, X. Palmer, L. Potter, C. Aliac, & L. Velasco, "ICT security tools and techniques among higher education institutions: a critical review", International Journal of Emerging Technologies in Learning (IJET), vol. 18, no. 15, p. 4-22, 2023. https://doi.org/10.3991/ijet.v18i15.40673

[47] N. Rahim, S. Hamid, M. Kiah, S. Shamshirband, & S. Furnell, "A systematic review of approaches to assessing cybersecurity awareness", Kybernetes, vol. 44, no. 4, p. 606-622, 2015. https://doi.org/10.1108/k-12-2014-0283

[48] S. Karagiannis, E. Magkos, E. Karavaras, A. Karnavas, M. Nikiforos, & C. Ntantogian, "Towards nice-by-design cybersecurity learning environments: a cyber range for SOC teams", 2022. https://doi.org/10.21203/rs.3.rs-1902186/v1

[49] G. Kabanda and T. Chingoriwo, "A cybersecurity culture framework for grassroots levels in Zimbabwe", Oriental Journal of Computer Science and Technology, vol. 14, no. 010203, p. 17-34, 2022. https://doi.org/10.13005/ojcst14.010203.03

The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data

Jinzhu Lin*, Tianwei Ni
School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, China

***Abstract*—Many existing systems struggle to strike a balance between global feature discrimination and local semantic understanding, despite the growing popularity of Self-Supervised Learning (SSL) for representation learning with unlabeled image data. This study introduces a novel SSL framework—Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)—which integrates contrastive learning mechanisms with auxiliary context-based pretext tasks, specifically rotation prediction and jigsaw puzzle solving. The proposed C2SRL enhances two leading constructive models, SimCLR and MoCo, by incorporating contextual modules and a unified multi-task loss function, thereby improving the robustness and generalizability of the learned representations. A lightweight ResNet backbone is employed for encoding, followed by a dual-view augmentation strategy and a projection head that maps features into a contrastive embedding space. The proposed C2SRL outperforms existing SSL approaches in terms of classification accuracy and clustering coherence on the STL-10 and CIFAR-10 datasets, two benchmark datasets. It demonstrates strong scalability, as evidenced by its 89.6% mAP and 0.81 NMI, achieved using only 10% labeled data for fine-tuning. These results highlight the potential of combining contextual and contrastive learning objectives to generate rich, transferable visual representations for low-label or label-free applications.**

Keywords*—*Self-supervised learning (SSL); unlabeled image data; representation learning; contrastive learning; convolutional neural network (CNN); image classification; feature embedding; label-efficient learning

I. Introduction

A. Background and Motivation

Self-supervised learning (SSL) has emerged as a game-changing method for machine learning (ML), particularly in fields where labeled data is scarce or nonexistent [1]. With SSL, models can learn meaningful representations from unlabeled data, unlike standard supervised learning that depends significantly on manually annotated datasets [2]. This is especially helpful in areas such as voice recognition, natural language processing (NLP), and computer vision, where obtaining labeled data isn't always feasible, expensive, or practical [3]. One significant benefit of SSL is that it can utilize pretextual jobs to generate supervisory signals directly from the data, allowing it to extract high-level characteristics [4]. Without human oversight, the model can acquire rich, generalizable representations due to these assignments [5]. Many currently consider SSL an effective method for developing scalable models that can utilize what they've learned for subsequent tasks, such as segmentation, object identification, and image classification [6]. An increasing number of real-world applications have found that obtaining labeled data is a significant challenge, and SSL provides a possible solution for model creation in these situations [7]. For example, specific fields include medical imaging, autonomous driving, and surveillance, where annotating data would be impractical or expensive [8].

*Corresponding Author

B. Problem Statement

Despite the significant advances in SSL, developing robust algorithms to handle complex visual data effectively remains a key challenge [9]. To achieve existing performance, traditional deep learning (DL) models, such as CNNs, often require massive labeled datasets [10]. Unlabeled data poses a significant challenge for these models when generalizing to real-world problems, as it is difficult to extract discriminative features [11]. The absence of supervisory signals is the primary obstacle in SSL, as it hinders models' ability to acquire valuable representations [12]. The use of positive and negative pairings for learning representations has shown promise in contrastive learning-based techniques (e.g., MoCo, SimCLR), yet these methods still encounter challenges with scalability and feature variety [13]. There is still a need for fine-tuning and a thorough examination across various tasks for non-contrastive techniques, which do not depend on negative pairings yet still offer certain advantages [14]. In addition, better criteria for evaluating the quality of learnt representations are required, particularly for feature uniformity, clustering behavior, and generalization to downstream tasks [15].

C. Motivation for the Proposed Framework

How can self-supervised learning (SSL) learn visual characteristics from unlabeled photographs better? Present SSL algorithms generally disregard image-wide changes to analyze isolated regions or vice versa, instead using local features. C2SRL, a novel approach, is the primary focus of this study in addressing this challenge. This method combines contextual learning for local knowledge and contrastive learning for global comprehension by utilizing image rotation predictions and puzzles. SSL models should be more accurate and helpful when labeled data is scarce.

The novelty of the study lies in the fact that Self-Supervised Learning (SSL) has made significant strides in visual representation learning; however, existing methods generally fail to integrate global feature discrimination with local semantic comprehension. Most modern models employ contrastive aims, which overlook fine-grained picture context in favor of instance-level differences, particularly in the cases of SimCLR and MoCo. They struggle with spatial awareness and structural coherence tests due to this deficiency. Although some have

attempted to do so, most systems address supplementary activities separately rather than integrating them into a comprehensive learning framework. Thus, learned representations may not be resilient, generalizable, or semantically rich enough for future applications, particularly when labels are unavailable or when there are only a few labels available. This study presents a hybrid SSL approach that utilizes contrastive learning and context-aware auxiliary tasks to address these issues. The model optimizes many tasks. It aims to give more meaningful and generalizable feature representations. Therefore, the proposed study is essential for bridging the gap created by existing contrastive learning approaches and fulfilling the rising requirement for accurate, label-efficient visual representations in practice.

D. Objectives and Scope

The primary objectives of this research are:

- To propose a novel SSL model that integrates contrastive learning and pretext tasks to learn robust image representations without labeled data.
- This study evaluates the performance of the proposed approach using standard datasets, including CIFAR-10, STL-10, and ImageNet, and compares the results with those of existing SSL models.
- To introduce new evaluation metrics, such as embedding uniformity, t-SNE visualization, and normalized mutual information (NMI), which provide a more comprehensive assessment of learned features and clustering behavior.

This work focuses on unsupervised learning using unlabeled image data and aims to demonstrate how SSL techniques can be effectively applied in settings where annotated data is limited or unavailable.

E. Contributions of the Study

The contributions of this research are as follows:

- Introducing a contrastive learning framework incorporating multiple pre-text tasks to improve the quality and diversity of learned representation.
- Evaluating the proposed Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) framework across several standard image datasets, using both traditional metrics (e.g., accuracy) and novel evaluation techniques (e.g., embedding uniformity score).
- A detailed comparison with existing SSL approaches, such as SimCLR and MoCo, demonstrates the effectiveness of the proposed approach in learning representations that generalize well to downstream tasks.

F. Structure of the Study

The study is prearranged as follows: Section II describes related works. Section III describes the suggested C2SRL model. Section IV offers experimental outcomes. Section V presents the discussion. Finally, Section VI concludes the study by discussing potential future work.

II. Literature Survey

Banafshe Felfeliyan et al. [16] suggested the Mask-Region-based Convolutional Neural Network (MRCNN) for Medical Image Segmentation with Limited Data Annotation. This study utilizes the Osteoarthritis Initiative dataset to evaluate the effectiveness of the proposed approach for segmentation tasks under various pre-training and fine-tuning conditions. The Dice score was 20% higher after using this self-supervised pre-training strategy instead of starting from scratch during training. Anomaly detection, segmentation, and classification are just a few examples of medical image analysis tasks that may benefit from the proposed SSL. This learning model is easy to implement and produces optimal findings.

Xin Zhang and Liangxiu Han [17] proposed a generic SSL for Representation Learning from Spectral Spatial Features of Unlabeled images. Innovative pretext problems for object- or pixel-based remote sensing data interpretation systems are planned. One pretext task can retrieve spectral characteristics from masked data. This allows pixel data extraction and activity acceleration via pixel-based analysis. Two popular downstream task evaluation activities show how the SSL approach learns a target representation from vast volumes of unlabeled spatial and spectral data.

Soroosh Tayebi Arasteh et al. [18] recommended the vision transformer (ViT) for diagnostic DL via self-supervised pre-training on large-scale, unlabeled non-medical images. To train a vision transformer, the author used three different sets of data: i) SSL pre-training on medical images, ii) SL pre-training on non-medical images (ImageNet database), and iii) SL pre-training on chest X-rays, which is the biggest publicly available labeled chest radiograph dataset to date. Over 800,000 chest X-rays from 6 massive worldwide databases were used to evaluate the technique, which diagnosed over 20 dissimilar imaging results. Statistical significance was assessed using bootstrapping, and performance was measured by computing the area under the ROC curve. Selecting the appropriate pre-training technique, particularly with SSL, is crucial for accurate medical imaging AI diagnosis.

Jiahe Shi et al. [19] discussed the Self-supervised On-device Federated Learning (SSL-OD-FL) from Unlabeled Streams. Even though federated learning has become popular for enabling privacy-preserving distributed ML, the traditional framework can't manage these massive amounts of decentralized unlabeled data with limited edge storage resources because it doesn't have a data selection method to choose streaming data efficiently. Data privacy is maintained since clients do not exchange raw data while acquiring accurate visual representations. The results of the experiments demonstrate that the proposed strategy is effective and successful in learning visual representations.

Chen Zhang et al. [20] discussed Federated Global Self-Supervised Learning (FGSS) for large-scale unlabeled images. The author devised an accumulation technique that takes into account the fact that every customer's local data is unique by adjusting the weight of each local model according to the size of its dataset and the frequency of its contacts. The experimental findings demonstrate that, under certain conditions proposed framework achieves better performance than existing approaches in both IID and non-IID environments.

M.A.F. Abdollah et al. [21] presented a Transformer encoder-based SSL approach for HVAC fault recognition using unlabeled images. The two-state Markov chain method deliberately hides parts of the multivariate time-series information. Predicting these hidden parts trains the model. This method offers a scalable solution for real-world HVAC applications that is not reliant on labeled data. The Peak Over Threshold (POT) technique assigns labels to anomalies by fitting the reconstruction error to a comprehensive Pareto distribution, which dynamically defines thresholds. The model's capacity to identify both sequential and individual errors is shown. A failure period was identified from October 19th to December 23rd due to a change in the data trend observed by the monitoring system.

Depeng Kong et al. [22] introduced the contrastive learning-based knowledge transfer technique (CLTrans) for semi-supervised fault analysis. Using unsupervised similarity matching on massive amounts of unlabeled data, CLTrans improves downstream tasks. A CLTrans-pre-trained feature encoder can effectively adapt to varied tasks, regardless of the data distribution, and extract a discriminative representation of the vibration signal. Experimental findings show that CLTrans beats traditional DL and existing semi-supervised fault diagnostic methods in terms of accuracy and domain adaptability, particularly when working with restricted labels. Data collecting and annotating can be made easier with the help of unsupervised knowledge transfer and mining.

Zhonglin Zuo et al. [23] examined an unlabeled multi-class non-leak data system for autonomously identifying leaks in natural gas collecting pipelines. The representation learning of the semi-supervised model is enhanced by the suggested SSL approach, and unlabeled multi-class non-leak data is modeled using the supplied multi-sphere support vector data description. Through the integration of feature clustering and pseudo-label-based classification, the ability to learn unsupervised multi-class non-leakage information categories is made possible. Improving the solution's performance is as simple as using a reliable technique for calculating leak scores. Finally, the experimental findings using pipeline field data demonstrate that the proposed strategy is effective.

Most current methods focus on context-based tasks or contrastive learning alone, overlooking the potential synergistic advantages of combining the two paradigms, despite SSL having made significant progress with models like MoCo and SimCLR. Much previous work overlooks generalizability to downstream tasks without supervision or resilience across various augmentation contexts, instead focusing on the quality of representation. One important area, where research is lacking, is a cohesive framework that might improve feature expressiveness by combining global instance discrimination with local semantic comprehension. To address this, the Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) model employs a hybrid learning approach that integrates context-based auxiliary tasks, such as jigsaw solving and rotation prediction, into a multi-task optimization framework. This model aims to close the gap between the two approaches. Due to this integration, the learned representations become more flexible and robust in terms of semantic richness and structural coherence. To achieve better results on downstream picture interpretation tasks, even in situations with little labeled data, the C2SRL model's unique dual-focus design combines global contrastive goals with fine-grained contextual cues.

III. Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)

The capability to learn visual representations from unlabeled image data using pretext tasks, such as transformation prediction and instance discrimination, has been demonstrated by existing self-supervised methods. Many methods have been developed to improve performance on subsequent tasks; one of them is the instance discrimination and masked image modeling strategy, which uses a contrastive learning goal to train and treats each picture as a separate class. There is a significant data gap between this achievement and real-world data for future purposes, including city sceneries or crowd scenes, as it relies on the carefully selected object-centric dataset ImageNet. Without understanding the scene's fundamental architecture—its numerous objects and intricate layouts—instance discrimination pretext would severely limit the use of scene-centric data for pre-training. Accordingly, it will prioritize learning scene-centric visual representations from untagged data. Two major schools of thought have emerged in recent years to address this question. Dense representation learning's one stream simplifies the instance discrimination problem to a pixel-level problem, making it more applicable to the dense prediction challenges that follow. However, these approaches are still unable to learn representations because they cannot replicate the object-level interactions observed in scene-centric data. Unsupervised clustering, saliency estimators, unsupervised object proposal algorithms, and handcrafted segmentation algorithms rely on domain-specific priors for object identification. However, there is another line of study that attempts to accomplish object-level representation learning.

Fig. 1 shows the proposed C2SRL Model. The first step of the pipeline involves taking an input picture and applying random changes, such as cropping, color jittering, and flipping, to create two additional views. Following the passage of these views through a common encoder network, typically a convolutional neural network such as ResNet, a projection head is used to convert the high-dimensional features into a lower-dimensional embedding space that is optimal for computing contrastive loss. Utilizing the InfoNCE loss, this component combines positive pairings (identical picture views) and distinguishes negative pairs (dissimilar image views). Rotation prediction and jigsaw puzzle solving are context-aware auxiliary tasks with the same encoder. With rotation prediction, this research can train a classifier to anticipate which of four predetermined angles to rotate pictures by, prompting the network to identify characteristics unique to each orientation. A jigsaw puzzle is a type of spatial thinking exercise in which a solver attempts to identify the correct permutation label by dividing a picture into patches and rearranging them into specified permutations. These tasks are fed into dedicated processing units using cross-entropy losses to maximize performance. Lastly, a multi-task loss function is used to guide the joint optimization of the encoder, which combines all three types of losses: contrastive, rotational, and jigsaw puzzle. The result is a strong, pre-trained encoder that can make sense of data

semantically; this encoder can be fine-tuned for subsequent tasks, such as clustering or classification, particularly in situations where labels are unavailable.

A. Multi-View Generation and Representation Embedding

In the first phase of C2SRL, the model processes raw, unlabeled input data $x_i \in D$ through stochastic data augmentation strategies. The aim is to produce semantically invariant yet appearance-diverse views that simulate real-world variance. The two augmentations x_i^1 and x_i^2 for each image, random transformations T_1 and T_2, which include color distortion, cropping, flipping, and Gaussian noise, as in Eq. (1).

$$x_i^{(1)}, x_i^{(2)} = T_1(x_i),\ T_2(x_i), \quad where\ T_1\ and\ T_2 \sim A \qquad (1)$$

Each augmented view is then passed through a shared convolutional encoder network $f: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$, such as ResNet-50, to extract high-level semantic features, as in Eq. (2):

$$h_i^{(1)} = f\left(x_i^{(1)}\right), \quad h_i^{(2)} = f\left(x_i^{(2)}\right) \qquad (2)$$

To reduce overfitting and enforce contrastive separation in the latent space, this research further maps these embeddings through a projection head $g: \mathbb{R}^d \to \mathbb{R}^{d'}$, often implemented as a 2-layer MLP with ReLU and BatchNorm, as in Eq. (3):

$$z_i^{(1)} = g\left(h_i^{(1)}\right), \quad z_i^{(2)} = g\left(h_i^{(2)}\right) \qquad (3)$$

Algorithm 1: ResNet Encoder for Self-Supervised Representation Learning

Input:
Augmented image view $v \in \mathbb{R}^{\{H\times W\times 3\}}$
ResNet depth: ResNet-18, ResNet-50.

Output:
Representation vector $h \in \mathbb{R}^d$

1: function ResNet_Encoder(v):
2: # Initial convolution and max pooling
3: $x \leftarrow Conv2D(v, kernel_size = 7\times 7, stride = 2, padding = 3)$
4: $x \leftarrow BatchNorm(x)$
5: $x \leftarrow ReLU(x)$
6: $x \leftarrow MaxPool2D(x, kernel_size = 3\times 3, stride = 2, padding = 1)$
7: # Residual blocks (based on depth)
8: $x \leftarrow ResBlock_Layer1(x)$ # e.g., 64 filters
9: $x \leftarrow ResBlock_Layer2(x)$ # e.g., 128 filters
10: $x \leftarrow ResBlock_Layer3(x)$ # e.g., 256 filters
11: $x \leftarrow ResBlock_Layer4(x)$ # e.g., 512 filters
12: # Global average pooling
13: $x \leftarrow GlobalAvgPool2D(x)$
14: # Flatten and normalize
15: $h \leftarrow Flatten(x)$
16: $h \leftarrow Normalize(h)$
17: return h

Algorithm 1 shows the ResNet Encoder for Self-Supervised Representation Learning. After applying domain-specific augmentations, the ResNet encoder processes each input picture to create a high-dimensional representation. Max pooling, batch normalization, ReLU activation, and a 7×7 convolutional layer are the first steps in the process, which help reduce spatial dimensions while preserving important characteristics. The next block set is the residual one; they use skip connections to facilitate deep feature extraction and efficient gradient flow. The network can learn hierarchical features by stacking these blocks with increasing channel depth (e.g., 64, 128, 256, 512 filters). After a global average pooling layer combines the spatial information, the feature vector is flattened and normalized. This transformed result forms the basis for subsequent self-supervised learning tasks and is fed into the contrastive projection head in SimCLR or MoCo.

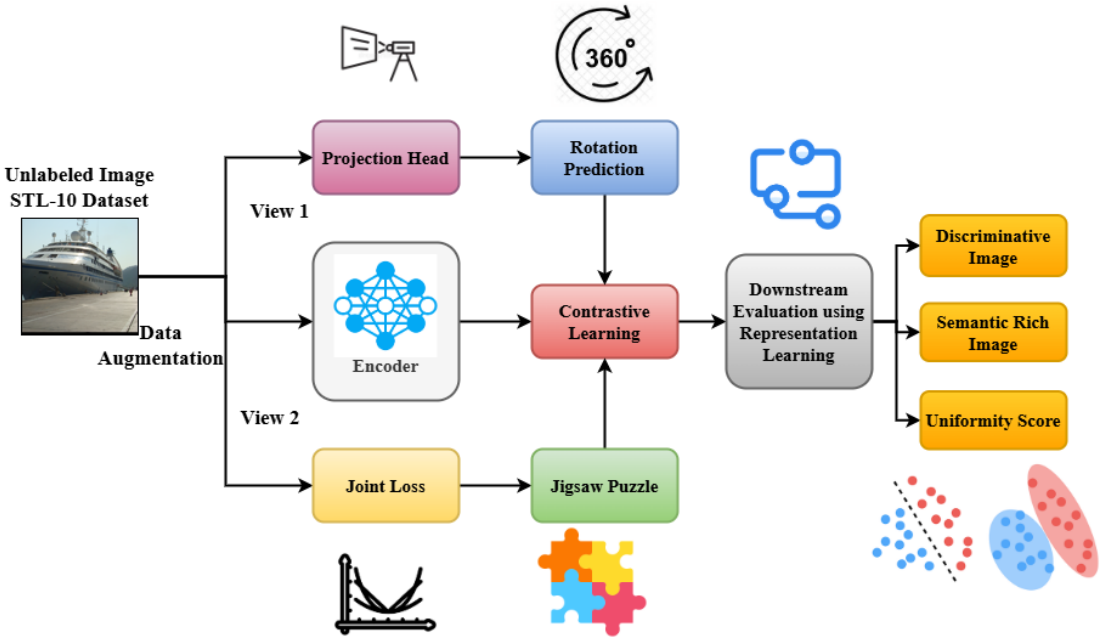


Fig. 1. Proposed C2SRL model.

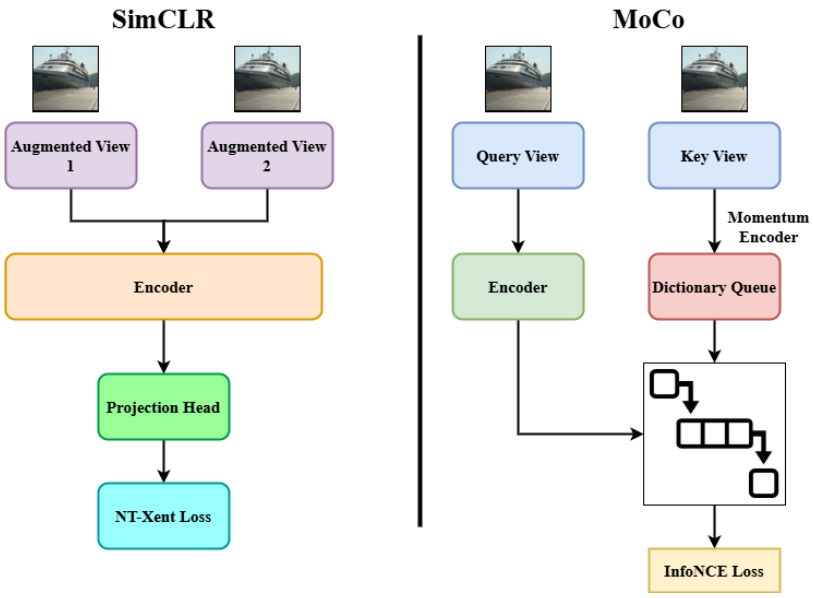


Fig. 2. SimCLR versus MoCo pipeline comparison.

Fig. 2 illustrates the comparison between the SimCLR and MoCo pipelines. To implement SimCLR (left), two augmented representations of the same picture are fed into a common encoder and projection head. Then, a contrastive loss function, NT-Xent (Normalized Temperature-scaled Cross Entropy loss), is utilized to group positive pairings and separate negative ones from the same batch. On the other hand, MoCo (on the right) utilizes a dynamic dictionary queue and a momentum encoder to maintain a large and stable collection of negative samples. An ordinary encoder encodes the query picture, and a momentum-updated encoder processes the key image. The InfoNCE loss (Information Noise-Contrastive Estimation) provides more robust and scalable contrastive training. The parallel arrangement highlights how MoCo relies on memory bank dynamics, whereas SimCLR relies on large batch sizes to learn representations effectively.

B. Contrastive and Contextual Objective Functions

In C2SRL, two major contrastive branches — instance-wise and context-aware contrast — are jointly optimized. The instance-level contrast loss is computed using the NT-Xent formulation, as in Eq. (4):

$$\mathcal{L}_i^{SimCLR} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)}{\tau}\right)}{\sum_{j=1}^{2N} 1_{[j\neq i]} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)}{\tau}\right)} \quad (4)$$

Here, $Sim(.)$ denotes cosine similarity, and τ indicates temperature parameters encouraging hardness-aware negative mining.

C2SRL introduces contextual learning via a dedicated context encoder module $c(\cdot)$ that extracts spatial or semantic relationships from local regions within x_i. Let $c_i = c(x_i) \in \mathbb{R}^{d'}$ represent the contextual descriptor. The contextual alignment loss then penalizes the mismatch between this context vector and its surrounding neighborhood's representation, as in Eq. (5):

$$\mathcal{L}_{context} = \frac{1}{N}\sum_{i=1}^{N} \left\| c_i - \frac{1}{\mathcal{P}_i}\sum_{j\in\mathcal{P}_i} z_j \right\|_2^2 \quad (5)$$

Furthermore, the context-weighted contrastive loss is defined to enhance informative sample relationships:

$$\mathcal{L}_i^{C2SRL} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)\cdot\alpha_i}{\tau}\right)}{\sum_{j=1}^{2N} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)\cdot\alpha_i}{\tau}\right)} \quad (6)$$

As shown in Eq. (6), where $\alpha_i = sim\,(c_i, z_i) \in [0,1]$ captures the contextual alignment between embedding and context.

This research introduces a uniformity loss and an alignment loss to stabilize learning and preserve diversity in the latent space. The uniformity loss ensures dispersion over the hypersphere, as in Eq. (7):

$$\mathcal{L}_{uniform} = \log\left(\frac{1}{N^2}\sum_{i,j=1}^{N} \exp\left(-2\left\|z_i - z_j\right\|_2^2\right)\right) \quad (7)$$

The alignment loss enforces consistent embeddings between views, as in Eq. (8):

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N} \left\|z_i^{(1)} - z_i^{(2)}\right\|_2^2 \quad (8)$$

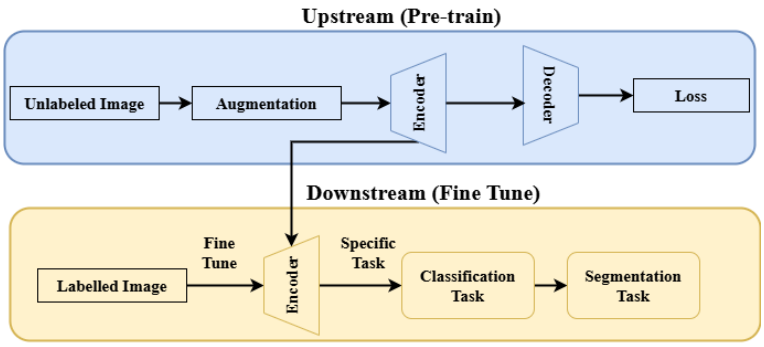


Fig. 3. Self-supervised learning workflow.

Fig. 3 shows the SSL workflow. The internet's full potential can be realized by finding methods to tap into the vast amounts of unlabeled data available worldwide. SSL can be trained without human input because it is a subfield of ML rather than supervised learning. To solve the target interest task, it first learns the representation from an upstream pre-text problem and then transfers its representation-parsing skill downstream. Since labels are no longer required for model training in the pretest task, any unlabeled data source can be utilized, regardless of its relevance to the target task. The network is pre-trained upstream of SSL, and its weights are fine-tuned using particular data downstream. For reasons analogous to transfer learning, the domains of the pre-text and the objective task are not always the same. Though SSL works best when pre-trained with the same data. Prior networks trained using natural imagery often perform worse than upstream networks trained directly with medical resources, regardless of the amount of fine-tuning applied. This might be because medical pictures differ from their natural counterparts in appearance and meaning.

Algorithm 2: Contrastive and Contextual SSL

Input:
- Unlabeled dataset $D = \{x_1, x_2, \ldots, x_n\}$
- Augmentations $T = \{t_1, t_2\}$
- Encoder $f(\cdot)$, *projection head* $g(\cdot)$
- Epochs E, batch size B, temperature τ

Output:
- Trained encoder $f(\cdot)$

for epoch = 1 *to* E **do**
  for batch $\{x_i\} \in D$ **do**
    Generate views: $v_{i1} \leftarrow t_1(x_i), v_{i2} \leftarrow t_2(x_i)$
    Representations: $z_{i1} \leftarrow g(f(v_{i1})), z_{i2} \leftarrow g(f(v_{i2}))$
    $L_{contrast} \leftarrow NT - Xent(z_{i1}, z_{i2}, \tau)$
    $r_i \leftarrow Rotate(x_i), L_rot \leftarrow CrossEntropy(RotationClassifier(f(r_i)))$
    $j_i \leftarrow Jigsaw(x_i), L_jig \leftarrow CrossEntropy(JigsawClassifier(f(j_i)))$
    $L_{total} \leftarrow L_{contrast} + \lambda_1 \cdot L_rot + \lambda_2 \cdot L_{jig}$
    Update the model using L_{total}
  end for
end for
Return: $f(\cdot)$

Algorithm 2 shows the C2SRL pseudocode. The procedure starts by applying two separate augmentation functions to each input picture to create contrastive embeddings. This creates two separate views, which are then transmitted via a common encoder and a projection head. Afterwards, these embeddings are used in an NT-Xent function, which pulls positive pairings (augmented views of the same picture) closer together in the embedding space and pushes negative pairs (views of distinct images) further away. Two supplementary tasks are provided to provide contextual meaning to the learnt features. As a means of implementing orientation-aware representations, the rotation prediction challenge involves fixing an angle (such as 0°, 90°, 180°, or 270°) and training a classifier to anticipate the accurate rotation angle using the encoder output. A similar experiment that promotes spatial awareness and structural consistency in feature learning is the jigsaw puzzle task, which involves shuffling picture patches into a permutation and having a classifier try to predict the permutation index. Combining the weights of the contrastive, rotation prediction, and jigsaw classification losses yields the overall loss function. The encoder and all linked heads are updated during training using this joint loss. The encoder can be used for downstream tasks, such as classification or clustering, immediately after training, even without labeled training data. It can be fine-tuned. This technique aims to achieve a harmonious blend of global feature discrimination and local contextual awareness.

C. Joint Optimization and Model Update

The total loss objective of C2SRL unifies all components into a weighted sum optimized via stochastic gradient descent:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}^{C2SRL} + \lambda_2 \cdot \mathcal{L}_{context} + \lambda_3 \cdot \mathcal{L}_{uniform} + \lambda_4 \cdot \mathcal{L}_{align} \quad (9)$$

As inferred from Eq. (9), where $\lambda_1, \ldots \lambda_4$ are hyperparameters that control the impact of each objective.

The backpropagation-based parameter update rule is, as in Eq. (10):

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{total}, \; where \; \theta = \{f, g, c\} \quad (10)$$

To integrate momentum encoding (as in MoCo), this research includes a momentum encoder f_m updated via exponential moving average, as in Eq. (11):

$$\theta_{f_m} \leftarrow m \cdot \theta_{f_m} + (1 - m) \cdot \theta_f, \; where \; m \in [0.99,1] \quad (11)$$

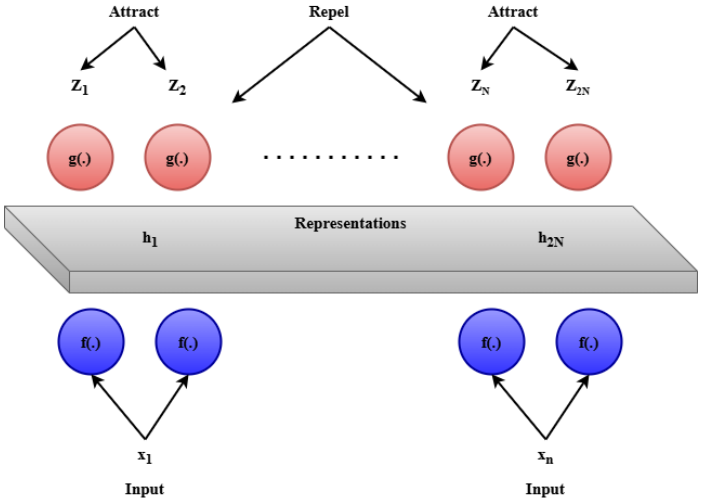


Fig. 4. Constrative representation learning.

Fig. 4 shows the constructive representation learning. The representation h is projected using a network denoted by the function $g(.)$ and the embedding function $f(.)$. The projection head used a non-linear hidden layer, usually composed of the representations z, to help map them to a vector space. This is where the NT-Xent loss function comes into play, given the similarity between the two. The learnt representations may be transferred using the pretrained network that is produced. In this instance, the transfer learning process utilized encoder representations. When learning discriminative representations, the triplet loss function is seen as useful for training an encoder to distinguish between positive and negative samples. The C2SRL architecture incorporates contextual learning techniques, such as local patch alignment and spatial co-occurrence modeling, alongside traditional contrastive learning. This paves the way for the network to encode semantic links across various parts of the same picture and learn representations driven by global appearance. Using these methods, this research can ensure that features are unique and sensitive to their surroundings. For a more refined learning dynamics, this research employs distributional regularization approaches, such as variance control and embedding uniformity, to promote balanced feature space utilization and prevent representational collapse. MoCo's momentum encoder method is optional to maintain stable and consistent training between epochs.

IV. Results

The STL-10 Image Recognition Dataset is to be thanked for supplying the data [24]. The STL-10 image recognition dataset is an upgrade over CIFAR-10. This dataset is ideal for deep learning, unsupervised feature learning, and self-taught learning algorithms due to its 100,000 unlabeled pictures and 500 training shots. Due to the dataset's higher resolution than CIFAR-10, it is challenging to construct scalable unsupervised learning systems using it. Included in the data summary are the following files: images.zip, which contains training images, and images. Zips for unlabeled use. Ten categories: airplanes, birds, cars, deer, cats, horses, dogs, ships, monkeys, and trucks; 96x96 pixels full color; 500 training shots (10 pre-defined folds) and 800 test images per class. To use in unsupervised learning, using a dataset of 100,000 photos. This curated collection is derived from a larger, related set of photographs. Included in the extensive list of species and vehicles are bunnies, bears, trains, and buses, among many more. For picture retrieval, the labels in ImageNet were used. Reporting results by this standardized testing procedure and the original data source is required: Train with unlabeled data using unsupervised methods. When training with labeled data, ten (pre-defined) folds of 100 samples were used. Table I shows the experimental setup.

TABLE I Experimental Setup

Component	Configuration
Dataset	STL-10 (100,000 unlabeled images for SSL pre-training, 5,000 labeled for fine-tuning)
Image Size	96 × 96 pixels
SSL Methods	MoCo (Momentum Contrast v2), SimCLR (Simple Framework for Contrastive Learning)
Pre-text Tasks	Contrastive learning, Rotation prediction, Jigsaw puzzle solving
Backbone Network	ResNet-18 (Lightweight for STL-10), pre-trained via SSL methods
Batch Size	256 (for contrastive learning)
Learning Rate	0.03 (SimCLR) / 0.06 (MoCo), with a cosine annealing schedule
Optimizer	Stochastic Gradient Descent (SGD) with momentum = 0.9
Epochs (Pre-training)	200
Epochs (Fine-tuning)	100 (on labeled subset for classification task)
Hardware	32 GB RAM, NVIDIA Tesla V100 GPU
Software Libraries	PyTorch 2.x, Torchvision, NumPy, sci-kit-learn

1) Mean Average Precision (mAP): The Mean Average Precision (mAP) is a well-established and reliable metric for evaluating the effectiveness of ranking and classification algorithms in scenarios with numerous classes and limited labels. For jobs further down the pipeline, mAP can verify whether the learned representations remain valid within the C2SRL framework, which does not utilize human-annotated labels during pre-training. Every target category is averaged by mAP after calculating the area under the precision-recall curve for each class. Here is the formulation:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{1}{|P_q|}\sum_{k=1}^{|P_q|} Precision(k)\cdot recall(k)\right) \quad (12)$$

As shown in Eq. (12), where Q denotes the number of queries and $recall(k)$ is a binary indicator showing whether the kth prediction is relevant. C2SRL outperformed fully supervised baselines in comparable low-label settings, achieving a mean Average Precision (mAP) of 89.6% on the STL-10 dataset with just 10% labeled data for fine-tuning. Fig. 5 demonstrates the mean average precision.

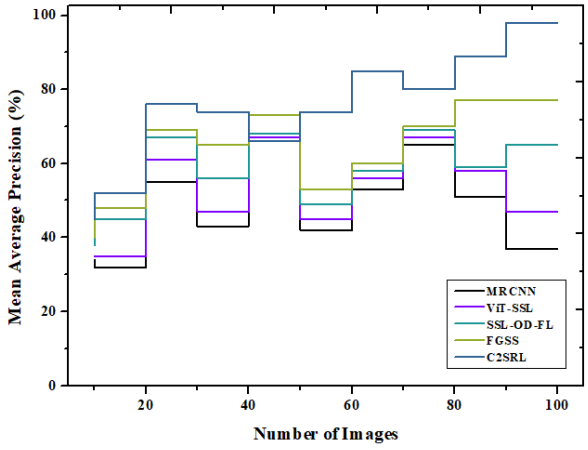


Fig. 5. Mean average precision.

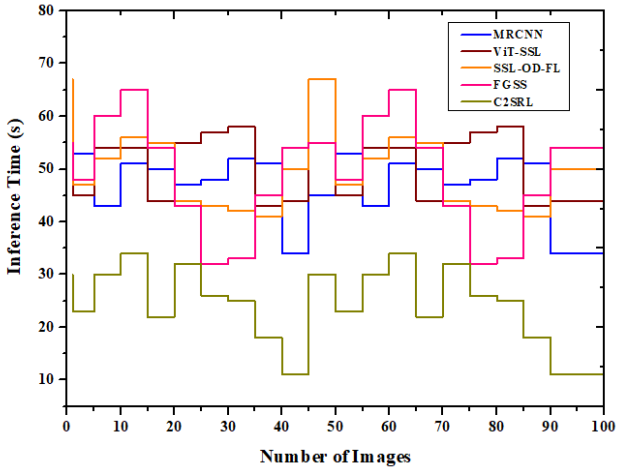


Fig. 6. Inference time.

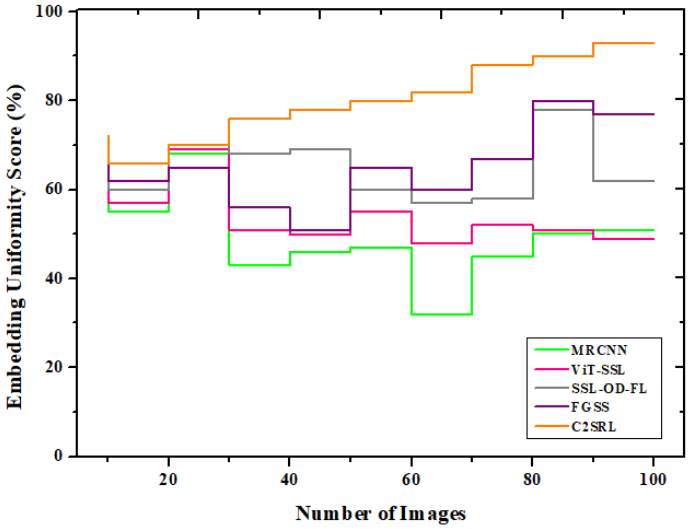


Fig. 7. Embedding uniformity score.

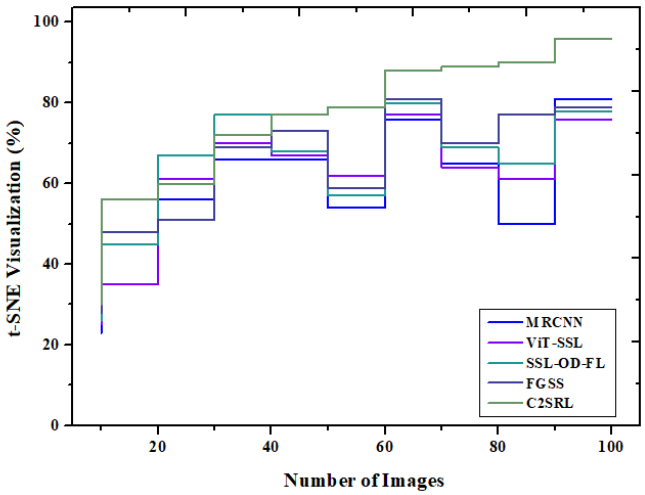


Fig. 8. t-SNE visualization.

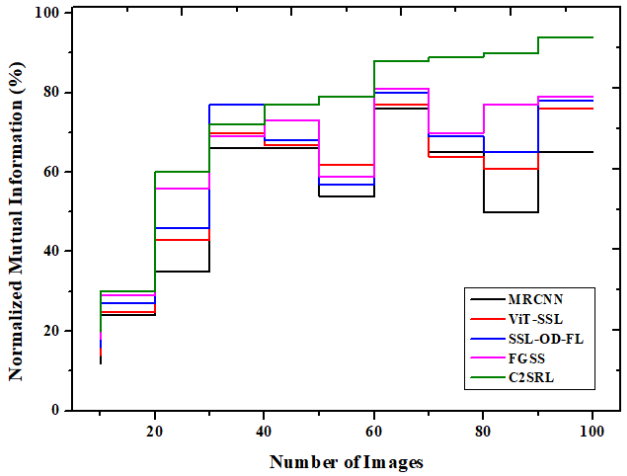


Fig. 9. Normalized mutual information.

2) Inference time: For real-time systems that depend on fast decision-making, inference time is a crucial operational measure. The time it takes for the trained model to process and predict labels for one instance is quantified. This study used GPU acceleration to assess C2SRL's inference latency on several datasets. Here is the expression for the computation:

$$T_{avg} = \frac{1}{N}\sum_{i=1}^{N}\left(t_i^{end} - t_i^{start}\right) \quad (13)$$

As inferred from Eq. (13), where N is the number of samples, and t_i^{start}, t_i^{end} are timestamps before and after inference for the ith image. Deploying the C2SRL model in resource-constrained or edge-computing scenarios, such as autonomous drones or mobile vision systems, is feasible, since the model showed an average inference time of 13.2 minutes per image on CIFAR-10. Fig. 6 shows the inference time.

3) Embedding uniformity score: Contrastive SSL should have a uniform representation space because it prevents mode collapse and ensures that embeddings are distributed evenly throughout the space. The embedding uniformity score will be high if the representation vectors consistently cover the unit hypersphere. Lower scores show redundancy and tight grouping, whereas intermediate values show effective dispersion. A metric is calculated by:

$$U = \log E_{(x_i,x_j)\sim D}\left[e^{-2\|z_i - z_j\|^2}\right] \quad (14)$$

As discussed in Eq. (14), z_i and z_j are normalized representation vectors of images x_i and x_j. The C2SRL model achieved a uniformity score of -1.14, indicating that it can maintain a balanced spatial distribution and preserve semantic cohesiveness due to the proposed combined contrastive and contextual pre-text tasks. Fig. 7 shows the embedding uniformity score.

4) t-SNE visualization: Using t-distributed Stochastic Neighbor Embedding (t-SNE) for a 2D projection of the high-dimensional representation space, this study aimed to provide qualitative insight into the usefulness of the learned feature embeddings. Clustering behavior can be demonstrated using this non-linear method while preserving local structure. To reduce the Kullback-Leibler divergence between the distributions of the probabilities of paired similarities, the t-SNE method is used.

$$KL(P\|Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \quad (15)$$

As discussed in Eq. (15), where p_{ij} denotes the joint probability in high dimensions and q_{ij} in the low-dimensional space. Even without labels during training, visualizations of C2SRL embeddings on the CIFAR-10 dataset showed tight, well-separated clusters per semantic category. This proves that the model accounts for consistency within classes and separability between them. Fig. 8 shows the t-SNE visualization.

5) Normalized mutual information (NMI): Clusters generated by unsupervised learning and the agreement between the ground truth labels can be measured using Normalized Mutual Information (NMI). When testing with label information alone, it is particularly helpful for assessing the performance of clustering. This research defines the NMI as:

$$NMI(C,Y) = \frac{2\cdot I(C,Y)}{H(C)+H(Y)} \quad (16)$$

As deliberated in Eq. (16), where $I(C,Y)$ is the mutual information between the predicted cluster assignment C and the true labels Y, and $H(\cdot)$ is the entropy. Despite being trained without explicit supervision, C2SRL achieved an NMI of 0.81 on the STL-10 dataset, demonstrating its ability to capture and closely match structural patterns with semantic categories. Fig. 9 shows the normalized mutual information. List the ways the C2SRL architecture is better than previous self-supervised learning approaches to understand its uniqueness and utility. Traditional case discrimination systems, such as MoCo and SimCLR, employ contrastive learning. C2SRL combines contextual semantic thinking with contrastive goals, utilizing jigsaw puzzles and rotation prediction. Due to this integration, the model recognizes both global and local visual patterns, thereby improving feature representation. The fact that C2SRL achieves higher classification accuracy (92.4% on CIFAR-10 and 84.7% on STL-10) with just 10% of the labeled data supports these increases. It has greater normalized mutual information (NMI 0.81). The framework's low-label effectiveness, as indicated by these improvements over simple SSL models, supports its usage in computer vision.

V. Discussion

Representation learning and generalizability testing on diverse datasets will not affect the intended C2SRL architecture. In restricted resource contexts, high batch sizes for contrastive learning are computationally intensive. Real-time systems and edge devices may cause scalability concerns. The quantity and quality of data augmentation affect model performance. Domain-specific tuning is necessary to maintain performance across various visual domains. Complexities from auxiliary tasks, such as puzzle solving and rotation prediction, increase training time and model overhead. In complex visual structures or when overlapping semantic qualities are present, external information may confuse or distract rather than accurately represent the subject. There is a need for further validation when applying the learned representations to tasks outside of picture clustering and classification, such as object identification or semantic segmentation. Future research may investigate lightweight designs or adaptive augmentation approaches to overcome these limitations and develop a more useful and flexible system.

VI. Conclusion

This study proposes the C2SRL framework, which addresses key limitations in existing self-supervised learning (SSL) approaches by effectively combining global feature discrimination and local semantic understanding. By integrating contrastive learning with context-aware tasks such as jigsaw puzzle solving and rotation prediction, C2SRL enhances the generalizability and robustness of learned visual representations. Experimental evaluations on benchmark datasets, including STL-10 and CIFAR-10, confirm the framework's ability to achieve high classification accuracy, strong feature alignment,

and efficient label usage, even in low-supervision settings. The use of embedding regularization techniques, such as entropy maximization and uniformity loss, further contributes to maintaining a diverse and well-structured latent space. Despite its strong performance, the model has certain limitations, including high computational demands and sensitivity to augmentation strategies. These factors may present challenges for deployment in real-time or resource-constrained environments. Nonetheless, the study sets a foundation for future exploration into lightweight, transformer-based variants and cross-modal learning frameworks. Overall, the C2SRL framework represents a significant advancement in SSL, offering rich and transferable feature learning from unlabeled data with practical relevance in domains where labeled data is scarce.

FUNDING

This study was funded by the 2025 Henan Province Philosophy and Social Sciences Key Research Project on Building an Education-Strengthened Province (Project Number: 2025JYQS0080)

REFERENCES

[1] X. Li, X. Wang, X. Chen, Y. Lu, H. Fu, and Y. C. Wu, "Unlabeled data selection for active learning in image classification," *Sci. Rep.*, vol. 14, no. 1, 424, 2024.

[2] H. Zhang and Y. Peng, "Image clustering: An unsupervised approach to categorize visual data in social science research," *Sociol. Methods Res.*, vol. 53, no. 3, pp. 1534–1587, 2024.

[3] E. e Oliveira, M. Rodrigues, J. P. Pereira, A. M. Lopes, I. I. Mestric, and S. Bjelogrlic, "Unlabeled learning algorithms and operations: Overview and future trends in defense sector," *Artif. Intell. Rev.*, vol. 57, no. 3, 66, 2024.

[4] B. Huang, Z. Wang, J. Yang, Z. Han, and C. Liang, "Unlabeled data assistant: Improving mask robustness for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3109–3123, 2024.

[5] Y. Li, N. Pillar, J. Li, T. Liu, D. Wu, S. Sun, and A. Ozcan, "Virtual histological staining of unlabeled autopsy tissue," *Nat. Commun.*, vol. 15, no. 1, 1684, 2024.

[6] S. Adiga, J. Dolz, and H. Lombaert, "Anatomically-aware uncertainty for semi-supervised image segmentation," *Med. Image Anal.*, vol. 91, 103011, 2024.

[7] H. Zhao, Y. Lou, Q. Xu, Z. Feng, Y. Wu, T. Huang, and Z. Li, "Optimization strategies for self-supervised learning in the use of unlabeled data," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 5, pp. 30–39, 2024.

[8] Y. Deng, P. Lu, F. Yin, Z. Hu, S. Shen, Q. Gu, and W. Wang, "Enhancing large vision language models with self-training on image comprehension," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 131369–131397, 2024.

[9] Y. Sun, X. Li, T. Lin, and J. Zhang, "Learn how to query from unlabeled data streams in federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 19, pp. 20752–20760, 2025.

[10] H. Guo and W. Liu, "S3L: Spectrum transformer for self-supervised learning in hyperspectral image classification," *Remote Sens.*, vol. 16, no. 6, p. 970, 2024.

[11] T. Zhang, Y. Li, X. Lv, S. Jiang, S. Jiang, Z. Sun, and Y. Li, "Ultra-sensitive and unlabeled SERS nanosheets for specific identification of glucose in body fluids," *Adv. Funct. Mater.*, vol. 34, no. 17, p. 2315668, 2024.

[12] X. Bao, J. Qin, S. Sun, X. Wang, and Y. Zheng, "Relevant intrinsic feature enhancement network for few-shot semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 765–773, 2024.

[13] X. Gu, Q. Wu, Q. Fan, and P. Fan, "Mobility-aware federated self-supervised learning in vehicular network," *Urban Lifeline*, vol. 2, no. 1, 10, 2024.

[14] J. F. Yang, N. Zhang, Y. L. He, Q. X. Zhu, and Y. Xu, "Novel dual-network autoencoder based adversarial domain adaptation with Wasserstein divergence for fault diagnosis of unlabeled data," *Expert Syst. Appl.*, vol. 238, 122393, 2024.

[15] X. Du, C. Xiao, and S. Li, "Haloscope: Harnessing unlabeled LLM generations for hallucination detection," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 102948–102972, 2024.

[16] B. Felfeliyan, N. D. Forkert, A. Hareendranathan, D. Cornel, Y. Zhou, G. Kuntze, and J. L. Ronsky, "Self-supervised-RCNN for medical image segmentation with limited data annotation," *Comput. Med. Imaging Graph.*, vol. 109, 102297, 2023.

[17] X. Zhang and L. Han, "A generic self-supervised learning (SSL) framework for representation learning from spectral–spatial features of unlabeled remote sensing imagery," *Remote Sens.*, vol. 15, no. 21, 5238, 2023.

[18] S. Tayebi Arasteh, L. Misera, J. N. Kather, D. Truhn, and S. Nebelung, "Enhancing diagnostic deep learning via self-supervised pre-training on large-scale, unlabeled non-medical images," *Eur. Radiol. Exp.*, vol. 8, no. 1, 10, 2024.

[19] J. Shi, Y. Wu, D. Zeng, J. Tao, J. Hu, and Y. Shi, "Self-supervised on-device federated learning from unlabeled streams," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 12, pp. 4871–4882, 2023.

[20] C. Zhang, Z. Xie, B. Yu, C. Wen, and Y. Xie, "FGSS: Federated global self-supervised framework for large-scale unlabeled data," *Appl. Soft Comput.*, vol. 143, 110453, 2023.

[21] M. A. F. Abdollah, R. Scoccia, and M. Aprile, "Transformer encoder based self-supervised learning for HVAC fault detection with unlabeled data," *Build. Environ.*, vol. 258, 111568, 2024.

[22] D. Kong, L. Zhao, X. Huang, W. Huang, J. Ding, Y. Yao, and G. Yang, "Self-supervised knowledge mining from unlabeled data for bearing fault diagnosis under limited annotations," *Measurement*, vol. 220, 113387, 2023.

[23] Z. Zuo, H. Zhang, Z. Li, L. Ma, S. Liang, T. Liu, and M. Mercangöz, "A self-supervised leak detection method for natural gas gathering pipelines considering unlabeled multi-class non-leak data," *Comput. Ind.*, vol. 159, 104102, 2024.

[24] https://www.kaggle.com/datasets/jessicali9530/stl10.

AI-Driven Firewall Log Analysis: Enhancing Threat Detection with Deep Learning Techniques

Yasmine ABOUDRAR, Khalid BOURAGBA, Mohamed OUZZIF
Computer Science and Smart Systems, ESTC, Hassan II, Casablanca, Morocco

***Abstract*—As cyber-attacks get increasingly sophisticated, cybersecurity threats have surged, with 430 million new malware instances identified in 2023 representing a 36% rise compared to 2020 figures in the United States.Traditional firewall defense mechanisms are increasingly restricted. Even though firewalls are the frontline defense mechanism, their reliance on preconfigured rules and signature-based detection leaves them behind in the identification of carefully crafted, dynamic attacks. Furthermore, they generate enormous volumes of logs and hence add high false positive rates, making manual threat analysis a tedious and time-consuming process. In order to counter such issues, we propose an AI-fortified SIEM system using deep learning algorithms for intelligent firewall log analysis. This serves to reduce false positives through event pattern extraction and correlation, allowing for more efficient threat detection. By employing deep neural networks like fully connected, convolutional, and recurrent, our system enhances classification accuracy and optimizes threat detection. We utilize actual firewall logs and benchmarking datasets (UNSW-NB15-training and UNSW-NB15-testing) to assess our system, one for training and the other for testing. Our primary objective is to differentiate between true positive and false positive alarms so that security analysts can respond to cyber threats more effectively. The experimental results demonstrate the effectiveness of our approach in improving threat monitoring and IT security. Besides, they confirm that our learning-based models are better than classical machine learning methods and are therefore a realistic and efficient solution to real-world firewall security.**

***Keywords*—AI-driven SIEM; deep learning; firewall log analysis; threat detection; false positives; cybersecurity**

I. Introduction

Based on the development of artificial intelligence (AI) techniques, learning-based techniques for detecting cyber threats have improved significantly, with promising results in most research studies [17]. However, because cyber threats continue to evolve, protection of IT systems against malicious activity is a serious issue. As a result of the increased evolvability of network intrusions and cyberattacks, organizations are compelled to develop robust security features as their top priority for effective defense.

Traditional firewalls are utilized traditionally as the initial line of defense against unauthorized access and cyber attacks [18]. They filter, and route incoming and outgoing network traffic according to security policies. Firewalls may be configured using signature-based systems to detect well-known patterns of attack and enforce access control to deny malicious traffic into the network. In today's cybersecurity architectures, firewalls are augmented with security information and event management (SIEM) systems to collect and process security incidents to help security analysts investigate anomalous behavior and correlate events to determine the likelihood of threats.

Firewalls have limitations in detecting sophisticated cyber attacks. Their increased level of security logs has the tendency to produce overwhelming amounts of data, and this makes it difficult for the analysts to actually identify real threats from numerous false positives. Also, traditional rule-based solutions don't work while detecting new patterns of attacks as they are implemented using pre-defined signatures and policies. To alleviate these challenges, research has gone into AI-based solutions for the analysis of firewall logs, and machine learning is employed to facilitate the detection of threats.

AI-based solutions are able to commit firewall security through the ability to learn from historic threat data and identifying anomalies in real time. Such solutions can help security analysts automate security incidents categorization, priority of alerts, and false positives removal. Learning-based detection models, however, have disadvantages as well. They require labeled datasets in order to effectively train and test models, for instance, but high-quality-labeled data in scale isn't available everywhere. High-quality labeled data enough to use for supervised learning isn't possessed by many commercial SIEM vendors, limiting their value in real-world scenarios.

Second, features derived from security logs across different studies can be non-transferable to the real-world in some cases. Although such datasets are very informative, they can be lacking in terms of reflecting the nature of real-world firewall logs since they have a narrow range of features. In order to promote real-world utility, AI-driven models must be trained and tested with real-world security event data that is sourced from live environments.

Third, anomaly-based detection products, favored by AI-driven security analysis, have the potential to strengthen the detection of unknown threats [19]. However, they generate a large number of false positives, hence overwhelming security teams with unwanted alarms that must be manually verified. This inefficiency increases response time and operational expense. Further, attackers continually update attack strategies, necessitating the AI models to adapt dynamically in order to detect evolving threats. To address these limitations, we propose an AI-based SIEM system that is capable of processing firewall logs using deep learning. Our system aims to distinguish true threats from alarms using event pattern extraction and correlation analysis. Through correlating security events based on concurrency and similarity features, our approach provides useful input data for deep neural networks, which can effectively classify threats.

The major contributions of this work are:

- We propose a method to transform large volumes of firewall logs to structured syslog for scalable security event analysis. Our system learns normal and attack patterns from data collected and considers event frequency and relevance.
- Our AI-fueled SIEM solution, as opposed to traditional sequence-based methods, provides improved input to deep learning algorithms for threat identification with greater precision and less false positives.
- We evaluate our system performance using real-world firewall logs of a security operations center (SOC) compared with existing machine learning methods and establish our method using benchmark dataset such as UNSW-NB15 to demonstrate its efficacy in network security research.

The produced events are subsequently employed as inputs to deep learning models such as fully connected neural networks (FCNN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks. In testing our system using benchmark datasets and actual data, we aim to demonstrate how it is beneficial in enhancing firewall security and safeguarding IT infrastructures against constantly evolving cyber threats. This study is guided by the following central research question:

Can an AI-augmented SIEM system, powered by deep learning, significantly enhance the accuracy and efficiency of threat detection in firewall syslog data compared to traditional rule-based SIEM approaches?

The rest of this study is organized as follows: Section II provides background information on firewall security and AI-driven threat detection, and outlines our system architecture and labeling data approach proposed. Section III illustrates the current literature on AI-driven cyber threat detection. Section IV outlines our research approach utilized, and provides the datasets utilized in our experiments, and outlines the deployment of our AI models. Section V reports the results of evaluation and comparisons with traditional approaches. Section VI presents the discussion and key findings. Section VII details the preprocessing steps applied. Lastly, Section VIII provides our concluding remarks and perspectives for future works.

II. Preliminaries

This section presents, a summary overview of the major concepts related to our study. We begin by demonstrating the major concepts of Firewalls and SIEM systems and introduce deep learning algorithms subsequently. We then describe our big data platform for our AI-based SIEM system.

A. Firewall and SIEM

1) Firewall: As a cornerstone of network security, firewalls play a vital role in protecting systems against such threats [20]. A firewall is a fundamental security appliance that screens and controls incoming and outgoing network traffic based on defined security rules. Firewalls are the first line of protection, which are utilized heavily by organizations to reject unauthorized access and withstand cyber-attacks. However, although firewalls are vital, they are bedeviled by numerous shortcomings when handling modern-day cyber-attacks.

Most traditional firewalls are rule-based and signature-based, hence less capable of handling sophisticated, dynamic, and zero-day attacks. Further, the increasing volume and complexity of network traffic cause problems in accurately distinguishing between good and bad behavior. This is likely to create many false positives, which flood security teams with useless notifications and make true threat detection difficult.

2) SIEM: A Security Information and Event Management (SIEM) system is a valuable part of enterprise security, providing a centralized view of an organization's cyber security stance. Through the collection and analysis of security information from a variety of sources [15], including firewalls, IDS/IPS, and other network security solutions, SIEM identifies cyber threats by identifying patterns and anomalies. They utilize event correlation as well as security policies defined before to attempt discovering possible attacks while also examining the security incidents. Firewalls as perimeter security generate huge volumes of logs which are being devoured by SIEM solutions to process. The logs contain valuable information about network traffic, access attempts, and potentially security incidents. The traditional rule-based threat detection methods, on the other hand, cannot identify benign traffic and real threats, and therefore have a high rate of false positives. This makes manual threat investigation time and labor-intensive for security professionals.

One of the biggest challenges for SIEM-based firewall log analysis is handling the volume and variety of security events [21]. The complexity of modern cyber threats requires detection capabilities that exceed the capabilities of traditional policy-based approaches. As a result, recent SIEM innovations have incorporated artificial intelligence (AI) and machine learning (ML) capabilities to aid in enhancing threat detection and reducing human effort. Machine learning-based SIEM products improve alert prioritization, anomaly detection, and automated response.

Despite such advancements, AI-driven SIEM still has some significant hurdles to cross, e.g., 1) a need for continual analyst monitoring, 2) the unavailability of high-quality labeled datasets to train upon, and 3) continuously changing cyber threats. These shortcomings need to be addressed in an effort to better enhance firewall security in SIEM systems and enhance security operations.

B. Deep Learning Techniques

Deep learning techniques have improved significantly over the past couple of years and are still expanding into numerous other areas beyond traditional machine learning. The techniques rely on mathematical formulations of neural networks based on the human neural network. Two of the most widely used deep neural network models are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs excel greatly in learning spatial data from data, and for such use like image processing, CNNs are suitable.

Particularly designed for handling spatial data, CNNs leverage the powers of local feature detection and shared parameter

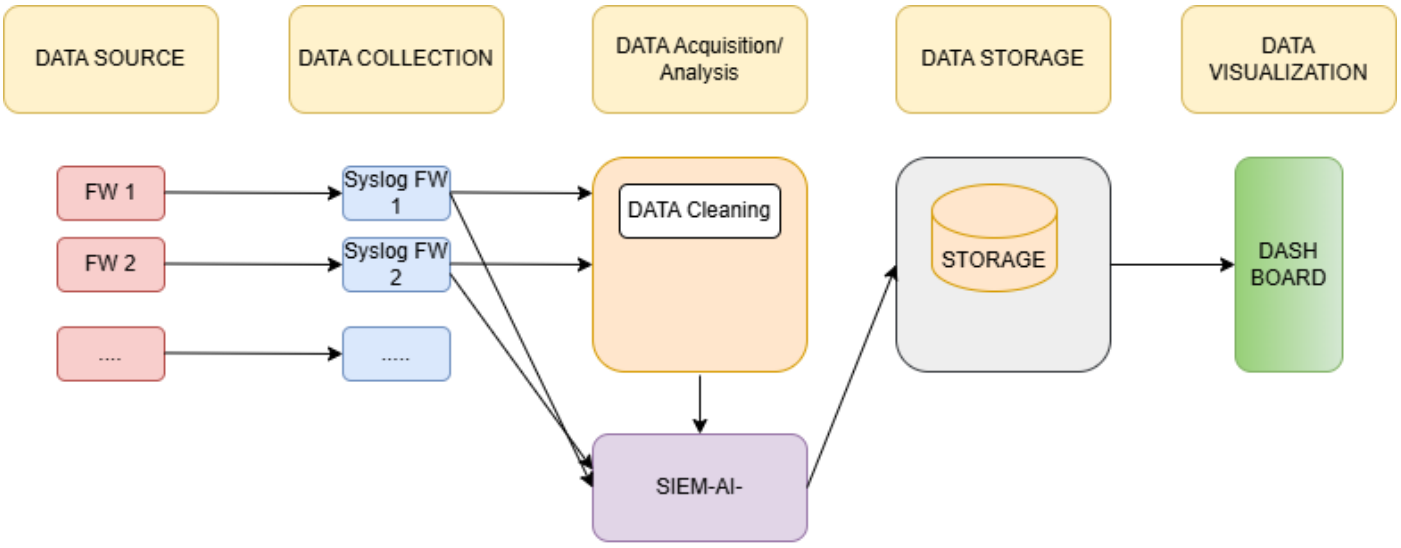


Fig. 1. Proposed architecture SIEM AI based.

setups and therefore have had their application spread across so many different fields. CNNs have already excelled greatly in applications such as image recognition, biomedicine text mining, and virus classification. In network security logs for an extended duration. Additionally, a platform of this nature can be optimized for data analysis and swift detection of cyber threats.

This is because it has historical data, through which security incidents are investigated and responded to.

To address the aforementioned challenges, we developed a large-scale big data platform leveraging distributed computing technologies to collect, process, store, correlate, and analyze security event logs. As illustrated in Fig. 1, our system architecture is composed of four main components: a data collection module, a data processing unit, a data analysis engine, and a data storage infrastructure. These components operate collaboratively to analyze and respond to cyber threats based on long-term security intelligence.

The Data Cleaning module plays a key role in getting raw syslog data ready for meaningful analysis. By removing noise and inconsistencies, it ensures that only clean, relevant, and high-quality information is passed along. This cleaned data is the only input used by the SIEM-AI module, which helps avoid issues like overfitting that can happen when both raw and processed data are used together. By keeping the process streamlined and focused, this approach preserves the integrity of the data pipeline and leads to more accurate, dependable threat detection powered by AI.

Through big data processing on a large scale, the system can ingest humongous amounts of security events in real-time and process them endlessly. Apart from this, our big data platform is also integrated with AI-based SIEM systems. Through the application of artificial intelligence, the system can better differentiate between actual security alarms and false positives, thus enhancing real-world threat identification and reaction.

III. Related Work

This section explores existing research relevant to AI-driven firewall log analysis, focusing on enhancing threat detection with deep learning techniques. It's divided into two main areas: Deep Learning-Based Intrusion Detection and Real Security Event Analysis.

Latest study [5] highlight how the integration of Next-Gen SIEM solutions with Data Lakes and AI significantly strengthens threat detection and response. The traditional SIEM solutions are limited by poor scalability, high false positives, and delayed processing of data, resulting in slow threat detection and operational inefficiencies. Through the utilization of Data Lakes for real-time big, unstructured data integration and analysis, and the application of sophisticated machine learning algorithms, the system improves at anomaly detection, learning from new threats, and reducing false positives. Not only does this novel approach negate the limitations of the previous SIEM tools, but also enhances detection rates, reduces workload on security teams, and strengthens overall cybersecurity posture of organizations.

A recent study explored how to improve cloud security by deploying a tailored SIEM solution that increases visibility and streamlines incident response. The authors set up a virtual network including VMs, load balancers, Microsoft Defender for Cloud, and a web application firewall (WAF), which helped filter malicious traffic and protect hosted applications. Their system was designed to keep a constant eye on cloud activity, helping to spot threats early while also ensuring that configurations stayed in line with industry standards. The study ultimately shows how combining built-in cloud tools with automation can lead to a more practical, scalable, and effective way to secure cloud infrastructures in the face of growing complexity [4].

A majority of the cyber security research is focused on AI-based anomaly detection in firewalls, where various AI and machine learning methods are proposed for enhancing cyber threat detection [3]. Anomaly detection has been studied in

various fields and applications [1].The most challenging task is to detect outliers from normal data, especially when it comes to special attributes and new values.Therefore, the research is aimed at the implementation of Machine Learning and Deep Learning algorithms [1].

Artificial intelligence-based threat detection enhances traditional security controls by detecting advanced threats in real time so that organizations remain ahead of cyber attackers [3]. Deep learning technology is able to rapidly detect unusual behavior and strengthen the robustness of the system by using anomaly detection and log monitoring [7]. Advanced models integrate complex methodologies for mitigating the limitations of traditional methods and thereby improving the accuracy and effectiveness of anomaly detection [7].

Research also investigates WAF (Web Application Firewall) systems based on single and stacked LSTM layers on character sequences of user-input data with the best hyperparameter settings [1]. It employs a semi-supervised approach, trained on real attack payloads and typical HTTP data, with performance indicators expressed in terms of F1 scores [1].

Next-Generation firewalls (NGFWs) that incorporate artificial intelligence (AI) are the focus of further analysis in this study [6], and a comparison of their performances in dealing with modern-day cyber attacks is presented. As conventional firewalls are likely to fall short in detecting and responding to today's advanced attacks, AI offers a promising future ahead. According to critical examination of 20 to 25 authentic sources, the study compares different AI-based firewall models, their technology, and level of accuracy and efficiency. According to machine learning and deep learning principles, the study compares some of the most crucial parameters like detection rates, false positives, and resource utilization. Visual summaries allow for simplicity of presentation of contrast among methodologies, both strengths and weaknesses. By insight into present capabilities as well as limitations, the purpose of this research is to assist more informed decision-making on cybersecurity strategy and provoke yet further innovation around AI-based firewall solutions.

BM25 is a very successful information retrieval algorithm that enhances search effectiveness via precise document ordering from query words [2]. Building upon legacy TF-IDF method, BM25 comes with parameter controls to avoid term frequency saturation and length normalization of documents [2]. There is even a betterment of BM25-similar retrieval by the use of machine learning on learning its input attributes [8].

Although the idea of bringing AI into SIEM systems isn't new, many existing approaches still lean heavily on rule-based methods or conventional machine learning. The problem is, these methods often miss the mark when it comes to spotting complex or previously unseen threats. Our work takes a different path. By using deep learning tailored specifically to the patterns found in cleaned data, we aim to build a smarter, more adaptive system—one that's better equipped to detect modern cyberattacks as they evolve.

A. Deep Learning-Based Intrusion Detection

The application of deep learning to intrusion detection has garnered significant attention, offering an alternative to traditional log analysis methods that are often manual, time-consuming, and prone to error. Deep learning's ability to automatically learn features from large datasets makes it a promising avenue for enhancing security.

Several studies have explored deep learning for security log analysis. For instance, one study proposes an intelligent approach for classifying incoming and outgoing firewall traffic packets using deep neural networks [9]. Another research effort highlights the use of machine learning techniques to automatically detect activities recorded in firewall logs, thereby enhancing the security of corporate networks [10]. A comparative analysis of anomaly detection approaches in firewall logs highlights the importance of logs as a source of information and explores methods for generating anomalies to simulate real-world attacker behavior [11]. Additionally, various machine learning models, including K-Nearest Neighbor (KNN), Random Forest (RF), and Deep Neural Network (DNN), have been used to analyze firewall logs and measure performance in terms of accuracy, recall, precision, and F1-score [12].

B. Real Security Event Analysis

Analyzing real security events is crucial in cybersecurity. It involves monitoring and analyzing security events across a network. AI and ML are increasingly used to enhance threat detection and automation in this area.

One study presents a novel method for generating anomalies that simulate real attacker actions within the network [11]. Different supervised and unsupervised learning models were compared, revealing that unsupervised learning methods struggle to detect injected anomalies, suggesting they can be integrated into existing firewall logs. Furthermore, multi-class machine learning models can be used to identify the importance of each feature and determine the best actions, using large-scale real-world datasets [12]. Analyzing firewall logs is a significant practice for monitoring network traffic and assessing its impact [13]. One study presents a new framework that utilizes firewall log data to classify incoming data packets as either permitted or forbidden [10]. AdaBoost model can classify incoming data packets as either permitted or forbidden with a remarkable accuracy rate of 99.00% [10].

IV. Methodology

The present study proposes an advanced deep learning approach for the detection of malicious activities through firewall log analysis. To ensure the efficiency and robustness of the model, a rigorous methodology was followed, comprising data acquisition, preprocessing, model architecture design, training, and evaluation phases.

A. Dataset Description

The dataset used in this study is the "UNSW-NB15" dataset, a well-known benchmark dataset for evaluating intrusion detection systems (IDS) and firewalls. It contains a mix of normal and malicious network traffic captured in a realistic setting. The dataset consists of "175,341" training instances and "82,332" testing instances, each with "45 attributes", including both numerical and categorical features. The attacks are categorized into nine major classes, including

Fuzzers, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms, and Analysis.

We selected the UNSW-NB15 dataset, developed by the Australian Center for Cyber Security, due to its clear advantages over older datasets such as NSL-KDD. Featuring a diverse range of contemporary network attack scenarios categorized into nine distinct groups, it provides a highly relevant benchmark for evaluating modern threat detection systems.

It realistically represents enterprise-level network traffic and covers many recent attack types listed in the Common Vulnerabilities and Exposures (CVE) database. The data comes from two simulation sessions—one lasting 16 hours with an attack happening roughly once every second, and another lasting 15 hours with a higher attack rate of about ten per second. Features were extracted using tools like Argus and BRO-IDS, capturing various aspects of the network traffic such as flow characteristics, content details, and timing information. The dataset includes labels for both normal traffic and nine different types of attacks, offering a comprehensive and challenging benchmark for testing detection models [16].

Fig. 2 illustrates the distribution of attack types within the dataset, highlighting the dominance of the Exploits category, followed by Fuzzers and DoS, while rare classes such as Worms and Shellcode are underrepresented [14].

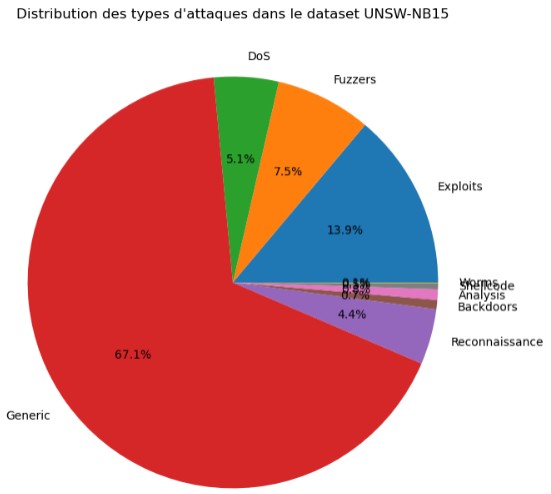


Fig. 2. Distribution of attack types in the UNSW-NB15 dataset.

Feature Selection and Preprocessing :

To enhance model efficiency, we performed feature engineering by:

- Dropping categorical and ID-based features: id, proto, service, state, attack_cat.
- Standardizing numerical features using StandardScaler.
- Encoding the labels using LabelEncoder, converting the binary classification task into a numerical format.
- Handling class imbalance via class weighting to ensure balanced training.

B. Model Selection Rationale

While Recurrent Neural Networks (RNNs), particularly LSTM architectures, are traditionally used for modeling sequential data such as firewall logs, recent studies have shown that one-dimensional Convolutional Neural Networks (1D CNNs) can also perform well in such tasks, especially when speed and scalability are priorities [22]. CNNs can efficiently extract local features from sequential data with significantly lower computational overhead compared to RNNs [23]. In our approach, we selected a 1D CNN due to its fast training time and suitability for real-time log analysis. To address concerns regarding model choice, we also conducted a comparative experiment using an LSTM-based model (see Table I).

C. Model Architecture

This study explores various Deep Learning architectures to improve threat detection. We designed and tested multiple architectures, including Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM models .

Proposed Hybrid CNN-LSTM Model :

The final selected architecture is a hybrid CNN-LSTM model, which leverages:

- CNN layers for spatial feature extraction.
- LSTM layers for capturing sequential dependencies.
- Batch Normalization to improve training stability and convergence.
- Dropout layers to prevent overfitting.

D. Model Structure

The proposed deep learning model follows a hybrid CNN-LSTM architecture, which is composed of the following layers:

- Input Layer: Reshaped input to $(39, 1)$.
- Convolutional Layers:
  - Conv1D (128 filters, kernel size = 3, ReLU activation).
  - Batch Normalization and Dropout (30%).
  - Conv1D (64 filters, kernel size = 3, ReLU activation).
  - Batch Normalization and Dropout (30%).
- LSTM Layer: Extracts temporal dependencies.
- Dense Layers: Fully connected layers for classification.
- Output Layer: Sigmoid activation for binary classification.

Model Architecture

The deep learning model designed for this study is a Deep Neural Network (DNN) composed of several densely connected layers. The architecture is detailed as follows:

- Input layer: 20 neurons corresponding to the selected features.

- Hidden layers:
  - First layer: 64 neurons, ReLU activation, Batch Normalization, Dropout (0.3).
  - Second layer: 32 neurons, ReLU activation, Batch Normalization, Dropout (0.3).
  - Third layer: 16 neurons, ReLU activation, Batch Normalization.
- Output layer: 9 neurons with Softmax activation for multiclass classification.

Training and Hyperparameters :

The models were trained using the following configuration:

- Optimizer: Adam (learning rate = 0.001)
- Loss Function: Binary Cross-Entropy
- Batch Size: 256
- Epochs: 100 (with Early Stopping, patience = 10)
- Validation Split: 20%
- Callbacks: ReduceLROnPlateau and EarlyStopping

E. Performance Metrics

To evaluate model performance, we used the following metrics:

- Accuracy
- Precision, Recall, and F1-score
- Confusion Matrix

V. Results

A. Model Performance Comparison

Table I presents the test accuracy obtained by each evaluated model, including MLP, CNN, LSTM, and the proposed CNN-LSTM architecture :

TABLE I. Test Accuracy Comparison of Different Models

Model	Test Accuracy
MLP	85.04%
CNN	81.95%
LSTM	83.17%
CNN-LSTM (Proposed)	**86.34%**

Fig. 3 illustrates the training progression of the proposed CNN-LSTM model in terms of accuracy over the epochs :

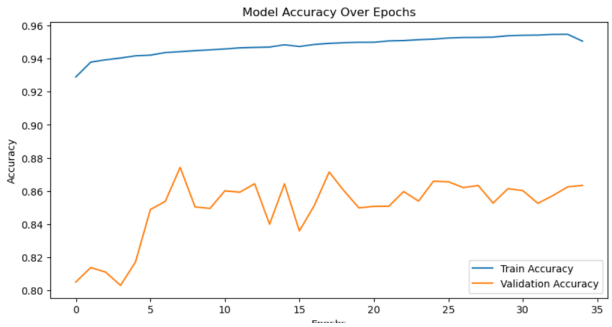


Fig. 3. Accuracy over epochs.

To further evaluate the model's classification performance, Fig. 4 shows the confusion matrix generated from the test dataset :

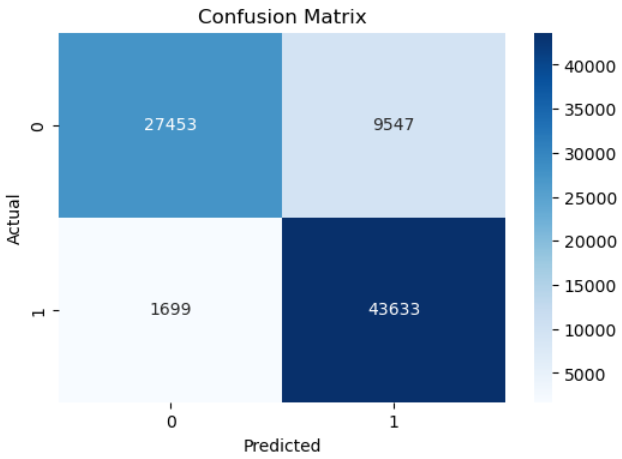


Fig. 4. Confusion matrix.

Additionally, the classification report displayed in Fig. 5 provides detailed metrics including precision, recall, and F1-score for each class.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.74      0.83     37000
           1       0.82      0.96      0.89     45332

    accuracy                           0.86     82332
   macro avg       0.88      0.85      0.86     82332
weighted avg       0.87      0.86      0.86     82332
```

Fig. 5. Classification report.

VI. Discussion

Compared to existing SIEM solutions, our proposed AI-SIEM demonstrates superior performance in detection accuracy, reduced false positives, and adaptability to evolving threats (see Table I). Unlike rule-based systems, our approach can learn complex attack patterns and generalize to novel data.

Table I presents the test accuracy results for the different deep learning models evaluated in our study. The Multi-Layer Perceptron (MLP) achieved a solid performance with an accuracy of 85.04%, while the Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) models reached 81.95% and 83.17%, respectively. Our proposed hybrid model, combining CNN and LSTM architectures, outperformed the others with the highest test accuracy of 86.34%. These results highlight the effectiveness of integrating spatial and temporal feature extraction for improving threat detection in firewall log analysis.

To verify the performance of our AI-SIEM model, we observed its accuracy at training epochs. As shown in Fig. 3, the model enhanced continuously with increasing epochs, and the curve of accuracy converged slowly, indicating steady learning and effective generalization.

In order to analyze the performance of the proposed AI-driven firewall detection model, we compared the confusion matrix, a common evaluation metric for classification issues. The matrix shown in Fig. 4 provides a detailed split of the predictions made by the model versus actual ground truth. It consists of four primary elements: True Positives (TP), where the model correctly identifies an attack; True Negatives (TN), where it correctly identifies normal traffic; False Positives (FP), where innocent traffic is wrongly labeled as malicious; and False Negatives (FN), where actual attacks are not detected. High TPs and TNs with low FP and FN show high detection capacity and reliability of the system. The results from the confusion matrix in this study confirm again that the model not only identifies with very high accuracy but also has a very low false positive rate, which is important to minimize unnecessary alarms in real-world security contexts. These findings vindicate the suitability of utilizing deep learning techniques for intelligent log analysis in future SIEM systems.

In addition to the confusion matrix assessment, a classification report with extensive information was utilized to evaluate the performance of the proposed model. The report in Fig. 5 provides significant metrics such as precision, recall, and F1-score for both classes: normal traffic (class 0) and attacks (class 1). For regular examples, the model was 94% accurate, so most predictions that were labeled as regular were indeed correct. However, the recall for the class was only slightly lower at 74%, so that some regular cases were being incorrectly labeled as attacks. In classifying attacks, though, the model was very good with a recall rate of 96% and precision of 82%, so it will catch most bad activity.

The F1-score, or the balance between precision and recall, was 0.83 for the normal class and 0.89 for the attack class, indicating balanced performance across classes. The overall accuracy of the model on the test set was 86% out of a total of 82,332 samples. Moreover, the macro average F1-score was 0.86, confirming well-balanced detection between the two classes, while the weighted average was also 0.86, confirming the reliability of the model in handling skewed data. The results confirm the effectiveness of the AI-augmented approach towards distinguishing normal and malicious traffic during real-world attacks.

TABLE II. TOP 5 SELECTED FEATURES BASED ON RELEVANCE SCORES

Rank	Feature	Score
1	sbytes	0.91
2	dbytes	0.88
3	ct_state_ttl	0.85
4	ct_src_dport_ltm	0.82
5	ct_dst_sport_ltm	0.80

A. Discussion and Key Findings

To gain a deeper understanding of the model's performance, we carried out a series of evaluation experiments supported by visual metrics. These analyses provide insight not only into the overall accuracy but also into how well the model handles different traffic types and attack categories. The key takeaways are as follows:

- The hybrid CNN-LSTM model achieved the highest accuracy (86.34%), outperforming traditional MLP and LSTM models.
- The confusion matrix shows that the model effectively classifies normal and attack traffic, though some misclassification remains between certain attack categories.
- Precision-Recall curves indicate a strong tradeoff, with high recall ensuring better attack detection.
- The ROC Curve AUC confirms the robustness of our model, achieving a high area under the curve.

VII. DATA PREPROCESSING

To ensure optimal model performance, several preprocessing techniques were applied:

- Data cleaning was performed to remove duplicate records and handle missing values.
- Categorical features were encoded using one-hot encoding.
- Feature selection was conducted via the SelectKBest algorithm, retaining the top 20 features based on their statistical relevance to the target variable. Notable selected features include:
  - sbytes
  - dbytes
  - ct_state_ttl
- Normalization of numerical attributes was performed using MinMaxScaler to harmonize feature scales between 0 and 1.
- The dataset was split into 70% training, 15% validation, and 15% testing subsets to assess generalization.

To boost the model's performance while avoiding unnecessary complexity, we used a feature selection process based on relevance scores. This allowed us to concentrate on the most meaningful variables in the dataset. The top five features identified—sbytes, dbytes, ct_state_ttl, ct_src_dport_ltm, and ct_dst_sport_ltm_are listed in Table II. These features capture

essential characteristics of network traffic and played a key role in helping the model accurately distinguish between normal and malicious activity.

A. Advantages of the Proposed Approach

Compared to existing SIEM solutions, our proposed AI-driven architecture demonstrates significant advantages. As shown in Table I, the CNN-LSTM model outperforms traditional approaches in terms of detection accuracy. In addition, it shows reduced false positive rates and stronger adaptability to evolving attack patterns.

Unlike conventional rule-based SIEM systems, which rely on manually defined signatures or heuristics, our model is capable of learning complex relationships within the data. This allows it to detect previously unseen threats and generalize to new types of network behavior. These strengths position the model as a valuable asset for next-generation threat detection in dynamic and large-scale environments.

VIII. Conclusion

In this study, we discussed an AI-enabled SIEM architecture designed to parse firewall logs intelligently and improve cyber threat detection by using deep learning models. Our main goal was to remove false alarms and improve overall threat detection accuracy using three types of models, namely, Fully Connected Neural Networks (FCNN), Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTM).

To give a comprehensive and realistic evaluation, we trained and tested our models with actual firewall log data as well as the benchmark datasets UNSW-NB15-training and UNSW-NB15-testing. Among the three, the LSTM model performed best, particularly, since it can learn temporal patterns and process sequential data providing higher accuracy and lower false alarms.

By integrating these powerful deep learning techniques into the SIEM solution, we unveiled a wiser and watchful solution that was capable of identifying suspicious behavior in complicated network environments. Our findings reinforce the ability of AI in broadening the functionality of SIEM and paving the way for proactive and better cybersecurity monitoring.

Future work will focus on improving real-time detection capabilities, expanding the system to incorporate additional data sources such as endpoint and operating system logs, and conducting live performance evaluations in enterprise production environments. Furthermore, issues such as model generalizability, dependency on labeled data, and minimizing false positives warrant ongoing attention. Addressing these limitations will be crucial to deploying robust, scalable AI-enhanced SIEM solutions in practice.

References

[1] Toprak, Sezer, and Ali Gökhan Yavuz. "Web application firewall based on anomaly detection using deep learning." Acta Infologica 6.2 (2022): 219-244.

[2] Lu, Meng, Catherine Chen, and Carsten Eickhoff. "Cross-Encoder Rediscovers a Semantic Variant of BM25." arXiv preprint arXiv:2502.04645 (2025).

[3] Sola, Rajendra Prasad, Nihar Malali, and Praveen Madugula. Cloud Database Security: Integrating Deep Learning and Machine Learning for Threat Detection and Prevention: 0. Notion Press, 2025.

[4] Tuyishime, Emmanuel, et al. "Enhancing cloud security—proactive threat monitoring and detection using a siem-based approach." Applied Sciences 13.22 (2023): 12359.

[5] Marri, Rahul, Sriram Varanasi, and Satwik Varma Kalidindi Chaitanya. "Integrating Next-Generation SIEM with Data Lakes and AI: Advancing Threat Detection and Response." Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 3.1 (2024): 446-465.

[6] Ahmadi, Sina. "Next generation ai-based firewalls: a comparative study." International Journal of Computer (IJC) 49.1 (2023): 245-262.

[7] Wang, Shuzhan, et al. "Deep learning-based anomaly detection and log analysis for computer networks." arXiv preprint arXiv:2407.05639 (2024).

[8] Shetty, Rahul. "Enhancing Context-Aware Search with Retrieval-Augmented Generation." Authorea Preprints (2025).

[9] Lillmond, Chandesh, and Geerish Suddul. "A Deep Neural Network Approach for Analysis of Firewall Log Data." (2021).

[10] et al. Afrah Fathima, "Cyber security Reinforcement through Firewall Log Analysis and Machine Learning", IJRITCC, vol. 11, no. 10, pp. 777–782, Nov. 2023.

[11] Comparative Analysis of Anomaly Detection Approaches in Firewall Logs: Integrating Light-Weight Synthesis of Security Logs and Artificially Generated Attack Detection.

[12] Aljabri, Malak, et al. "Classification of firewall log data using multiclass machine learning models." Electronics 11.12 (2022): 1851.

[13] AL-Tarawneh, Batool A., and Hani Bani-Salameh. "Classification of firewall logs actions using machine learning techniques and deep neural network." AIP Conference Proceedings. Vol. 2979. No. 1. AIP Publishing, 2023.

[14] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

[15] Tiwari, Anamika, et al. "Implementing Robust Cyber Security Strategies to Protect Small Businesses from Potential Threats in the USA." Journal of Ecohumanism 4.3 (2025): 322-333.

[16] Li, Yueyang. "Detection for Malicious Network Traffic Based on Convolutional Neural Networks."

[17] Conti, Mauro, Tooska Dargahi, and Ali Dehghantanha. "Cyber threat intelligence: challenges and opportunities." Cyber threat intelligence (2018): 1-6.

[18] Liang, Junyan, and Yoohwan Kim. "Evolution of firewalls: Toward securer network using next generation firewall." 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2022.

[19] Rangaraju, Sakthiswaran. "Secure by intelligence: enhancing products with AI-driven security measures." EPH-International Journal of Science And Engineering 9.3 (2023): 36-41.

[20] Kolli, RAJA KUMAR, E. Priyanshi, and S. Gupta. "Palo Alto Firewalls: Security in Enterprise Networks." International Journal of Engineering Development and Research, 12 (3), 1-13. rjwave ijedr/viewpaperforall. php? paper= IJE DR200A001 (2024).

[21] Shirazi, Patrick, and Ali Padyab. "Discerning Challenges of Security Information and Event Management (SIEM) Systems in Large Organizations." International Symposium on Human Aspects of Information Security and Assurance. Cham: Springer Nature Switzerland, 2024.

[22] Kim, Jiyeon, et al. "CNN-based network intrusion detection against denial-of-service attacks." Electronics 9.6 (2020): 916.

[23] Umair, Muhammad Basit, et al. "A network intrusion detection system using hybrid multilayer deep learning model." Big data 12.5 (2024): 367-376.

Automated Dried Fish Classification Using MobileNetV2 and Transfer Learning

Rajmohan Pardeshi[1], Rajermani Thinakaran[2], Sanjay Kharat[3]
Department of Computer Science, Karnatak Arts, Science and Commerce College, Bidar, Karnataka, India[1]
Faculty of Data Science and Information Technology (FDSIT),
INTI International University, Nilai, Negeri Sembilan, Malaysia[2]
Department of Zoology, Modern College of Arts, Science and Commerce, Ganeshkhind, Pune, India[3]

***Abstract*—India, the second largest fish producer globally, contributes significantly to food security, nutrition, and economic development. Dried fish is a vital component of the fisheries value chain, especially in South Asia, yet current classification methods are manual, inconsistent, and labor-intensive. This study aims to automate dried fish classification using MobileNetV2 through transfer learning, enabling real-time, lightweight deployment on edge devices. We trained and evaluated the model across four diverse publicly available datasets using single, bulk, head, and tail image modalities. Our experiments demonstrated high accuracy (up to 100%) and strong generalization across datasets. The proposed model offers a practical, scalable, and efficient solution to modernize dried fish processing and enhance productivity and traceability in fisheries.**

***Keywords*—Dried fish classification; MobileNetV2; transfer learning; edge deployment; fisheries automation**

I. Introduction

India is the second largest fish producer globally, with an extensive coastline of 8,118 km, making fisheries a crucial sector for food security, nutrition, and economic development. The industry contributes 6.56% to global fish production and over 5.37% to the country's agricultural Gross Value Added (GVA), providing livelihoods to millions [1]. With increasing export earnings and government initiatives such as the Blue Revolution, there is a significant potential for sustainable growth. Integrating advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML) can enhance productivity, modernize infrastructure, and strengthen the fisheries value chain, ultimately maximizing its economic and nutritional benefits.

Among the various seafood products, dried fish is one of the most well-known and widely consumed fish commodities. It is also recognized as an affordable and rich source of protein [2]. Advances in fish processing technologies, including preservation, waste utilization, and safety assurance, have expanded applications from the food industry to pharmaceuticals [3] [4]. The current methods for dry fish processing are mostly manual, time consuming and labor-intensive. Given the industrial and economic significance of dried fish processing, it is essential to develop technological solutions for automating processing, storage, and sorting. As a step toward this objective, we propose an AI-based solution to optimize these processes, improving efficiency and sustainability in the fisheries sector by introducing the deep learning based dried fish classification system [5].

Despite the growing use of AI in fish classification, prior studies have focused predominantly on fresh or underwater fish species, often employing complex, resource intensive models unsuitable for real-time deployment in rural fish markets. There is limited research addressing dried fish, particularly those species endemic to India and Bangladesh. Furthermore, previous works rarely compare multiple imaging modalities (e.g., head, tail, bulk), which are critical for robust classification under practical conditions. This study fills these gaps by applying MobileNetV2 for cross-dataset evaluation under various scenarios, targeting deployment in constrained environments.

The remainder of this study is organized as follows: Section II presents a review of existing literature related to fish classification techniques. Section III details the proposed methodology, including the transfer learning strategy using MobileNetV2. Section IV discusses the experimental results, including dataset descriptions and performance metrics. It also presents a discussion of key findings and their implications. Section V concludes the study and outlines future research directions.

II. Literature Review

Automatic classification of fish has gained attention for more than four decades and considered image classification and shape analysis as a problem of computer vision. The necessity for automatic sorting of fish by species has been addressed in [6] [7]. These researchers characterized fish shape using length measurements (defined as a horizontal straight line from the tip of the head to the base of the tail) and width measurements (defined as the distance from the top to the bottom of the body) taken at several equidistant points along the length. They discovered that they could sort four fish species with a reliability of 95% and seven species with an accuracy of 90%.

The study presented in [8] is for fish species classification. Authors developed three techniques namely, invariant moments, optimization of the mismatch of shape, and shape descriptors for differentiating between fish species. The discriminant analysis method was used for classification and achieved the classification rate of 73%, 63%, and 90%, respectively. In [9], the authors presented the work based on color and shape features such as aspect ratio and area. Discriminant analysis was applied for classification of 18 fish species achieving 98% of accuracy.

A novel artificial neural network-based technique for autonomously counting fish is introduced in [10]. To count artificial fish populations, a back propagation of error feed-forward neural network has been trained. It is then demonstrated that trained networks generalize effectively to scenes of fish tanks that have never been seen before, with 94% success rate on scenarios with up to 100 fish in a range of overlaps and orientations. Authors presented a classifier combination approach for classification of five species of fishes. They applied moment invariants, Fourier descriptors and shape features with combination of local vector quantization, naïve Bayes and MLP classifier and achieved 70% accuracy [11].

An automatic fish species recognition system [12] was designed using shape recognition methods, comparing test fish contours with a database. A power cepstrum-based matching technique was developed and evaluated alongside other methods. Despite a 60% recognition accuracy, the system significantly reduces manual effort and holds potential for biological research. A study explored the identification of a suitable feature set for fish classification [13]. It utilized four size measurement features, 19 shape measurement features, eight color signature features, and 16 texture measurement features. Using a Bayesian classifier, the study achieved 87% accuracy in classifying six fish species. Trials of a computer vision system, The CatchMeter, for identifying and measuring fish species are described [14]. Fish are transported along a conveyor under a digital camera, where image processing algorithms determine orientation using a moment-invariant method. The system achieves 100% accuracy in distinguishing flatfish from roundfish, measures length with a standard deviation of 1.2 mm, and identifies species with up to 99.8% sorting reliability for seven fish species. This study presents an automatic fish species classification framework using image analysis and artificial immune systems. It integrates Scale-Invariant Feature Transform and Principal Component Analysis for feature extraction, Artificial Immune Network and Adaptive Radius Immune Algorithm for clustering, and a nearest neighbor classifier. Validated on economically significant fish species, the framework achieved 92% accuracy, offering a robust alternative to manual classification methods [15].

A state-of-the-art computer vision method for fine-grained fish species classification using deep learning is presented [16]. The approach employs a cross-layer pooling algorithm with a pre-trained convolutional neural network as a generalized feature extractor, reducing the need for extensive training data. Classification is performed using an SVM on extracted features, achieving 94.3% accuracy on fish species from underwater video imagery captured off the coast of Western Australia. This study introduces a novel approach for automatic remote acoustic identification of fish using pattern recognition techniques [17]. Fish calls are extracted from background noise and parameterized using LFCC, MFCC, Shannon Entropy, and Syllable Length. Three machine learning algorithms—KNN, Random Forest, and SVM—were used for classification, successfully identifying 102 fish species with average accuracies of 95.24%, 93.56%, and 95.58%, respectively. This study proposes a fish species classification method using a 32-layer VGGNet with a deep hierarchical supervision mechanism [18]. Five controlled subnetworks with additional convolutional layers, varied kernels, and hidden layer features were introduced to optimize training and enhance feature extraction. The approach achieved 96% accuracy on the Pak Fish dataset.

An automated real-time deep learning framework was developed using convolutional neural networks and Kalman filters for fish species classification. YOLOv3 and Mask-RCNN achieved mAP scores of 0.73 and 0.62, respectively [19]. The adapted YOLO model successfully detected and classified eight fish species from a high-resolution DIDSON sonar dataset captured in the Ocqueoc River, Michigan, USA. This study presents a dataset of fish tray images [20] from a local wholesale fish market, featuring pixel-wise (mask) labeled specimens, species information, and size measurements. The dataset includes 1,291 labeled images, comprising 7,339 specimens from 59 species across 60 class labels. It is valuable for evaluating fish instance segmentation and size estimation methods, which are crucial for automated stock management systems and can help to sustain fish populations over time. This image dataset is designed for benchmarking automated fish detection and classification algorithms. The dataset contains 69,917 fish tags representing 30 different taxa, with the image content extracted through tagging for accurate identification and analysis [21].

This study highlights the use of artificial intelligence in fish classification and taxonomy, demonstrating its ability to surpass traditional methods in identifying fine-grained morphological features and advancing aquatic biodiversity research [22]. AI's role in conservation, citizen science initiatives, and education is emphasized, showcasing its potential to enhance species protection and environmental preservation. More details about the fish classification can be found in [23]. This study presents a deep learning and convolutional neural network (CNN) model [24] for classifying various types of dried fish, including locally recognized species like Bashpata, Chanda, Ilish, and Tengra. The dataset [29] was augmented and segmented, achieving an accuracy of 97.72%. In the recent study given by authors, the database of Indian dried fishes having 5 common categories and several deep learning models were evaluated for accuracy before and after training. EfficientNetB0 achieved the highest accuracy (95.23%), followed by ResNet50 (94.03%) and Xception (93.89%). InceptionV3 and VGG16 showed relatively lower final accuracies of 89.56% and 87.56%, respectively. These results highlight EfficientNetB0's superior performance in feature extraction and classification.

Despite extensive research on fish classification using image processing and machine learning techniques, most studies primarily focus on fresh fish or underwater species. There is a notable lack of dedicated research on dried fish classification, particularly for regionally significant species found in India and Bangladesh. While deep learning models such as convolutional neural networks (CNNs) have been employed for fish classification, limited studies have explored transfer learning approaches, specifically MobileNetV2, for dried fish classification. Moreover, existing research predominantly relies on full-body fish images, with minimal investigation into classification performance based on different perspectives, such as single images, head, tail, and bulk images. This gap is critical, as considering multiple perspectives could enhance the robustness of classification models in real-world applications, such as wholesale markets and automated sorting systems. Addition-

ally, many studies utilize computationally expensive models like YOLO, Mask-RCNN, and large CNN architectures, which require substantial resources. There is a pressing need for lightweight and efficient models capable of performing real-time dried fish classification on edge devices. Another significant gap lies in dataset availability. While datasets for fresh fish classification exist, comprehensive datasets for dried fish are scarce. The available datasets often lack diversity in terms of species representation, image quality, and environmental variations, which can impact model generalizability. Although our study attempts to address these limitations by evaluating classification performance across multiple perspectives and conducting detailed experiments, further research is needed to develop specialized, transfer-learning-based models that ensure improved accuracy, computational efficiency, and adaptability to real-world deployment.

III. Transfer Learning using MobileNetV2

Transfer learning is a deep learning technique that uses a pre-trained model on a large dataset and fine-tunes it for a specific task with a smaller dataset [25]. This approach is particularly effective in scenarios with limited training data, as it allows the model to use previously learned features to improve performance in new tasks.

Mathematically, if $f(x;\theta)$ represents a deep learning model with parameters θ, transfer learning can be expressed as:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i;\theta)) \tag{1}$$

where:

- x_i and y_i are input-output pairs in the new dataset,
- $\mathcal{L}$ is the loss function,
- θ^* represents the optimized parameters adapted from the pre-trained model.

MobileNetV2, a lightweight deep convolutional neural network (CNN), is widely used for transfer learning in image classification tasks due to its efficiency in mobile and embedded applications [26]. MobileNetV2 is a convolutional neural network (CNN) architecture designed for efficient performance in resource-constrained environments. It employs inverted residuals and linear bottlenecks, enhancing both accuracy and computational efficiency [27] [36]. The core of MobileNetV2 consists of depthwise separable convolutions, which decompose standard convolutions into two operations: depthwise convolution, where a single filter is applied per input channel, and pointwise convolution, a 1×1 convolution that increases the channel dimension. This approach significantly reduces computational complexity compared to conventional convolutional layers [28]. Additionally, MobileNetV2 introduces inverted residual blocks, where shortcut connections exist between narrow bottleneck layers instead of wide layers, thereby improving feature reuse [26]. Each block follows a sequence: an expansion layer, which scales up the input dimension; a depthwise convolution, which reduces spatial dimensions while maintaining depth; and a projection layer (linear bottleneck), which compresses features into a lower-dimensional space.

Algorithm 1 Dried Fish Image Classification Using MobileNetV2 and Transfer Learning

Require: Image dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{h\times w\times c}$ and y_i is the label
Ensure: Trained classification model $f_\theta : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^K$
Step 1: Data Preprocessing
Load dataset $\mathcal{D}$ from directory
Resize images: $x_i \leftarrow \text{resize}(x_i, (224, 224)) \quad \forall x_i \in \mathcal{D}$
Normalize images: $x_i \leftarrow \frac{x_i}{255} \quad \forall x_i \in \mathcal{D}$
Split dataset into training set $\mathcal{D}_{train}$ (80%) and testing set $\mathcal{D}_{test}$ (20%)
Step 2: Model Construction
Initialize pretrained MobileNetV2 backbone $\phi_{\text{MBV2}} : \mathbb{R}^{224\times224\times3} \to \mathbb{R}^d$
Freeze backbone layers: $\phi_{\text{MBV2}} \leftarrow$ frozen
Define classifier head:
    $h_\theta \leftarrow \text{Dense}(128, \text{ReLU})$
    $h_\theta \leftarrow \text{Dense}(128, \text{ReLU})$
    $h_\theta \leftarrow \text{Dense}(K, \text{Softmax})$
Construct final model: $f_\theta = h_\theta \circ \phi_{\text{MBV2}}$
Compile model with Adam optimizer and categorical cross-entropy loss
Step 3: Model Training
for $t = 1$ to T (epochs) **do**
    Train model on $\mathcal{D}_{train}$ using backpropagation
    if Validation loss does not improve for 5 epochs **then**
        Stop Training (Early Stopping)
    end if
end for
Step 4: Model Evaluation
Compute predictions: $\hat{y}_i = f_\theta(x_i) \quad \forall x_i \in \mathcal{D}_{test}$
Compute evaluation metrics:
    Accuracy: $\frac{1}{|\mathcal{D}_{test}|} \sum_i \mathbf{1}(\hat{y}_i = y_i)$
    Precision, Recall, and F1-score from confusion matrix
Generate classification report
End

For dried fish classification, MobileNetV2 is fine-tuned using a dataset containing various species of dried fish images. The preprocessing pipeline includes image resizing, data augmentation (rotation, flipping, normalization), and dataset splitting (80% training, 20% validation) [25]. Transfer learning is applied by leveraging a pre-trained MobileNetV2 model on ImageNet, replacing its fully connected layers with a custom classifier, as given in Fig. 1. The classification function is modeled as a softmax function, where extracted features from the pre-trained layers are fed into a newly trained classification head [26]. The training process involves minimizing cross-entropy loss, optimizing with the Adam optimizer, fine-tuning only the later layers while freezing earlier ones, and evaluating performance using accuracy, precision, recall, and F1-score. This approach ensures efficient and accurate classification of dried fish species while maintaining computational efficiency.

The key parameters used in the model include image size (224×224), learning rate (0.001), batch size (32), and number of dense layers in the classifier head. The choice of MobileNetV2 as a backbone minimizes the number of

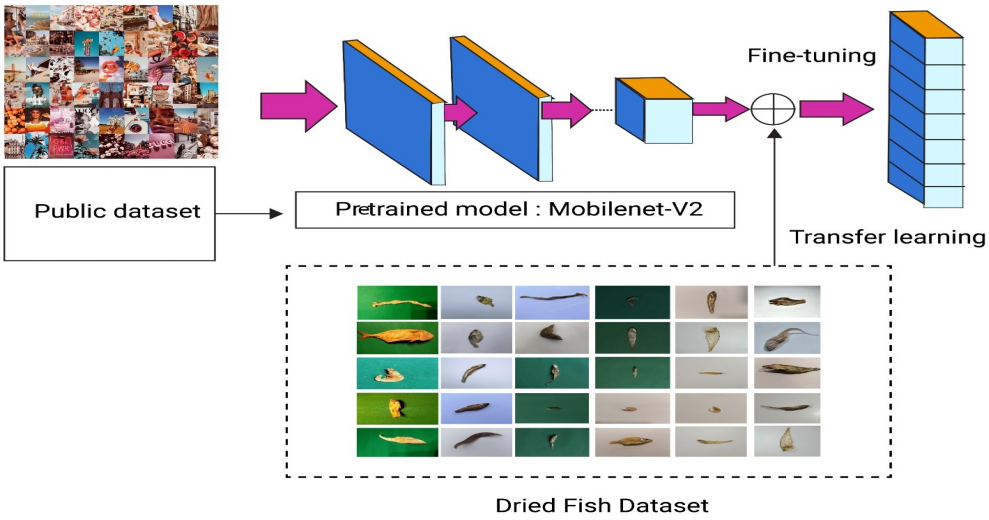


Fig. 1. Transfer learning-based dried fish image classification framework. [35]

trainable parameters (3.4M), supporting fast training. Early stopping was used to prevent overfitting. Sensitivity analysis showed that increasing the number of dense layers beyond two led to overfitting without significant accuracy improvement. Lowering the learning rate increased training time but did not yield significantly better results, justifying the current choice. Algorithm 1 details the dried fish image classification using MobileNetV2 and Transfer learning.

IV. Results and Discussion

A. Dataset Description

The datasets for this research were selected based on their relevance, public availability, and diversity of dried fish species from South Asian regions, especially India and Bangladesh, where dried fish is commercially significant. These datasets—Indian Dried Fish, BDDryFish, DriedFishBD, and SAR-DFD-2025—offer varying image perspectives (single, bulk, head, tail) and environmental conditions, providing a robust testbed for evaluating classification performance.

1) Indian dried fish image dataset [29]: This dataset consists of 8,290 images of dried fish, including both single and bulk images. It covers five fish categories commonly found in India: Prawns (Shrimp), Small Anchovi (Tingle), Golden Anchovi (Mandeli), Mackerel (Bangada), and Bombay Duck (Bombil). The dataset provides 6,784 single images and 1,506 bulk images, making it useful for fish classification and quality assessment studies (see Table I).

2) BDDryFish an image dataset of dry fishes found in Bangladesh [30]: This dataset includes 1,251 images of seven different dried fish species commonly found in Bangladesh. The fish categories include Shundori Shutki Aluva, Bele Fish, Boicha Shutki, Chingri Fish, Churi Shutki, Faisha Shutki, and Mola Fish. The dataset is useful for species identification and classification research in the domain of dried fish processing and trade (see Table II).

3) DriedFishBD a detailed image dataset of common small-sized dried fish varieties in Bangladesh [31]: This dataset contains 3,288 images of eight different small-sized dried fish species. Each fish category is represented through multiple perspectives, including single, bulk, head, and tail images, making it a rich dataset for detailed morphological analysis. The categories include Chanda, Chapila, Chela, Chepa, Guchi, Kachki, Loitta, and Tengra (see Table III).

4) SAR-DFD-2025 a high-quality and diverse dry fish dataset from South Asia [32]: SAR-DFD-2025 is a comprehensive, high-quality dataset designed for dry fish species identification in South Asia. It features a diverse collection of 2,138 images across 11 fish species, captured under varying lighting conditions and angles. The dataset includes species such as Boro Chingri, Chanda, Chela, Kachki, Kocho Chingri, Mola, Tengra, Poti, Shundori, Taki Actual, and Faissa. This dataset serves as a crucial resource for image-based fish species recognition and classification, aiding research in fisheries and related fields (see Table IV).

TABLE I. Dataset1 : Indian Dried Fish Image Dataset [29]

Sr. No	Category	No. of Single Images	No. of Bulk Images
1	Prawns (Shrimp)	1021	245
2	Small Anchovi (Tingle)	1051	408
3	Golden Anchovi (Mandeli)	977	499
4	Mackerel (Bangada)	2633	114
5	Bombay Duck (Bombil)	1102	240
Total		**6784**	**1506**

B. Experimental Results

The experimental results for the classification of dried fish species using deep learning models are presented in this section. The classification was conducted on multiple datasets, including Indian Dried Fish, BDDryfish, DriedFishBD, and

TABLE II. BDDRYFISH: AN IMAGE DATASET OF DRY FISHES FOUND IN BANGLADESH [30]

Sr. No	Category	Total Images
1	Shundori Shutki Aluva	208
2	Bele Fish	202
3	Boicha Shutki	209
4	Chingri Fish	219
5	Churi Shutki	101
6	Faisha Shutki	104
7	Mola Fish	208

TABLE III. DATASET 3: DRIEDFISHBD - A DETAILED IMAGE DATASET OF COMMON SMALL-SIZED DRIED FISH VARIETIES IN BANGLADESH [31]

Sr. No	Category	Single	Bulk	Head	Tail	Total
1	Chanda	100	100	100	100	400
2	Chapila	120	100	102	100	422
3	Chela	100	100	100	100	400
4	Chepa	100	100	100	100	400
5	Guchi	105	108	103	111	427
6	Kachki	121	107	103	104	435
7	Loitta	100	100	100	100	400
8	Tengra	100	100	100	104	404

SARDFD2025. The models were evaluated based on precision, recall, and F1-score, which measure the effectiveness of the classification. Accuracy is also reported to provide an overall performance assessment.

Each dataset was tested using different imaging approaches, including single images, bulk images, head-only images, tail-only images, and full-body images. The following subsections provide a detailed breakdown of the classification results for each dataset.

Table V presents the classification results for Indian dried fish species using a single image per sample. The model achieved high precision, recall, and F1-score across all classes, with an overall accuracy of 100%. This indicates that the model correctly identified the species without any misclassification.

The slight variation in recall for Mandeli and precision for Tingali suggests that the model had minor confusion when classifying these species. However, the F1-score remained nearly perfect, demonstrating robust performance.

In the bulk image classification setting Table VI, multiple fish were present in a single image for classification. The overall accuracy remained high at 99%, with some minor variations in precision and recall. Here, Bombayduck showed a slight drop in precision (0.95), suggesting some misclassification in bulk image scenarios. However, the model still maintained strong performance across all classes.

The classification results for BDDryfish, another dried fish dataset, are shown in Table VII. The model achieved an overall accuracy of 99%, with minor variations in precision and recall across some classes. Here, Alua and Chingri fish had slight variations in recall (0.98 and 0.96, respectively), suggesting some misclassification. However, the overall accuracy remained strong, indicating that the model performed well across different fish species.

TABLE IV. SAR-DFD-2025: A HIGH-QUALITY AND DIVERSE DRY FISH DATASET FROM SOUTH ASIA [32]

Sr. No	Fish Category	No. of Images
1	Boro Chingri	214
2	Chanda	240
3	Chela	252
4	Kachki	282
5	Kocho Chingri	257
6	Mola	241
7	Tengra	230
8	Poti	156
9	Shundori	120
10	Taki Actual	84
11	Faissa	62

TABLE V. CLASSIFICATION RESULTS FOR INDIAN DRIED FISH SPECIES USING SINGLE IMAGES

Class	Precision	Recall	F1-score
Bombayduck	1.00	1.00	1.00
Mackeral	1.00	1.00	1.00
Mandeli	1.00	0.99	1.00
Prawns	0.99	0.99	0.99
Tingali	0.98	1.00	0.99
Accuracy	1.00 (Overall)		

TABLE VI. CLASSIFICATION RESULTS FOR INDIAN DRIED FISH SPECIES (BULK IMAGES)

Class	Precision	Recall	F1-score
Mandeli	1.00	1.00	1.00
Prawn	1.00	1.00	1.00
Tingali	0.99	0.98	0.98
Bombayduck	0.95	0.98	0.97
Mackeral	1.00	1.00	1.00
Accuracy	0.99 (Overall)		

TABLE VII. CLASSIFICATION RESULTS FOR BDDRYFISH DATASET

Class	Precision	Recall	F1-score
Alua	1.00	0.98	0.99
Bele fish	1.00	1.00	1.00
Boicha shutki	1.00	1.00	1.00
Chingri fish	1.00	0.96	0.98
Churi shutki	1.00	1.00	1.00
Faisha shutki	0.96	1.00	0.98
Mola fish	0.95	1.00	0.97
Accuracy	0.99 (Overall)		

The classification results for the DriedFishBD dataset are evaluated under four different settings: Full Image, Head Only, Tail Only, and Bulk. Each setting analyzes how well the model classifies various dried fish species based on different input

TABLE VIII. CLASSIFICATION RESULTS FOR DRIEDFISHBD DATASET

Class	Precision	Recall	F1-score
Chanda	1.00	1.00	1.00
Chapila	0.96	1.00	0.98
Chela	1.00	1.00	1.00
Chepa	1.00	1.00	1.00
Guchi	0.91	0.95	0.93
Kachki	1.00	1.00	1.00
Loitta	1.00	1.00	1.00
Tengra	1.00	0.89	0.94
Accuracy	0.98 (Overall)		

TABLE IX. CLASSIFICATION RESULTS FOR DRIEDFISHBD (HEAD ONLY)

Class	Precision	Recall	F1-score
Chanda	1.00	1.00	1.00
Chapila	1.00	1.00	1.00
Chela	1.00	1.00	1.00
Chepa	0.95	1.00	0.98
Guchi	1.00	1.00	1.00
Kachki	1.00	0.95	0.97
Loitta	1.00	1.00	1.00
Tengra	1.00	1.00	1.00
Accuracy	0.99		

TABLE X. CLASSIFICATION RESULTS FOR DRIEDFISHBD (TAIL ONLY)

Class	Precision	Recall	F1-score
Chanda	1.00	1.00	1.00
Chapila	1.00	1.00	1.00
Chela	1.00	1.00	1.00
Chepa	1.00	1.00	1.00
Guchi	0.90	1.00	0.95
Kachki	1.00	1.00	1.00
Loitta	1.00	0.95	0.97
Tengra	1.00	0.88	0.94
Accuracy	0.98		

TABLE XI. CLASSIFICATION RESULTS FOR DRIEDFISHBD (BULK)

Class	Precision	Recall	F1-score
Chanda	1.00	1.00	1.00
Chapila	1.00	1.00	1.00
Chela	1.00	1.00	1.00
Chepa	1.00	1.00	1.00
Guchi	1.00	1.00	1.00
Kachki	1.00	1.00	1.00
Loitta	1.00	1.00	1.00
Tengra	1.00	1.00	1.00
Accuracy	1.00		

conditions.

Table VIII presents the classification performance when using the full image of the dried fish. The model achieved a high accuracy of 98%, indicating strong performance in distinguishing the eight fish species. Several species, such as Chanda, Chela, Chepa, Kachki, and Loitta, were classified with perfect precision, recall, and F1-score values of 1.00. However, minor classification errors were observed for Chapila (F1-score = 0.98), Guchi (F1-score = 0.93), and Tengra (F1-score = 0.94), suggesting that these species may have overlapping visual characteristics.

Table IX illustrates the classification results when using only the head portion of the dried fish. The overall accuracy slightly improved to 99%. Most species achieved an F1-score of 1.00, except Kachki (F1-score = 0.97). The improved performance compared to full image classification suggests that the head region contains distinctive features that aid in species identification.

Table X reports the classification performance when using only the tail portion of the fish. The overall accuracy remained at 98%, similar to full image classification. While most species were classified perfectly, slight performance drops were observed for Guchi (F1-score = 0.95), extitLoitta (F1-score = 0.97), and Tengra (F1-score = 0.94). This suggests that the tail alone may not be as distinctive as the head for classification purposes.

Table XI presents the classification results when using bulk images, where multiple fish appear in a single image. The model achieved perfect accuracy (100%) across all classes, indicating that it successfully identified every species without error. This suggests that bulk classification benefits from multiple views of the fish, allowing the model to generalize effectively.

The experimental results indicate that the model performs best on bulk images (100% accuracy), followed by head-only images (99%), full images (98%), and tail-only images (98%).

The head region provides more distinguishing features than the tail, leading to better classification performance.

The species Guchi and Tengra exhibit slightly lower F1-scores, indicating potential visual similarity with other species.

The consistently high accuracy ($\geq 98\%$) across all settings validates the robustness of the classification model for the DriedFishBD dataset.

These results confirm that the proposed classification model is highly effective in distinguishing dried fish species under different conditions.

Table XII presents the classification results obtained for the SARDFD2025 dataset. Precision measures the proportion of correctly identified instances for a given class among all instances predicted as that class. Recall quantifies the proportion of correctly identified instances among all actual instances of that class. The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of classification performance. The model achieves an overall accuracy of 97%, indicating a high level of classification performance across all classes. The individual class-wise results reveal strong performance across various categories.

The model demonstrates excellent classification performance for multiple fish species, including Faissa and Taki

TABLE XII. CLASSIFICATION RESULTS FOR SARDFD2025

Class	Precision	Recall	F1-score
Boro Chingri	0.95	0.95	0.95
Chanda	0.97	1.00	0.99
Chela	0.92	0.98	0.95
Faissa	1.00	1.00	1.00
Kachki	0.98	0.97	0.97
Kocho Chingri	0.98	0.98	0.98
Mola	0.98	0.92	0.95
Poti	1.00	0.97	0.99
Shondori	0.96	0.96	0.96
Taki Actual	1.00	1.00	1.00
Tengra	0.98	1.00	0.99
Accuracy		0.97	

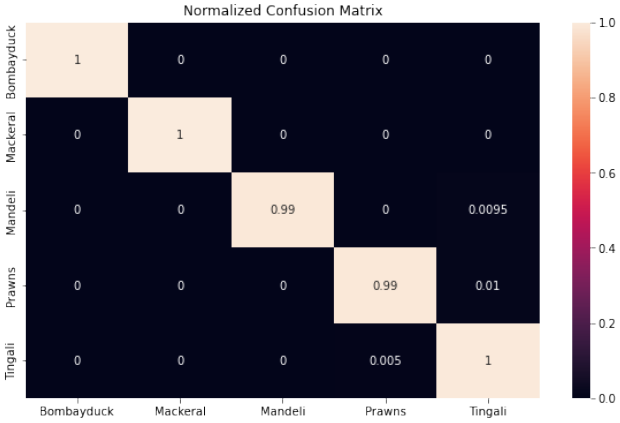


Fig. 2. Confusion matrix for Indian dried fish species using single images.

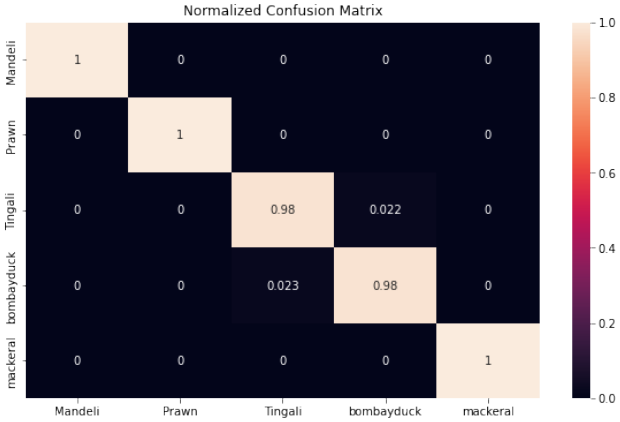


Fig. 3. Confusion matrix for Indian dried fish species (bulk images).

Actual, where both precision and recall reach 1.00, implying perfect classification. The majority of fish classes, such as Boro Chingri, Chela, Kocho Chingri, Mola, and Tengra, achieve F1-scores above 0.95, highlighting the reliability of the model in distinguishing between different species. A few classes exhibit minor differences between precision and recall values. For instance, Mola has a slightly lower recall (0.92) compared to its precision (0.98), suggesting some instances were not correctly

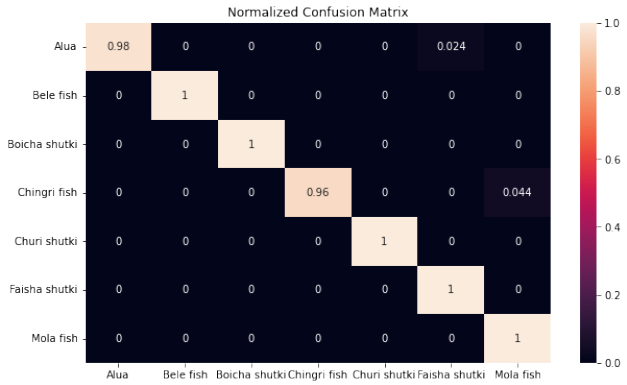


Fig. 4. Confusion matrix for BDDryfish dataset.

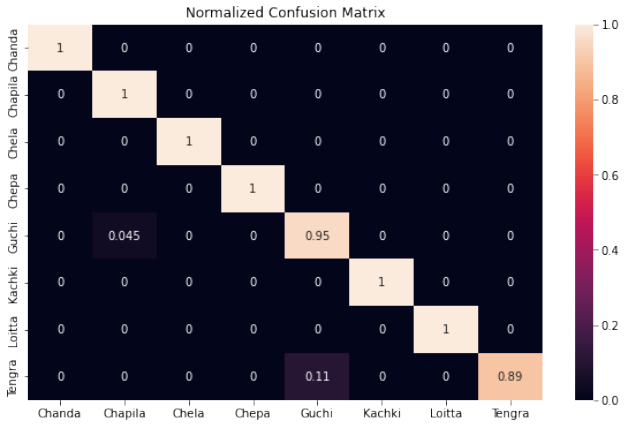


Fig. 5. Confusion matrix for DriedFishBD dataset.

identified.

To evaluate the classification performance of the proposed model, we computed the accuracy and its 95% confidence interval (CI) [33] [34] for different dried fish image datasets. The confidence intervals were derived using the standard binomial proportion method to assess the statistical reliability of the classification results.

The model achieved high classification accuracy across all datasets, indicating its robustness in distinguishing dried fish categories. The Indian Dried Fish (single images) dataset recorded the highest accuracy of 100%, with a 95% CI of (0.9994, 1.0000), confirming the model's near-perfect performance.

For the Indian Dried Fish (bulk images) dataset, the accuracy was 99%, with a 95% CI of (0.9863, 0.9951). Similarly, the BDDryfish dataset achieved 99% accuracy, with a 95% CI of (0.9847, 0.9971). The DriedFishBD dataset showed an accuracy of 98%, with a 95% CI of (0.9721, 0.9903), demonstrating slightly lower but still highly reliable classification performance. We have also given the confusion matrices for all datasets in Fig. 2 to Fig. 9 and a comparison between the predicted and actual labels for a sample from the DriedFishBD (Single Images) dataset [31] in Fig. 10 for better understanding of the results.

These results suggest that the model maintains high generalizability across different datasets, ensuring effective clas-

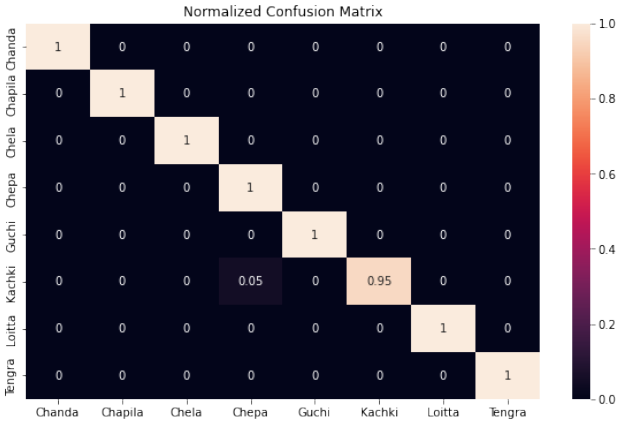


Fig. 6. Confusion matrix for DriedFishBD (head only).

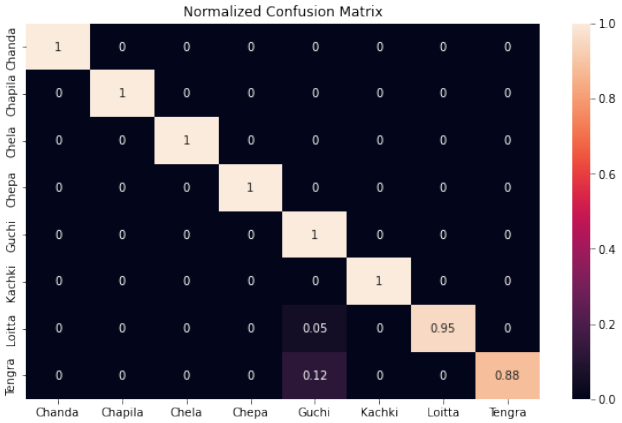


Fig. 7. Confusion matrix for DriedFishBD (tail only).

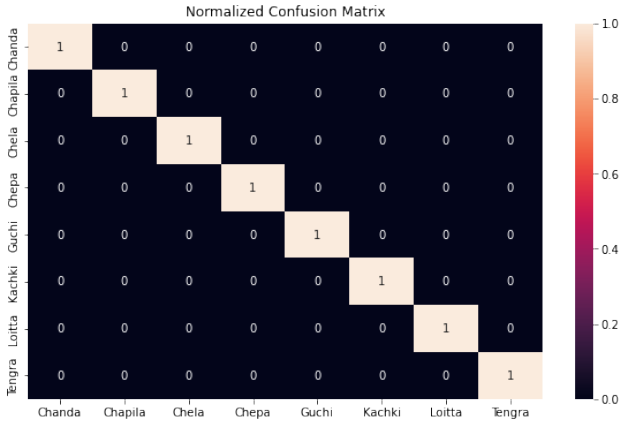


Fig. 8. Confusion matrix for DriedFishBD (bulk).

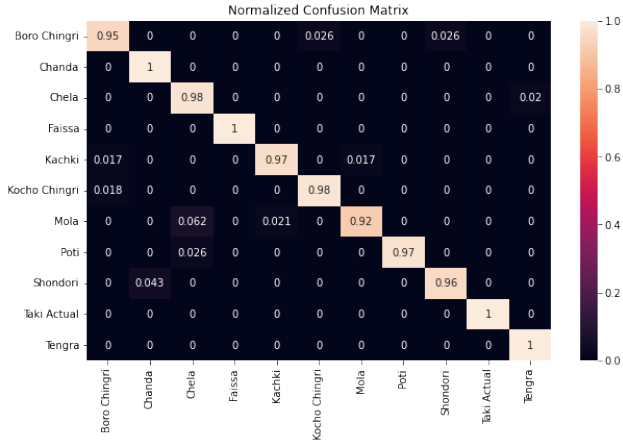


Fig. 9. Confusion matrix for SARDFD2025.

sification regardless of variations in dried fish images. The marginal differences in accuracy across datasets may be attributed to variations in image quality, lighting conditions, and dataset-specific characteristics.

The narrow confidence intervals further validate the consistency and reliability of the proposed model. The findings indicate that the model can be effectively used in real-world applications, such as automated fish species identification in fisheries and quality control processes.

While the proposed approach achieves high accuracy across diverse datasets, there are some limitations. First, the datasets used have relatively small sample sizes for certain classes, potentially affecting model generalization in highly imbalanced real-world settings. Second, despite using transfer learning, the model requires moderate GPU resources during training, which may limit its applicability for institutions with limited infrastructure. Third, classification errors occurred in species with similar morphological features, especially under varying lighting conditions. These factors suggest opportunities for further refinement through dataset augmentation and hybrid models.

V. Conclusion and Future Work

This study demonstrates the effectiveness of a MobileNetV2-based transfer learning framework for automated dried fish classification across diverse South Asian datasets. Theoretically, the results validate the feasibility of using lightweight CNNs for species-level classification in low-data domains. Practically, the model supports real-time classification on mobile and edge devices, enabling scalable deployment in local markets and fisheries. The research contributes a novel evaluation of dried fish classification using different image modalities and introduces benchmarks on unexplored dried fish datasets from India and Bangladesh. The proposed method reduces manual labor, increases classification consistency, and supports digitization and traceability in fisheries processing. Its low computational cost suits resource limited environments. However, class imbalance and morphological similarity between species limit precision in specific cases. Performance under real-time video streams and environmental variability also needs validation.

Future studies should focus on integrating Vision Transformers or hybrid CNN-transformer architectures for improved feature extraction, developing a larger, balanced dataset of dried fish species across regions, and implementing explainable AI techniques for model transparency in commercial settings.

Acknowledgment

The authors would like to thank INTI International University Persiaran Perdana BBN, Putra Nilai 71800 Nilai, Negeri

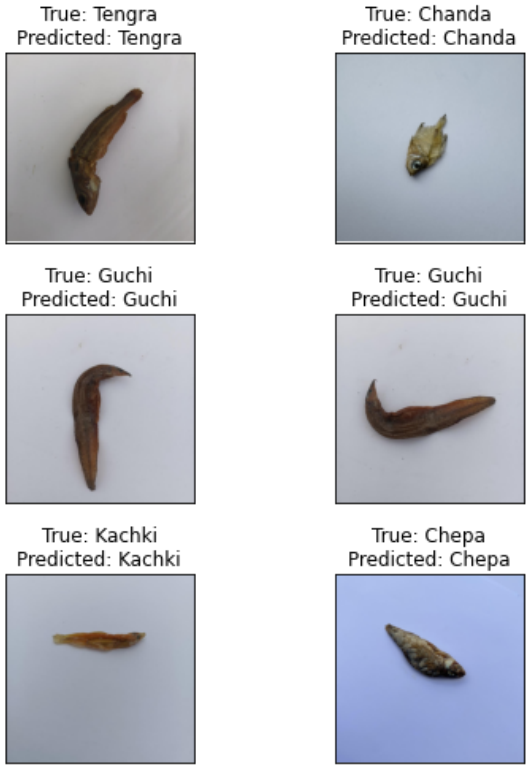


Fig. 10. Few examples of predicted and actual labels for DriedFishBD [31] (single images).

Sembilan for supporting this work.

REFERENCES

[1] Marine Products Export Development Authority. 2018. *Annual Report 2017–18.* Ministry of Commerce and Industry, Government of India. Accessed March 20, 2025. https://mpeda.gov.in/wp-content/uploads/2020/11/1550120514MPEDAAR201718.pdf.

[2] Kallon, A., A. Lebbie, B. Sturm, T. Garnett, and R. Wadsworth. 2017. "Comparative Studies of Fish Smoking and Solar Drying in the Sierra Leone Artisanal Fishing Industry." *Journal of Stored Products and Postharvest Research* 8 (3): 40–48. https://doi.org/10.5897/JSPPR2016.0217.

[3] Majumder, Ranendra K., and Amjad K. Balange, eds. 2023. *Advances in Fish Processing Technologies: Preservation, Waste Utilization, and Safety Assurance.* CRC Press.

[4] Madan, M. S., Radhakrishnan Kalidoss, L. Ranjith, R. Narayanakumar, N. Aswathy, and K. P. Kanthan. 2018. "Economics and Marketing of Dry Fish Production in Thoothukudi District, Tamil Nadu, India." *Indian Journal of Fisheries* 65 (4): 135–141.

[5] Subash, Abhirami, Hareesh N. Ramanathan, and Marko Šostar. 2024. "Market-Driven Mapping of Technological Advancements in the Seafood Industry: A Country-Level Analysis." *Economies* 12 (11): 313. https://doi.org/10.3390/economies12110313.

[6] Tayama, I., Shimdate, M., Kubuta, N., and Nomura, Y., "Application of optical sensor for fish sorting," *Reito (Tokyo) Refrigeration*, vol. 57, no. 661, pp. 1146-1150, 1982.

[7] Shimdate, M., "Automatic fish species and fish size sorting equipment using pattern recognition," *Mitsubishi Denki Hiko (Tokyo)*, vol. 56, no. 3, pp. 44-48, 1982.

[8] Strachan, N. J. C., Nesvadba, P., and Allen, A. R., "Fish species recognition by shape analysis of images," *Pattern Recognition*, vol. 23, no. 5, pp. 539-544, 1990.

[9] Strachan, N. J. C., "Recognition of fish species by colour and shape," *Image and Vision Computing*, vol. 11, no. 1, pp. 2-10, 1993.

[10] Newbury, P. F., Culverhouse, P. F., and Pilgrim, D. A., "Automatic fish population counting by artificial neural network," *Aquaculture*, vol. 133, no. 1, pp. 45-55, 1995.

[11] Cadieux, S., Michaud, F., and Lalonde, F., "Intelligent system for automated fish sorting and counting," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, vol. 2, pp. 1279-1284, 2000.

[12] Lee, D. J., Schoenberger, R. B., Shiozawa, D., Xu, X., and Zhan, P., "Contour matching for a fish recognition and migration-monitoring system," in *Proc. SPIE Two-and Three-Dimensional Vision Systems for Inspection, Control, and Metrology II*, vol. 5606, pp. 37-48, 2004.

[13] Nery, M. S., Machado, A. M., Montenegro Campos, M. F., Pádua, F. L. C., Carceroni, R., and Queiroz-Neto, J. P., "Determining the appropriate feature set for fish classification tasks," in *Proc. XVIII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 173-180, 2005.

[14] White, D. J., Svellingen, C., and Strachan, N. J. C., "Automated measurement of species and length of fish by computer vision," *Fisheries Research*, vol. 80, no. 2-3, pp. 203-210, 2006.

[15] Rodrigues, M. T. A., Pádua, F. L. C., Gomes, R. M., and Soares, G. E., "Automatic fish species classification based on robust feature extraction techniques and artificial immune systems," in *Proc. IEEE Fifth Int. Conf. Bio-Inspired Computing: Theories and Applications (BIC-TA)*, pp. 1518-1525, 2010.

[16] Siddiqui, S. A., Salman, A., Malik, M. I., Shafait, F., Mian, A., Shortis, M. R., and Harvey, E. S., "Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data," *ICES Journal of Marine Science*, vol. 75, no. 1, pp. 374-389, 2018.

[17] Noda, J. J., Travieso, C. M., and Sánchez-Rodríguez, D., "Automatic taxonomic classification of fish based on their acoustic signals," *Applied Sciences*, vol. 6, no. 12, p. 443, 2016. DOI: https://doi.org/10.3390/app6120443.

[18] Rauf, H. T., Lali, M. I. U., Zahoor, S., Shah, S. Z. H., Rehman, A. U., and Bukhari, S. A. C., "Visual features-based automated identification of fish species using deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 167, p. 105075, 2019.

[19] Kandimalla, V., Richard, M., Smith, F., Quirion, J., Torgo, L., and Whidden, C., "Automated detection, classification and counting of fish in fish passages with deep learning," *Frontiers in Marine Science*, vol. 8, p. 823173, 2022.

[20] Garcia-d'Urso, N., Galan-Cuenca, A., Pérez-Sánchez, P., Climent-Pérez, P., Fuster-Guillo, A., Azorin-Lopez, J., Saval-Calvo, M., Guillén-Nieto, J. E., and Soler-Capdepón, G., "The DeepFish computer vision dataset for fish instance segmentation, classification, and size estimation," *Scientific Data*, vol. 9, no. 1, p. 287, 2022.

[21] Francescangeli, M., Marini, S., Martínez, E., Del Río, J., Toma, D. M., Nogueras, M., and Aguzzi, J., "Image dataset for benchmarking automated fish detection and classification algorithms," *Scientific Data*, vol. 10, no. 1, p. 5, 2023.

[22] Wasik, S., and Pattinson, R., "Artificial intelligence applications in fish classification and taxonomy: Advancing our understanding of aquatic biodiversity," *FishTaxa-Journal of Fish Taxonomy*, vol. 31, 2024.

[23] Saleh, A., Sheaves, M., and Azghadi, M. R., "Computer vision and deep learning for fish classification in underwater habitats: A survey," *Fish and Fisheries*, vol. 23, no. 4, pp. 977-999, 2022.

[24] Laboni, M. A., Rimi, I. F., Deb, S., Afrin, F., and Hena, M. H., "Dried Fish Classification Using Deep Learning," in *Proc. Second Int. Conf. Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1-6, IEEE, 2022.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[26] M. Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *CVPR*, 2018.

[27] A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[28] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *CVPR*, 2017.

[29] Paygude, Priyanka, Milind Gayakwad, Dhanashri Wategaonkar, Rajendra Pawar, Ramchandra Pujeri, and Rahul Joshi. "Dried fish dataset for Indian seafood: A machine learning application." Data in Brief 55 (2024): 110563. Accessed in January 2025 from PAYGUDE, PRIYANKA (2024), "Dry Fish Dataset", Mendeley Data, V1, doi: 10.17632/b35yrbh762.1

[30] Islam, Mohammad Manzurul; Akash, Md. Anik Khan ; Suhita, Sabira Musfirat (2025), "BDDryFish: An image dataset of dry fishes found in Bangladesh", Mendeley Data, V1, doi:10.17632/4xt4pjwk9k.1 Accessed on February 2025

[31] Ferdaus, Md Hasanul; Prito, Rizvee Hassan; Ahmed, Masud; Haider, Lamia; Hossain, Shajia; Nuzhat, Subah; Shifa, Sadia Tasnim; Islam, Mohammad Manzurul; Ali, Sawkat; Islam, Maheen; Rasel, Ahmed Abdal Shafi; Rahoman, Md Mizanur (2024), "DriedFishBD: A Detailed Image Dataset of Common Small-sized Dried Fish Varieties in Bangladesh for Identification and Classification.", Mendeley Data, V2, doi:10.17632/vz4xgybxr2.2 Accessed on February 2025

[32] Sarower, Afjal (2025), "SAR-DFD-2025: A High-Quality & Diverse Dry Fish Dataset from South Asia", Mendeley Data, V1, doi:10.17632/9ktwk5yh5r.1 Accessed in March 2025

[33] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

[34] Bouckaert, R. R. (2004). Estimating replicability of classifier learning experiments. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 15–22.

[35] Wu, Xiaoxin & He, Yigang & Duan, Jiajun. (2020). A Deep Parallel Diagnostic Method for Transformer Dissolved Gas Analysis. Applied Sciences. 10. 1329. 10.3390/app10041329.

[36] Diana, D., Kurniawan, T. B., Dewi, D. A., Alqudah, M. K., Alqudah, M. K., Zakari, M. Z., and Fuad, E. F. B. E. (2025). *Convolutional Neural Network Based Deep Learning Model for Accurate Classification of Durian Types*. Journal of Applied Data Sciences, 6(1), 101–114.

Enhancing Portfolio Optimization with Weighted Scoring for Return Prediction Through Machine Learning and Neural Networks

Ruili Sun, Qiongchao Xia, Shiguo Huang
Zhengzhou University of Light Industry
School of Mathematics and Information Science, Zhengzhou, China

***Abstract*—Accurately predicting stock return can enhance the effectiveness of portfolio optimization models. Many previous studies typically divide machine learning algorithms and portfolio optimization into two separate stages: the first step leverages the powerful modeling capabilities of machine learning algorithms to select stocks, and the second step optimizes weights using traditional portfolio models. This separation means that the modeling strengths of machine learning are only utilized in the stock selection phase and not fully exploited during weight optimization. Therefore, this study proposes a portfolio construction method based on Return Prediction Weighted Scoring (RPWS). RPWS generates a stock ranking by assigning weighted scores to each stock, cleverly maps this ranking to weight biases, and then optimizes actual weights using a traditional covariance matrix. This process successfully integrates the modeling capabilities of machine learning into the weight optimization phase, ensuring its full utilization throughout the portfolio construction process. Backtesting experiments are conducted using the U.S. stock market, A-share market, and major cryptocurrencies as datasets, with Support Vector Regression (SVR), Transformer, and other machine learning algorithms as prediction models. Empirical results from these three markets show that the SVR-RPWS and Transformer-RPWS models significantly outperform mainstream funds and traditional portfolio models in terms of annualized returns, sharpe ratio, and drawdown control.**

***Keywords*—Machine learning; stock return prediction; portfolio optimization; support vector regression; transformer; NASDAQ stock market; a-share stock market; cryptocurrency market**

I. Introduction

The optimization of a portfolio selection entails the strategic distribution of assets to maximize returns and minimize risk, two pivotal elements in this process. The endeavor to enhance profits while mitigating risk is a primary goal for investors. The genesis of portfolio selection as a domain of study is attributed to the introduction of the Mean-Variance model (Markowitz [1]), which evaluates return in terms of mean and risk in terms of variance. Subsequently, variance emerges as a commonly used risk measure, and studies related to models such as the Mean-Variance model gain popularity within the academic community, drawing significant attention from scholars such as Björk et al. [2], Liagkouras and Metaxiotis [3], Rosadi et al. [4] and Katsikis et al. [5].

Mean-Variance model relies on the anticipated return and risk associated with different assets to generate optimal portfolios corresponding to various levels of expected return and risk (Beheshti [6]). Consequently, through the careful selection of assets for inclusion in the optimization process, the Mean-Variance model has the potential to enhance overall portfolio performance (Thakur et al. [7]). In recent years, numerous prediction-based portfolio selection models have been introduced in the realm of portfolio management. When confronted with equivalent expected returns, a portfolio selection model boasting a more favorable efficient frontier can effectively reduce risk exposure. Therefore, enhancing the efficacy of prediction-based portfolio selection models remains a crucial endeavor in maximizing portfolio performance.

In financial market, individual investors are generally interested in understanding the fluctuations in the returns of their investment assets today, the potential trends in returns for tomorrow, and the strategies that should be implemented to optimize their portfolio composition (Zhang et al. [8]). Consequently, integrating forecasting theory into portfolio construction holds significant promise for financial investment (Kolm et al. [9]). Forecasting financial time series is consistently seen as one of the most formidable tasks due to the dynamic, nonlinear, unstable, and complex characteristics, coupled with the long-term fluctuations of the financial market (Chen and Hao [10]; Paiva [11]). Nevertheless, reliable investment decisions should be based on long-term observations and behavioral patterns of asset data rather than short-term data (Chong et al. [12]). Accordingly, it is imperative to examine the changes and volatility of financial data over an extended historical period to adequately prepare for future trend forecasts and investment decisions. Numerous broadly accepted empirical studies suggest that financial time series exhibit memory of past periods, indicating that financial markets are, to some extent, predictable. The long-term behavior of an asset significantly influences the risks and returns of a portfolio, thereby further impacting investment decisions (Liu and Loewenstein [13]).

To date, the predominant focus of prior research has been on the utilization of statistical methodologies and machine learning techniques in the predictive analysis domain. Statistical methodologies endeavor to anticipate forthcoming trends by scrutinizing historical pricing attributes such as autoregressive integrated moving average (ARIMA) (Anderson et al. [14]), auto-regressive conditional heteroscedasticity (ARCH) (Engle [15]; Shephard [16]), and generalized auto-regressive conditional heteroscedasticity (GARCH) (Bollerslev [17]; Garcia et al. [18]), whereas prevalent machine learning methodologies encompass support vector machines (Villegas et al. [19]), random forests (Kavzoglu and Teke [20]), and neural networks (Ma [21]). Notwithstanding the capacity of statistical methods

to contribute to stock price forecasting, empirical investigations have revealed that machine learning exhibits superior efficacy in addressing challenges associated with non-stationarity and non-linearity when compared to traditional statistical models (Boulesteix and Schmid [22]; Zhang et al. [23]). Recently, several studies (Freitas et al. [24], Jiang et al. [25], Kaczmarek [26], Padhi et al. [27]) have incorporated the pre-selection of assets into portfolio selection models utilizing machine learning techniques. For example, Deng and Min [28] implemented a linear regression model with ten factors to select stocks from both U.S. and international markets, subsequently constructing a mean-variance-based portfolio that accounted for practical risk tolerance, error monitoring, and turnover constraints. Their results demonstrated that the risk-adjusted return of the global equities model surpassed that of the domestic equity universe, with portfolio returns increasing in conjunction with systemic monitoring errors and risk tolerance. Similarly, Ma [21] employed two machine learning models—random forest (RF) and support vector regression—alongside three deep learning models: convolutional neural network (CNN), long short-term memory (LSTM) neural network, and deep multilayer perceptron (DMLP), to forecast returns for portfolio construction. They subsequently utilized these forecasts to enhance omega portfolio optimization and Mean-Variance model. Their findings revealed that the return predictions for Mean-Variance model and omega models using the RF algorithm outperformed those produced by the other models.

The utilization of machine learning and neural networks in the realm of stock forecasting has garnered considerable success, as evidenced by various studies (Ballings et al. [29]; Sezer and Ozbayoglu [31]; Tsai and Hsiao [32]). These findings advocate for an integration of machine learning methodologies with traditional models of portfolio optimization for the creation of investment portfolios. Numerous academics have been delving into this field of study, primarily branching out into three distinct avenues. Specifically, the first approach involves certain researchers employing Machine Learning models solely during the initial stock selection phase, subsequently utilizing the chosen stocks to construct portfolios through traditional portfolio optimization techniques (Paiva [11]; Vo et al. [33]; Wang et al. [34]). Evidence from these studies suggests that the utilization of stocks selected via machine learning algorithm can enhance the efficacy of conventional portfolio optimization frameworks. In the second vein, several scholars adopt machine learning algorithm for forecasting future stock returns, thereafter incorporating these forecasts to formulate novel objective functions. These newly developed functions aim to refine and augment the objective functions inherent in classical portfolio optimization models (Ma et al. [35]; Ustun and Kasimbeyli [36]; Yu et al. [37]; Yu et al.). Findings from these inquiries indicate that objective functions, enriched by predictive insights from machine learning algorithm, can significantly improve upon traditional portfolio optimization methods. Lastly, a third group of researchers leverages machine learning algorithm to forecast stock returns, opting to utilize the forecast errors in lieu of historical returns when constructing traditional portfolio optimization models (i.e., prediction-based portfolio optimization models). This preference stems from the observation that the normality of forecast errors surpasses that of historical returns (Freitas et al. [39]; Freitas et al. [24]; Hao et al. [40]; Ma et al. [41]). Insights from these studies reveal that forecast errors derived from machine learning algorithm prove more adept than historical returns for the purpose of enhancing portfolio optimization models.

Many previous studies primarily focus on using machine learning algorithms and neural network models to predict stock returns for stock selection, followed by optimizing weights through traditional portfolio models (for example, Minimum Variance Model, Mean VaR Model, Mean Absolute Deviation Model, and others) to enhance investment decisions. This study introduces a novel approach to portfolio construction that combines return predictions from machine learning and neural network models with historical covariance matrix, employing a weighted scoring method to enhance portfolio optimization. This method consists of two main steps, aimed at optimizing portfolio construction and risk management. Specifically, the first step involves using a trained machine learning model or neural network to predict the expected future returns of each asset in the backtest dataset. Once the expected returns are obtained, the assets are ranked according to these return values, and different weights are assigned to each asset through squared weighting. A higher weight corresponds to a higher expected return. The squared weighting method helps to increase the investment proportion in assets with higher expected returns, while relatively reducing the weight of assets with lower expected returns, thereby enhancing the overall potential return of the portfolio. The second step is to calculate the historical covariance matrix. The covariance matrix of the portfolio reflects the correlation and volatility of returns between different stocks. Then the covariance matrix is used to constrain the optimal portfolio weights. This method uses the covariance matrix optimization mechanism to improve expected returns while reasonably controlling portfolio risk, thereby achieving a more stable and optimized asset allocation in a complex market environment. This proposed approach will undergo backtesting in the constituents of the CSI 300 Index, the NASDAQ Index, and major cryptocurrencies, with comprehensive comparisons to traditional portfolio models based on the backtest results.

The rest of this study is organized as follows: Section II provides an overview of the machine learning algorithms, neural network models, and traditional portfolio models used in this study. Section III introduces the novel portfolio construction method proposed in this study, as well as the data selection and experimental design process. Section IV and Section V evaluate the performance of the predictive models and analyze the empirical results. Finally, Section VI concludes the work.

II. Methodology

This section will introduce the methods used in this study, including prediction models and traditional portfolio models.

A. Machine Learning Methods

XGBoost, which stands for "eXtreme Gradient Boosting" was introduced by Chen and Guestrin [42]. This algorithm is characterized by its low computational complexity, rapid execution speed, and high accuracy. The objective function of XGBoost integrates the standard penalty term with the loss function to derive the optimal solution. The inclusion of the regularization term serves to minimize the model's variance, thereby mitigating the risk of overfitting.

The k-nearest neighbors (KNN) algorithm, also known as k-NN, is a non-parametric, supervised learning method applied to both classification and regression tasks. In the context of KNN regression, the algorithm functions by identifying a subset of points in the dataset that are nearest to the "query" point, as determined by a specified distance metric, such as Euclidean, Minkowski, or Manhattan distance [43]. KNN also employs a hierarchical tree-based data structure to manage the dataset efficiently [44].

Support Vector Regression (SVR) represents a supervised learning algorithm specifically designed for regression tasks, as initially proposed by Cortes and Vapnik in 1995 [45]. The underlying principle of SVR is Vapnik's Structural Risk Minimization (SRM), which is adept at tackling various regression problems. The incorporation of kernel functions facilitates the handling of diverse types of input data, thereby enhancing the model's flexibility and applicability. Given its robustness and accuracy, SVR has emerged as a quintessential machine learning technique, extensively utilized in the field of stock market prediction (Emir [46]; Matías and Reboredo [47]; Rasel et al. [48]).

B. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM), a variant of Recurrent Neural Network (RNN), was first proposed by Hochreiter and Schmidhuber [49]. The LSTM model, distinct from traditional RNNs, possesses a memory function, enabling it to retain data over an extended period [50]. This unique feature is facilitated by gate structures, namely an input gate, a forget gate, and an output gate, which filter incoming information to enhance and sustain memory cells. LSTMs are particularly favored in financial time-series prediction due to their ability to effectively manage redundancy in historical data [51].

The operational formulas of LSTM are as follows:

Forget gate:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

Input gate:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c)$$

Output gate:

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t$$

$$h_t = o_t \tanh(c_t)$$

In these formulas, f_t, i_t, and o_t signify the forget gate, input gate, and output gate, respectively.

The weight of the matrix is represented by w, while b_t, b_i, and b_o denote the bias of the forget gate, input gate, and output gate, respectively.

The sigmoid function is denoted by σ, and x_t and h_t indicate the input and the current output at time t, respectively. The value from the input gate at time t is represented by c_t, and the hyperbolic function is represented by $\tanh$.

C. Transformer

The Transformer model, introduced by Vaswani et al. [52], represents a significant advancement in natural language processing and sequence modeling. Unlike recurrent architectures such as LSTMs, the Transformer relies entirely on self-attention mechanisms to process input sequences, enabling parallelization and enhancing efficiency. This self-attention mechanism allows the model to assess the importance of different words in a sequence when encoding information, effectively capturing long-range dependencies without the constraints of sequential processing.

The Transformer architecture comprises an encoder and a decoder, each consisting of multiple layers. The encoder processes an input sequence to generate a continuous representation, while the decoder utilizes this representation to produce an output sequence. The core operations within the encoder can be described by the following formulas:

First, input embeddings are augmented with positional encodings to preserve the order of the sequence:

$$z_i = x_i + PE(i)$$

where, z_i denotes the input embedding at position i, x_i is the original embedding, and $PE(i)$ represents the positional encoding for position i.

The self-attention mechanism computes attention scores for each position in the input sequence:

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

where, Q, K, and V are the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the keys, used to scale the dot-product attention scores.

Subsequently, the output from the attention mechanism is processed through a feed-forward neural network:

$$\text{FFN}(z) = \text{ReLU}(W_1 z + b_1)W_2 + b_2$$

where, W_1 and W_2 are weight matrices, and b_1 and b_2 are bias terms.

Each layer of the encoder includes residual connections and layer normalization to stabilize training:

$$\text{Output} = \text{LayerNorm}(z + \text{FFN}(\text{Attention}(Q, K, V)))$$

The decoder follows a similar structure, incorporating masked self-attention to prevent attention to future tokens, ensuring autoregressive output generation.

The Transformer has demonstrated exceptional performance across various tasks, including machine translation and text generation, owing to its efficiency in capturing complex dependencies in data. Its flexibility and scalability have established it as a cornerstone of modern deep learning architectures, significantly influencing subsequent models in the field [53].

D. Informer

Informer is a novel model designed for long sequence time-series forecasting, as proposed by Zhou et al. [54]. Unlike traditional time-series models, Informer employs a self-attention mechanism specifically optimized for efficient handling of long sequences. Its key innovation is an adaptive attention mechanism that focuses on relevant input features while discarding less relevant information, thus alleviating the computational burden associated with standard attention mechanisms.

The Informer architecture consists of an encoder and a decoder, similar to the Transformer model, but with enhancements for long-sequence performance. The core components of Informer are as follows:

The adaptive attention mechanism generates a sparse attention matrix using a probability distribution D:

$$D_{ij} = \frac{\exp(\text{score}(Q_i, K_j))}{\sum_{k \in \mathcal{N}} \exp(\text{score}(Q_i, K_k))}$$

where, $\text{score}(Q_i, K_j)$ represents the score function computed between query Q_i and key K_j, and $\mathcal{N}$ denotes the neighborhood of keys attended to by the query. This ensures that each query focuses on a subset of keys, significantly reducing computational complexity.

The encoder processes the input sequence to generate context-aware representations through multiple layers of self-attention and feed-forward networks:

$$Z^{(l)} = \text{LayerNorm}\left(Z^{(l-1)} + \text{Attention}(Z^{(l-1)}, Z^{(l-1)}, Z^{(l-1)})\right)$$
$$H^{(l)} = \text{LayerNorm}\left(Z^{(l)} + \text{FFN}(Z^{(l)})\right)$$

where, $Z^{(l)}$ is the output of layer l, and $H^{(l)}$ represents the result after applying the feed-forward network FFN.

In the decoder, the predicted output is generated based on the encoder's output and previous outputs. The decoder employs masked self-attention to ensure autoregressive prediction:

$$\hat{Y}_t = \text{LayerNorm}\left(Y_{t-1} + \text{MaskedAttention}(Y_{t-1}, Z)\right)$$

where, $\hat{Y}_t$ is the predicted output at time t, and Y_{t-1} denotes the previous outputs.

Informer performs well in forecasting tasks by effectively managing long-term dependencies and reducing computational overhead. Compared to previous methods, it shows significant improvements in accuracy and efficiency, Lu et al. [55].

E. Mean-Variance Model

The selection of investment strategies has always been a vital area in financial research. Amongst a wide array of investment strategies, the Mean-Variance Model, put forth by Harry Markowitz in 1952, has gained substantial recognition [1].

The model's theoretical rigor and practicality have led to its broad utilization in the practical realm of investment decisions. Its main idea is to find the investment proportion that minimizes the total risk of the portfolio at a given expected return level; or at a given risk level, find the investment proportion that maximizes the total return of the portfolio. Let $R = (R_1, R_2, \ldots, R_n)^T$, where $R_i = E(r_i)$ represents the expected return of the i-th asset; $X = (x_1, x_2, \ldots, x_n)^T$ is the weight vector of the portfolio; $\Sigma = (\sigma_{ij})_{n \times n}$ is the covariance matrix between n assets; $E(r_p)$ and σ_p^2 are the expected return and variance of the portfolio, respectively.

The mathematical formula for the Mean-Variance Model is as follows:

$$\begin{aligned} \min \quad & \sigma_p^2 = X^T \Sigma X \\ \max \quad & E(r_p) = X^T R \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = 1 \end{aligned}$$

Minimizing the variance (risk) of the portfolio and maximizing the expected return of the portfolio are two objectives that usually cannot be achieved simultaneously, so a trade-off between risk and return is needed. To compare with the portfolio methods proposed in this study, we will solve two special solutions of the mean-variance model through numerical methods.

The first extended model is the **Minimum Variance Portfolio Model (MVM)**, which aims to find the portfolio weight that minimizes the risk of the portfolio among all possible portfolio weight. The formula for this model is as follows:

$$\begin{aligned} \min \quad & \sigma_p^2 = X^T \Sigma X \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = 1 \end{aligned}$$

The second extended model is the **Maximum Sharpe Ratio Portfolio Model (MSRM)**, which aims to find the weight distribution that maximizes the Sharpe ratio (the ratio of expected return to risk) of the portfolio among all possible weight distributions. The formula for this model is as follows:

$$\max \quad \frac{X^T R - r_f}{\sqrt{X^T \Sigma X}}$$
$$\text{s.t.} \quad \sum_{i=1}^{n} x_i = 1$$

where, r_f stands for the risk-free rate, which is set to 0 in this study.

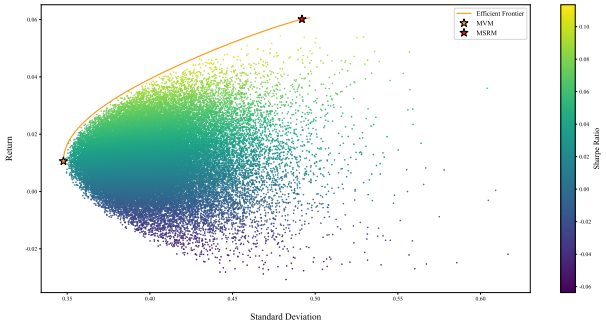


Fig. 1. Efficient frontier with random portfolios.

In Fig. 1, we can see a scatter plot composed of many random investment portfolios. The picture shows a curve called the efficient frontier. Efficient frontier points represent portfolios offering the optimal risk-return trade-off, either maximizing return for a given risk or minimizing risk for a given return. Among these points, two points are specially marked, one is the point with the minimum variance, and the other is the point with the maximum Sharpe ratio. These two points correspond to the solutions of the two models we mentioned above.

F. Risk Parity Model (RPM)

The risk parity strategy, proposed by Dr. Edward Qian, the Chief Investment Officer of Pan Agora, is renowned for its innovative approach to investment. This model was later applied in practice by Bridgewater Associates and achieved significant success for a time. From the perspective of risk budgeting, risk parity involves evenly distributing the overall risk of the portfolio among different assets, meaning that each asset contributes equally to the overall risk of the portfolio. In other words, in the risk parity model, assets with higher volatility are allocated lower weights, while those with lower volatility receive higher weights.

The marginal risk contribution (MRC) of an asset to a portfolio is defined as the partial derivative of the portfolio volatility with respect to the weight of the asset:

$$MRC_i = \frac{\partial \sigma_p}{\partial x_i} = \frac{\partial \sqrt{X^T \Sigma X}}{\partial x_i}$$

The contribution of the asset to the total risk (RC) is the product of the asset's weight and its marginal risk contribution:

$$RC_i = x_i \cdot MRC_i$$

For a risk parity portfolio, the contributions to total risk from each asset are equal, i.e.:

$$RC_i = RC_j$$

Calculating the weights for a risk parity portfolio essentially involves solving a quadratic optimization problem, which can be mathematically described as:

$$\min \quad \sum_{i=1}^{n} \sum_{j=1}^{n} (RC_i - RC_j)^2$$
$$\text{s.t.} \quad \sum_{i=1}^{n} x_i = 1$$
$$0 \leq x_i \leq 1$$

III. Newly Proposed Strategy for Portfolio Selection

A. Proposed Method: Return Prediction Weighted Scoring Strategy (RPWS)

In the field of portfolio optimization, Markowitz's (1952) mean-variance model stands as the most classical theoretical framework, constructing optimal portfolios by balancing expected returns and risk (volatility) [1]. However, the model's reliance on historical data assumptions may fail in rapidly changing markets. With advancements in machine learning and neural network technologies, numerous studies have validated the effectiveness of these techniques in stock prediction. For instance, Chaweewanchon et al. (2022) proposed a portfolio strategy combining the CNN-BiLSTM prediction model with the Markowitz Mean-Variance method. Backtesting experiments show that it outperforms other benchmark models in terms of Sharpe ratio, average return, and risk control [30].

Behera et al. (2023) proposed a portfolio construction method based on machine learning algorithms and the Mean-Value at Risk (Mean-VaR) model. First, stock returns are predicted using algorithms including XGBoost and AdaBoost to select stocks with higher expected returns; then, the Mean-VaR model is used for portfolio optimization. Based on stock data from the Bombay Stock Exchange in India, the Tokyo Stock Exchange in Japan, and the Shanghai Stock Exchange in China, experimental results show that the Mean-VaR model combined with AdaBoost prediction outperforms other combinations in performance [38].

Previous studies typically divide machine learning algorithms and portfolio optimization into two separate stages: the first step leverages the powerful modeling capabilities of machine learning algorithms to select stocks, and the second step optimizes weights using traditional portfolio models. This separation means that the modeling strengths of machine learning are only utilized in the stock selection phase and not fully exploited during weight optimization. Therefore, this study proposes a portfolio construction method based on Return Prediction Weighted Scoring (RPWS). RPWS generates a stock ranking by assigning weighted scores to each stock, maps this ranking to weight biases, and then optimizes actual

weights using a traditional covariance matrix. This process successfully integrates the modeling capabilities of machine learning into the weight optimization phase, ensuring its full utilization throughout the portfolio construction process.

The RPWS algorithm's mathematical model achieves dynamic weight optimization through two core components: 1) constructing a reference weight vector based on return prediction and rank weighting, and 2) covariance-constrained weight allocation optimization. The specific steps are as follows: Let $P = (p_1, p_2, p_3, \ldots, p_n)^T$, where p_i represents the ranking of the i-th stock among all stocks based on the expected return predicted by machine learning and neural network models. Let $\Sigma = (\sigma_{ij})_{n\times n}$ be the covariance matrix between n assets. The reference weight vector X_0 is constructed as:

$$X_0 = \frac{p_j^{-2}}{\sum_{j=1}^{n} p_j^{-2}}$$

The objective of this reference weight vector is to allocate higher weights to higher-ranked stocks. The optimal weights are then solved through convex optimization with risk budget constraints:

$$\begin{aligned} \min \quad & X^T \Sigma X + \lambda \|X - X_0\|_2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = 1 \end{aligned}$$

where,

- λ: Regularization parameter controlling the emphasis on the reference matrix (larger λ forces weights closer to X_0, degenerating to minimum variance model when $\lambda = 0$)
- $\|X - X_0\|_2$: L2 norm measuring the deviation between optimal weights and reference weights.

This formulation integrates machine learning predictions through X_0 while maintaining risk constraints through covariance matrix optimization. Considering that this is a convex optimization problem, in order to simplify the hyperparameter tuning, this study sets $\lambda = 1$ to solve the problem in the experiment.

B. Experimental Process

In financial markets, investors consistently aim to optimize their portfolios using scientific methods and strategies to achieve the optimal balance between risk and return.

However, the uncertainty and complexity of the market make constructing an efficient portfolio a challenging task. This study focuses on prediction as its core research direction, leveraging machine learning and neural network techniques to model and assign higher weights to stocks with higher potential returns and manageable risks, thereby constructing investment portfolios. By adopting this method, not only can the overall return expectations of the portfolio be improved, but the reliability and guidance of the decision-making process for investors can also be significantly enhanced. This methodology involves two main phases:

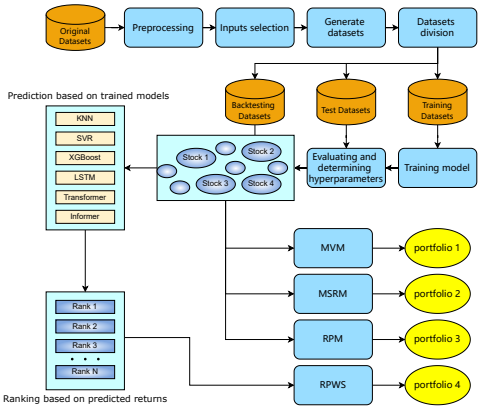


Fig. 2. Experimental process of the study.

1) Model training and performance evaluation: In this experimental approach, various models, including XGBoost, KNN, SVR, LSTM, Transformer, and Informer, are employed for stock return prediction. To evaluate the accuracy and performance of these models, three metrics are used: Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The training in this stage is conducted on a dataset separate from the backtesting dataset. The primary goal at this stage is to determine appropriate model parameters and assess the models' performance. The hyperparameters for each model are detailed in Table I.

2) Portfolio backtesting: In this phase, portfolio backtesting is conducted to compare the performance of different portfolio models. The models include the Minimum Variance Portfolio Model (MVP), Maximum Sharpe Ratio Portfolio Model (MSR), Risk Parity Model (RP), and the Return Prediction Weighted Scoring Model (RPWS). For the A-share and U.S. stock markets, portfolio allocations are adjusted every 20 trading days. In contrast, for the cryptocurrency market—which operates year-round without interruptions—allocations are adjusted every 7 days. When using the Return Prediction Weighted Scoring Model, a rolling training approach is applied, utilizing the 500 most recent historical samples prior to each current date. This ensures the model adapts to the trends and dynamics of changing market conditions.

The overall experimental workflow is illustrated in Fig. 2.

IV. Data and Performance Evaluation

A. Datasets and Pre-processing

The selection of the dataset is crucial for model evaluation, and this experiment will use index component stocks as the data source. Indices are reliable financial indicators of high-quality stocks and reflect mainstream investment trends in the market. Index component stocks are of high quality, with minimal occurrences of suspension or delisting, ensuring the

TABLE I. PARAMETERS FOR PREDICTION MODELS

Model	KNN	XGBoost	SVR	LSTM	Transformer	Informer
Parameter	n_neighbors	n_estimators	C	epochs	epochs	epochs
(Value)	(3)	(150)	(10)	(50)	(50)	(50)
		max_depth	gamma	batch_size	batch_size	batch_size
		(10)	(0.1)	(64)	(64)	(64)
		learning_rate		optimizer	optimizer	optimizer
		(0.01)		(adam)	(adam)	(adam)
		gamma		dropout_rate	dropout_rate	dropout_rate
		(2)		(0.1)	(0.1)	(0.1)
				units	embed_dim	embed_dim
				(64)	(8)	(8)
					num_heads	num_heads
					(2)	(2)
					ff_dim	ff_dim
					(64)	(64)

TABLE II. RANDOMLY SELECTED TICKERS AND MAJOR CRYPTOCURRENCIES

U.S. Stock Market		A-share Market		Crypto Market
AAPL	MCHP	600028	601398	ADA
ADBE	MDLZ	600031	601601	AVAX
ADP	MNST	600036	601628	BNB
ALGN	MSFT	600050	601633	BTC
AMD	NFLX	600089	601668	DOGE
AMZN	NTES	600111	601669	ETH
CHTR	NXPI	600150	601857	LINK
CMCSA	PAYX	600276	601899	SHIB
CSCO	PEP	600309	601919	SOL
CSX	QCOM	600406	601988	XRP
CTSH	ROST	600436		
DLTR	SBUX	600438		
EA	SWKS	600519		
EBAY	TTWO	600809		
GOOGL	TXN	600893		
IDXX	VRSK	601088		
INTU	WDAY	601166		
ISRG	WDC	601288		
LULU	XEL	601318		
MAR	ZBRA	601390		

continuity and completeness of the experimental data samples. This quality is advantageous for model construction, as portfolios based on index component stocks tend to be more robust and offer better potential returns. Due to the immaturity and high volatility of the cryptocurrency market, only a few selected major currencies are included in the dataset.

To validate the robustness of the model, this study will conduct backtesting across multiple markets, including the A-share market, the U.S. stock market, and the cryptocurrency market. The updated dataset includes the following components:

- CSI 300 Index Component Stocks: Daily data from January 2017 to January 2024, with a random selection of 30 stocks.
- NASDAQ Index Component Stocks: Daily data from January 2014 to January 2024, with a random selection of 40 stocks.
- Cryptocurrency Data: Daily data from January 2023 to October 2024, covering 10 major cryptocurrencies.

The dataset includes various factors such as opening price, closing price, highest price, lowest price, trading volume, and three technical indicators: the Relative Strength Index (RSI) with a 12-period, the Moving Average Convergence Divergence (MACD) with a 12-period, and the On-Balance Volume (OBV). This comprehensive datasets aims to enhance the reliability and stability of the experimental results, providing a solid foundation for validating the model's performance across different financial markets. The specific stocks selected are detailed in Table II.

Data normalization involves transforming data with different scales and units into a uniform format, allowing for comparability across different dimensions and facilitating feature extraction during the model training process. In this experiment, we use the Max-Min normalization method to process each feature factor in a dimensionless manner. The specific formula for normalization is shown below:

$$y_i = \frac{x_i - \min\limits_{1 \le j \le n} \{x_j\}}{\max\limits_{1 \le j \le n} \{x_j\} - \min\limits_{1 \le j \le n} \{x_j\}}$$

where, x_i represents the original data, y_i denotes the normalized dimensionless data, and j is the j-th element in the sequence.

B. Performance Evaluation

This study employs six metrics, namely MSE, RMSE and MAE, to provide a comprehensive analysis of various models' performance in predicting stock returns. These metrics are

TABLE III. PERFORMANCE OF DIFFERENT MODELS ACROSS THREE MARKETS

		U.S. Stock Market			A-share Market			Crypto Market		
Model	Metric	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
KNN	Mean	0.0338	0.1725	0.1421	0.0360	0.1646	0.1377	0.0916	0.2769	0.2461
	Std	0.0302	0.0651	0.0595	0.0416	0.0962	0.0881	0.0714	0.1291	0.1309
	Var	0.0009	0.0042	0.0035	0.0017	0.0093	0.0078	0.0051	0.0167	0.0171
SVR	Mean	0.0702	0.2355	0.2014	0.0613	0.2201	0.1927	0.1041	0.2705	0.2478
	Std	0.0687	0.1239	0.1268	0.0572	0.1158	0.1063	0.1202	0.1855	0.1900
	Var	0.0047	0.0154	0.0161	0.0033	0.0134	0.0113	0.0144	0.0344	0.0361
XGBoost	Mean	0.0377	0.1767	0.1481	0.0416	0.1712	0.1461	0.0989	0.2707	0.2443
	Std	0.0374	0.0818	0.0788	0.0530	0.1131	0.1052	0.1058	0.1687	0.1738
	Var	0.0014	0.0067	0.0062	0.0028	0.0128	0.0111	0.0112	0.0285	0.0302
LSTM	Mean	0.0112	0.1004	0.0806	0.0167	0.1131	0.0907	0.0273	0.1485	0.1210
	Std	0.0086	0.0337	0.0292	0.0195	0.0641	0.0545	0.0243	0.0767	0.0672
	Var	0.0001	0.0011	0.0009	0.0004	0.0041	0.0030	0.0006	0.0059	0.0045
Transformer	Mean	0.0412	0.1879	0.1588	0.0864	0.2402	0.2126	0.1339	0.3295	0.3105
	Std	0.0383	0.0782	0.0699	0.1096	0.1730	0.1656	0.1037	0.1677	0.1686
	Var	0.0015	0.0061	0.0049	0.0120	0.0299	0.0274	0.0107	0.0281	0.0284
Informer	Mean	0.0389	0.1818	0.1529	0.1509	0.3003	0.2715	0.0951	0.2786	0.2434
	Std	0.0359	0.0779	0.0713	0.2395	0.2515	0.2524	0.0773	0.1396	0.1344
	Var	0.0013	0.0061	0.0051	0.0573	0.0633	0.0637	0.0060	0.0195	0.0181

commonly utilized as evaluative indicators due to their comprehensive portrayal of predictive capabilities (Freitas [24]; Gandhmal and Kumar [56]; Ma et al. [41]).

1) Mean Squared Error (MSE): MSE calculates the average squared differences between predicted and actual values, emphasizing larger errors due to the squaring operation. The formula for MSE is:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

2) Root Mean Squared Error (RMSE): RMSE represents the square root of the average squared differences between predicted and actual values, providing a measure of the standard deviation of residuals. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

3) Mean Absolute Error (MAE): MAE measures the average magnitude of errors between predicted and actual values, providing a straightforward assessment of model performance. The formula for MAE is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

In this study, we evaluate the performance of six models—KNN, SVR, XGBoost, LSTM, Transformer, and Informer—in predicting stock returns. Table III presents the results for each model based on the performance metrics MSE, RMSE, and MAE.

In the U.S. stock market, the LSTM model clearly outperforms the others, exhibiting the lowest error and highest stability. XGBoost and KNN follow closely behind, while Transformer and Informer show similar performances, though they lag behind XGBoost and KNN. The SVR model performs the worst overall. In the Chinese A-share market, the performance patterns of the models are similar to those observed in the U.S. market. The LSTM model again stands out, with the smallest MSE, RMSE, and MAE, and good stability. SVR exhibits relatively large errors and instability, while Transformer and Informer perform the worst, with significantly higher errors. In the cryptocurrency market, the errors of all models increase due to the immature nature and high volatility of the market, making prediction a particularly challenging task. However, the overall performance trends of the models remain similar to those in the U.S. and A-share markets, with LSTM still showing the best performance and SVR and Transformer showing the worst.

It is important to note that the performance metrics presented here are based solely on the results from the test dataset. The data used for the rolling backtest, which is entirely separate from this test dataset, may not reflect the actual predictive performance of each model throughout the backtesting process.

V. EMPIRICAL RESULTS

After comparing the performance of different models in stock return prediction, a backtesting simulation is conducted to study the investment performance of various portfolios, with an initial capital of 100,000. In the A-share and U.S. stock markets, portfolio weights are adjusted every 20 days, while in the cryptocurrency market, portfolio weights are adjusted every 7 days. For simplicity, dividends and transaction fees are disregarded, and leverage and short-selling are not considered during the portfolio rebalancing process. In this

TABLE IV. DIFFERENT PORTFOLIO OPTIMIZATION MODELS FOR COMPARISON

Abbreviation	Description
QQQ (ETF)	Invesco QQQ Trust (Nasdaq: QQQ).
CSI 300 (ETF)	CSI 300 Index Securities Investment Fund.
BTC	Bitcoin: A mainstream cryptocurrency.
1/N	Equal Weight Strategy.
RPM	Risk Parity Model.
MVM	Minimum Variance Portfolio Model.
MSRM	Maximum Sharpe Ratio Portfolio Model.
SVR-RPWS	SVR for prediction and ranking, RPWS for portfolio construction.
KNN-RPWS	KNN for prediction and ranking, RPWS for portfolio construction.
XGBoost-RPWS	XGBoost for prediction and ranking, RPWS for portfolio construction.
LSTM-RPWS	LSTM for prediction and ranking, RPWS for portfolio construction.
Informer-RPWS	Informer for prediction and ranking, RPWS for portfolio construction.
Transformer-RPWS	Transformer for prediction and ranking, RPWS for portfolio construction.
SVR-MVM(N=10)	SVR for prediction and screening, MVM for portfolio construction.

Notes:
1. To more clearly highlight the performance of RPWS, the SVR-MVM model is introduced as a comparative experiment. This model uses the best-performing SVR algorithm to select the top 10 stocks based on predicted returns. We select 10 stocks for the following reasons: Evans and Archer (1968) have demonstrated that the traditional rule of thumb suggesting 8 to 10 stocks, established by pioneering studies, is indeed sufficient to achieve optimal diversification effects [58]. Alexeev and Dungey (2015) assert that an equally weighted portfolio of 10 stocks would suffice for an average investor aiming to diversify away 90% of the risk [59]. Recognizing that individual investors may struggle to track too many assets, we have decided that the number of stocks we select will be 10. Optimizes portfolio weights using the Minimum Variance Model (MVM).

2. Cryptocurrencies market, due to limited samples, are no longer included in the comparison.

TABLE V. METRICS FOR MEASURING PORTFOLIO PERFORMANCE

Metrics	Formula	
Annualized Return	$R = \left(\prod_{i=1}^{N}(1+r_i)\right)^{\frac{252}{N}} - 1$	
Annualized Volatility	$V = \sqrt{\frac{252}{N}\sum_{i=1}^{N}(r_i-\mu)^2}$	$\mu = \frac{1}{N}\sum_{i=1}^{N} r_i$
Sharpe Ratio	Sharpe Ratio $= \frac{R - r_f}{V}$	
Sortino Ratio	Sortino Ratio $= \frac{R - r_f}{D}$	$D = \sqrt{\frac{252}{N}\sum_{i=1}^{N}\min(r_i - r_f, 0)^2}$
Maximum Drawdown	$\text{MDD} = \max_{1 \le i \le N}\left(\frac{P_{\text{peak},i} - P_i}{P_{\text{peak},i}}\right)$	
VaR(95%)	VaR(95%) $= \mu - 1.645 \times \sigma$	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i-\mu)^2}$

Notes:

1. r_i denotes the return in the i-th day, N is the total number of days, and 252 is the number of trading days in a year.

2. r_f is the risk-free rate, which is set to 0 in this study.

3. P_i represents the return on the i-th day, while $P_{\text{peak},i}$ denotes the historical maximum return up to the i-th day.

4. 1.645 is the quantile of the standard normal distribution at the 95% confidence level.

study, six key metrics are used to evaluate the performance of different portfolios: annualized return, Sharpe ratio, maximum drawdown, Sortino ratio, annualized volatility, and Value at Risk (VaR) at the 95% confidence level. Table IV shows different portfolio optimization models for comparison. The details of each indicator can be found in Table V.

From 2018 to 2024, the U.S. stock market experienced dramatic fluctuations, including a bull market recovery, a bear market adjustment, and dramatic fluctuations caused by the epidemic. The market was affected by multiple factors such as interest rate hikes, inflation, and stimulus policies, and generally showed a fluctuating upward trend. In the backtest, the SVR-RPWS model performs best overall, followed by the Transformer-RPWS model. The SVR-RPWS model achieves the highest annualized return (28.11%), the highest Sharpe ratio (1.1436) and the Sortino ratio (1.6704), and is also excellent in drawdown control, with the maximum drawdown (-22.74%) second only to the LSTM-RPWS model (-21.86%). The SVR-RPWS and Transformer-RPWS models outperform QQQ (QQQ is an exchange-traded fund that tracks the Nasdaq 100 Index, which includes the 100 largest non-financial companies listed on the Nasdaq stock market) in all indicators and perform significantly better than other models overall. In addition, all models using the RPWS algorithm also outperform QQQ overall, especially in terms of seeking returns. The traditional portfolio model performs well in risk control, with annualized volatility and value at risk (VaR) slightly better than the model using the RPWS algorithm at the 95% confidence level, but performs poorly in terms of annualized returns and Sharpe ratio, and failed to surpass QQQ. The SVR-MVM model performs quite well overall, with a Sharpe ratio second only to SVR-RPWS and Transformer-RPWS, and it also demonstrates good risk control. Compared to the MVM model without any stock screening, it shows significant improvement, reflecting the effectiveness of stock screening. However, in terms of pursuing returns, it still lags considerably behind models using the RPWS algorithm. The detailed backtest performance and cumulative returns are available in Table VI and Fig. 3.

From 2021 to 2024, affected by factors such as the slowdown in economic growth, the Sino-US trade war and the external environment, the A-share market was dominated by a bear market, continued to adjust, and continued to be sluggish. Although there was a short-term rebound, the overall trend was weak. The three-year annualized return of the CSI 300 ETF (the CSI 300 ETF tracks the CSI 300 Index, which represents the 300 largest and most liquid stocks listed on the Shanghai and Shenzhen Stock Exchanges) was -13.25%. In the backtest, the SVR-RPWS and Transformer-RPWS models perform similarly to the US stock market, both perform well and significantly outperform all other models. Among them, the SVR-RPWS model has the highest annualized return (25.30%), the highest Sharpe ratio (1.2401), and the highest Sortino ratio (2.0532). At the same time, while maintaining high returns, the Transformer-RPWS model performs best in risk control, with a maximum drawdown of only -12.58%, an annualized volatility of 15.29% and a VaR (95%) of -1.29%. It is worth noting that although the traditional models perform well in terms of risk control, the overall returns are disappointing. The MVM model is the only traditional model that achieves positive returns in the bear market, but the return is only 1.73%. In contrast, all models using the RPWS algorithm achieve positive returns and are better than the MVM model. The SVR-MVM model performs similarly in the A-share market as in the U.S. stock market. While maintaining good risk control, the annualized return of the MVM model, after SVR stock selection, increases from 1.73% to 3.75%, representing a 105.84% improvement in the original annualized return metric. Although this is already a significant improvement, it still significantly underperforms portfolios using the RPWS model in terms of annualized return, Sharpe ratio, and Sortino ratio. The detailed backtest performance and cumulative returns are available in Table VII and Fig. 4.

From January 2023 to October 2024, the cryptocurrency market experienced significant growth, with major cryptocurrencies such as Bitcoin and Ethereum seeing sharp price increases, with Bitcoin breaking through its all-time high. The gradual acceptance of cryptocurrencies by institutional investors and the traditional financial industry has played a key role in driving mainstream market applications. Investor sentiment remains optimistic, and a large amount of capital has flowed into the market, jointly driving a strong bull market trend. Bitcoin dominates with a market share of 67.3%, and in this bull market, it has achieved an annualized return of 42.87% in just over a year. However, compared with traditional stock markets, the cryptocurrency market is relatively immature, extremely volatile, and greatly affected by market sentiment, making modeling and prediction particularly difficult. As a result, most models based on the RPWS algorithm perform poorly. Surprisingly, the Transformer-RPWS model, which performs well in both the US and A-share markets, performs poorly in the cryptocurrency market, with an annualized return of only 28.85%. In contrast, the SVR-RPWS model continues to show excellent performance, achieving an ultra-high annualized return of 75.01%, and its Sharpe ratio (1.2522) and Sortino ratio (2.1192) are the highest among all models. Whether in terms of returns or risk control, the performance of traditional models is only above average at best, and fails to outperform Bitcoin. The detailed backtest performance and cumulative returns are available in Table VIII and Fig. 5.

While the LSTM model demonstrates excellent predictive performance on the test set, its performance in backtesting is relatively mediocre. In contrast, models such as SVR and Transformer, which perform poorly on the test set, show exceptional results when combined with the RPWS algorithm in the backtest. I hypothesize two potential reasons for this discrepancy. First, the time-series nature of financial market data makes overfitting particularly pronounced. Models may learn incidental patterns or short-term fluctuations in historical data that do not persist in the future, thereby limiting their generalization ability. As a result, models that perform well on the test set may fail to produce satisfactory results in subsequent backtests. Second, models that underperformed on the test set may exhibit stronger robustness—such as the SVR model, which demonstrated good stability across multiple markets. These models may be better equipped to adapt to varying market conditions or data volatility, leading to superior performance in the backtest phase.

Overall, in mature markets such as the U.S. and A-share stock markets, models incorporating the RPWS algorithm have demonstrated strong investment performance, often outper-

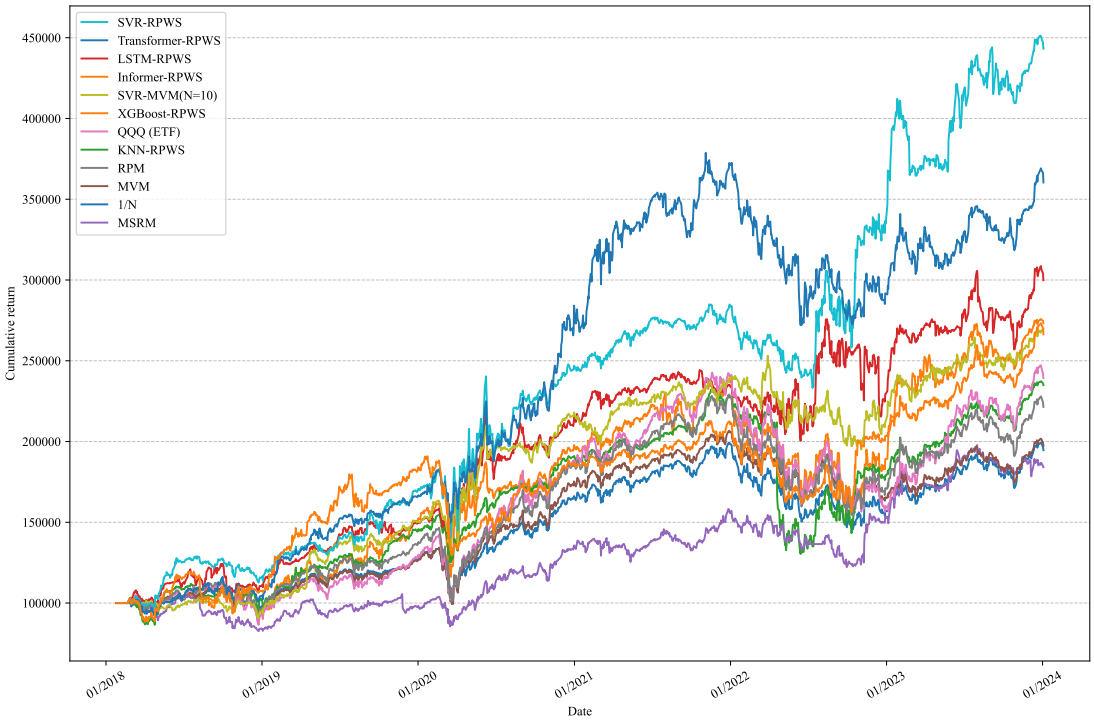


Fig. 3. Backtest cumulative return for the U.S. stock market.

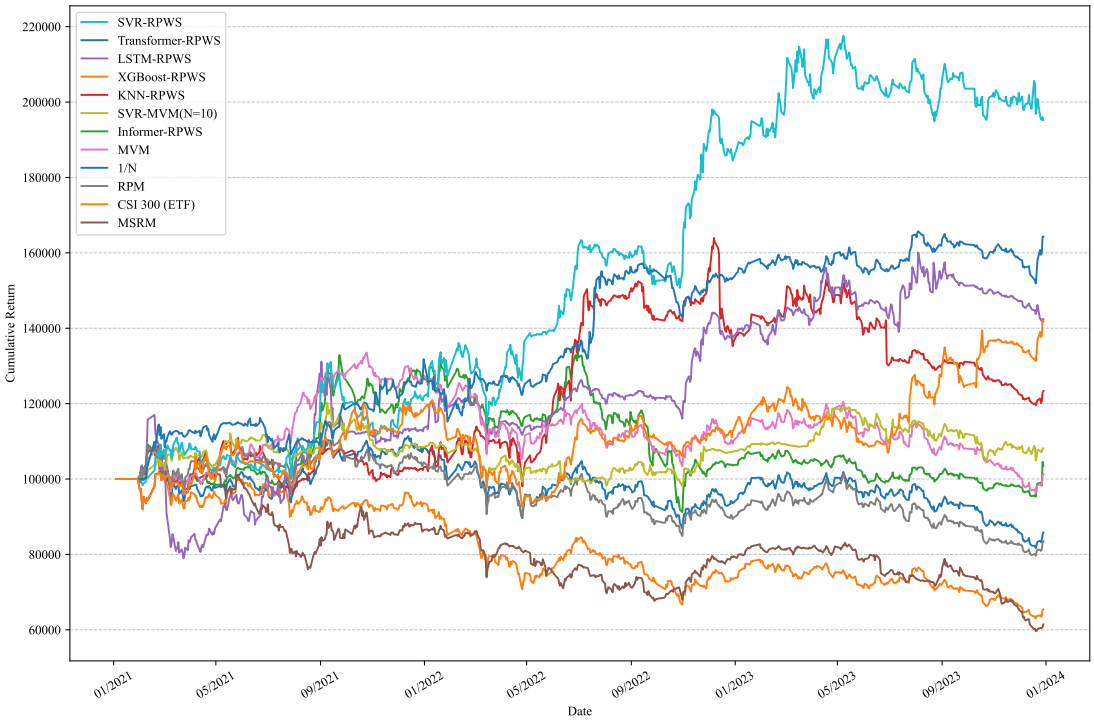


Fig. 4. Backtest cumulative return for A-share market.

TABLE VI. BACKTEST PERFORMANCE OF THE U.S. STOCK MARKET

U.S. Stock Market	Annualized Return	Sharpe Ratio	Maximum Drawdown	Sortino Ratio	Annualized Volatility	VaR (95%)
SVR-RPWS	**28.11%**	**1.1436**	-22.74%	**1.6704**	24.58%	-1.89%
Transformer-RPWS	23.97%	1.1069	-28.18%	1.5858	21.66%	-2.00%
SVR-MVM(N=10)	18.76%	0.9319	-22.95%	1.3659	20.13%	-1.72%
LSTM-RPWS	21.60%	0.8684	**-21.86%**	1.1458	24.87%	-2.13%
Informer-RPWS	20.95%	0.8629	-33.60%	1.1562	24.28%	-2.26%
XGBoost-RPWS	18.99%	0.8572	-31.99%	1.0688	22.16%	-1.97%
KNN-RPWS	16.88%	0.7691	-43.88%	0.8381	21.95%	**-1.62%**
QQQ (ETF)	17.82%	0.7176	-35.57%	0.9339	24.83%	-2.56%
1/N	12.86%	0.7143	-26.94%	0.9386	**18.01%**	-1.67%
RPM	16.04%	0.7013	-32.69%	0.9049	22.87%	-2.19%
MVM	13.33%	0.6916	-25.95%	0.9071	19.27%	-1.75%
MSRM	5.71%	0.3059	-27.09%	0.3795	18.65%	-1.89%

TABLE VII. BACKTEST PERFORMANCE OF THE A-SHARE MARKET

A-share Market	Annualized Return	Sharpe Ratio	Maximum Drawdown	Sortino Ratio	Annualized Volatility	VaR (95%)
SVR-RPWS	**25.30%**	**1.2401**	-15.66%	**2.0532**	20.40%	-1.87%
Transformer-RPWS	18.40%	1.2038	**-12.58%**	1.8676	15.29%	-1.29%
XGBoost-RPWS	14.05%	0.7150	-23.44%	1.1837	19.65%	-1.73%
LSTM-RPWS	15.19%	0.6591	-31.86%	0.9548	22.59%	-1.91%
KNN-RPWS	9.45%	0.4556	-27.06%	0.5923	20.74%	-1.87%
SVR-MVM(N=10)	3.56%	0.2727	-18.49%	0.3941	**13.05%**	**-1.26%**
Informer-RPWS	2.76%	0.1551	-31.82%	0.2244	17.78%	-1.57%
MVM	1.73%	0.1076	-27.92%	0.1625	16.09%	-1.63%
1/N	-4.01%	-0.2500	-26.70%	-0.3778	16.05%	-1.65%
RPM	-4.96%	-0.3077	-30.80%	-0.4523	16.13%	-1.65%
CSI 300 ETF	-13.25%	-0.7743	-41.97%	-1.1289	17.11%	-1.70%
MSRM	-15.26%	-0.8414	-45.16%	-1.1381	18.13%	-1.96%

TABLE VIII. BACKTEST PERFORMANCE OF THE CRYPTOCURRENCY MARKET

Cryptocurrency Market	Annualized Return	Sharpe Ratio	Maximum Drawdown	Sortino Ratio	Annualized Volatility	VaR (95%)
SVR-RPWS	**75.01%**	**1.2522**	-39.61%	**2.1192**	59.90%	-4.41%
BTC	42.87%	1.2131	-26.24%	1.7970	35.34%	-3.34%
MSRM	50.13%	1.1605	-45.00%	1.4531	43.20%	-4.06%
MVM	44.98%	1.1389	-40.72%	1.6393	39.49%	-3.54%
1/N	52.46%	1.1196	-45.36%	1.5988	46.85%	-4.39%
KNN-RPWS	53.59%	1.1018	-45.33%	1.6127	48.64%	-4.37%
Informer-RPWS	34.04%	1.0674	**-24.01%**	1.6081	**31.89%**	**-2.72%**
RPM	41.62%	0.9026	-43.60%	1.2611	46.11%	-4.32%
Transformer-RPWS	28.85%	0.7420	-43.38%	0.9349	38.88%	-3.50%
LSTM-RPWS	33.59%	0.6434	-34.55%	1.0374	52.21%	-4.54%
XGBoost-RPWS	21.48%	0.4881	-47.38%	0.6637	44.00%	-4.02%

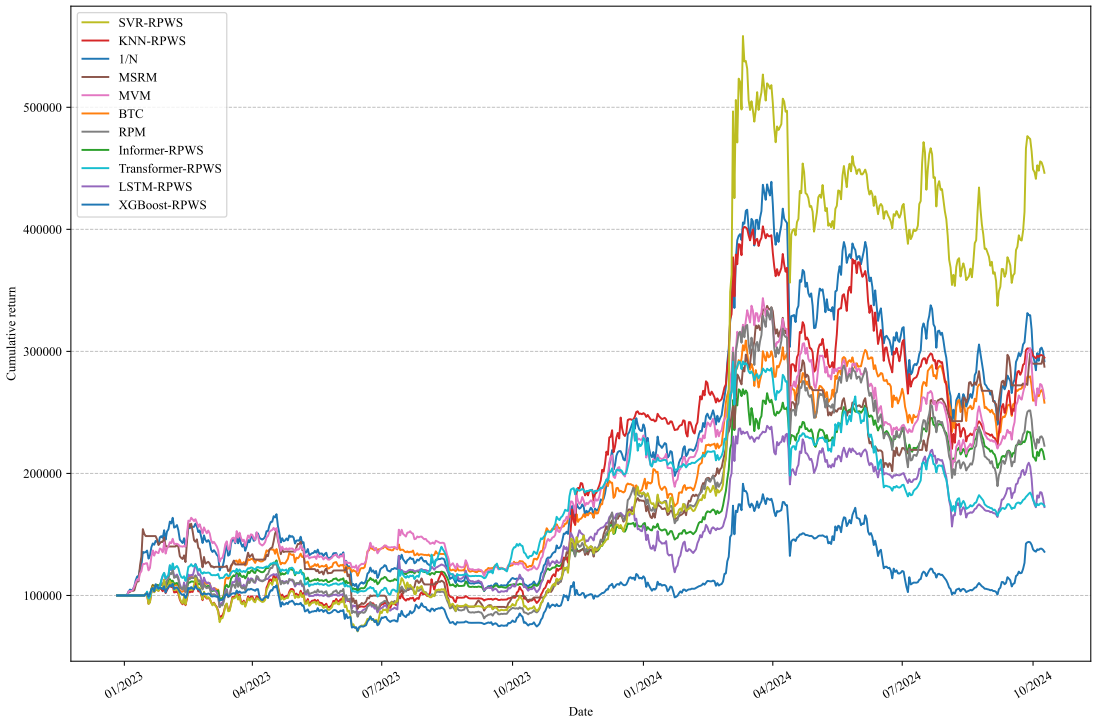


Fig. 5. Backtest cumulative return for cryptocurrency market.

forming traditional portfolio models and major ETFs. This is particularly evident in terms of generating higher returns, thereby validating the fundamental principle of high risk leading to high reward. The experiment highlights the substantial advantages of employing machine learning and neural network models in constructing better-performing investment portfolios. It further reinforces the effectiveness of machine learning and neural network models in portfolio optimization.

VI. Conclusion and Future Work

Machine learning algorithms and neural network models have become essential tools in various fields, including financial market investment, to make more advantageous and informed decisions. This study applies machine learning algorithms and neural network models to portfolio construction, introducing a prediction-based weighted scoring method to build optimized investment portfolios. The research enriches portfolio theory by offering novel insights. The key findings are as follows:

Firstly, this study provides a detailed introduction to six models: K-Nearest Neighbors (KNN), Support Vector Regression (SVR), XGBoost, Long Short-Term Memory (LSTM), Transformer, and Informer. These models are evaluated based on their stock return prediction capabilities on a test dataset. Secondly, this paper presents a portfolio construction method based on return prediction and weighted scoring (RPWS). The effectiveness of this method is validated through the combination with the aforementioned predictive models. Among the models tested, SVR-RPWS and Transformer-RPWS perform well in terms of both maximizing returns and controlling risk and drawdowns. Overall, their performance outperforms mainstream funds and traditional portfolio models. Finally, in practical backtesting, the training time of the models also requires attention. For instance, the Transformer model takes nearly a week to complete a backtest on the US stock market, whereas the SVR model only requires about one hour. In the future, further validation of the RPWS technique will require expanding the dataset, making the SVR model's faster convergence speed a notable advantage for further exploration.

Previous studies typically divide machine learning algorithms and portfolio optimization into two separate stages, which means that the modeling capabilities of machine learning are only utilized in the stock screening phase and not fully exploited during weight optimization. The method proposed in this study, which uses RPWS to construct portfolios, provides a new perspective for the interdisciplinary field of machine learning and portfolio optimization. The key aspect of this method lies in mapping the prediction results of machine learning algorithms to weight biases, successfully integrating the predictive capabilities of machine learning into the weight optimization phase, ensuring its full utilization throughout the portfolio construction process.

In the future, we will explore more weighting methods to reasonably and effectively map the prediction results of machine learning algorithms to weight biases. We have considered several directions to further optimize the initial weight biases, including logarithmic weighting and enforcing log uniform distribution. In particular, enforcing log uniform distribution adjusts the initial weight biases of stocks based on their rankings to conform to a log uniform distribution, and its actual performance will be tested in future backtesting experiments. Additionally, this study uses historical data and technical indi-

cators as input features for predicting future returns. However, many external factors, such as government policies, interest rates, and public events, also impact financial markets, so future work may consider incorporating more sentiment factors into the dataset. We believe that by continuously optimizing weight mapping methods and integrating more influencing factors, the stability and practical value of portfolio strategies will be further enhanced.

REFERENCES

[1] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, pp. 77–91, 1952, doi: 10.2307/2975974.

[2] T. Björk, A. Murgoci, and X. Y. Zhou, "Mean-variance portfolio optimization with state-dependent risk aversion," *Math. Finance*, vol. 24, pp. 1–24, 2014, doi: 10.1111/j.1467-9965.2011.00515.x.

[3] K. Liagkouras and K. Metaxiotis, "Multi-period mean–variance fuzzy portfolio optimization model with transaction costs," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 260–269, 2018, doi: 10.1016/j.engappai.2017.10.010.

[4] D. Rosadi, E. P. Setiawan, M. Templ, and P. Filzmoser, "Robust covariance estimators for mean-variance portfolio optimization with transaction lots," *Oper. Res. Perspect.*, vol. 7, p. 100154, 2020, doi: 10.1016/j.orp.2020.100154.

[5] V. N. Katsikis, S. D. Mourtas, P. S. Stanimirović, S. Li, and X. Cao, "Time-varying mean-variance portfolio selection under transaction costs and cardinality constraint problem via beetle antennae search algorithm (BAS)," *Oper. Res. Forum*, vol. 2, 2021, doi: 10.1007/s43069-021-00060-5.

[6] B. Beheshti, "Effective stock selection and portfolio construction within US, international, and emerging markets," *Front. Appl. Math. Stat.*, vol. 4, 2018, doi: 10.3389/fams.2018.00017.

[7] G. S. M. Thakur, R. Bhattacharyya, and S. Sarkar (Mondal), "Stock portfolio selection using Dempster–Shafer evidence theory," *J. King Saud Univ.- Comput. Inf. Sci.*, vol. 30, pp. 223–235, 2018, doi: 10.1016/j.jksuci.2016.07.001.

[8] Y. Zhang, X. Li, and S. Guo, "Portfolio selection problems with Markowitz's mean–variance framework: a review of literature," *Fuzzy Optim. Decis. Making*, pp. 125–158, 2018, doi: 10.1007/s10700-017-9266-z.

[9] P. N. Kolm, R. Tütüncü, and F. J. Fabozzi, "60 years of portfolio optimization: Practical challenges and current trends," *Eur. J. Oper. Res.*, pp. 356–371, 2014, doi: 10.1016/j.ejor.2013.10.060.

[10] Y. Chen and Y. Hao, "Integrating principle component analysis and weighted support vector machine for stock trading signals prediction," *Neurocomputing*, vol. 321, pp. 381–402, 2018, doi: 10.1016/j.neucom.2018.08.077.

[11] F. D. Paiva, "Decision-making for financial trading: A fusion approach of machine learning and portfolio selection," *Expert Syst. Appl.*, vol. 115, pp. 635–655, 2019, doi: 10.1016/j.eswa.2018.08.003.

[12] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Syst. Appl.*, pp. 187–205, 2017, doi: 10.1016/j.eswa.2017.04.030.

[13] H. Liu and M. Loewenstein, "Optimal portfolio selection with transaction costs and finite horizons," *Rev. Financ. Stud.*, vol. 15, pp. 805–835, 2002, doi: 10.1093/rfs/15.3.805.

[14] O. D. Anderson, G. E. P. Box, and G. M. Jenkins, "Time series analysis: Forecasting and control," *The Statistician*, p. 265, 1978, doi: 10.2307/2988198.

[15] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, pp. 987–1007, 1982, doi: 10.2307/1912773.

[16] N. Shephard, "Statistical aspects of ARCH and stochastic volatility," in *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67, 1996.

[17] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econom.*, pp. 307–327, 1986, doi: 10.1016/0304-4076(86)90063-1.

[18] R. C. Garcia, J. Contreras, M. van Akkeren, and J. B. C. Garcia, "A GARCH forecasting model to predict day-ahead electricity prices," *IEEE Trans. Power Syst.*, vol. 20, pp. 867–874, 2005, doi: 10.1109/TPWRS.2005.846044.

[19] M. A. Villegas, D. J. Pedregal, and J. R. Trapero, "A support vector machine for model selection in demand forecasting applications," *Comput. Ind. Eng.*, vol. 121, pp. 1–7, 2018, doi: 10.1016/j.cie.2018.04.042.

[20] T. Kavzoglu and A. Teke, "Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," *Arab. J. Sci. Eng.*, vol. 47, pp. 7367–7385, 2022, doi: 10.1007/s13369-022-06560-8.

[21] Y. L. Ma, R. Z. Han, and W. Z. Wang, "Portfolio optimization with return prediction using deep learning and machine learning," *Expert Syst. Appl.*, p. 113973, 2021, doi: 10.1016/j.eswa.2020.113973.

[22] A.-L. Boulesteix and M. Schmid, "Machine learning versus statistical modeling," *Biom. J.*, vol. 56, pp. 588–593, 2014, doi: 10.1002/bimj.201300226.

[23] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," *IEEE Access*, vol. 6, pp. 60079–60087, 2018, doi: 10.1109/ACCESS.2018.2874979.

[24] F. D. Freitas, A. F. de Souza, and A. R. de Almeida, "Prediction-based portfolio optimization model using neural networks," *Neurocomputing*, vol. 72, pp. 2155–2170, 2009, doi: 10.1016/j.neucom.2008.08.019.

[25] Z. Jiang, R. Ji, and K.-C. Chang, "A machine learning integrated portfolio rebalance framework with risk-aversion adjustment," *J. Risk Financ. Manag.*, vol. 13, p. 155, 2020, doi: 10.3390/jrfm13070155.

[26] T. Kaczmarek and K. Perez, "Building portfolios based on machine learning predictions," *Econ. Res.-Ekon. Istraž.*, pp. 19–37, 2022, doi: 10.1080/1331677X.2021.1875865.

[27] D. K. Padhi, N. Padhy, A. K. Bhoi, J. Shafi, and S. H. Yesuf, "An intelligent fusion model with portfolio selection and machine learning for stock market prediction," *Comput. Intell. Neurosci.*, pp. 1–18, 2022, doi: 10.1155/2022/7588303.

[28] S. Deng and X. Min, "Applied optimization in global efficient portfolio construction using earning forecasts," *J. Invest.*, vol. 22, pp. 104–114, 2013, doi: 10.3905/joi.2013.22.4.104.

[29] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, pp. 7046–7056, 2015, doi: 10.1016/j.eswa.2015.05.013.

[30] A. Chaweewanchon and R. Chaysiri, "Markowitz mean-variance portfolio optimization with predictive stock selection using machine learning," *Int. J. Financ. Stud.*, vol. 10, no. 3, Art. no. 64, 2022, doi: 10.3390/ijfs10030064.

[31] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," *Appl. Soft Comput.*, pp. 525–538, 2018, doi: 10.1016/j.asoc.2018.04.024.

[32] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches," *Decis. Support Syst.*, pp. 258–269, 2010, doi: 10.1016/j.dss.2010.08.028.

[33] N. N. Vo, X. He, S. Liu, and G. Xu, "Deep learning for decision making and the optimization of socially responsible investments and portfolio," *Decis. Support Syst.*, vol. 124, p. 113097, 2019, doi: 10.1016/j.dss.2019.113097.

[34] W. Wang, W. Li, N. Zhang, and K. C. Liu, "Portfolio formation with preselection using deep learning from long-term financial data," *Expert Syst. Appl.*, vol. 143, p. 113042, 2020, doi: 10.1016/j.eswa.2019.113042.

[35] Y. L. Ma, R. Z. Han, and W. Z. Wang, "Portfolio optimization with return prediction using deep learning and machine learning," *Expert Syst. Appl.*, vol. 165, p. 113973, 2021, doi: 10.1016/j.eswa.2020.113973.

[36] O. Ustun and R. Kasimbeyli, "Combined forecasts in portfolio optimization: A generalized approach," *Comput. Oper. Res.*, vol. 39, pp. 805–819, 2012, doi: 10.1016/j.cor.2010.09.008.

[37] J. R. Yu, W. J. P. Chiou, W. Y. Lee, and S. J. Lin, "Portfolio models with return forecasting and transaction costs," *Int. Rev. Econ. Finance*, vol. 66, pp. 118–130, 2020, doi: 10.1016/j.iref.2019.11.002.

[38] J. Behera, A. K. Pasayat, H. Behera, and P. Kumar, "Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets," *Eng. Appl. Artif. Intell.*, vol. 120, pp. 105843, 2023, doi: 10.1016/j.engappai.2023.105843.

[39] F. D. Freitas, A. F. de Souza, and A. R. de Almeida, "Autoregressive neural network predictors in the Brazilian stock markets," in *IEEE Latin American Robotics Symposium*, Brazil, 2005.

[40] C. Y. Hao, J. Q. Wang, and W. Xu, "Prediction-based portfolio selection model using support vector machines," in *6th International Conference on Business Intelligence and Financial Engineering*, pp. 567–571, 2013, doi: 10.1109/BIFE.2013.118.

[41] Y. L. Ma, R. Z. Han, and W. Z. Wang, "Prediction-based portfolio optimization models using deep neural networks," *IEEE Access*, vol. 8, pp. 115393–115405, 2020, doi: 10.1109/ACCESS.2020.3003819.

[42] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[43] G. Lin, A. Lin, and D. Gu, "Using support vector regression and k-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient," *Inf. Sci.*, pp. 517–531, 2022, doi: 10.1016/j.ins.2022.06.090.

[44] Y. Qiu, D. Garg, S.-M. Kim, I. Mudawar, and C. R. Kharangate, "Machine learning algorithms to predict flow boiling pressure drop in mini/micro-channels based on universal consolidated data," *Int. J. Heat Mass Transf.*, p. 121607, 2021, doi: 10.1016/j.ijheatmasstransfer.2021.121607.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, pp. 273–297, 1995, doi: 10.1023/A:1022627411411.

[46] S. Emir, "Predicting the Istanbul stock exchange index return using technical indicators," *Int. J. Finance Bank. Stud.*, vol. 2, pp. 111–117, 2016, doi: 10.20525/ijfbs.v2i3.158.

[47] J. M. Matías and J. C. Reboredo, "Forecasting performance of nonlinear models for intraday stock returns," *J. Forecast.*, vol. 31, pp. 172–188, 2012, doi: 10.1002/for.1218.

[48] R. I. Rasel, N. Sultana, and P. Meesad, "An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators," *Int. J. Comput. Intell. Stud.*, vol. 4, p. 134, 2015, doi: 10.1504/IJCISTUDIES.2015.071180.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[50] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, pp. 654–669, 2018, doi: 10.1016/j.ejor.2017.11.054.

[51] Y. Gao, R. Wang, and E. Zhou, "Stock prediction based on optimized LSTM and GRU models," *Sci. Program.*, pp. 1–8, 2021, doi: 10.1155/2021/4055281.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, doi: 10.5555/3295222.3295349.

[53] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[54] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, pp. 11106–11115, 2021, doi: 10.1609/aaai.v35i12.17325.

[55] Y. Lu, H. Zhang, and Q. Guo, "Stock and market index prediction using informer network," *arXiv preprint arXiv:2305.14382*, 2023, doi: 10.48550/arXiv.2305.14382.

[56] D. P. Gandhmal and K. Kumar, "Systematic analysis and review of stock market prediction techniques," *Comput. Sci. Rev.*, vol. 34, p. 100190, 2019, doi: 10.1016/j.cosrev.2019.08.001.

[57] W. Chen, M. Jiang, W.-G. Zhang, and Z. Chen, "A novel graph convolutional feature based convolutional neural network for stock trend prediction," *Inf. Sci.*, vol. 556, pp. 67–94, 2021, doi: 10.1016/j.ins.2020.12.068.

[58] Evans, J.; Archer, S. Diversification and the reduction of dispersion: An empirical analysis. J. Financ. 1968, 23, 761–767.

[59] Alexeev, V.; Dungey, M. Equity portfolio diversification with high-frequency data. Quant. Financ. 2015, 15, 1205–1215.

CeC-SMOTE: A Clustering and Centroid-Based Adaptive Oversampling Method for Imbalanced Data

Xiaoling Gao[1], Marshima Mohd Rosli[2]*, Muhammad Izzad Ramli[3], Nursuriati Jamil[4]
Xinhua College of Ningxia University, Yinchuan, Ningxia, 750021, China[1,2]
College of Computing, Informatics and Mathematics
Universiti Teknologi MARA, Shah Alam 40450, Selangor, Malaysia[1,2,3,4]

***Abstract*—Class imbalance is a common challenge in real-world datasets, leading standard classifiers to perform poorly on underrepresented classes. Traditional oversampling techniques, such as SMOTE and its variants, often generate synthetic samples without fully considering the local data structure, resulting in increased noise and class overlap.This study introduces CeC-SMOTE, an adaptive oversampling method that integrates clustering and centroid-based strategies to enhance the quality of synthetic minority samples. By first partitioning minority instances using K-means clustering, CeC-SMOTE identifies safe and boundary regions, selectively generating new samples where they are most needed while filtering out noise. This targeted approach preserves the underlying distribution of the minority class and minimizes the risk of overfitting. Extensive experiments on artificial and benchmark UCI datasets demonstrate that CeC-SMOTE consistently delivers competitive or superior results compared to established oversampling techniques, particularly in cases with complex or ambiguous class boundaries. Sensitivity analysis confirms that the method is robust to parameter settings, enabling strong performance with minimal tuning.**

***Keywords*—Imbalanced data classification; synthetic oversampling; k-means clustering; centroid-based neighbor**

I. Introduction

Imbalanced data, where some classes are significantly underrepresented compared to others, is a widespread issue in real-world applications. This imbalance often causes models to favor majority classes while neglecting minority categories [1][2]. For example, in healthcare, class imbalance can hinder the early detection of rare diseases, leading to missed diagnoses and delayed treatment. Similarly, in financial systems, fraudulent transactions are often overshadowed by legitimate ones, making it difficult for conventional models to identify and prevent fraud [3].

Traditional classification algorithms, which typically assume a balanced class distribution or seek to minimize overall error, tend to underperform on imbalanced datasets. These models frequently exhibit a bias toward the majority class, resulting in high accuracy for common categories but poor performance for the minority class—often the primary focus in critical applications [4]. This bias leads to minority instances being overlooked or misclassified, sometimes even treated as noise, which compromises both the reliability and generalizability of predictive models. To address these limitations, data-level solutions such as oversampling have been developed to enhance the representation of minority classes by generating synthetic samples.

*Corresponding authors.

The introduction of the SMOTE algorithm marked a significant advancement in oversampling by creating synthetic minority samples through interpolation between existing neighbors. This approach has inspired many variants, each aiming to overcome specific limitations. For example, Borderline-SMOTE [5] generates synthetic samples near decision boundaries, while Safe-Level-SMOTE [6] focuses on safer regions to reduce noise amplification. Other methods, such as SVM combine SMOTE [7] and cluster-based approaches, refine the placement of synthetic samples using support vectors or cluster centroids [8]. Further advancements, like LD-SMOTE [9] and Simplicial SMOTE [10], leverage local density, information entropy, and topological structures to ensure that synthetic samples align more closely with the actual data distribution.

Despite these advances, imbalanced learning continues to evolve as researchers seek new methods that improve sample quality, minimize noise, and enhance classifier robustness, particularly in complex data environments. In this context, we propose CeC-SMOTE—an adaptive oversampling technique that combines clustering and centroid-based analysis. CeC-SMOTE systematically categorizes minority instances by utilizing both global and local cluster characteristics, applies a Nearest Centroid Neighbour strategy for more effective neighbor selection, and adaptively limits synthetic data generation to safe and boundary regions to reduce noise and overfitting. This approach addresses important gaps in existing oversampling methods. The primary contributions of this study are as follows:

- We introduce an adaptive oversampling algorithm that integrates centroid-based clustering with safety-aware sample selection.
- We propose a systematic method for filtering noisy samples and directing oversampling to the most informative minority regions.
- We present extensive experiments on multiple benchmark datasets, demonstrating that CeC-SMOTE outperforms established oversampling techniques.

Accordingly, the research question guiding this study is: How can synthetic minority sample generation be improved to better preserve the structure of imbalanced data, minimize noise, and enhance classification accuracy, especially in challenging datasets with complex class boundaries?

The remainder of this study is organized as follows: Section II reviews related work on imbalanced data learning and oversampling. Section III details the methodology of the proposed

CeC-SMOTE algorithm. Section IV describes the experimental setup and presents the results. Section V discusses the findings and their implications, and Section VI concludes the study with directions for future research.

II. Related Works

Imbalanced data remains a significant challenge in classification problems, often resulting in poor performance for minority classes. To address this issue, researchers have developed various techniques to improve the detection and representation of these underrepresented groups. This section reviews the most relevant methods, highlighting advances in oversampling and clustering-based strategies designed to enhance minority class representation.

A. SMOTE and Variants

Oversampling techniques are widely used to address class imbalance by generating additional synthetic instances of the minority class [11]. This strategy is generally preferred over undersampling, which reduces the number of majority class samples and may result in the loss of valuable information crucial for accurate learning [12] [13]. Among these approaches, the Synthetic Minority Over-sampling Technique (SMOTE) [14] is particularly well-known. SMOTE creates synthetic samples by linearly interpolating between existing minority instances and their nearest minority neighbors [15][16]. Its main goals are to expand the decision region of the minority class and reduce the risk of overfitting that often arises from simply duplicating existing minority data [17].

Despite its popularity, SMOTE's linear interpolation mechanism introduces several notable limitations, especially when applied to complex datasets. First, SMOTE often generates synthetic samples without considering the distribution of the majority class [18]. This can lead to overgeneralization, where new samples encroach into majority class regions or inadvertently bridge distinct minority sub-clusters, increasing class overlap and complicating subsequent classification tasks [19]. Second, SMOTE is susceptible to noise and outliers in the minority class. When outliers or borderline samples are present, the algorithm may create synthetic instances around them, unintentionally amplifying noise and reducing class separability [20]. Third, SMOTE assumes linear relationships among minority samples, which is not always appropriate for datasets with non-linear structures or multiple, distinct minority clusters. In such cases, synthetic samples may not accurately capture the true distribution of the minority class [21]. Additionally, SMOTE faces challenges in high-dimensional data [22]. The nearest neighbor concept, central to its function, becomes less meaningful and more computationally intensive in higher dimensions, which can result in less diverse or even misleading synthetic samples [23]. Finally, SMOTE can struggle with small, isolated minority regions [24]. The algorithm may fail to adequately represent these unique data pockets, leading to their persistent underrepresentation or mischaracterization [25].

B. Enhancing Oversampling through Clustering

To overcome the limitations of traditional oversampling, researchers have incorporated clustering as a preparatory step. In clustering-based SMOTE variants, the minority class is first clustered, and then SMOTE is applied within each cluster [26]. This approach preserves local data structures and addresses imbalances within the minority class itself [27].

Combining clustering with oversampling offers several advantages. First, it reveals underlying subgroups within the minority class, enabling more targeted and nuanced oversampling. By recognizing these sub-clusters, oversampling can be focused on sparser or more critical regions rather than treating all minority samples equally, as standard SMOTE does. Second, clustering helps to identify and manage noise or outliers. Instances that do not fit well into any cluster can be excluded from synthetic sample generation or handled separately, reducing the risk of amplifying noise. Finally, clustering allows for differentiated oversampling strategies which more synthetic data can be generated in sparse but well-defined clusters, or efforts can be focused on safe regions to minimize overlap with the majority class.

Various clustering methods guide the oversampling process. K-means clustering [28]is a commonly used technique that partitions the data into distinct groups based on their centroids. K-Means SMOTE integrates this method by generating synthetic samples within each minority cluster, ensuring that new instances reflect the local distribution and preventing the artificial connection of unrelated groups [29][30] [31]. Other approaches, like LD-SMOTE, adjust the number of synthetic samples based on cluster density, focusing efforts on underrepresented regions and improving sample representativeness [9].

Density-based clustering algorithms, such as DBSCAN, are also employed for their ability to handle irregularly shaped clusters and detect noise [32]. For instance, DBSMOTE uses DBSCAN to cluster minority samples and place new synthetic points towards the center of each cluster, strengthening their core representation [33]. Similarly, adaptive clustering SMOT within clusters formed by DBSCAN [34], and clustering and optimization-based G-mean iteratively applies SMOTE in K-means clusters while optimizing performance metrics such as G-mean [35]. Some methods also use information entropy to monitor and control ambiguity in overlapping regions created by synthetic samples [36].

Despite significant advances in oversampling techniques, several common shortcomings remain across the reviewed methods. Many approaches, such as SMOTE and its variants, tend to generate synthetic samples without fully accounting for the underlying data distribution, leading to increased noise and class overlap. Methods that rely on linear interpolation may struggle with datasets exhibiting complex or nonlinear minority class structures. Additionally, many approaches are also sensitive to clustering quality and may perform poorly on high-dimensional or highly imbalanced datasets. These limitations underscore the need for more adaptive solutions.

III. CeC-SMOTE Methodology

This section delineates the technological process of CeC-SMOTE in greater detail. The flowchart illustrating imbalanced learning using CeC-SMOTE is presented in Fig. 1. The CeC-SMOTE algorithm is principally concerned with the following processes: the capture of the local structure of the minority

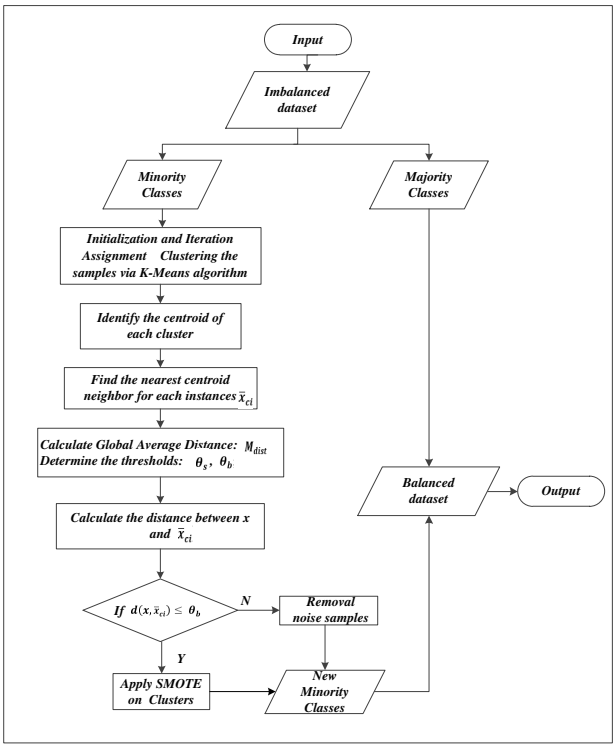


Fig. 1. Block diagram of CeC-SMOTE.

class; the identification of "safe" minority points, with outliers that are too close to majority territory being discarded; and the expansion of the minority class, followed by the rebuilding of the full training set.

The process begins by applying K-means clustering to group the minority class samples, and each cluster's centroid is determined. The algorithm then calculates the global average distance within each cluster to categorize samples as "safe", "boundary", or "noise". Outliers and samples near the majority class are discarded to reduce potential overlap and noise. Next, SMOTE is selectively applied to the safe and boundary regions within each cluster, generating synthetic minority samples where they are most needed. Finally, the new synthetic samples are combined with the filtered original data, resulting in a more balanced dataset that is ready for training the classifier. While LD-SMOTE [9] and Simplicial SMOTE [10] have introduced local density estimation and topological techniques, CeC-SMOTE further advances the field by integrating centroid-based clustering with safety-aware filtering, ensuring that synthetic samples are generated primarily in well-defined, safe or boundary minority regions, as opposed to uniform interpolation.

A. Construct Minority Clusters

To capture the local structure of the minority class, this step involves initialising centroids and iteratively assigning samples to the nearest cluster based on their features. According to the step of the K-means clustering algorithm, until the change in the cluster centre is very small or the preset number of iterations is reached, $k$cluster centres are created to ensure that all samples are stable.

STEP 1: Randomly assign the initial point as the initial cluster centre point;

STEP 2: Calculate the Euclidean distance between all individuals in the sample set and the cluster centre, then assign the samples to the nearest cluster. The Euclidean distance between a point x and a centroid x_{c_i} in cluster c_i is given by Eq. (1):

$$d(x, x_{c_i}) = \|x - x_{c_i}\|_2 = \sqrt{(x_1 - x_{c1})^2 + (x_2 - x_{c2})^2 + \cdots + (x_n - x_{cn})^2} \quad (1)$$

STEP 3: Recalculate the cluster centre according to the samples in the class;

STEP 4: Iterate Steps 2 and 3 until the centroids no longer change significantly or the iteration limit is reached.

Step 2 is critical as it groups similar points together based on their proximity to the centroids. After all points have been assigned to clusters, the centroids are updated to reflect the means of their respective clusters.

B. Safety Assessment and Noise Cleaning

In order to improve class balance and accuracy while limiting noise and overfitting, After clusters the minority class and then identify "safe" minority points and discard outliers that sit too close to majority territory.

STEP 1: Apply the ensemble Nearest centroid neighborhood (NCN) strategy [37]for minority classes. For each data point, the algorithm identifies the nearest cluster centroid $\overline{x}_{ci}$. This involves calculating the distance from the point to each centroid and selecting the smallest one. This NCN helps to assess each data point's proximity to its cluster centre, which is pivotal for classifying the data points in later steps.

The idea of NCN is to find the nearest neighbor by the centroid. For a set of points $X_i = \{x_1, x_2, \cdots, x_n\}$, the centroid is calculated as Eq. (2):

$$\overline{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \quad (2)$$

Neighbors should be distributed in geographical areas, and the idea of NCN requires that proximity be fully considered. First, NCN believes that the centroid neighbors are as close as possible to the test sample in terms of distance and are distributed as evenly as possible around the test sample. The NCN of sample p in X should be obtained by querying the following steps and the pseudocode in Algorithm 1.

STEP 2: Calculate the global average distance. Determine the global average distance by calculating the average distance of all samples to their respective centroids, which will serve as a benchmark for defining safe areas and boundary areas, as in Eq. (3):

$$M_{dist} = \frac{1}{m}\sum_{j=1}^{m} \left\|x_j - \overline{x}_{l_j}\right\|_2 \quad (3)$$

where, m is the total number of points in the dataset, $\overline{x}_{l_j}$ is the centroid of the cluster to which point x_j belongs, and l_j is the index pointing to the cluster to which point x_j belongs.

STEP 3: Classify regions within each cluster;

1) Define the thresholds for safe areas θ_s and boundary areas θ_b [see Eq. (4) and Eq. (5)];

$$\theta_s = \alpha \times M_{dist}, \quad (\alpha < 1) \tag{4}$$
$$\theta_b = \beta \times M_{dist}, \quad (\beta > \alpha) \tag{5}$$

2) Mark each sample as belonging to the safe zone, boundary area, or noise zone based on the comparison of the distance from the sample to the centroid $\mathbf{d}(x, \overline{x}_{c_i})$ and the global average distance M_{dist}:

- If $\mathbf{d}(x, \overline{x}_{c_i}) \leq \theta_s$, x is considered 'safe';
- if $\theta_s < \mathbf{d}(x, \overline{x}_{c_i}) \leq \theta_b$, x is considered 'boundary';
- and if $\mathbf{d}(x, \overline{x}_{c_i}) > \theta_b$, x is considered 'noise'.

By following these steps, the algorithm ensures that only points in well-defined, safe, or boundary regions are considered for oversampling, reducing the impact of noise and outliers.

Algorithm 1 Nearest Centroid Neighborhood (NCN)

Input: $X_i = \{x_1, x_2, \cdots, x_n\}$: input dataset; k: number of neighbors to search; p: query point
Output: $Q = \{q_1, q_2, \cdots, q_k\}$: set of k nearest centroid neighbours
Initialize $Q = \emptyset$
Find the first NCN of p as its nearest neighbour, q_1
$Q \leftarrow Q \cup \{q_1\}$
for $i = 2$ to k **do**
    Select the ith neighbor q_i such that the centroid of q_i and all previously selected neighbours $(q_1, q_2, \cdots, q_{i-1})$ is the closest to p
    $Q \leftarrow Q \cup \{q_i\}$
end forreturn Q

C. Cluster-Aware Oversampling and Merge

After each sample is marked as belonging to the safe area, boundary area, or noise area. Within each cluster, apply SMOTE only to the safe points to create synthetic samples that stay inside the minority manifold.

STEP 1: First, for each sample $\mathbf{x}_i$ in the positive sample set, Euclidean distance is calculated between it and each other sample in the positive sample set, and k nearest neighbor samples are found, marked as $\mathbf{x}'_i$, $i \in \{1, 2, 3, \ldots, k\}$;

STEP 2: For each randomly selected neighbor sample $\mathbf{x}'_i$, a new sample is constructed according to Eq. (4) with $\mathbf{x}_i$, respectively [see Eq. (6)].

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \text{rand}(0, 1) \times (\mathbf{x}'_i - \mathbf{x}_i) \tag{6}$$

STEP 3: Combine the generated samples to form the final dataset.

The CeC-SMOTE achieves a more balanced dataset through targeted oversampling of minority classes and removal of noisy, outlier samples. The pseudocode of the oversampling process of CeC-SMOTE is illustrated in Algorithm 2.

Algorithm 2 CeC-SMOTE

Input: minority imbalanced dataset X^P, K: maximum number of clusters
Output: D: the dataset with augmented minority class samples
Initialize;
Apply the K-means algorithm to the minority class;
Identify the centroid of each cluster: $C = \{c_1, c_2, \ldots, c_K\}$;
Nearest centroid neighbour: $x_{ci} = []$;
for each x_i in X^P **do**
    Initialize *min_distance*; Initialize $x_{ci} = []$;
    for each c_j in C **do**
        Calculate the Euclidean distance from x_i to each centroid c_j;
        $d(x_i, c_j) = \|x_i - c_j\|_2$;
        Keep track of the centroid c_j that has the minimum distance to x_i, storing the index of this nearest centroid c_j;
        if $d(x_i, c_j) <$ *min_distance* **then**
            min_distance $= d(x_i, c_j)$; update the index of the nearest centroid;
        end if
    end for
    Append the index of the nearest centroid to x_{ci};
end forreturn x_{ci}
Calculate the global average distance between all samples in X^P and their respective nearest centroids.

$$M_{dist} = \frac{1}{m} \sum_{j=1}^{m} \|x_j - x_{c_j}\|_2$$

Classify regions within each cluster:
    if $d(x, \overline{x}_{ci}) \leq \theta_s$, mark x_i as "Safe"
    if $\theta_s < d(x, \overline{x}_{ci}) \leq \theta_b$, mark x_i as "Boundary"
    else mark x_i as "Noise"
Apply SMOTE to samples classified as "Safe" and "Boundary" to generate synthetic samples X_G^P;
Merge the original minority class samples X^P with the newly generated synthetic samples X_G^P;

IV. Experimental Analysis

In this section, CeC-SMOTE is compared with four commonly used re-sampling methods on four metrics across seven UCI datasets and two artificial datasets under three quality evaluation measures, the G-mean, F1 score, AUC. We performed 5-fold cross-validation on all datasets to ensure the reliability of the results, averaging metrics across independent runs to reduce variance.

A. Dataset

To showcase the efficacy of CeC-SMOTE, the experimental data in this section are divided into three parts. The first are the two artificially generated two-dimensional (2D) datasets as

shown in Table I. According to the definition of the imbalanced degree, dataset A is highly imbalanced, while dataset B is almost extremely imbalanced. Table II provides a detailed overview of 7 benchmark binary datasets. The IR of binary datasets selected from the UCI data repository varies from a minimum of 1.9 to a maximum of 129.5, with an average of 23.13. According to the imbalance ratio, Pima, ecoli3, Cleveland, vehicle, page-blocks, and Breast datasets are lowly imbalanced. The alone is extremely imbalanced.

TABLE I. ARTIFICIAL DATASETS

Dataset	Imbalance Ratio	Number of Datasets	Number of Minority Classes	Number of Majority Classes	Dimension
A	48.47	1484	30	1454	2
B	94.46	2673	28	2645	2

B. Experimental Metrics

In this study, the performance of CeC-SMOTE was validated using several metrics commonly used in imbalanced data learning: F1-score, geometric mean(G-Mean), and the area under the receiver operating characteristic curve (AUC) to systematically compare and analyze the efficacy of various methodologies. These metrics were selected due to their ability to balance the precision-recall trade-off and capture the model's discriminative power across both classes. The G-Mean is calculated to assess the balance between classification accuracies of the positive and negative classes. It is given by the following Eq. (7):

$$G\text{-}Mean = \sqrt{Specificity \times Recall} \tag{7}$$

where, recall (True Positive Rate), specificity (True Negative Rate) and precision (successfully identified positives out of all positives predicted) are respectively defined as Eq. (8) to Eq. (10):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

TABLE II. DETAILS OF UCI IMBALANCED DATASET

Dataset	Instances	Positive	IR	Attributes
Pima	758	258	1.9	8
ecoli3	336	35	8.6	7
Cleveland	297	35	7.49	13
vehicle	846	199	3.25	19
page-blocks	5472	28	8.79	10
Breast	286	85	2.36	9
Abalone	4176	32	129.5	8

Here, TP denotes the true positives, TN denotes the true negatives, FP denotes the false positives, and FN denotes the false negatives. The G-mean evaluates how well the model performs across both positive and negative classes by considering their accuracies. Furthermore, F1-score is determined as the harmonic mean of the method's precision and recall [see Eq. (11)].

$$\text{F1}_{\text{score}} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

AUC quantifies the overall ability of the model to discriminate between negative and positive classes, and is defined as the area under the receiver operating characteristic curve (ROC), which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). AUC is used to evaluate the model performance. A higher AUC indicates a model with better discriminatory ability.

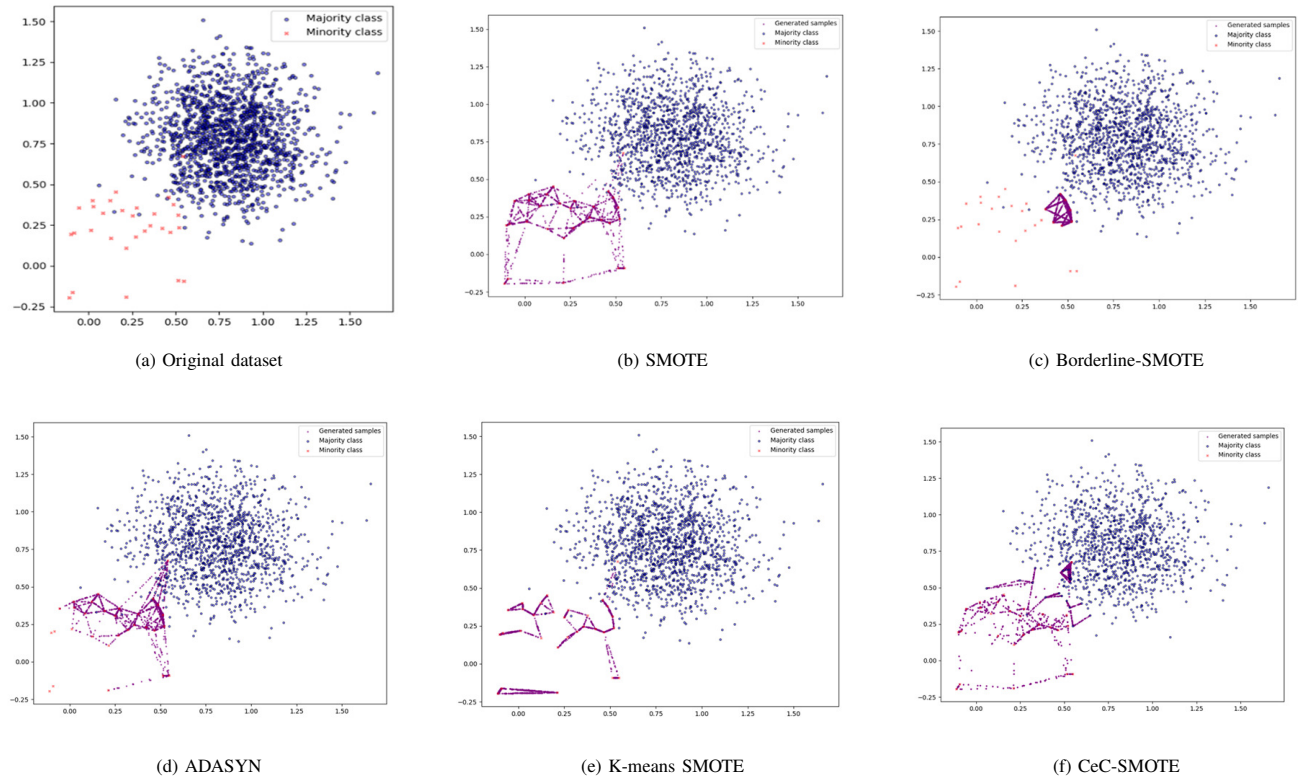
C. Comparison Methodology

In all experiments, CeC-SMOTE is compared with SMOTE and three variants of SMOTE, Borderline-SMOTE, ADASYN and K-Means SMOTE. The comparative methods involved in the experiments are all run with default parameters. The MLP is chosen as the classifier to evaluate the effectiveness of rebalancing on the balanced datasets that have been adjusted using the different comparison methods. The MLP implementation is available in the SciKit-Learn package is provided by the Python implementation, and configured with various layers and neurons using the ReLU activation function. The number of nearest neighbours K involved in all the compared algorithms is five in all experiments.

D. Experimental Results

1) The results on artificial datasets: To demonstrate the efficacy of CeC-SMOTE, we conduct a visualization experiment on artifical datasets. Fig. 2 and Fig. 3 presents the visualization outcomes of samples after being respectively rebalanced by Cec-SMOTE and other comparative methods, followed by dimensionality reduction using t-SNE. The blue, red, and purple points correspond to the negative, positive, and synthetic samples, respectively.

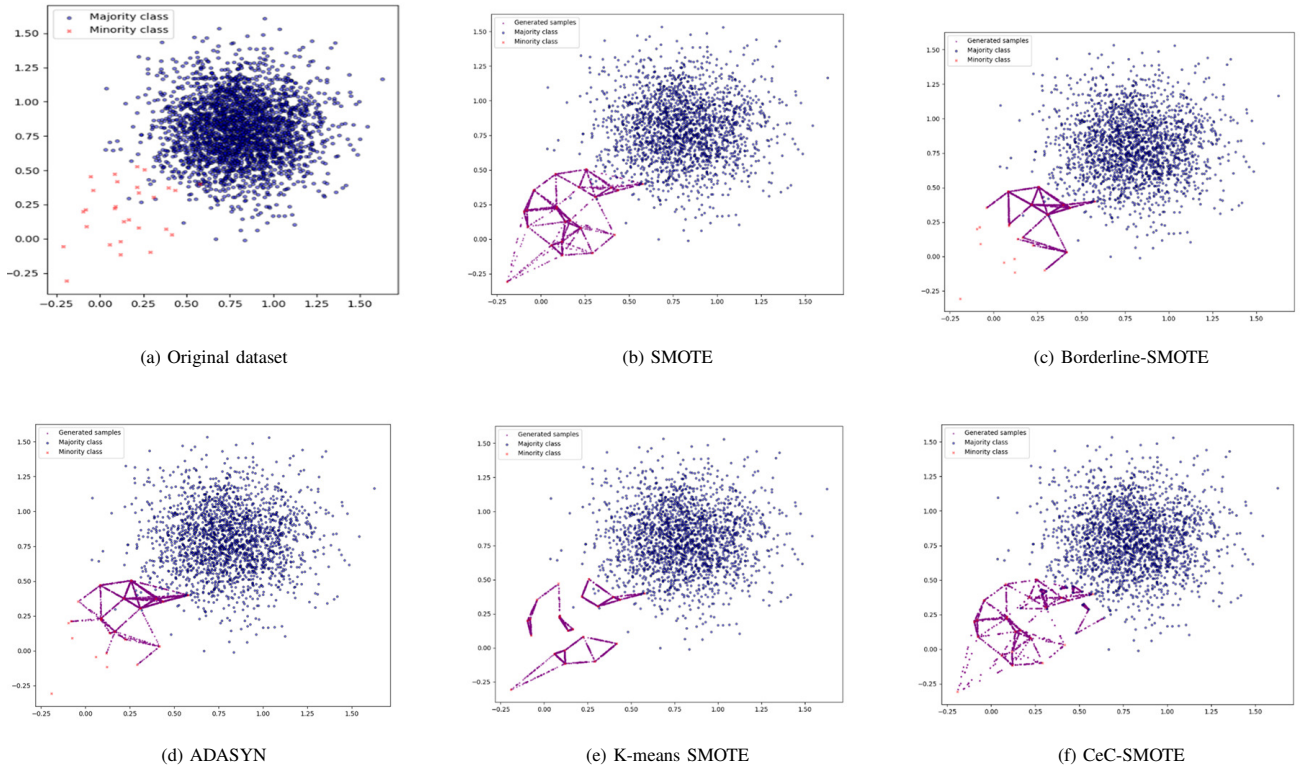
The figures show that the SMOTE algorithm treats all sample points equally. This results in new sample points being generated that overlap with the majority class. CeC-SMOTE first clusters the minority class samples, and then generates new samples near the cluster centroids. This results in the generation of concentrated, pink synthetic samples that adhere closely to the existing cluster structures of the minority class. The sampled data is more consistent with the original data distribution than data sampled using the SMOTE algorithm. Fig. 2(f) shows more minority class data points closer to the boundary than Fig. 2(e). This is because CeC-SMOTE oversample each cluster and the data at the boundary are more likely to be assigned to one cluster after clustering. Thus, the synthesised data are concentrated near the boundary, creating a clear distinction between the minority and majority classes.



(a) Original dataset (b) SMOTE (c) Borderline-SMOTE

(d) ADASYN (e) K-means SMOTE (f) CeC-SMOTE

Fig. 2. Sampling effects of different methods on dataset A.



(a) Original dataset (b) SMOTE (c) Borderline-SMOTE

(d) ADASYN (e) K-means SMOTE (f) CeC-SMOTE

Fig. 3. Sampling effects of different methods on dataset B.

2) The results on UCl datasets: To evaluate the classification performance of the CeC-SMOTE algorithm, the experimental setup consists of seven datasets that present different imbalance ratios, which can be found in the UCI Machine Learning repository. The UCI datasets are used to conduct experiments on the MLP classifier and compared with the other four algorithms.

Table III shows the classification performance of various synthetic oversampling algorithms across multiple UCI datasets. The effectiveness of each method varies by dataset. For instance, in the vehicle dataset, all methods show high G-Mean and AUC, indicating robust performance. However, in datasets like Cleveland and Abalone, the variations in performance metrics are more pronounced, suggesting challenges related to specific characteristics of these datasets, such as feature distributions or class separability. In most cases, AUC scores are high, indicating good discriminative ability of the models post-oversampling. However, the F1-score and G-Mean sometimes show significant variation, highlighting the impact of these methods on precision-recall balance and class-specific accuracy. The performance of the CeC-SMOTE algorithm demonstrates a substantial enhancement in metric values compared to K-means SMOTE. Across the datasets, CeC-SMOTE often shows competitive or superior performance in terms of F1-score and AUC, suggesting its effectiveness in handling border cases and dataset specificities.

Fig. 4 presents a radar chart that compares the experimental results of CeC-SMOTE with those of the other four methods, providing more intuitive experimental results. The CeC-SMOTE line in the chart extends further in most datasets for the G-Mean and AUC metrics, suggesting that this technique might be more effective in distinguishing between classes.

E. Sensitivity Analysis of Critical Parameters

CeC-SMOTE involves several key parameters, including the number of clusters (k) in K-means, and the threshold coefficients α and β used to define "safe", "boundary", and "noise" regions within clusters. The number of clusters k influences the granularity of the minority class partitioning; too few clusters may overlook local data structure, while too many can result in overfitting or fragmented synthetic sampling. The parameters α and β directly affect which samples are classified as safe or boundary points, thereby controlling where new synthetic samples are generated.

Fig. 5 presents the effects of varying the parameters α and β on the F1 and G-mean scores across four datasets: Pima, ecoli3, vehicle, and page-blocks. Each plot illustrates how different (α, β) pairs influence classification performance, with F1 and G-mean used as evaluation metrics.

Across all datasets, the results show that the choice of α and β significantly impacts both F1 and G-mean. The most favorable performance is consistently observed when α is set between 0.65 and 0.80, and β is chosen to be 0.20 to 0.30 higher than α. In this range, both F1 and G-mean reach their peak or maintain stable, high values. This trend suggests that these parameter settings provide an optimal balance between generating a sufficient number of synthetic samples and maintaining the quality of those samples by targeting "safe" and "boundary" regions in the data.

When α is set too low, the algorithm becomes overly permissive, allowing many borderline points to be treated as "safe". This results in more synthetic data but can reduce overall precision due to the inclusion of ambiguous samples. Conversely, when α approaches 0.90, especially in datasets with small minority classes, recall tends to drop. This is likely because the "safe" region becomes too restrictive, limiting the creation of synthetic minority samples and, therefore, failing to sufficiently balance the dataset.

Notably, the (α, β) pair of (0.80, 1.20) performs among the best across all tested datasets, making it a reliable default setting when dataset-specific tuning is not practical. This combination provides a robust trade-off, consistently supporting both strong recall and precision.

For situations involving very scarce minorities or a high number of borderline samples, lowering α to 0.65 and increasing β by 0.10 to 0.20 further improves recall with only a minor reduction in precision. In contrast, for large and highly imbalanced datasets where high precision is more important than recall, setting α near 0.90 and β above 1.30 can be more advantageous, even if recall drops slightly.

In summary, Fig. 5 demonstrates that CeC-SMOTE is robust across a range of α and β values, with optimal performance achieved in moderate settings. The method's flexibility allows practitioners to tailor the balance between recall and precision according to the needs of specific datasets, while also providing a strong default setting for general use.

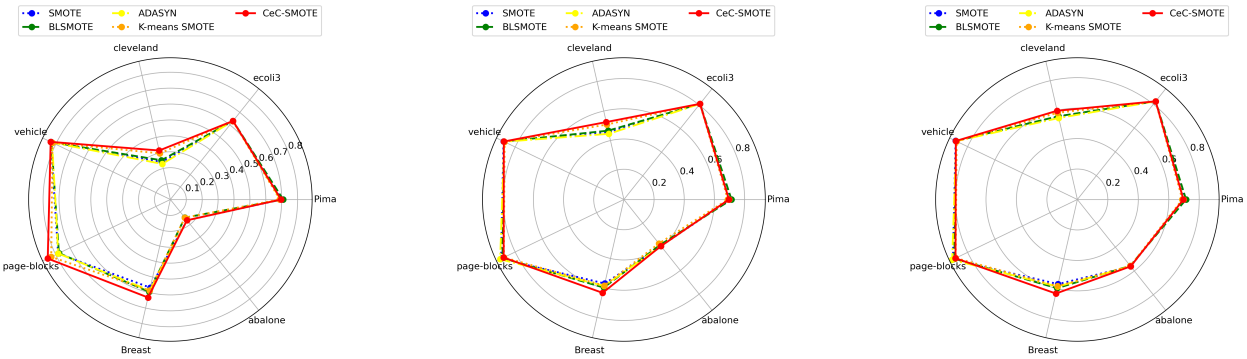
V. Discussion

The comparative results vary across datasets due to differences in characteristics such as imbalance ratio, feature count, class overlap, and minority class distribution. For example, datasets with well-separated classes and moderate imbalance, like "vehicle" or "page-blocks", typically show strong, consistent performance across most oversampling methods, including CeC-SMOTE. However, CeC-SMOTE excels in datasets with smaller minority classes or unclear class boundaries, such as "Cleveland" and "Abalone". Its clustering and noise filtering strategies help preserve meaningful minority samples while reducing the impact of outliers and class overlap.

CeC-SMOTE is particularly effective for datasets, where the minority class is well-clustered or where class boundaries are ambiguous. This observation is consistent with previous findings in the literature, where advanced oversampling techniques that incorporate clustering or density estimation have shown improved performance in handling complex imbalanced datasets. Recent works such as LD-SMOTE [9] and Cluster-Based Reduced Noise SMOTE [15] demonstrate that considering local density or clustering information can enhance the representativeness of synthetic samples and minimize noise. CeC-SMOTE builds upon these principles by combining centroid-guided clustering with adaptive sampling, resulting in synthetic data that better preserves local data structure and reduces class overlap. As shown in Table III and Fig. 4, CeC-SMOTE consistently achieves competitive or superior F1 and AUC scores, particularly in datasets with higher complexity, demonstrating its robustness and adaptability in real-world applications.

TABLE III. METRIC VALUES FOR DIFFERENT IMBALANCE RATIOS

Dataset	Metric	SMOTE	BLSMOTE	ADASYN	K-means SMOTE	CeC-SMOTE
Pima	F1	0.6994	**0.7088**	0.6922	0.6881	0.6959
	G-Mean	0.7015	**0.7133**	0.6972	0.6891	0.6962
	AUC	0.7014	**0.7133**	0.6971	0.6891	0.6961
ecoli3	F1	0.6316	0.6316	0.6316	0.6300	**0.6316**
	G-Mean	0.8094	0.8094	0.8094	0.8084	**0.8094**
	AUC	0.8245	0.8245	0.8245	0.8234	**0.8245**
Cleveland	F1	0.2410	0.2529	0.2295	0.2975	**0.3158**
	G-Mean	0.4557	0.4670	0.4457	0.5085	**0.5522**
	AUC	0.5570	0.5609	0.5489	0.5841	**0.5976**
vehicle	F1	0.8340	0.8298	0.8252	0.8225	**0.8352**
	G-Mean	**0.8859**	0.8832	0.8806	0.8778	0.8836
	AUC	**0.8876**	0.8850	0.8825	0.8798	0.8855
page-blocks	F1	0.7833	0.7776	0.7827	0.8314	**0.8556**
	G-Mean	0.9043	0.8931	**0.9105**	0.8828	0.8848
	AUC	0.9068	0.8961	**0.9125**	0.8864	0.8881
Breast	F1	0.5685	0.5955	0.5865	0.5872	**0.6319**
	G-Mean	0.5699	0.5974	0.5865	0.5840	**0.6333**
	AUC	0.5698	0.5974	0.5865	0.5839	**0.6333**
abalone	F1	0.1672	0.1456	0.1672	0.1460	**0.1672**
	G-Mean	0.3940	0.3844	0.3940	0.3721	**0.3940**
	AUC	0.5612	0.5567	0.5612	0.5578	**0.5613**

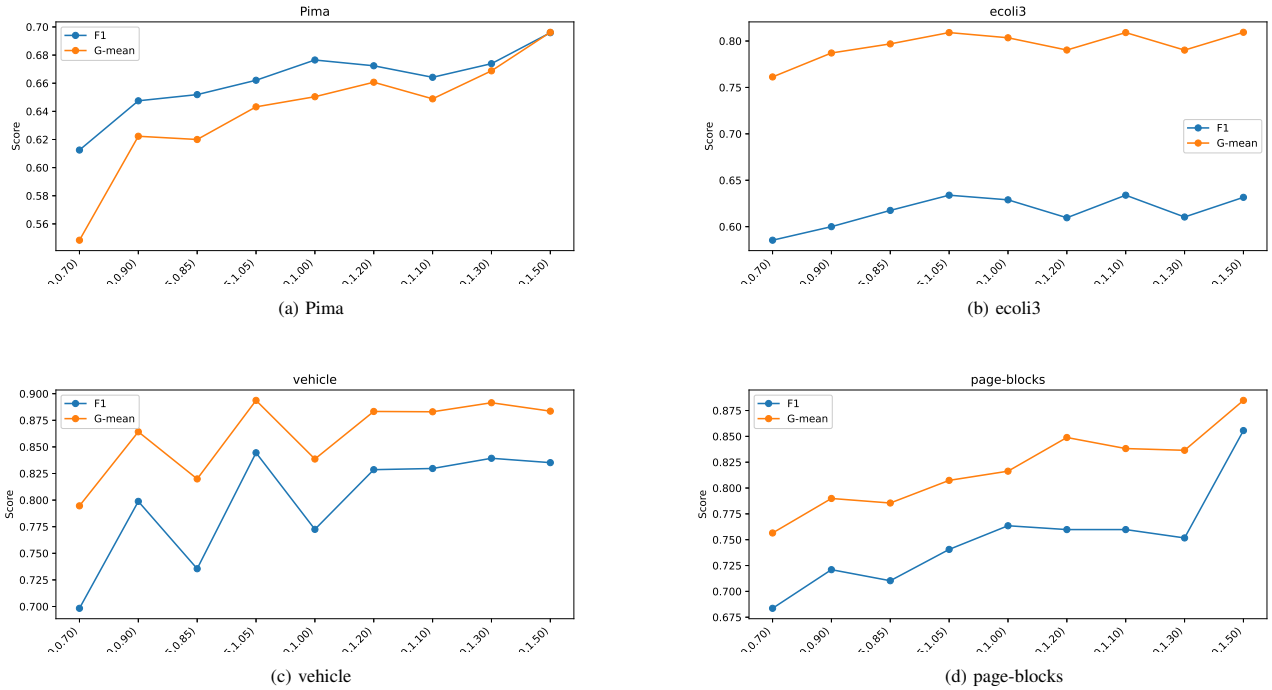


(a) F1-score (b) G-Mean (c) AUC

Fig. 4. Radar chart of experimental results of various algorithms.

CeC-SMOTE's robust performance can be attributed to its combination of centroid-guided clustering and adaptive sampling within safe and boundary regions. By focusing on generating synthetic data near cluster centroids, the method preserves local data structures and avoids excessive overlap with the majority class, which is a common pitfall in standard SMOTE. The visualizations on artificial datasets further demonstrate that CeC-SMOTE produces synthetic samples that align well with the minority class distribution, resulting in clearer class separations.

Parameter sensitivity analysis shows that CeC-SMOTE remains stable across a wide range of (α, β) values. The method performs best when α is between 0.65 and 0.80, with β exceeding α by 0.20 to 0.30. These results suggest that CeC-SMOTE can be deployed with minimal parameter tuning and still deliver reliable improvements in both recall and precision. Moreover, the recommended default setting $(\alpha = 0.80,\ \beta = 1.20)$ consistently ranked among the top configurations for most datasets. For datasets with extremely rare minorities, lowering α and widening β further enhances recall, while highly imbalanced datasets with abundant data



(a) Pima (b) ecoli3

(c) vehicle (d) page-blocks

Fig. 5. Effects of different α and β on performance.

benefit from higher α and β, favoring precision.

VI. Conclusion

In summary, CeC-SMOTE introduces a robust approach to class-imbalance learning by integrating K-means clustering, centroid-based neighbor selection, and safety-aware sampling. This combination enhances the representativeness of synthetic minority samples, preserves the geometric structure of the data, and reduces the risk of amplifying noise. Across multiple real-world and synthetic datasets, CeC-SMOTE demonstrates competitive or superior performance relative to established oversampling techniques, particularly in challenging scenarios with complex class boundaries. The method's stability across different parameter settings makes it practical for real-world deployment, requiring minimal fine-tuning for strong results.

While CeC-SMOTE demonstrates robust performance across multiple real-world and synthetic datasets, several limitations should be acknowledged to provide a more comprehensive review. First, the algorithm's effectiveness relies on the quality of clustering, which may be sensitive to the choice of cluster number and initial centroid selection. Inappropriate clustering could lead to suboptimal synthetic sample generation or noise amplification. Second, this study is limited to binary classification tasks. The adaptation of CeC-SMOTE to multi-class imbalanced datasets and its performance in high-dimensional feature spaces require further investigation. Finally, the current experiments utilize a limited set of benchmark datasets. Future research could extend CeC-SMOTE to multi-class and high-dimensional datasets, and explore automated parameter selection based on data characteristics to fully establish the generalizability of the approach. Overall, CeC-SMOTE offers a valuable and reliable solution for improving classifier performance on imbalanced data.

Acknowledgment

This research was supported by the Natural Science Foundation of Ningxia (grant number: 2024AAC03326).

References

[1] S. Alahyari and M. Domaratzki, Local distribution-based adaptive oversampling for imbalanced regression, vol. 1, no. 1. Association for Computing Machinery, 2025. [Online]. Available: http://arxiv.org/abs/2504.14316

[2] D. Annie, M. Swetha, and S. Sarawagi, "Synthetic Tabular Data Generation for Imbalanced Classification: The Surprising Effectiveness of an Overlap Class".

[3] K. Kitova, "Improving Financial Distress Prediction through Clustered SMOTE," vol. 5, no. 2, pp. 3663–3680, 2025.

[4] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," IEEE Access, vol. 13, no. January, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[5] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," Lecture Notes in Computer Science, vol. 3644, no. PART I, pp. 878–887, 2005, doi: 10.1007/11538059_91.

[6] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5476 LNAI, pp. 475–482, 2009, doi: 10.1007/978-3-642-01307-2_43.

[7] D. T. Utari, "Integration of SVM and SMOTE-NC for classification of heart failure patients," BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 17, no. 4, pp. 2263-2272, 2023.

[8] B. Fish, "SMOTE Variants for Imbalanced Binary Classification: Heart Disease Prediction," vol. 2507, no. February, pp. 1–9, 2020.

[9] J. Lyu, J. Yang, and Z. Su, "LD-SMOTE: A Novel Local Density Estimation-Based Oversampling Method for Imbalanced Datasets," pp. 1–21, 2025.

[10] O. Kachan, A. Savchenko, and G. Gusev, "Simplicial SMOTE: Oversampling Solution to the Imbalanced Learning Problem," 2025, doi: 10.1145/3690624.3709268.

[11] S. He, H., Bai, Y., Garcia, E., & Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning.," IJCNN 2008.(IEEE World Congress on Computational Intelligence) (pp. 1322– 1328), no. 3, pp. 1322– 1328, 2008.

[12] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," Applied Soft Computing Journal, vol. 83, p. 105662, 2019, doi: 10.1016/j.asoc.2019.105662.

[13] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," Inf Sci (N Y), vol. 501, pp. 118–135, 2019, doi: 10.1016/j.ins.2019.06.007.

[14] 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, no. Sept. 28, pp. 321–357, 2002, [Online]. Available: https://arxiv.org/pdf/1106.1813.pdf

[15] J. Hemmatian, R. H. Id, and F. Nazari, "Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE," pp. 1–20, 2025, doi: 10.1371/journal.pone.0317396.

[16] D. Lee and H. Kim, "Adaptive Oversampling via Density Estimation for Online Imbalanced Classification," Information , vol. 16, no. 1, pp. 1–17, 2025, doi: 10.3390/info16010023.

[17] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," Applied Soft Computing, vol. 143, p. 110415, 2023.

[18] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," Machine Learning, vol. 113, no. 7, pp. 4903–4923, 2024.

[19] Y. Chen, W. Pedrycz, and J. Yang, "A new boundary-degree-based oversampling method for imbalanced data," pp. 26518–26541, 2023, doi: 10.1007/s10489-023-04846-4.

[20] X. Yuan, S. Chen, H. Zhou, C. Sun, and L. Yuwen, "CHSMOTE: Convex hull-based synthetic minority oversampling technique for alleviating the class imbalance problem," Inf Sci (N Y), vol. 623, pp. 324–341, 2023, doi: 10.1016/j.ins.2022.12.056.

[21] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning," IEEE Trans Knowl Data Eng, vol. 26, no. 2, pp. 405–425, 2014, doi: 10.1109/TKDE.2012.232.

[22] A. Fern and S. Garc, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," vol. 61, pp. 863–905, 2018.

[23] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," Artificial Intelligence Review, vol. 57, no. 6, p. 137, 2024.

[24] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: range-controlled synthetic minority over- sampling technique for handling the class imbalance problem," Inf Sci (N Y), 2020, doi: 10.1016/j.ins.2020.07.014.

[25] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," Mach Learn, no. 0123456789, 2023, doi: 10.1007/s10994-022-06296-4.

[26] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Inf Sci (N Y), vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[27] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," Inf Sci (N Y), vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[28] A. Alam, M. Muqeem, M. K. Ahamad, and K. O. Mohammed Aarif, "K-means clustering hybridized with nature inspired optimization algorithm: A review," AIP Conf Proc, vol. 2935, no. 1, 2024, doi: 10.1063/5.0202041.

[29] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," pp. 1–19, 2017, doi: 10.1016/j.ins.2018.06.056.

[30] S. Bhavani and N. Subhash Chandra, Histogram Based Initial Centroids Selection for K-Means Clustering, vol. 137. Springer Nature Singapore, 2023. doi: 10.1007/978-981-19-2600-6_38.

[31] J. Fonseca and G. Douzas, "Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures," 2021.

[32] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5059–5074, 2022, doi: 10.1016/j.jksuci.2022.06.005.

[33] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DB-SMOTE: Density-based synthetic minority over-sampling technique," Applied Intelligence, vol. 36, no. 3, pp. 664–684, 2012, doi: 10.1007/s10489-011-0287-y.

[34] C. Bunkhumpornpat, E. Boonchieng, V. Chouvatut, and D. Lipsky, "FLEX-SMOTE: Synthetic over-sampling technique that flexibly adjusts to different minority class distributions," Patterns, vol. 5, no. 11, p. 101073, 2024, doi: 10.1016/j.patter.2024.101073.

[35] J. G. B, A Novel Oversampling Technique for Imbalanced Learning Based on SMOTE and Genetic Algorithm. Springer International Publishing, 2021. doi: 10.1007/978-3-030-92238-2.

[36] L. Tao et al., "Newton cooling theorem-based local overlapping regions cleaning and oversampling techniques for imbalanced datasets," Neurocomputing, vol. 616, no. November 2024, p. 128959, 2025, doi: 10.1016/j.neucom.2024.128959.

[37] L. J. Cohen, "On the use of neighbourhood-based non-parametric classifiers," Philosophy, vol. 30, no. 112, pp. 7–14, 1955, doi: 10.1017/S0031819100036329.

Towards Robust IoT Security: The Impact of Data Quality and Imbalanced Data on AI-Based IDS

Hiba El Balbali, Anas Abou El Kalam
Cadi Ayyad University, National School of Applied Sciences, LaRTID Laboratory, Marrakech, Morocco

***Abstract*—The increased number of connected devices and the rise of Big Data have revolutionized industries and triggered a surge in cyberattacks, making security a top priority. Machine learning and Deep Learning algorithms are crucial in intrusion detection and classification, enabling systems to identify and respond to threats with precision. However, the success of these algorithms is directly related to the quality of the data they process, underscoring the critical importance of robust and well-prepared datasets. Furthermore, despite their potential in detecting and classifying attacks, some algorithms are susceptible to imbalanced datasets, struggling to accurately classify minority classes, while others demonstrate resilience to such challenges. Hence, this study presents a comprehensive analysis of the impact of data quality and imbalanced data on different classification problems, particularly binary, 8-class, and 34-class classification in an intrusion detection context. Our work extensively evaluates six ML and DL algorithms using a novel IoT dataset. Unlike existing research, we use a diverse set of metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and other visual tools, to provide a robust and reliable algorithm performance assessment. This unique analysis underscores the critical importance of addressing data quality and the impact of different balancing techniques on the type of algorithms and type of classification.**

***Keywords*—Machine learning; intrusion detection; internet of things; data quality; big data**

I. Introduction

The proliferation of interconnected devices, often called the Internet of Things (IoT), has led to an unprecedented increase in the volume of data generated and exchanged daily, and has revolutionized how we live and work. These devices, from smart homes to industrial control systems and autonomous vehicles, are increasingly integrated into our daily lives, creating a vast digital ecosystem. While the IoT has brought numerous benefits, it has also introduced new vulnerabilities that cybercriminals are eager to exploit; the surge in connected devices has created a complex attack surface, making it increasingly difficult to identify and mitigate cyber threats. In fact, according to the mid-year update to the 2023 SonicWall Cyber Threat Report, IoT malware globally increased by 37% in the first six months of 2023, resulting in 77.9 million attacks, up from 57 million in the first six months of 2022 [1].

As the number of devices and the complexity of networks grow, so does the sophistication of cyberattacks. These attacks can range from relatively simple exploits, like unauthorized access to personal devices, to large-scale data breaches that compromise sensitive information. More concerning are the cyberattacks that target critical infrastructure, such as power grids, transportation networks, and healthcare systems, where IoT devices play a pivotal role. These attacks, often leveraging vulnerabilities in IoT ecosystems, can lead to severe consequences, including disruptions in critical services, financial losses, and threats to public safety.

Among these, intrusion detection has become a crucial area of research due to the increasing complexity and frequency of cyberattacks. These systems form a critical line of defense against cyber threats by identifying unauthorized access or malicious activities. However, the accuracy of intrusion detection algorithms is highly sensitive to the quality of the data they rely on [2], [3]. Low-quality or imbalanced datasets can lead to misleading results, poor model performance, and a higher risk of false positives or false negatives, which can undermine the detection system's effectiveness.

With the rise of Big Data, the challenge of handling vast and complex datasets has become even more prominent. In the context of intrusion detection, Big Data presents both an opportunity and a challenge. The enormous volume, velocity, and variety of data generated from various sources—such as network traffic logs, user behavior, and IoT device interactions—offer a wealth of information for developing sophisticated and accurate detection mechanisms. This abundance of data enables Machine Learning models to identify subtle patterns and anomalies that could indicate malicious activity, thereby improving the precision and adaptability of intrusion detection systems.

However, leveraging Big Data effectively requires overcoming several challenges. The quality and structure of data are crucial, as cybersecurity datasets are often plagued with challenges such as noise, missing values, and, most notably, class imbalance. Class imbalance is often significant in the field of intrusion detection, where normal network behavior vastly outnumbers instances of malicious activity. In a typical dataset, attacks may represent only a tiny fraction of the total data, leading to a highly skewed distribution that complicates the training of machine learning models. If not addressed properly, this imbalance can cause the models to be skewed toward the majority class, effectively ignoring minority classes that represent actual threats.

Interestingly, not all algorithms are equally affected by data imbalance. As we will discover in this study, some algorithms, like tree-based models, are inherently more robust to class imbalance and can perform well without extensive preprocessing. On the other hand, other algorithms may struggle with imbalanced data unless additional steps are taken to balance the dataset or adjust decision thresholds.

In this study, we present a comprehensive analysis of the impact of class imbalance and data quality on intrusion detection, utilizing a variety of Machine Learning and Deep Learn-

ing algorithms, namely Random Forest, XGBoost, Logistic Regression, Deep Neural Networks, K-Nearest Neighbors, and LSTM. Our research aims to explore the critical importance of addressing class imbalance, which is a common challenge in cybersecurity datasets, and to evaluate how different techniques influence the performance of various algorithms. We investigate both binary and multiclass classification scenarios to provide a broader understanding of how algorithm sensitivity to imbalance classes varies depending on the nature of the classification task.

To rigorously assess the impact of class imbalance, we employ a range of rebalancing techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), SMOTE-Tomek Links, SMOTE with Random Undersampling, and SMOTE-Tomek Links with Random Undersampling. These techniques allow us to examine how oversampling and hybrid methods affect different classifiers. In addition, we utilize a diverse set of evaluation metrics, including precision, recall, F1-score, AUC-ROC, and others, to provide a detailed and nuanced assessment of model performance. By combining these metrics with visualization tools, such as confusion matrices, ROC curves, and PR curves, we aim to deliver a robust comparison and analysis of the effectiveness of each algorithm and technique.

Our study presents several key contributions that distinguish it from previous research in the field:

- The use of a novel dataset with several classes, enabling a more comprehensive evaluation of intrusion detection in complex scenarios where the number of attack types is varied and extensive.
- Hybrid feature selection, combining Chi-squared test and random Forest.
- The evaluation of the impact of class imbalance and data quality on intrusion detection using various Machine Learning algorithms; Random Forest, XGBoost, Logistic Regression, Deep Neural Networks, K-Nearest Neighbors, and LSTM. Our approach provides a broader understanding of how different algorithms are affected by data challenges.
- The creation of several sub-datasets adapted for different classification and balancing scenarios, providing a useful resource for researchers.
- The comparison of multiple balancing methods, including SMOTE, SMOTE-Tomek Links, and combinations with Random Undersampling, to explore their impact across different classifiers in an intrusion detection context.
- The use of a diverse set of evaluation metrics and visual tools, such as precision, recall, F1-score, AUC-ROC, and confusion matrices, ROC curve..., to deliver a detailed analysis of the model's performance. This multidimensional evaluation provides a more complete view compared to traditional existing studies.

Subsequently, this study is organized as follows: Section II will be devoted to discuss some of the related works. In Section III, we will discuss class imbalance and the importance of data quality especially in an Intrusion Detection context. Next, we will present the dataset we used to perform our analysis in Section IV. Section V will be dedicated to present our comprehensive study and workflow. Section VI presents the results, analysis, and findings. Finally, Section VII will conclude our study.

II. Related Work

The authors of [4] conducted a study on the impact of data balancing techniques for SCADA-based intrusion detection systems, utilizing the Morris Power dataset and the CICIDS2017 dataset. They evaluated several balancing methods: Random Sampling, One-Side Selection, Near-Miss, SMOTE, and ADASYN. The authors employed a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks in a binary classification problem. The performance of these techniques was assessed using accuracy, precision, recall, and F1-score metrics. Their findings revealed that the unbalanced data yielded better results for the Morris Power dataset than the balanced versions. Conversely, for the CICIDS2017 dataset, the SMOTE oversampling technique demonstrated better performance. The best accuracy achieved by the model is 99.47% using the SMOTE technique with the CICIDS2017 dataset and 73.63% using the unbalanced Morris Power dataset.

The authors of [5] conducted a comparative analysis of several Deep Learning algorithms, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for binary classification of network intrusion detection tasks. The UNSW-NB15, KDDCup'99, and NSL-KDD datasets were employed to evaluate the performance of these algorithms under both balanced and imbalanced data conditions. The models were assessed using standard metrics such as accuracy, precision, recall, and F1-score. The results indicate that, in general, balanced datasets yield superior performance compared to imbalanced datasets. Notably, the LSTM model achieved the highest accuracy of 95.4% on the balanced KDDCup'99 dataset.

In [6], the authors presented a study on the influence of resampling techniques on the performance of the ANN model. They compared different methods, namely, random undersampling (RU), random oversampling (RO), random undersampling and random oversampling (RURO), random undersampling with Synthetic Minority Oversampling Technique (RU-SMOTE), and random undersampling with Adaptive Synthetic Sampling Method (RU-ADASYN) using four datasets; KDD99, UNSW-NB15, UNSW-NB17 and UNSW-NB18, and they used macro precision, macro recall and macro F1-score to evaluate the results. From this study, we note that the training time increases with oversampling methods and decreases with undersampling, however, the precision scores were almost unchanged.

In IoT intrusion detection contexts, there is a notable gap in the literature regarding the comparative evaluation of different methods for addressing class imbalance using different classifiers to assess the impact of these techniques on each Machine Learning or Deep Learning algorithm.

So in another context, the authors of [7] compared several sampling techniques, namely, Random Under Sampling,

Random Oversampling, and the combination of the SMOTE-NC and RUS as a hybrid resampling technique. They used the Random Forest algorithm for binary classification and the accuracy, precision, recall, F1-score, and ROC-AUC metrics to evaluate each resampling technique. This study showed that sampling techniques improve the model's performance, especially the hybrid method. However, the Random Oversampling method presented an overfitting problem.

The authors [8] performed a comparative study on imbalanced learning techniques for both classification and regression tasks. In the classification part, they employed Random Forest, SVM, and K-NN algorithms to address binary classification. Their analysis compared the performance of models trained on the original dataset and models enhanced with Over-sampling, Under-sampling, SMOTE, and a combination of Over-sampling and Under-sampling techniques. Accuracy was used to evaluate the performance of the algorithms used. The results highlighted that the over-sampling approach, particularly with the Random Forest algorithm, produced the best performance, achieving an accuracy of 98.29%.

The analysis of existing studies that examine the impact of imbalanced data on the performance of machine learning and deep learning algorithms reveals a significant gap in the literature. Very few studies have conducted a comprehensive comparative analysis of various resampling techniques in the context of intrusion detection. Existing works are primarily limited to binary classification, employ a narrow range of algorithms, and rely on a limited set of evaluation metrics, which, while useful, are often insufficient to assess performance fully. Furthermore, the absence of multiclass classification in these studies leaves a critical gap in understanding how resampling techniques perform in more complex scenarios.

Our study aims to address these limitations by comparing multiple resampling techniques for intrusion detection in binary and multiclass classification tasks. This approach will enable us to explore the impact of different resampling methods on each algorithm and classification type. Additionally, we will evaluate a broad range of machine learning and deep learning algorithms using diverse metrics and graphical analysis to provide a more comprehensive performance assessment.

III. Data Quality and Imbalanced Data Challenges

A. Data Quality Concept

Data quality is a critical aspect of any data-driven process, including the degree to which data meets the needs of its intended use. High-quality data drives accurate and reliable insights, fosters trust in analytical outcomes, and supports decision-making processes across various domains. In traditional data systems, ensuring data quality primarily involved managing structured datasets and addressing issues such as missing values, redundancy, and inconsistencies.

The main part of data quality management is data quality assessment, and data quality is generally assessed using several dimensions. A dimension is a set of attributes that indicate a specific element of data quality [9]. The authors [10] described dimension as a quantitative feature of data quality that describes an aspect of data, such as accuracy, precision, consistency, and so on. Dimensions are assessed using metrics, and a metric is a quantifiable instrument that specifies how a dimension is measured. Regarding the number of existing dimensions of data quality, there is no exact agreed-upon count. The authors [9] conducted a comprehensive study that identified 179 distinct dimensions of data quality. In a subsequent analysis focused on the significance of these dimensions, they refined the list to 20 key dimensions, including Completeness, Accuracy, and Consistency, among others.

Over the years, a wide variety of dimensions have been proposed, and the most used data quality dimensions in literature are:

- Completeness: Ensures that all relevant data points are present. Missing values in critical fields, such as timestamps or identifiers, can hinder analyses and lead to misinterpretations. No data should be missing in the dataset.
- Accuracy: Ensures that the data is free from errors and reflects the real-world entities or processes it represents.
- Uniqueness: It guarantees that no duplicates are present in the data.
- Consistency: Emphasizes uniformity within and across datasets, ensuring that data from all systems within an institution is synced and reflects the same information.
- Timeliness: Ensures that data is up-to-date, relevant to the current context, and accessible when it is needed.
- Validity: Ensures that the data aligns with the specific requirements.

The advent of Big Data—characterized by its volume, velocity, variety, veracity, and value—has introduced new complexities to data quality management. Challenges such as handling heterogeneous data sources (structured, semi-structured, and unstructured) like logs, social media, and IoT sensors ..., ensuring real-time processing, and managing data veracity are amplified [11]. Additionally, the rapidly growing nature of big data makes it difficult to ensure timeliness and relevance, as outdated data can quickly lose its value. Furthermore, the presence of noise, biases, and anomalies across massive datasets necessitates advanced techniques for data cleaning, integration, and preprocessing. Big data quality management, therefore, requires scalable approaches that combine traditional quality metrics with modern techniques such as anomaly detection and automated validation mechanisms to ensure that large-scale data remains reliable and useful.

B. Importance of Data Quality in Intrusion Detection Systems

As we have seen in our previous studies [2], [3], the effectiveness of numerous security mechanisms, tools, and modern machine learning (ML) and deep learning (DL) models is heavily reliant on the quality of the data they process. These systems leverage data at various stages to make critical decisions aimed at ensuring system integrity and protection against cyber threats. For instance, authentication systems analyze user credentials and behavior to decide whether to grant or deny access, while intrusion detection and prevention systems

(IDS/IPS) generate alerts based on patterns and anomalies within network traffic. Similarly, Security Information and Event Management (SIEM) systems aggregate and correlate logs from multiple sources to detect potential threats or unusual activities. Hence, poor data quality can severely undermine their performance, leading to inaccurate threat detection and increased false alarms [12]; the rule "Garbage in garbage out" relates here.

Additionally, ML and DL techniques have further revolutionized security mechanisms by enabling the development of predictive models that can dynamically adapt to evolving threats. High-quality data ensures that both traditional and ML/DL-based security mechanisms can differentiate between normal and malicious activities with greater precision, thereby reducing false positives and negatives. Conversely, poor data quality can lead to weak models, decreased detection accuracy, and unreliable predictions.

Thus, maintaining robust data quality through preprocessing techniques such as cleaning, normalization, and feature extraction is crucial. These measures not only enhance the performance of traditional security tools but also ensure the reliability and adaptability of ML and DL models in complex cyber environments. Ensuring data quality is therefore a cornerstone of building resilient, data-driven security systems capable of addressing the ever-evolving landscape of cyber threats.

C. Imbalanced Data

In many real-world applications, including intrusion detection, data is often imbalanced, meaning that certain classes are significantly underrepresented compared to others. The initial research on imbalanced data originated from binary classification issues; which involves the presence of both majority and minority classes, with a certain imbalance ratio [13]. This imbalance poses a significant challenge for some ML and DL models, as it can lead to biased predictions favoring the majority class while neglecting the minority class. Balancing a dataset simplifies model training by preventing the model from becoming biased toward one class. In other words, the model will no longer favor the majority class simply because it has more data [14]. Consequently, accurate detection of rare but critical events, such as intrusions, becomes difficult.

In intrusion detection systems, imbalanced data is a persistent challenge, as malicious activities, which generally represent the minority class, are significantly less frequent than normal activities. In some cases, this class imbalance can have severe implications for the performance of IDS, as it can lead to models being biased towards the majority class, thus failing to detect rare but critical intrusions. Moreover, it can result in an increased number of false alarms, where benign activities are incorrectly classified as intrusions. However, the imbalanced data issue is still very context-dependent, determined by the type of ML and DL algorithms, dataset features, and the classification task. Some models are significantly affected by this imbalance, leading to biased predictions and poor generalization performance. In contrast, other models can adapt more readily as they exhibit a certain level of robustness to class imbalance.

Several techniques have been developed to address the imbalanced data issue, including SMOTE, SMOTE combined with random undersampling (SMOTE RUS), and SMOTE-Tomek Links combined with random undersampling (SMOTE-TL RUS). Each of these techniques operates with unique mechanisms to balance datasets, enabling a better representation of the minority class during model training.

1) SMOTE (Synthetic Minority Over-sampling Technique) [15]: This method generates synthetic samples for the minority class. For each instance in the minority class, SMOTE selects one or more nearest neighbors and creates new examples by interpolating between these points. The process of generating synthetic samples involves selecting random data from the minority class, calculating the Euclidean distance to its k nearest neighbors, multiplying the difference by a random number between 0 and 1, and then adding the result to the minority class as a synthetic sample [16]. This procedure is repeated until the desired proportion of the minority class is achieved.

For a given minority sample x, a synthetic sample x_{new} is created using the formula:

$$x_{new} = x + \lambda(x_{nearest} - x) \tag{1}$$

where, $x_{nearest}$ is one of the k-nearest neighbors of x and λ is a random number in the range [0, 1].

Fifteen years after its introduction, the SMOTE method has become a foundation for the research community in addressing imbalanced classification challenges. Its innovative approach to generating synthetic samples for the minority class has inspired numerous extensions and adaptations, each building upon its foundation to tackle specific limitations or enhance its applicability [17].

2) SMOTE-TL (Synthetic Minority Over-sampling Technique – Tomek Links): It is a hybrid technique that combines the SMOTE technique with Tomek Links to address class imbalance in datasets.

Tomek Links is a variant of the Condensed Nearest Neighbors undersampling approach invented by [18]. Unlike the CNN method, which randomly selects samples along with their k-nearest neighbors from the majority class to be removed, the Tomek Links method employs a rule to identify and eliminate specific pairs of observations. A pair of observations (a, b) is classified as a Tomek Link if it satisfies the following conditions [16]:

*Observation a's nearest neighbor is b.

*Observation b's nearest neighbor is a.

*Observations a and b belong to different classes, with one observation from the minority class and the other from the majority class.

Mathematically, for two instances x_a from the majority class and x_b from the minority class, a Tomek Link exists if they are the nearest neighbors of each other, and the distance $d(x_a, x_b)$ is minimal.

SMOTE-TL combines the SMOTE capability to generate synthetic data for the minority class and the Tomek Links

ability to remove data recognized as Tomek links from the majority class.

The overall process of SMOTE-TL can be stated as follows:

*Apply SMOTE to create synthetic samples for the minority class.

*Identify Tomek Links by checking the nearest neighbor relationships.

*Remove the majority class instances that form Tomek Links with the minority class instances.

3) SMOTE with random under sampling: It is a hybrid resampling technique that combines the SMOTE technique with undersampling. While SMOTE effectively increases the number of minority class samples by generating synthetic examples, it can also increase the risk of overfitting, especially if the majority class is significantly larger. To counter this, undersampling is employed simultaneously, which involves randomly removing instances from the majority class to achieve a more balanced dataset and potentially improve the model's performance.

4) SMOTE-TL with random under sampling: This technique consists in applying SMOTE-TL followed by an additional Random Undersampling step. It is an advanced resampling technique that addresses class imbalance while further refining the dataset. This approach leverages the strengths of both methods to ensure balanced class distributions, remove noisy samples, and reduce the dataset's size, making it computationally efficient for subsequent model training, especially for large datasets.

In this study, we will explore several resampling techniques, namely SMOTE, SMOTE with random undersampling, and SMOTE-TL with random undersampling, and their impacts on the performance of different ML and DL models. Furthermore, we will evaluate how these techniques affect binary and multiclass classification tasks and compare these resampling methods with models trained on imbalanced data to highlight their impact.

IV. Use Case: The CICIOT 2023 Dataset

To conduct our study, we used the CICIoT2023 dataset [19], a recent and extensive IoT attacks dataset. It was created at the Canadian Institute for Cybersecurity, University of New Brunswick. This dataset includes 33 attacks that are classified into seven types of attacks. Table I represents the types of attacks and their description.

The authors extracted over 68 million records and 47 features using an IoT topology regrouping 105 devices.

To optimize computing time, we randomly selected 20% of the original dataset, constituting a subset of 13,584,750 records. Fig. 1 represents the class distribution of our selected data.

This selection was made to facilitate efficient analysis and ensure the feasibility of processing the data within the available computational resources.

V. Workflow and Execution

Our comprehensive study focuses on a multifaceted approach to intrusion detection. It begins with data selection and extraction, followed by a rigorous data cleaning process to prepare the dataset for further analysis. The study then involves creating three distinct datasets: a binary dataset, an 8-class dataset, and a 34-class dataset.

For each dataset, feature engineering is used to identify relevant features and scale the data appropriately. Subsequently, we apply multiple Machine Learning and Deep Learning algorithms to classify attacks under various scenarios of imbalanced data treatment, including using SMOTE, SMOTE-RUS, SMOTE-TL RUS, and unbalanced data. This holistic approach ensures a robust benchmarking and evaluation of the proposed classification techniques across different data representations and balancing strategies.

A. Data Preprocessing

The first phase of our work focused on data preparation, which is a critical step in ensuring the quality and relevance of our analysis.

In this initial phase of our research, we chose the CICIOT 2023 dataset, a novel Internet of Things dataset that aligns well with our study's objectives. The original dataset comprises approximately 68 million records; however, to optimize our computational resources, we randomly extracted a subset of around 13 million records. The extracted data contained approximately 2% benign traffic, compared to the original dataset's 1.6%, ensuring consistency in data representation.

Following, we performed a comprehensive data cleaning process, as illustrated in Fig. 2.

This step was crucial due to the presence of duplicate entries, which we identified and removed. We also standardized the formatting of our columns to ensure uniformity in value representation and corrected data types where necessary; for instance, some columns initially labeled as Float were incorrectly designated as String. Additionally, we conducted a thorough check for missing values to maintain data integrity. To facilitate further analysis, we encoded categorical features using the label encoder technique.

By the end of this phase, we successfully obtained a clean dataset from which we derived three distinct datasets: a binary dataset, an 8-class dataset, and a 34-class dataset, to evaluate our classification models under varying scenarios and complexities.

Following the data preparation phase, we conducted feature engineering, a crucial step in improving the performance of machine learning models by selecting the most relevant features and transforming the data into a more informative format.

As shown in Fig. 3, for each dataset created (binary, 8-class, 34-class), this phase began with feature selection, where we employed a hybrid method that integrates the Chi-Squared test and Random Forest techniques. The Chi-Squared test is a statistical method used to evaluate the independence of categorical variables, helping us identify features that have a

TABLE I. TYPES OF ATTACKS

Type of attack	Attacks	Description
DoS	TCP flood, HTTP flood, SYN flood, UDP flood	An attempt to overload a machine or network, with the aim of weakening its performance or making it completely inaccessible.
DDoS	ACK fragmentation, UDP flood, SlowLoris, ICMP flood, RSTFIN flood, PSHACK flood, HTTP flood, UDP fragmentation, TCP flood, SYN flood, SynonymousIP flood	A DoS attack where multiple machines flood the target.
Brute Force	Dictionary brute force	A hacking method that cracks passwords, encryption keys... using trial-and-error.
Reconnaissance	Ping sweep, OS scan, Vulnerability scan, Port scan, Host discovery	An attack where the actor gathers all the information of his target for exploit.
Spoofing	ARP spoofing, DNS spoofing	An attempt to gain access to a system by hiding the real identity or location.
Mirai	GRE IP flood, GRE Ethernet flood, UDP plain flood	A botnet designed to take over IoT devices and turn them into controlled bots.
Web Based	SQL injection, Command injection, Backdoor malware, Uploading attack, Cross-Site Scripting (XSS), Browser hijacking	A method used by cybercriminals to compromise computer systems, steal data, or cause damage by exploiting vulnerabilities in applications or services accessible via the Internet.

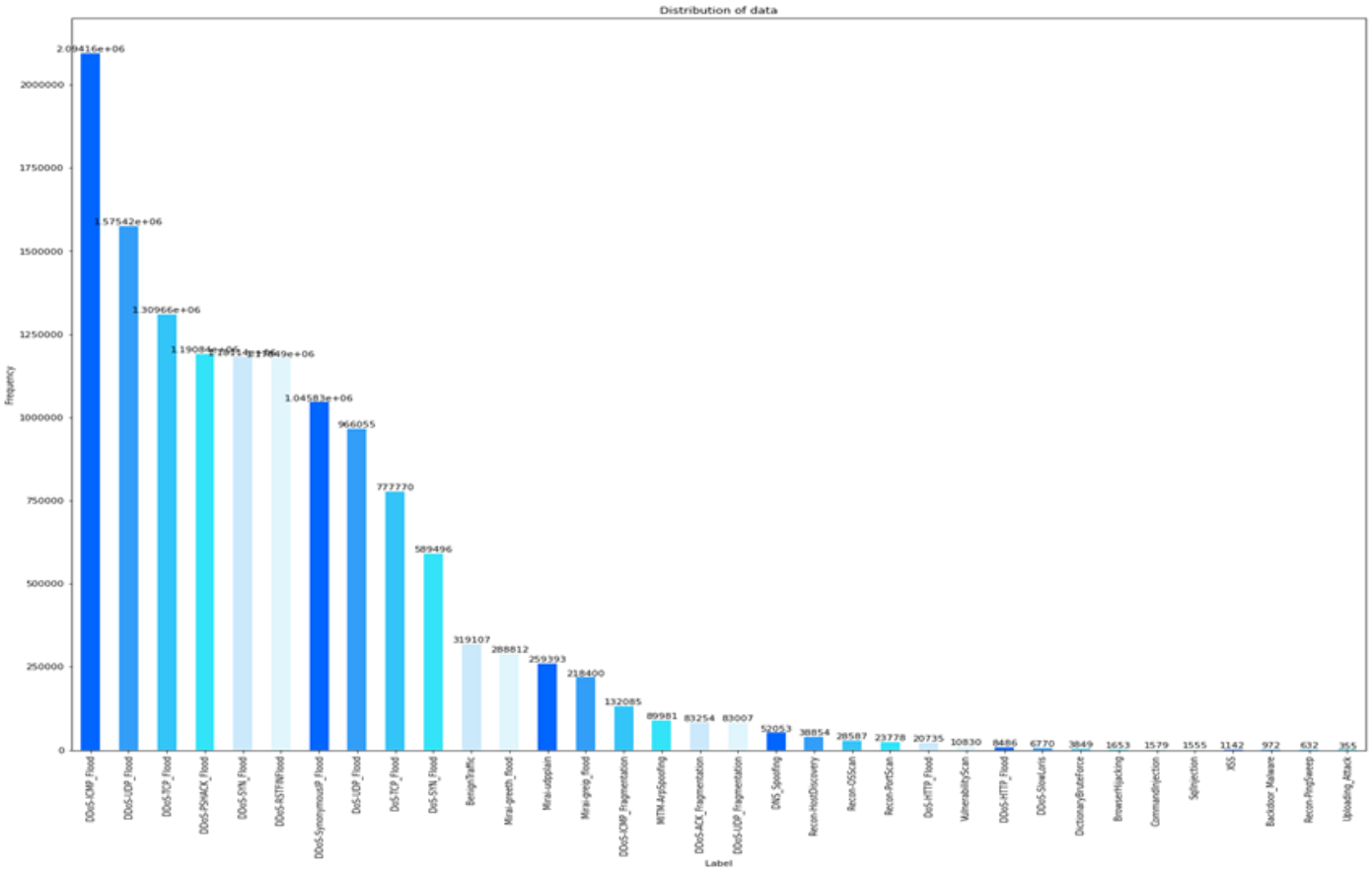


Fig. 1. Class distribution of our selected data.

significant relationship with the target variable. A high Chi-Square score indicates that the independence hypothesis is incorrect, which means the greater the Chi-Square value, the more reliant the feature is on the response and hence suitable for model training [20].

This initial filtering process efficiently narrows the feature set by highlighting the most relevant features based on their statistical significance. Subsequently, we applied the Random Forest algorithm, an ensemble learning method that assesses the importance of features based on their contribution to the model's predictive accuracy. By combining these two techniques, we benefit from the strengths of both: the Chi-Squared test provides a quick and interpretable means of initial selection, while Random Forest offers a robust evaluation of feature importance. This hybrid approach enhances the efficiency of our feature selection process and ensures that we retain the most informative features, ultimately leading to improved model performance.

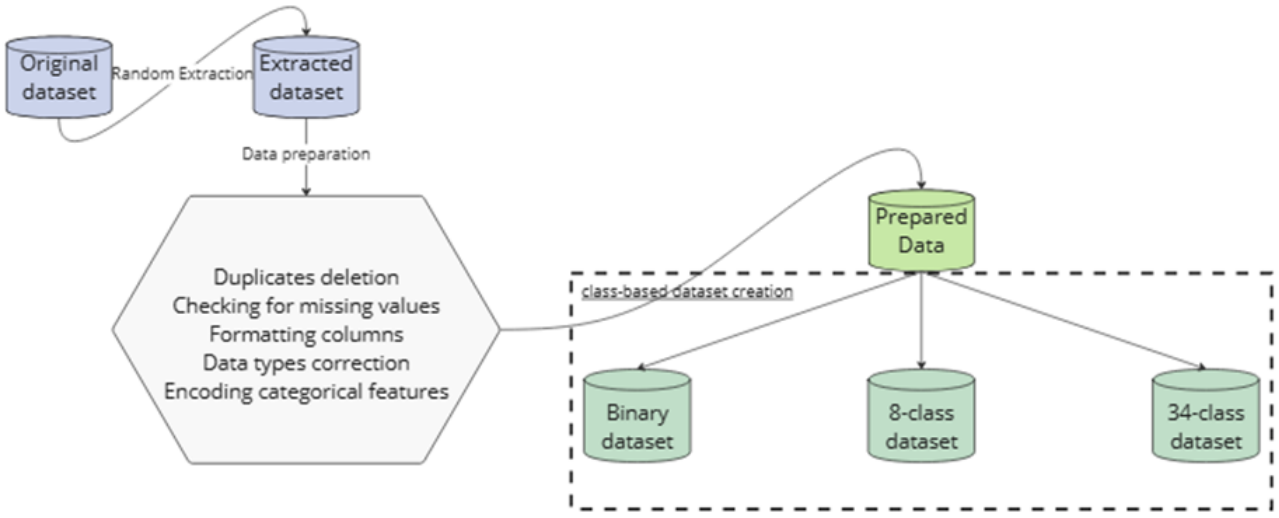


Fig. 2. Our data preparation process.

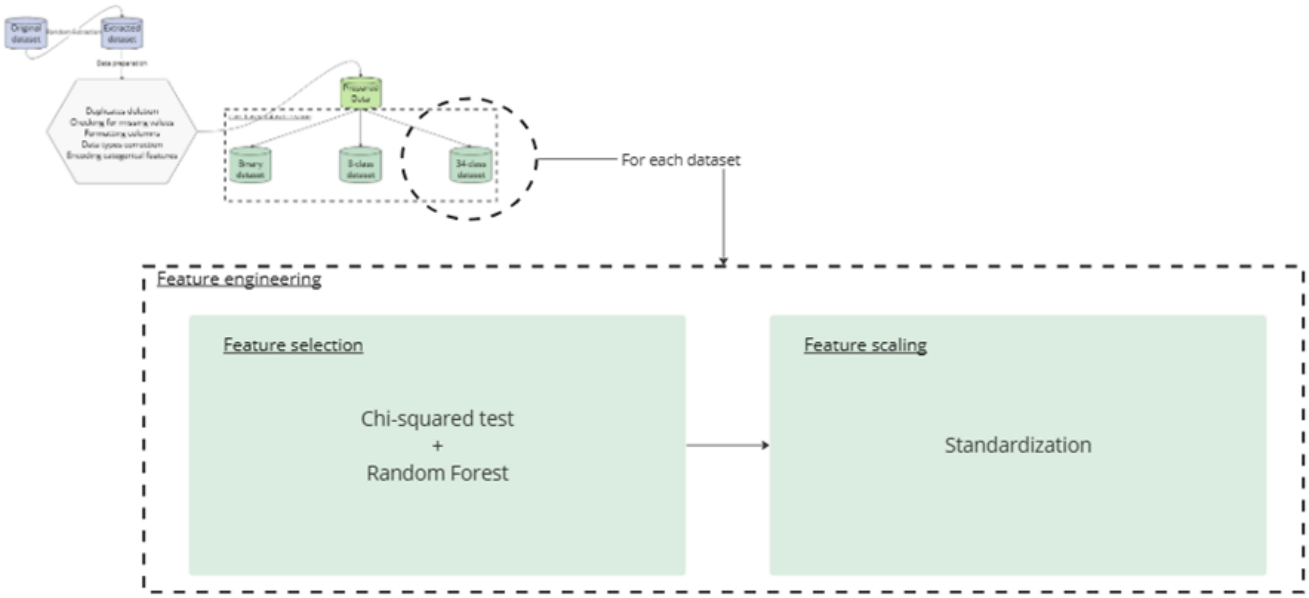


Fig. 3. Our feature engineering process.

Next, we implemented feature scaling using standardization, a crucial process that transforms our features to have a mean of zero and a standard deviation of one. This technique is particularly beneficial in machine learning, as it ensures that all features contribute equally to the distance calculations and optimization algorithms used in various models, regardless of their original scale.

B. Machine Learning and Deep Learning Modeling Under Different Balancing Scenarios

In this phase of our research, we focused on modeling various Machine Learning and Deep Learning algorithms to classify attacks under multiple scenarios: imbalanced data, balanced data using SMOTE, balanced data using SMOTE-RUS, and balanced data using SMOTE-TL RUS, for each dataset (binary, 8-class, and 34-class); which results in 12 scenarios for each algorithm.

As shown in Fig. 4, we chose different algorithms, namely XGBoost, Random Forest, Logistic Regression, Deep Neural Networks, Long Short-Term Memory (LSTM), and K-Nearest Neighbors (K-NN). These algorithms were trained on the respective training datasets, and hyperparameter optimization was conducted using the GridSearch technique.

GridSearch systematically evaluates a given set of hyperparameter combinations to identify the optimal configuration for each algorithm. This approach ensures that the models are fine-tuned to their best possible performance, reducing the risk of underfitting or overfitting. GridSearch offers the advantage of exhaustive search across the hyperparameter space, ensuring no potential configuration is missed.

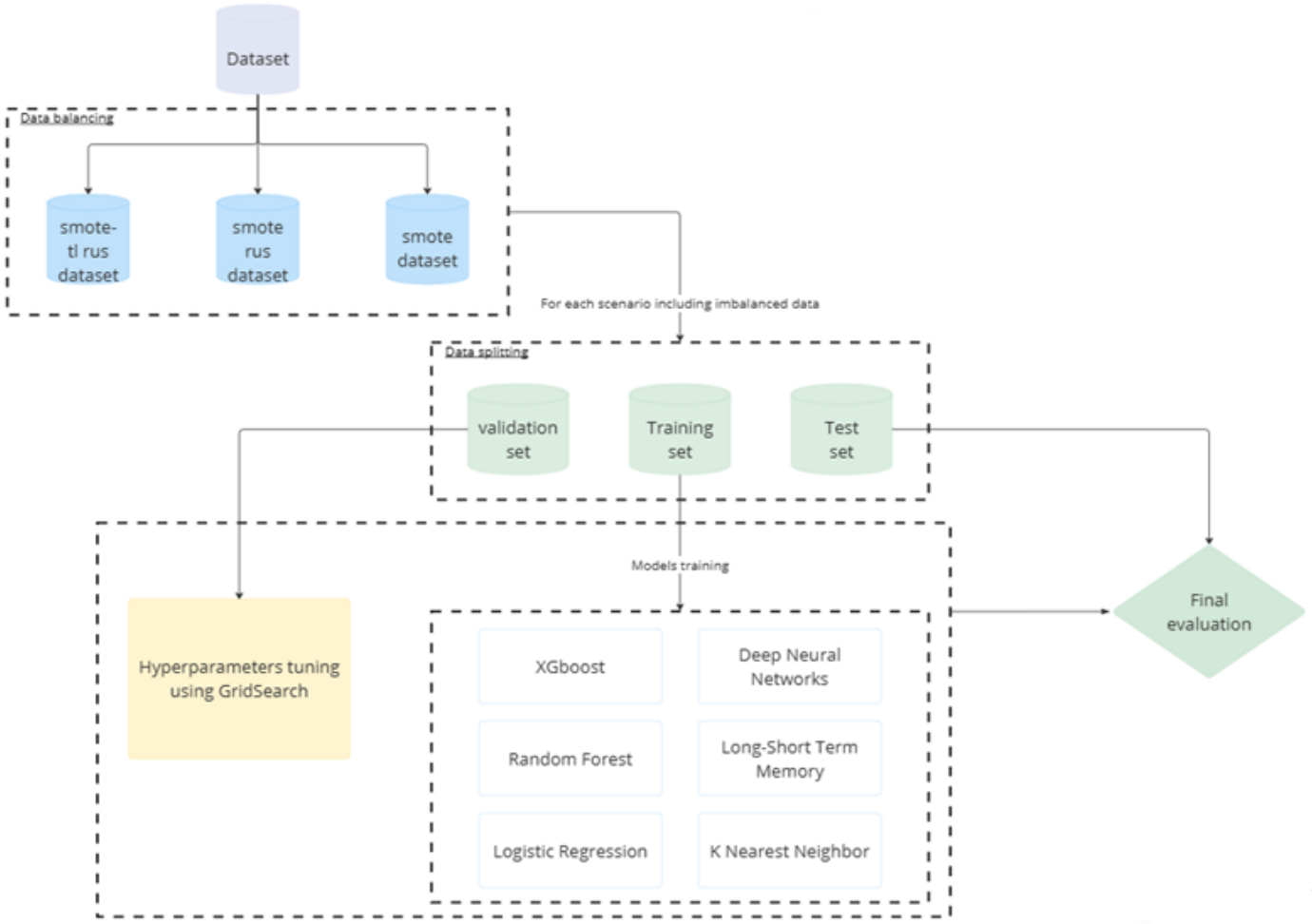


Fig. 4. Our ML and DL modeling workflow.

Once the optimal hyperparameters were identified through validation, the final models were evaluated on the test set. This evaluation step aimed to provide a robust and unbiased assessment of the model's performance.

To ensure a comprehensive and reliable evaluation of our models, we used various performance metrics, namely accuracy, precision, recall, F1-score, AUC-ROC, confusion matrix, PR curve, and ROC curve. These metrics were chosen to capture different aspects of model performance, such as overall accuracy, ability to minimize false positives, sensitivity to detecting true positives, and the balance between precision and recall. Furthermore, the AUC-ROC, PR curve, and confusion matrix provide thorough insights into the model's behavior, offering a deeper understanding of each classifier's strengths and weaknesses.

By leveraging diverse balancing techniques, a wide range of classification algorithms, and multiple datasets, our study provides a detailed and thorough evaluation of the models, ensuring that the results are reliable and generalizable across different attack classification scenarios. This comprehensive approach represents a significant improvement over existing studies that rely on fewer metrics or scenarios.

VI. Results and Discussion

As we mentioned before, we present and analyze the performance of several Machine Learning and Deep Learning algorithms for intrusion detection using a binary dataset, an 8-class dataset, and a 34-class dataset under different balancing scenarios, namely imbalanced data, balanced with SMOTE, and balanced with SMOTE-RUS and SMOTE-TL RUS.

Our experiment was conducted on a Windows 10 Operating System with 16 GB RAM and an Intel(R) Core(TM) i7-7600U CPU @ 2.80 GHz, 2904 MHz processor.

As shown in Table II, for the binary dataset, the results reveal notable differences across balancing strategies.

XGBoost achieved consistent results across all metrics, particularly with the original imbalanced dataset, where it performs slightly better in accuracy, precision, recall, and F1-score. This suggests that XGBoost handles imbalanced data effectively. On the other hand, K-NN demonstrates an excellent performance, especially with the SMOTE-TL RUS

TABLE II. PERFORMANCE METRICS OF OUR MODELS FOR BINARY CLASSIFICATION ACROSS DIFFERENT BALANCING SCENARIOS

		Binary classification			
		SMOTE-TL RUS	**SMOTE RUS**	**SMOTE**	**Imbalanced data**
XGBoost	**accuracy**	98.78	98.20	98.68	**99.28**
	precision	98.85	99.07	98.96	99.57
	recall	98.71	97.33	98.38	99.69
	f1-score	98.78	98.19	98.67	99.63
	AUC ROC	98.78	98.20	98.67	90.99
KNN	**accuracy**	**99.35**	99.07	99.27	99.32
	precision	99.78	99.77	99.86	99.75
	recall	98.92	98.36	98.68	99.55
	f1-score	99.35	99.06	99.27	99.65
	AUC ROC	99.64	99.3	99.61	99.44
Logistic Regression	**accuracy**	95.75	95.44	95.54	**98.54**
	precision	93.61	93.48	93.60	99.26
	recall	98.21	97.71	97.78	99.25
	f1-score	95.86	95.54	95.64	99.25
	AUC ROC	98.61	98.48	98.59	97.26
Random Forest	**accuracy**	95.16	95.60	98.45	**99**
	precision	91.77	92.94	99.66	99.26
	recall	99.23	98.72	97.24	99.70
	f1-score	95.35	95.74	98.43	99.49
	AUC ROC	99.71	99.45	99.68	99.77
LSTM	**accuracy**	**98.78**	97.75	98.66	97.66
	precision	99.01	95.55	99.26	97.66
	recall	98.55	97.96	98.04	100
	f1-score	98.78	97.76	98.65	98.82
	AUC ROC	93.83	97.91	98.76	79.74
DNN	**accuracy**	98.75	98.76	**99.24**	99.22
	precision	98.66	98.75	99.70	99.65
	recall	98.84	98.76	98.78	99.55
	f1-score	98.75	98.75	99.24	99.60
	AUC ROC	99.81	99.82	99.90	99.79

and SMOTE scenarios, highlighting high precision and recall values. It also maintains strong performance on the imbalanced dataset, indicating that K-NN can benefit from both balanced and imbalanced cases.

In addition, Logistic Regression shows moderate performance compared to other models, with significant improvements in the imbalanced dataset. Random Forest shows considerable strength in precision and recall, particularly on the imbalanced dataset and SMOTE scenarios. Its ability to achieve high AUC-ROC scores across all scenarios demonstrates its robustness and reliability for binary classification tasks.

Regarding Deep Learning models, LSTM demonstrates robust performance across most metrics, with high accuracy, precision, recall, and F1-score in all balancing scenarios. However, the AUC-ROC value varies significantly, being lower with imbalanced data, indicating that the model's ability to distinguish between classes is less effective without balancing techniques. SMOTE-TL RUS and SMOTE show relatively consistent and strong results, suggesting these techniques enhance the model's discrimination ability while maintaining other performance metrics. Furthermore, for DNN, the performance is consistently strong across all scenarios, with minimal differences in metrics. The DNN model achieves its highest AUC-ROC with SMOTE, reflecting its remarkable ability to differentiate between classes with imbalanced data. Additionally, its precision and F1-score are notably high, indicating effective handling of class imbalance.

Table III shows the performance of our models across different balancing techniques on an 8-class dataset.

As we can see, XGBoost performs best with imbalanced data, achieving the highest accuracy and AUC ROC. However, its recall and F1-score are low, indicating poor performance with minority classes. Furthermore, K-NN exhibits strong performance across all techniques, with particularly high accuracy and AUC ROC on imbalanced data. Recall and F1-score are low however, SMOTE improves its recall and F1-score, highlighting its suitability for balanced datasets.

In addition, Logistic Regression's performance is relatively stable, with moderate accuracy and AUC ROC. Balancing techniques improve recall and F1-score slightly, but it is clear that the model struggles to capture complex relationships in the data. Random Forest performs poorly on imbalanced data,

TABLE III. PERFORMANCE METRICS OF OUR MODELS FOR 8-CLASS CLASSIFICATION ACROSS DIFFERENT BALANCING SCENARIOS

		8-class classification			
		SMOTE RUS	**SMOTE-TL RUS**	**SMOTE**	**Imbalanced data**
XGBoost	**accuracy**	51.46	59.59	44.18	**81.22**
	precision	56.25	61.42	69.52	76.59
	recall	50.46	58.75	44.15	54.87
	f1-score	49.16	55.05	38.37	52.52
	AUC ROC	88.43	92.3	91.93	96.64
KNN	**accuracy**	64.51	70.59	78.92	**88.07**
	precision	65.62	70.48	79.66	64.82
	recall	64.53	69.10	78.93	57.13
	f1-score	64.55	69.06	78.32	58.56
	AUC ROC	92.20	92.47	96.93	92.54
Logistic Regression	**accuracy**	55.20	60.50	59.12	**80.55**
	precision	57.27	61.22	61.41	63.03
	recall	55.58	58.23	59.11	41.15
	f1-score	53.15	57.32	58.06	42.81
	AUC ROC	88.73	89.68	90.56	94
Random Forest	**accuracy**	49.1	63.11	58.04	**73.30**
	precision	55.08	68.53	76.70	77.16
	recall	48.45	61.82	58.06	56.78
	f1-score	46.55	61.62	57.25	57.57
	AUC ROC	89.39	92.42	95.32	92.69
LSTM	**accuracy**	51.46	58.18	59.61	**78.47**
	precision	69.72	67.29	79.83	27.30
	recall	52.42	57.12	59.59	32.35
	f1-score	47.35	56	59.21	27.72
	AUC ROC	88.04	89.76	89.29	86.02
DNN	**accuracy**	55.83	61.18	73.22	**88.13**
	precision	60.55	65.55	75.96	72.96
	recall	56.30	58.88	73.22	57.3
	f1-score	54.22	58.08	73.03	59.73
	AUC ROC	89.54	90.06	96.15	97.52

with accuracy and F1-score trailing other models. Smote-TL RUS significantly boosts RF's recall and F1-score, indicating its reliance on balanced datasets to handle minority classes effectively.

Concerning Deep Learning models, LSTM shows inconsistent performance, with high precision on SMOTE but very low recall on imbalanced data. Balancing techniques like SMOTE-TL RUS and SMOTE enhance its recall and F1-score. Moreover, DNN performs well overall, with the best accuracy and F1-score on imbalanced data. Its AUC ROC is also the highest among all models. SMOTE significantly enhances its recall and F1-score, highlighting its ability to leverage balanced data effectively.

Table IV shows the performances of our models across different balancing techniques on a 34-class dataset.

As observed, XGBoost performs quite well with imbalanced data, achieving moderate accuracy and a high AUC ROC, indicating its ability to separate classes probabilistically. However, its low recall and F1-score reveal poor performance on minority classes. Balancing techniques such as SMOTE and SMOTE RUS provide slight improvements in recall and F1-score but result in lower overall accuracy. Furthermore, K-NN demonstrates robust performance, particularly with Sample SMOTE, achieving high accuracy and AUC ROC. While its recall and F1-score are relatively low on imbalanced data, balancing techniques like Sample SMOTE significantly boost these metrics, showing its adaptability to rebalanced datasets.

Moreover, Logistic Regression maintains moderate performance across all scenarios, with high AUC ROC values. However, its accuracy and F1-score are low, reflecting its inability to model complex relationships. SMOTE slightly improves recall and F1-score, but the model struggles to achieve competitive performance compared to others. Random Forest shows high accuracy and precision on imbalanced data but low recall and F1-score, indicating challenges in handling minority classes. SMOTE noticeably improves RF's recall and F1-score, demonstrating its reliance on balanced datasets to perform effectively.

Among Deep Learning models, LSTM shows inconsistent performance. It achieves a high AUC ROC on imbalanced data but has low recall and F1-score. SMOTE enhances its recall and F1-score, indicating its potential with balanced

TABLE IV. PERFORMANCE METRICS OF OUR MODELS FOR 34-CLASS CLASSIFICATION ACROSS DIFFERENT BALANCING ACENARIOS

		34-class classification			
		Imbalanced data	**SMOTE-TL RUS**	**SMOTE RUS**	**SMOTE**
XGBoost	**accuracy**	**76.28**	39.08	40.18	40.44
	precision	56.55	35.65	41.06	44.55
	recall	53.63	35.86	40.82	40.46
	f1-score	49.64	30.83	35.80	38.46
	AUC ROC	95.01	90.73	92.38	69.83
KNN	**accuracy**	**79,03**	50.76	46.72	74,6
	precision	58,78	46.43	47.51	74,72
	recall	48,89	44.50	47.15	74,62
	f1-score	48,78	39.65	45.12	73,48
	AUC ROC	88,14	85.06	84.31	97,77
Logistic Regression	**accuracy**	**76.03**	46.27	42.81	51.86
	precision	44.94	43.04	43.10	51.40
	recall	36.50	44.17	45.86	51.82
	f1-score	35.37	38.19	39.76	48.97
	AUC ROC	97.83	92.91	92.64	94.67
Random Forest	**accuracy**	**79.93**	30.35	30.46	66.54
	precision	64.34	23.72	26.35	72.97
	recall	51.74	29.90	31.95	66.54
	f1-score	52.18	22.06	24.12	64.52
	AUC ROC	95.55	86.83	90.99	98.39
LSTM	**accuracy**	**80.63**	42.04	35.94	55.75
	precision	38.07	31.12	31.88	59.62
	recall	35.33	38.24	40.57	55.69
	f1-score	33.77	30.16	30.34	51.67
	AUC ROC	98.07	92.51	92.40	96.11
DNN	**accuracy**	**83.14**	36.81	34.68	63.55
	precision	50.35	24.52	28.29	67.12
	recall	45.99	31.55	37.93	63.49
	f1-score	45.46	22.82	29.02	61.49
	AUC ROC	98.98	89.47	90.31	97.48

datasets. Similarly, DNN achieves the best overall performance on imbalanced data, with the highest accuracy and AUC ROC. While its recall and F1-score are initially low, SMOTE and Smote-TL RUS significantly enhance these metrics.

Due to space limitations, we only show ROC curves, PR curves, and confusion matrices of the best models across all balancing scenarios on each dataset.

Fig. 5, Fig. 6, and Fig. 7 show the ROC curve of the best model in each dataset.

Fig. 8, Fig. 9, and Fig. 10 show the PR curve of the best model in each dataset.

Fig. 11, Fig. 12, and Fig. 13 show the confusion matrices of the best model in each dataset.

As we can deduce, our curves and matrices further confirm and support the reported scores and the accompanying discussion. They visually illustrate the variations in performance across models and balancing techniques.

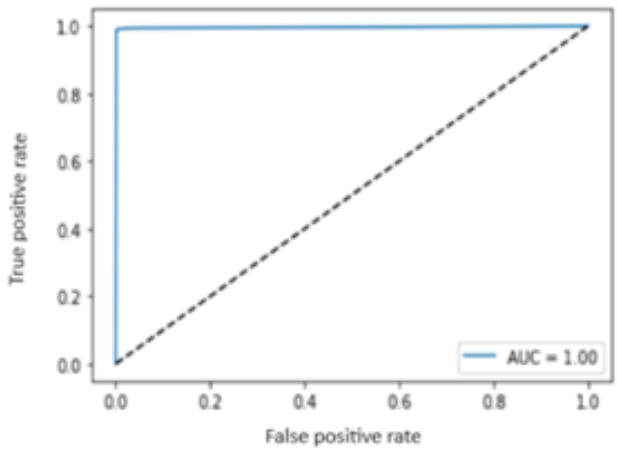


Fig. 5. ROC Curve of the DNN model using imbalanced data and the binary dataset.

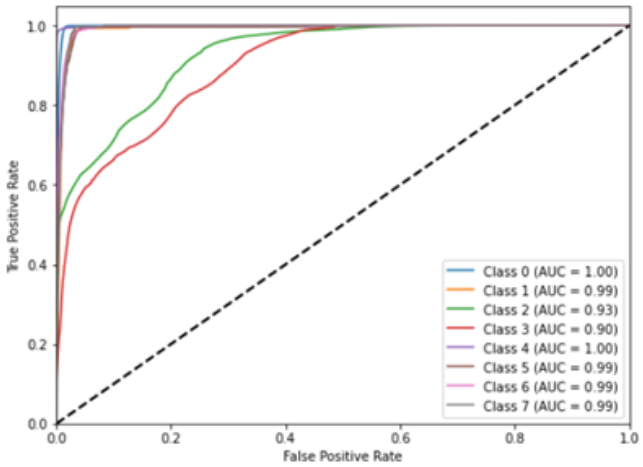


Fig. 6. ROC Curve of the DNN model using imbalanced data and the 8-class dataset.

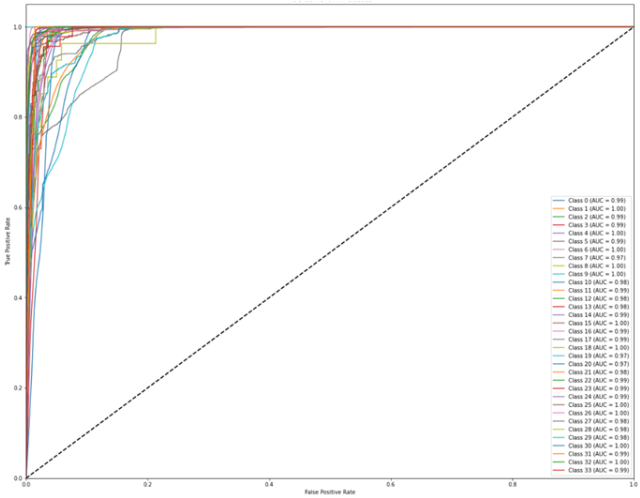


Fig. 7. ROC Curve of the DNN model using imbalanced data and the 34-class dataset.

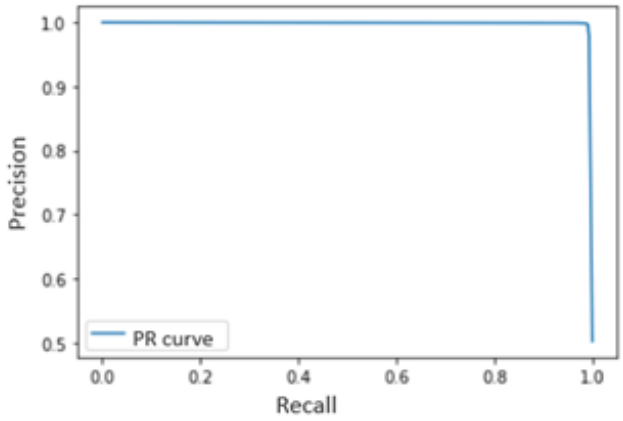


Fig. 8. PR Curve of the DNN model using imbalanced data and the binary dataset.

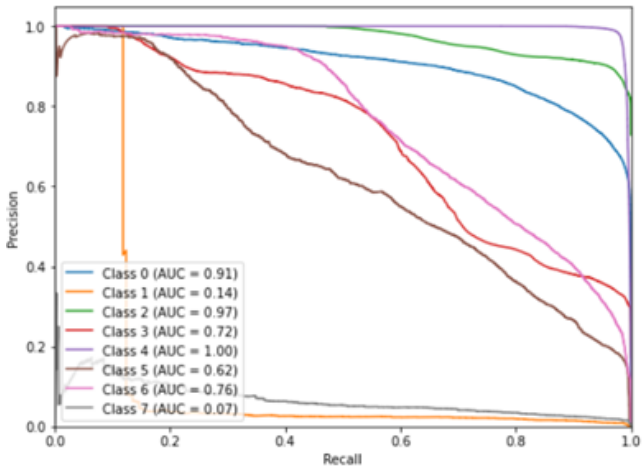


Fig. 9. PR Curve of the DNN model using imbalanced data and the 8-class dataset.

VII. Conclusion

In this study, we conducted a comprehensive analysis of the impact of data quality and class imbalance on classification problems, focusing on binary, 8-class, and 34-class classification in the context of intrusion detection. Using a novel IoT dataset, we thoroughly evaluated the performance of various Machine Learning and Deep Learning algorithms in different balancing scenarios and with different datasets. Our work goes beyond traditional evaluations by incorporating a diverse set of performance metrics; accuracy, precision, recall, F1-score, AUC-ROC, and visual tools such as confusion matrices, PR curves, and ROC curves. This approach ensures a robust and reliable assessment of algorithmic performance.

Additionally, our study demonstrates how results vary depending on each balancing technique and classification type, with certain models excelling in binary tasks but showing limitations in multiclass settings, while others are susceptible to imbalanced data and handle it differently. Another key outcome of this research is the creation of multiple sub-datasets tailored for different classification and balancing scenarios, offering a valuable resource for researchers seeking to extend this analysis or apply these datasets for their research.

In our future work, we intend to investigate the strategic role of data quality in reinforcing the resilience of Intrusion Detection Systems against sophisticated adversarial threats. Moreover, our findings also emphasize the trade-off between the performance of our models and computational cost, as these robust models often require significantly greater computational resources. Finally, extending this approach to other critical domains such as automotive or healthcare systems could provide further insights into the generalization of our findings.

References

[1] A. Marton, "IoT malware attacks up by 37% in the first half of 2023," iotac.eu, 2023. https://iotac.eu/iot-malware-attacks-up-by-37-in-the-first-half-of-2023/#: :text=According to the mid-year,first six months of 2022. (accessed Dec. 14, 2024).

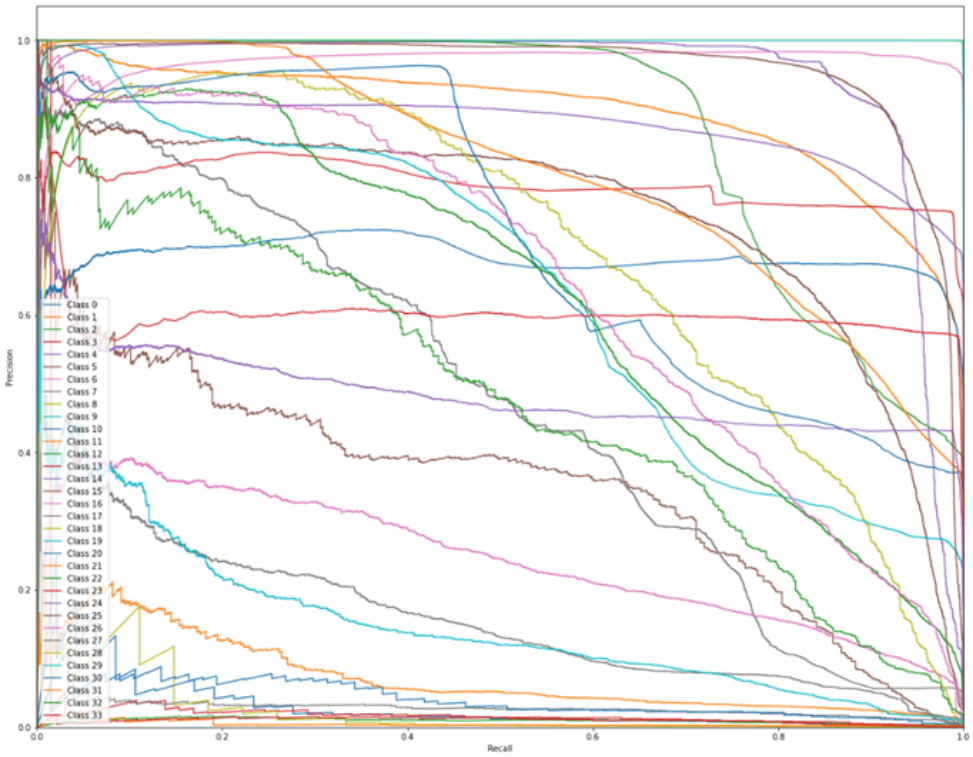


Fig. 10. PR Curve of the DNN model using imbalanced data and the 34-class dataset.

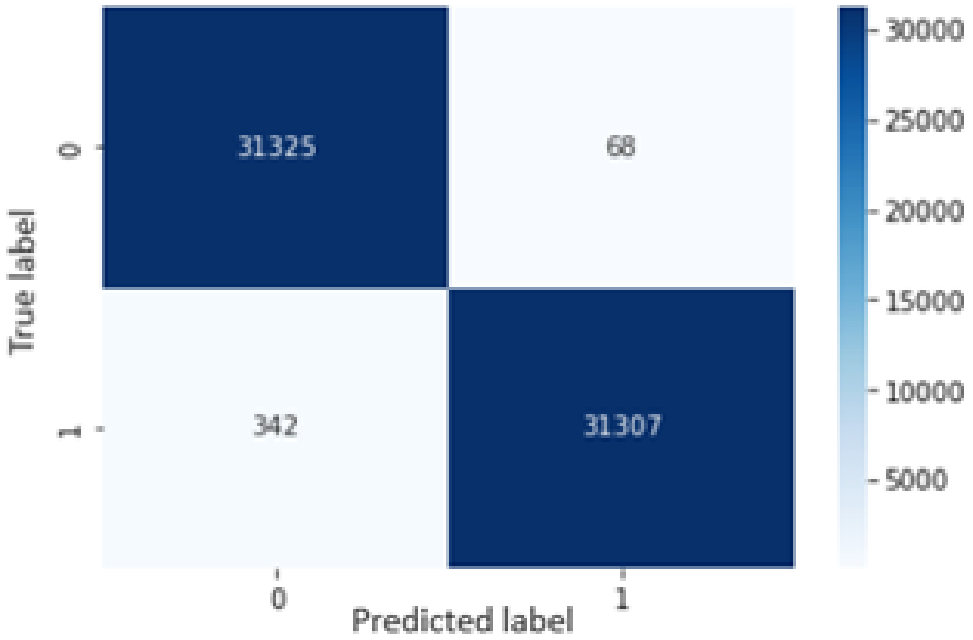


Fig. 11. Confusion matrix of the DNN model using imbalanced data and the binary dataset.

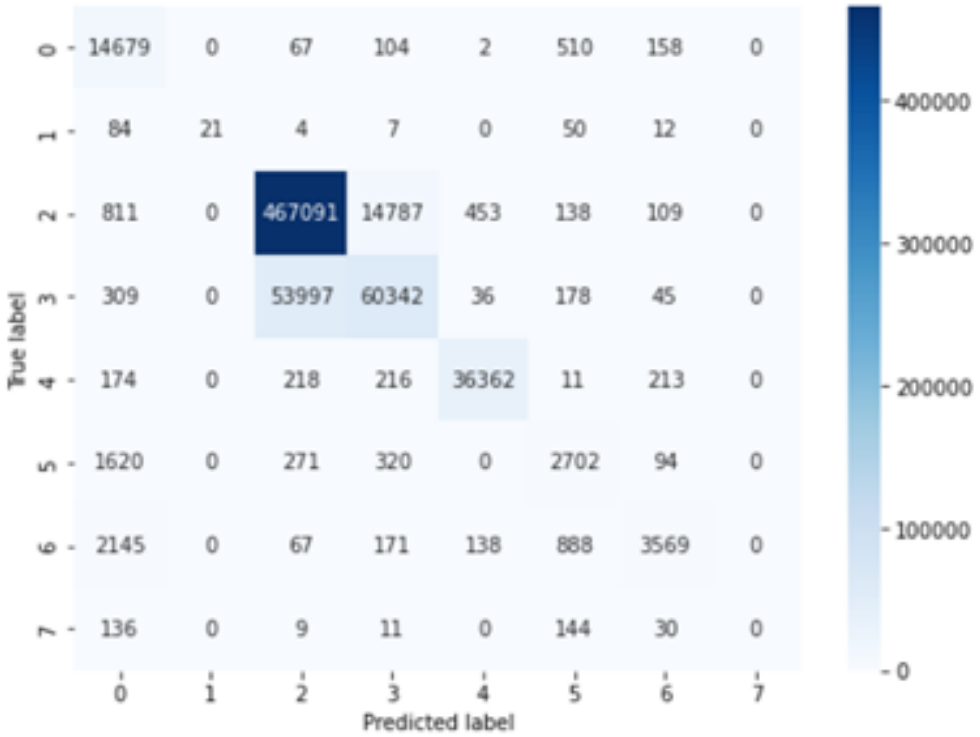


Fig. 12. Confusion matrix of the DNN model using imbalanced data and the 8-class dataset.

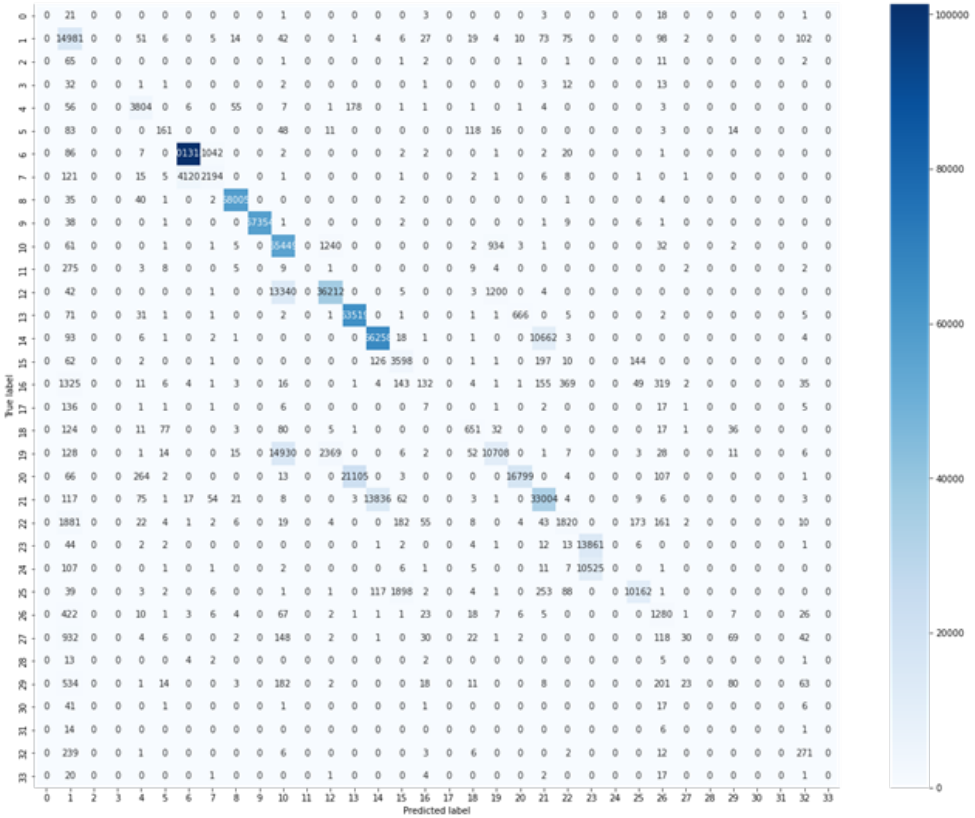


Fig. 13. Confusion matrix of the DNN model using imbalanced data and the 34-class dataset.

[2] H. El Balbali and A. Abou El Kalam, "AI-Driven Big Data Quality Improvement for Efficient Threat Detection in Agricultural IoT Systems," 2023. doi: https://doi.org/10.1007/978-3-031-54318-0_5.

[3] H. El Balbali, A. Abou El Kalam, and M. Talha, "Big Data Between Quality and Security," pp. 1315–1326, 2023, doi: 10.1007/978-3-031-27409-1_120.

[4] A. Balla, M. H. Habaebi, E. A. A. Elsheikh, M. R. Islam, and F. M. Suliman, "The Effect of Dataset Imbalance on the Performance of SCADA Intrusion Detection Systems," Sensors, vol. 23, no. 2, 2023, doi: 10.3390/s23020758.

[5] A. Meliboev, J. Alikhanov, and W. Kim, "Performance Evaluation of Deep Learning Based Network Intrusion Detection System across Multiple Balanced and Imbalanced Datasets," Electron., vol. 11, no. 4, 2022, doi: 10.3390/electronics11040515.

[6] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00390-x.

[7] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," Inf., vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[8] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," Appl. Soft Comput., vol. 143, 2023, doi: 10.1016/j.asoc.2023.110415.

[9] R. Y. Wang and D. M. Stong, "Beyond accuracy: What data quality means to data consumers," J. Manag. Inf. Syst., vol. 12, no. 4, pp. 5–34, 1996, doi: 10.1080/07421222.1996.11518099.

[10] M. Talha, A. Abou El Kalam, and N. Elmarzouqi, "Big data: Trade-off between data quality and data security," Procedia Comput. Sci., vol. 151, pp. 916–922, 2019, doi: 10.1016/j.procs.2019.04.127.

[11] M. Talha, N. El Marzouqi, and A. Abou El Kalam, "Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution," J. Ubiquitous Syst. Pervasive Networks, vol. 12, no. 1, pp. 09–15, 2019, doi: 10.5383/juspn.12.01.002.

[12] N. Tran, H. Chen, J. Bhuyan, and J. Ding, "Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection," IEEE Access, vol. 10, pp. 121900–121923, 2022, doi: 10.1109/ACCESS.2022.3211313.

[13] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," IEEE Trans. Neural Networks Learn. Syst., vol. 34, no. 9, pp. 6390–6404, 2023, doi: 10.1109/TNNLS.2021.3136503.

[14] N. Buhl, "Introduction to Balanced and Imbalanced Datasets in Machine Learning," encord, 2022. https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/ (accessed Dec. 11, 2024).

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[16] R. A. A. Viadinugroho, "Imbalanced Classification in Python: SMOTE-Tomek Links Method," Towards Data Science, 2021. https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc (accessed Dec. 24, 2024).

[17] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," J. Artif. Intell. Res., vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[18] I. Tomek, "A generalization of the k-NN rule," IEEE Trans. Syst. Man. Cybern., vol. 6, no. 2, pp. 121–126, 1976, doi: 10.1109/TSMC.1976.5409182.

[19] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," Sensors, vol. 23, no. 13, p. 5941, 2023, doi: 10.3390/s23135941.

[20] sampath kumar Gajawada, "Chi-Square Test for Feature Selection in Machine learning," Towards Data Science, 2019. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

Automated Anatomical Analysis of Wood Cross Sections Using Macroscopic Images

Khanh Nguyen-Trong[1], Thanh Nhan Nguyen-Thi[2]
Intelligent Computing for Sustainable Development Laboratory (IC4SD)[1]
Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[1,2]

***Abstract*—Wood anatomical features are crucial in forestry science, traditionally relying on manual inspection of wood cross-sections. This conventional method is time-consuming, subjective, and dependent on expert experience. Recent advancements in deep learning offer high accuracy but often operate as black-box models, lacking interpretability and struggling with out-of-distribution challenges under real-world variations. To address these limitations, we propose a two-stage framework combining deep-learning-based image classification and explicit anatomical feature analysis, directly extracting expert-recognized morphological attributes such as pore size, frequency, and spatial arrangement from macroscopic images. By quantifying these anatomical descriptors, our framework yields transparent, OOD-robust features that can be directly fed into downstream species-identification models, thereby enhancing future classification accuracy while preserving interpretability. An end-to-end implementation integrates data acquisition, automated feature extraction, and interactive visualization, making the methodology practically applicable in both laboratory and field settings.**

***Keywords*—Wood species identification; wood anatomical analysis; segmentation; Mask R-CNN; DenseNet**

I. Introduction

Wood anatomical analysis is fundamental to forestry management, biodiversity conservation, and accurate identification of timber species. In routine practice, specialists inspect anatomical features on the transverse surface—such as vessel (pore) arrangement, ray structure, and axial parenchyma—to determine the taxon. Especially in many developing countries, samples for wood-species identification are shipped to a central laboratory, where trained analysts view the cross-sections under optical magnifiers or microscopes and match their observations to standardized references (e.g., the IAWA list [1]). This manual workflow is time-consuming, labour-intensive, and inherently subjective because accuracy depends on the expert's experience.

In recent years, numerous research efforts have explored automated or semi-automated machine learning approaches for wood species analysis, often reporting high accuracy on benchmark datasets [2], [3], [4]. For example, Pratondo *et al.* [4] studied and compared various machine learning techniques for classifying three similar wood types—jeungjing, puspa, and suren—using k-nearest neighbors, support vector machines, decision trees, random forests, and Inception-v3. Their experiments showed that the random forest classifier achieved the highest accuracy at 90.67%.

Bello *et al.* [2] proposed combining Mask R-CNN with a modified residual network, attaining a notable 92% accuracy in wood species identification. Such advancements may be surprising to lay readers, as they adapt state-of-the-art deep learning methods—commonly seen in broader computer vision tasks—to the specialized context of wood identification. Nguyen-Trong [3] further introduced an approach based on convolutional neural networks to analyze wood images at multiple magnification levels, providing a comprehensive discussion of the research framework, methodology, and findings for deeper insight into its effectiveness.

Despite these promising results, many of these solutions operate as black-box image classifiers with limited interpretability regarding which anatomical features underlie the final decision. To address these limitations, our work proposes a strategy to analyze and extract important anatomical features commonly used by experts, such as pore size, pore arrangement, and pore frequency, directly from macroscopic images. Traditional wood anatomy is grounded in systematically categorizing xylem anatomical features, which often facilitates identification at the genus level. However, fully capturing every detail in the xylem structure using only visual judgment and accumulated expertise can be challenging. Consequently, several approaches now focus on quantifying anatomical markers such as cell wall thickness and cell count, in contrast to conventional inspection methods that remain somewhat subjective [5].

Therefore, this approach not only aids in identification by leveraging anatomically meaningful descriptors but also enhances explainability, as the model outputs explicit morphological attributes that correlate with standard wood anatomy references.

The main contributions of this study are as follows:

- We propose a two-stage methodology that first classifies macroscopic images into wood versus non-wood categories, and then segments Vessel-Centric Regions (VCRs), defined as pore lumina along with immediately adjacent perivascular parenchyma, if present.
- We employ DenseNet121-based feature extraction for robust wood image classification and utilize Mask R-CNN for precise segmentation of anatomical features, enabling detailed quantitative analysis.
- We provide an end-to-end implementation integrating mobile-based image classification, remote anatomical segmentation, and visualization, demonstrating practical utility for both laboratory and field applications.

The remainder of this study is organized as follows: Section II reviews recent literature on automated wood identification. Section III details our proposed two-stage framework, including data preprocessing and anatomical feature extraction.

Section IV presents experimental setups and quantitative evaluations. In Section V, we demonstrate an end-to-end application that integrates the proposed method into a real-world workflow. Finally, Section VI concludes the study and discusses future research directions.

II. Related Works

Recent research in wood species analysis has explored various approaches using machine learning and computer vision. Regarding methodologies, two main categories can be distinguished: i) traditional machine learning, including k-NN [6], SVM [7], and ANN [8], which often require handcrafted features, and ii) deep learning methods, such as CNN-based architectures [3] (VGG [9], ResNet [10], SqueezeNet [11], DenseNet [12], etc.), which automatically extract features from image data. Studies have utilized both microscopic and macroscopic images, with a trend towards more accessible macroscopic imaging techniques [3], [13].

Many recent works favor deep learning because of its higher performance when sufficient data is available. Silva et al. [14] demonstrated high success rates (88 to 90%) in classifying 77 Central-African timber species using microscopic images and texture-based features. Advancements in deep learning have shown promise, with He et al. [15] achieving 100% accuracy on a dataset of 8 wood species using an ensemble of convolutional neural networks. Herrera-Poyatos et al. [13] proposed a patch-based classification method for high-resolution macroscopic images, addressing challenges in capturing fine-grained patterns. Nguyen-Trong [3] evaluated various CNN architectures across different magnification levels, finding DenseNet121 to perform best on 20x magnified images.

Despite these advances, most studies primarily focus on classification tasks (i.e., assigning a species label to an image). The anatomical analysis of wood, particularly the segmentation and measurement of pores (e.g., solitary pores, pore multiples, pore clusters), parenchyma arrangements, or other microfeatures, remains relatively under, explored at the macroscopic scale. However, these features are crucial for a detailed understanding of wood anatomy and can bolster both automated classification accuracy and the broader morphological characterization of new wood samples. For example, identifying whether pores are solitary or clustered (and quantifying pore size, frequency, and distribution) directly contributes to more accurate species identification and deeper insight into the wood's anatomical structure.

In literature, anatomical analysis has been widely applied across medical imaging to enhance classification accuracy and provide deeper insights into structural characteristics. This approach is crucial for various applications, including surgical planning, disease progression tracking, and treatment planning [16]. Recent advancements in image segmentation techniques have been employed to define anatomical structures in CT images, particularly for COVID-19 diagnosis [17]. Image analysis techniques are essential for improving the quality of medical images and extracting features that lead to better classification results and more accurate diagnoses [18]. Deep learning methods, such as CNNs, have shown remarkable performance in anatomy-specific classification of medical images. By training on large datasets of CT scans and employing data augmentation techniques, CNNs can achieve high accuracy in classifying anatomical structures, with error rates as low as 5.9% [19].

Techniques like CNNs have demonstrated remarkable success in automatically delineating complex anatomical features from medical images, thereby improving both the efficiency and accuracy of medical assessments [20].

Despite the proven efficacy of anatomical analysis in fields like medical imaging, its application within the domain of wood identification remains relatively underexplored, particularly at the macroscopic level. Traditional wood identification methods often rely on manual examination of features such as pore arrangement, parenchyma patterns, and other microstructures. However, these manual approaches are time-consuming and subject to human error. While some studies have employed computer vision techniques for wood species classification, they predominantly focus on assigning species labels without delving into the segmentation and quantitative analysis of anatomical features [21].

Integrating anatomical analysis into wood identification holds significant potential. By employing advanced image segmentation techniques to automatically identify and measure key features, such as distinguishing between solitary pores and pore clusters, and quantifying their size, frequency, and distribution, researchers can achieve more accurate species identification and gain deeper insights into wood anatomy. This approach mirrors the successes seen in medical imaging and suggests a promising avenue for enhancing wood identification processes through the application of sophisticated computer vision methodologies.

Building upon this research context, our work aims to develop a two-stage approach for automated anatomical analysis of wood cross-sections using macroscopic images. First, a classifier (trained with a DenseNet121 feature extractor and an GMM) determines whether an input image indeed represents a magnified wood cross-section. Next, images confirmed to be wood cross-sections proceed to a detection and segmentation model to identify and segment pores, followed by computing various anatomical descriptors (i.e., average pore size, pore distribution, solitary versus multiple pore frequency). The overall objective is to integrate these methods into a system deployable on resource-constrained devices, such as smartphones, thereby facilitating affordable, efficient, and accurate wood anatomical analysis in both laboratory and field settings.

III. Materials and Methods

In this section, we present the materials and methods underlying our automated wood anatomical analysis. First, we describe our approach for classifying macroscopic images into wood versus non-wood categories. Next, we outline the segmentation of Vessel-Centric Regions (VCRs) using Mask R-CNN. Finally, we detail the extraction of anatomical descriptors, including VCR size, frequency, and spatial arrangement, essential for characterizing different wood species.

A. Wood Macroscopic Image Classification

Binary classification is arguably the most widely documented branch of supervised machine learning, especially in

domains where both positive and negative classes are naturally bounded by the same medium (e.g., "ham versus spam" in email). However, complications emerge when applying this framework to image classification tasks.

Consider the challenge of building a model to determine whether an image shows a magnified wood cross-section. We may have a substantial collection of wood cross-section images for the "wood" class, but "not wood" examples can be drawn from virtually any image category—making it nearly infinite in scope. This is unlike email filtering, where even "spam" emails are still emails, sharing a common format and feature space with "ham".

There are multiple ways to tackle this. One approach is one-class classification, in which a model is trained primarily (or exclusively) on positive-class data—in this case, images of wood cross-sections. When such a model encounters a novel input, it determines whether the input sufficiently resembles the positive class ("wood") or should be flagged as an outlier ("not wood"). Techniques such as autoencoders, One-Class SVM, or Isolation Forest can be used in this context.

However, if we do have access to certain negative examples—i.e., images known to contain no wood cross-sections—then more traditional binary classification methods can be applied. A common strategy is to perform feature extraction via a pretrained CNN (e.g., DenseNet121), reduce the dimensionality (e.g., with PCA), and then train a conventional machine learning model (e.g., SVM, Gaussian Mixture with Isotonic Regression) on these features. This latter approach benefits from explicitly learning to discriminate between "wood" and "not wood", but requires a representative set of negative examples.

In our work, we adopt transfer learning for feature extraction, followed by testing multiple classification schemes. This allows us to balance performance needs with the availability of negative data. By leveraging proven CNN architectures trained on vast image datasets (e.g., ImageNet), we can efficiently generate high-level embeddings that are then used by simpler classifiers. This technique proves especially practical in real-world scenarios, where the negative class (all images that are "not wood") is both diverse and difficult to comprehensively represent.

1) Feature extraction with DenseNet: To handle this classification task, we adopt a transfer learning approach that relies on DenseNet121, a well-known convolutional neural network pretrained on large-scale image datasets (e.g., ImageNet). Specifically, we remove the original classification layer (i.e., the fully connected top) and leverage the network's final feature maps (global average pooled) as a 512-dimensional feature vector for each input image.

All images are first resized to 224×224 and standardized (e.g., pixel values divided by 255) to match the input layer requirements of DenseNet121. Next, the pretrained model processes each image to produce the condensed feature representation. This step provides a robust and discriminative embedding of the original imagery without the need to train a full CNN from scratch.

2) Dimensionality reduction and scaling: Following feature extraction, we scale and apply Principal Component Analysis to reduce dimensionality. Dimensionality reduction serves two main purposes:

- It helps mitigate noise or redundant information in the extracted feature space.
- It substantially decreases training time for subsequent classification models, as high-dimensional feature vectors are compressed into a more manageable number of principal components.

3) Classification approaches: Once the feature vectors are extracted and reduced via PCA, we evaluate several classification methods, including Gaussian Mixture Models (GMM) with Isotonic Regression, One-Class SVM, and Isolation Forest.

GMM with Isotonic Regression can leverage a small validation set containing both positive (wood) and negative (non-wood) examples. By modeling the distributions of each class, the GMM can estimate the likelihood that a new image belongs to the learned "wood" distribution, and Isotonic Regression refines the decision threshold for improved calibration.

In scenarios with little or no labeled negative data, one-class classifiers such as One-Class SVM and Isolation Forest are valuable. Trained primarily (or exclusively) on images known to be wood, these models learn a boundary around the "wood" manifold in the feature space. At inference time, any sample that falls outside this boundary is considered "not wood".

One-Class SVM constructs a decision function tailored to capture the dense region of positive samples in the high-dimensional space. Isolation Forest isolates anomalies by randomly selecting a feature and splitting values between them, effectively "isolating" outliers—i.e., images that do not resemble wood. These one-class methods excel when assembling a truly representative negative dataset is impractical. However, whenever domain knowledge or partial negative examples are available, combining them with positive samples for a more traditional binary classification (e.g., SVM, Random Forest, or GMM with a mixed training set) can boost performance and interpretability.

B. Wood-Pore (Vessel-Centric Region) Segmentation

In diffuse- and ring-porous hardwoods the most conspicuous anatomical element is the vessel (the pore lumen). Depending on species, the lumen may be surrounded by a narrow or wide sheath of perivascular parenchyma; in other cases the lumen is embedded directly in fibre tissue with little or no parenchyma contact. Because both situations occur within the same slide, a consistent segmentation target is essential. Throughout this study, instead of segmenting pore lumina alone—which is challenging due to the close adjacency of lumina and perivascular tissues in many species—we predict vessel-centric regions (VCRs), defined as a pore lumen together with any immediately adjacent perivascular parenchyma, if present.

If two lumina are in contact or share a common PP envelope the region is labelled as a double-VCR; otherwise it is considered a single-VCR. This strategy—which we abbreviate P+PP—allows one annotation protocol to cover solitary pores,

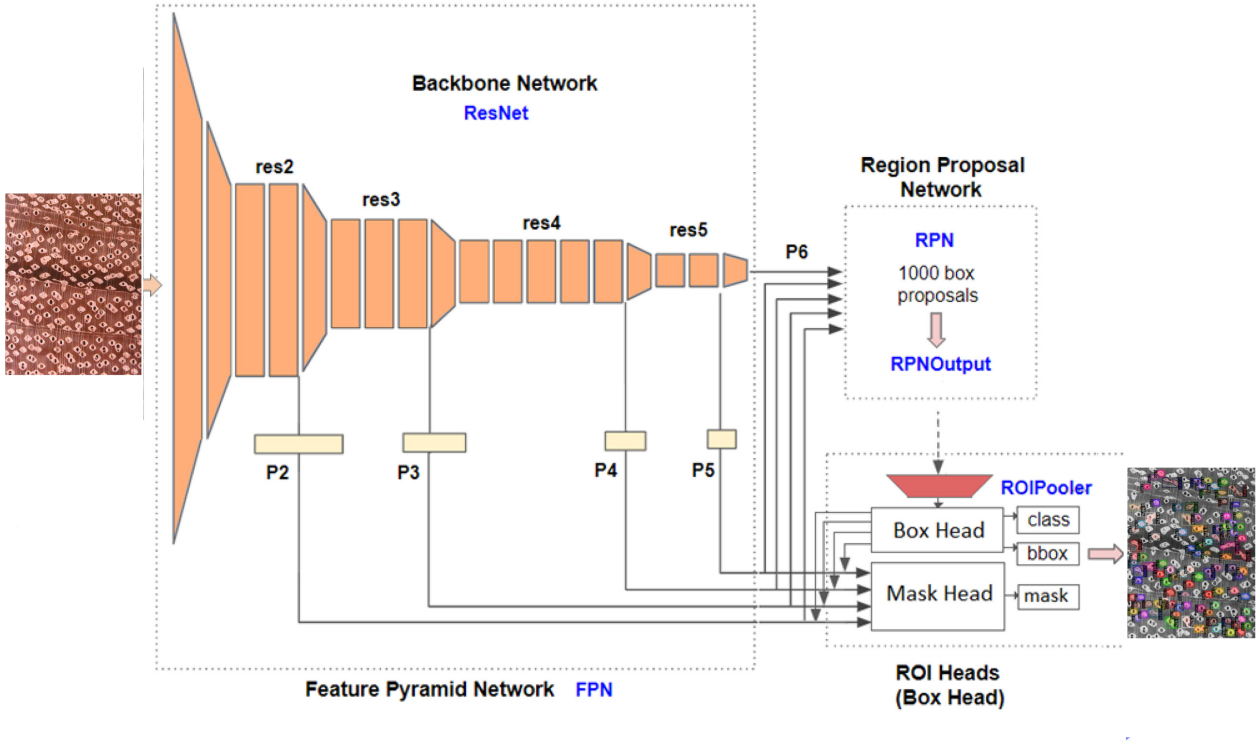


Fig. 1. Mask R-CNN architecture.

radial multiples, and pores embedded in aliform or confluent parenchyma without having to redraw masks when the surrounding tissue changes.

To segment these VCRs in 50× macroscopic images we employ Mask R-CNN [22], which couples a region-proposal network with a fully convolutional mask head and thus yields, in a single forward pass, bounding boxes, class labels (single vs. double) and pixel-accurate masks for every VCR detected.

Fig. 1 presents the architecture of the Mask R-CNN using in this study. The model employs a ResNet-50 backbone, which extract hierarchical features from an input RGB image at multiple scales. These features are processed by a Feature Pyramid Network, which generates multi-scale feature maps. This multi-scale representation captures anatomical details of varying sizes, such as small and large pores, by combining high-resolution features from early stages (e.g., P2 with H/4 × W/4) with semantically rich features from deeper stages (e.g., P6 with H/64 × W/64).

The Region Proposal Network leverages these FPN feature maps to generate candidate regions (box proposals) that may contain pores. Using an anchor-based approach, the RPN applies convolutional operations to propose regions of interest (RoIs). These proposals are then aligned with the feature maps using the RoI Pooler, which employs RoI Align—a precise pooling mechanism that preserves spatial accuracy by avoiding quantization errors.

The aligned features are fed into the RoI Heads, which consist of two components:

- Box head: Performs object classification (e.g., identifying a region as a pore) and bounding box regression to refine the coordinates of each detected object.
- Mask head: Generates binary instance masks for each pore, delineating their exact boundaries through convolutional layers.

This dual-head approach enables the model to detect, classify, and segment individual pores in a single pass. The final output is a segmented image where each pore is identified with a bounding box and a corresponding mask, facilitating detailed analysis.

Table I provides a detailed summary of the Mask R-CNN architecture employed in this study.

TABLE I. SUMMARY OF MASK R-CNN ARCHITECTURE

Component	Description
Backbone	ResNet-50 with stages res1 to res4, extracting hierarchical features at multiple scales
Feature Aggregation	Feature Pyramid Network (FPN) generating multi-scale feature maps P2 to P6
Region Proposal Network	Generates candidate regions (box proposals) from FPN feature maps using anchor-based method
RoI Align	Precise pooling with RoI Align, aligning region proposals to feature maps
Classification Head	Box Head with fully connected layers for object classification and bounding box regression
Mask Head	Convolutional layers generating binary instance masks for each detected pore

C. Wood Anatomical Feature Analysis

After segmenting VCRs from macroscopic wood cross-section images, we compute a series of anatomical descriptors to characterize and differentiate wood species. These descriptors—such as size, frequency, and spatial arrangement—are

well-established in wood anatomy and provide key diagnostic information. Because our VCR masks may include either the pore lumen alone or the lumen together with immediately adjacent perivascular parenchyma (when visually inseparable). All subsequent analysis is performed over these vessel-centric regions rather than the pore lumen alone.

1) VCR classification and counting: Each segmented VCR is categorized based on proximity and grouping patterns to reflect anatomical arrangements: solitary, multiple, or clustered. This classification follows adjacency checks, where two or more VCRs are considered grouped if they share a boundary or are positioned in close radial or tangential proximity.

a) Solitary VCR: A vessel-centric region that is spatially isolated.

b) Multiple/Clustered VCRs: Two or more VCRs appearing together, either in radial multiples or compact clusters.

Counting each category provides a basic anatomical fingerprint. For example, the predominance of solitary vessels versus radial multiples can suggest different taxonomic characteristics.

2) VCR Size measurement: As images are acquired at a fixed magnification (50×), we apply a known pixel-to-micrometer conversion factor to translate geometric measurements into physical units.

From each segmented VCR mask, we compute:

- Area: Total pixels within the VCR mask, converted into square micrometers.
- Equivalent Circular Diameter (ECD):

$$d = \sqrt{\frac{4A}{\pi}} \tag{1}$$

  where, A is the area of the VCR mask.

- Optional Shape Descriptors: Including elliptical fits (major or minor axis lengths) when more detailed morphology is needed.

These measurements provide a robust morphological overview of the segmented regions. By using VCRs—especially in diffuse-porous species where parenchyma halos often co-occur with vessels—we ensure more consistent and anatomically meaningful descriptors, even when vessel and parenchyma are difficult to distinguish in 50× imagery.



Fig. 2. Example images used in the classification task. The first row shows wood cross-section images at various magnifications. The second row contains non-wood images collected from general sources such as magnified surfaces of flowers, fruits.

IV. Experiments

A. Datasets

We employed two separate datasets for wood versus non-wood classification and segmentation tasks. For the wood/non-wood classification task, as illustrated in Fig. 2, we created two distinct subsets:

1) Wood dataset: This dataset aggregates macroscopic wood images collected from five distinct sources detailed in [3], including: i) VN_26, our self-collected dataset, ii) WRD_21, a Southeast Asian wood dataset published by Sun [23], iii) BFS_46, the Brazilian flora species dataset [24], iv) BD_11, a Brazilian dataset from 2021 [25], and v) PCA_11,

featuring wood species from the Pacific and Colombian Amazon regions [26]. In total, this dataset contains 20,000 images captured at magnifications ranging from 10x to 50x.

2) Non-Wood dataset: Approximately 9,000 images were gathered from online sources, depicting various subjects such as vehicles, landscapes, humans, and magnified images of non-wood objects.

The combined dataset was split into three subsets: 50% images for training, 20% for validation, and 30% for testing. Notably, we allocated a larger portion of the data to the test set to allow for more stable and representative performance evaluation, particularly given the diversity of both wood and non-wood images. This strategy was chosen to minimize variance in evaluation results and to better reflect real-world deployment conditions, where unseen data is common.

For the segmentation task, we utilized our self-collected subset (VN_26) exclusively at 50x magnification. Due to the extensive annotation required for segmentation, where the Labelme tool was employed to draw polygons around each pore, the labeling process was notably labor-intensive.

As presented, we clarify that every polygon corresponds to a VCR:

- Pore-only VCR: the lumen and its wall when no perivascular parenchyma collar is visible at the chosen magnification.
- Pore + PP (P+PP) VCR: the pore lumen together with the contiguous PP sheath if such a sheath clearly encircles the vessel(s).[1]

Hence, our masks deliberately "over-segment" to include PP whenever it is morphologically inseparable from the vessel on 50 × imagery, thereby preserving the original mask geometry without the need to re-annotate.

In total, 1,012 images were annotated, comprising approximately 34,500 single-VCR annotations and about 11,000 double-VCR annotations. Typically, each macroscopic wood image contained around 40 to 60 VCRs, highlighting the complexity and meticulous nature of the annotation work.

Given the substantial annotation effort involved, we adopted an iterative labeling strategy, progressively annotating smaller subsets of images while continuously evaluating the segmentation model's performance. Annotation was halted when model accuracy stabilized, ensuring efficient use of annotation resources without sacrificing model effectiveness. Additionally, the emphasis on single and double VCR facilitated focused improvements in segmentation accuracy, ultimately enhancing the robustness and practical usability of the resulting segmentation model.

B. Wood Macroscopic Image Classification

Because our objective is to detect if an image represents a magnified wood cross-section, we split data into training/validation/test sets at the feature level. In many real-world scenarios, it can be beneficial to have a small, curated set of "likely negatives" (e.g., images of plant bark or random textures) to help tune decision thresholds.

For each candidate model, hyperparameters (e.g., the number of PCA components, kernel parameters for SVM, and the number of estimators in Isolation Forest) are tuned based on performance metrics. In this study, we evaluate classification performance by standard metrics, accuracy, precision, recall, and F1-score, when negative examples are available. For one-class approaches, we track metrics such as the false-positive rate on non-wood samples to gauge how well the model excludes irrelevant data.

By combining DenseNet121, based feature extraction with robust classification techniques, be the binary or one-class, we ensure a flexible pipeline capable of handling the open-ended nature of "wood versus non-wood" classification. High-quality embeddings, coupled with dimensionality reduction and tailored classification algorithms, allow us to accurately filter and flag images for subsequent segmentation and anatomical feature analysis.

Fig. 3, Fig. 4, and Fig. 5 show the confusion matrices obtained from these models. In each matrix, the rows (actual labels) and columns (predicted labels) correspond to two classes: "0" (non-wood image) and "1" (wood cross-section image).

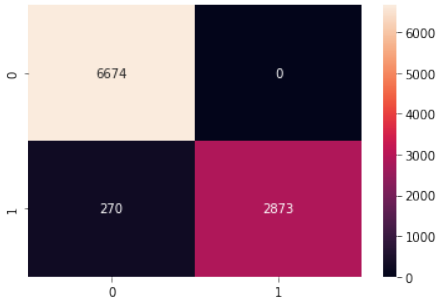


Fig. 3. Confusion matrix - OneSVM.

1) OneSVM: Although OneSVM achieves an impressive outcome of zero false positives (i.e., no non-wood images were misclassified as wood), it still has 270 false negatives. In other words, it fails to correctly identify 270 wood images as wood. This indicates a tendency toward under-classification of wood samples, which can be problematic in scenarios where missing a true wood specimen is undesirable.

2) Isolation forest: Isolation Forest exhibits a similar pattern to OneSVM, with zero false positives and a slightly higher number of false negatives (279). This small increase in false negatives yields a slight reduction in overall accuracy, but the difference between Model OneSVM and Isolation Forest performance is minimal in most practical contexts.

3) GMM: GMM stands out in that it achieves a significantly lower number of false negatives (only 2), thus correctly identifying nearly all true wood images. However, it does have a small number of false positives (1). From an overall performance standpoint, GMM demonstrates a much stronger capability to detect wood images without sacrificing non-wood accuracy by any large margin.

[1]In diffuse-porous and ring-porous taxa where parenchyma halos are diagnostic, this choice is anatomically meaningful because the pore + PP structure functions as a single macro-texture cue.

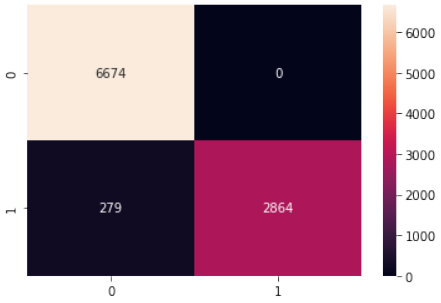


Fig. 4. Confusion matrix - IsolationForest.

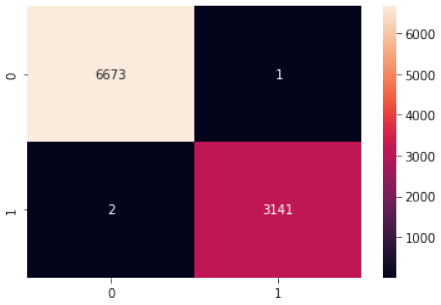


Fig. 5. Confusion matrix - GMM.

TABLE II. COMPARISON OF CLASSIFICATION METRICS FOR THE THREE MODELS ON THE TESTSET

Model	Accuracy	Precision	Recall	F_1-score
Isolation Forest	97%	98%	95.5%	96.5%
OneSVM	97%	98%	95.5%	97%
GMM	99%	99%	99%	99%

Table II compares the performance of three models. Isolation Forest and OneSVM have similar accuracy but struggle with higher false negatives, which drags down their recall. GMM, on the other hand, has almost no false negatives, boosting its recall and giving it a better F_1-score. In wood identification, missing a true wood sample (a false negative) is a bigger deal than accidentally labeling a non-wood image as wood (a false positive). So, keeping false negatives low is critical. With only 2 false negatives out of 3141 true positives, GMM stands out as the best option.

The results prove that combining transfer learning with solid classification can nail the task of separating wood from non-wood images. Isolation Forest and OneSVM play it safe (no false positives, but they miss more wood samples), while GMM strikes a great balance between recall and precision. In a two-step system, where classification feeds into pore segmentation, GMM makes sure most true wood images move forward. Looking ahead, tweaking feature extraction or anomaly detection could cut down errors even more and boost generalization.

C. Wood VCR Segmentation

Accurately segmenting individual wood pores from macroscopic images is inherently challenging due to the close adjacency of pore lumina and surrounding perivascular parenchyma. As discussed previously, we therefore adopted the concept of Vessel-Centric Regions, defined as a pore lumen together with any immediately adjacent PP tissue if present. Annotating each VCR required meticulously drawing polygons around each region using the Labelme tool, making this a particularly labor-intensive process. Each macroscopic wood image typically contained around 40 to 60 VCRs, leading to an estimated total of approximately 32,500 annotated VCRs across our 1,012-image dataset. These annotations comprised about 23,500 single-VCR and 9,000 double-VCR instances.

Given the substantial annotation effort involved, we adopted an iterative annotation strategy, incrementally annotating subsets of images and evaluating model performance continuously. Annotation was halted when the model's performance stabilized, thus efficiently balancing annotation cost and model accuracy.

Additionally, we augmented the dataset by cropping each annotated image into two smaller segments, effectively enhancing training-data diversity without additional annotation work. The numbers of annotated images and corresponding experimental results are summarized in Table III.

Two experimental scenarios were conducted to evaluate the performance of our segmentation model under different conditions, detailed as follows.

1) Scenario 1 Effect of training set size: In Scenario 1, we assessed accuracy improvements resulting from incremental increases in annotated training data. We employed the iterative annotation strategy outlined above, progressively adding labeled subsets until segmentation performance stabilized.

Six dataset sizes were evaluated, summarized in Table III, with each set split into training, validation, and testing subsets at a 70:15:15 ratio. Models were trained for 10,000 epochs, with a batch size of 512 and a learning rate of 10^{-4}.

Table III presents the accuracy achieved at various stages of this iterative process. Fig. 6 shows training progress, including accuracy and loss trends for classification and segmentation on both training and validation sets. Key observations include:

TABLE III. SCENARIO 1: SEGMENTATION ACCURACY WITH INCREASING TRAINING SET SIZES

No.	Train Images	Val Images	Test Images	Accuracy (%)
1	78	16	16	69.55
2	388	84	84	93.378
3	500	108	108	93.923
4	528	114	114	91.193
5	682	146	146	84.629
6	708	152	152	91.393

- Incremental Annotation versus Accuracy: Initially, the accuracy was low (69.55%) with only 78 annotated training images. Accuracy increased significantly to 93.378% and 93.923% as annotations reached 388 and 500 images, respectively, indicating the effectiveness of additional labeled data in early stages.

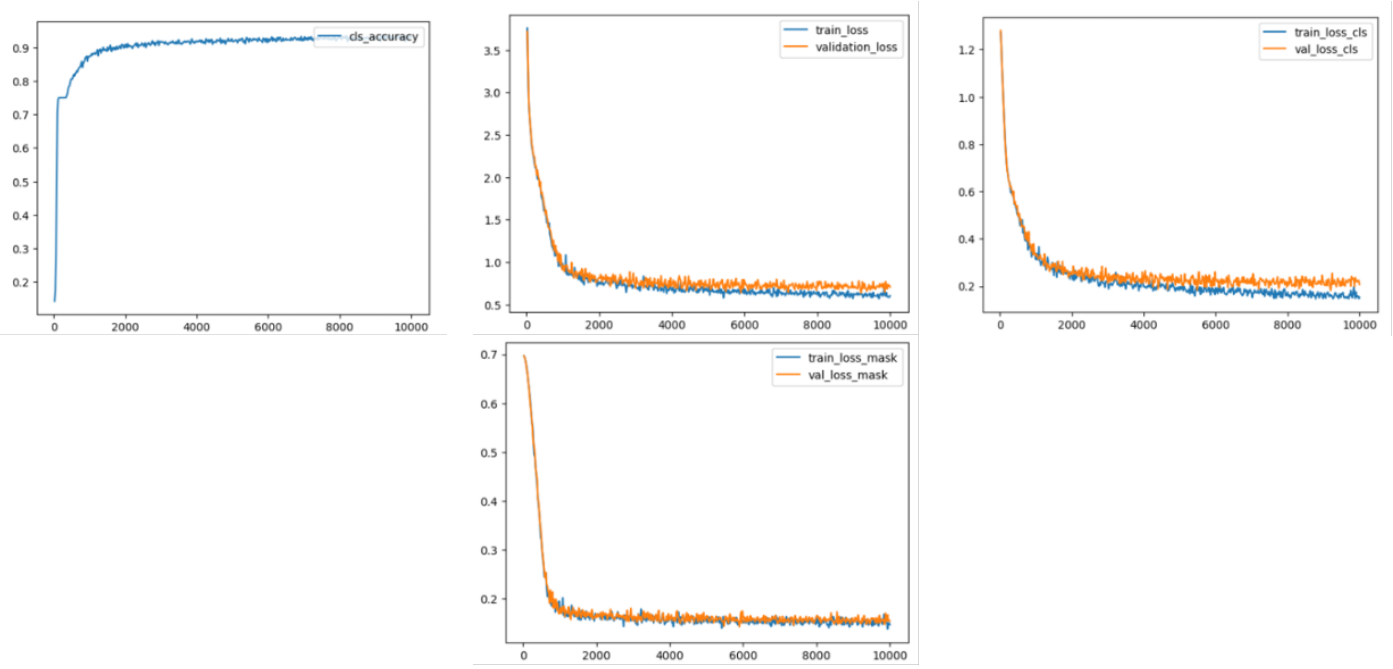


Fig. 6. Experiment 1 – Training loss and accuracy progresses with 500 training images.

- Incremental Annotation and Accuracy: Initially, accuracy was relatively low (69.55%) with just 78 annotated images. However, significant improvements were observed when annotations increased to 388 (93.378%) and 500 images (93.923%), demonstrating clear early-stage benefits of additional labeled data.
- Performance Saturation and Overfitting: Accuracy peaked at 500 annotated images. Beyond this point, accuracy fluctuated, declining to 84.629% at 682 images before improving slightly again at 708 images (91.393%). This indicates potential saturation and overfitting when adding excessively diverse annotations beyond an optimal point.
- Data Augmentation Effectiveness: The augmentation strategy of cropping annotated images successfully enhanced data diversity, mitigating accuracy declines observed at larger dataset sizes.

These results emphasize the practical effectiveness of iterative annotation and augmentation strategies for maximizing segmentation accuracy while minimizing annotation overhead. Future work may investigate adaptive annotation and augmentation strategies to further enhance model performance.

2) Scenario 2 Effect of number of epochs: In the second scenario, we investigated the impact of training duration on segmentation accuracy by varying the number of epochs (5,000; 10,000; and 15,000). This experiment employed 500 training images, 108 validation images, and 108 test images. Other training parameters remained constant (image size: 480×640, batch size: 512, learning rate: 10^{-4}). Table IV summarizes the accuracy results.

The results highlight several key insights:

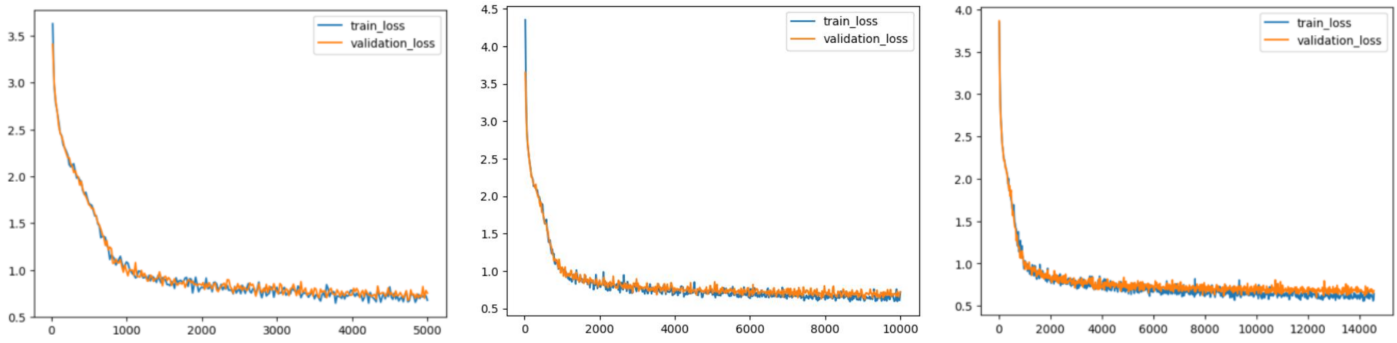
- Optimal performance at 10,000 epochs: Peak accuracy (93.923%) occurred at 10,000 epochs, slightly outperforming 5,000 epochs (93.677%). Further training to 15,000 epochs led to decreased accuracy (91.877%), indicating potential overfitting.
- Loss curve behavior: Analyzing loss curves (Fig. 7) revealed a consistent downward trend in loss, with diminishing accuracy returns beyond 10,000 epochs. Continued training beyond this point tended toward overfitting noise rather than learning meaningful anatomical patterns.
- Practical training recommendations: Balancing computational efficiency and accuracy, training around 10,000 epochs offers optimal segmentation performance without unnecessary resource expenditure.

TABLE IV. SCENARIO 2: TEST ACCURACY BY NUMBER OF TRAINING EPOCHS

Epochs	Accuracy (%)
5,000	93.677
10,000	93.923
15,000	91.877

Combining the findings of Scenarios 1 and 2, we observe the following:

- Optimal annotation size: A moderate dataset size (e.g., 500 images) yields optimal performance. Larger datasets may introduce complexity, potentially reducing accuracy without careful management.
- Balanced epoch selection: Training duration of approximately 10,000 epochs best balances accuracy gains against risks of overfitting.



(a) Loss progress at 5,000 epochs (b) Loss progress at 10,000 epochs (c) Loss progress at 14,000 epochs

Fig. 7. Training loss curves of segmentation models at different epoch milestones (5k, 10k, and 15k).

- Model robustness and generalization: Despite varying training conditions, accuracy consistently ranged between 84% and 94%, demonstrating robustness. Future improvements may involve advanced augmentation, regularization techniques, or specialized segmentation architectures.

These results confirm that strategic adjustments of annotation quantity and training epochs significantly affect segmentation performance, offering practical guidance for efficient annotation and training protocols.

V. Demonstration of End-to-End Wood Anatomical Analysis

To illustrate the practical value and usability of our models, we have implemented an end-to-end system for wood anatomical analysis. This system combines a lightweight Android mobile application with a remote server for advanced segmentation tasks, thereby demonstrating how classification and segmentation can be integrated into a real-world workflow.

A. System Overview

1) Mobile application - Frontend: The user-facing component is an Android application that performs on-device image classification to determine whether an input image depicts a magnified wood cross-section. Users can capture new images via the device camera or select existing images from the phone's gallery. For smartphones lacking sufficient native magnification, an external macro lens attachment can be used to acquire clear macroscopic wood images.

Upon pressing the "Analyze" button, the app uses a lightweight classifier trained as described in Section IV-B to check locally if the input represents a magnified wood cross-section. If the classification result is positive, the app forwards the image to the remote server for detailed segmentation and vessel-centric anatomical analysis.

2) Server segmentation - Backend: On the server side, our Mask R-CNN model (described in Section IV-C) receives the uploaded image and segments VCRs —each defined as a pore lumen plus its adjacent perivascular parenchyma, when PP is present. Specifically, the model segments and categorizes each identified region into:

- Single VCRs (solitary pore lumen $\pm$ PP),
- Double VCRs (two adjacent pore lumina sharing a common PP sheath or directly contacting each other).

From these segmented regions, the server computes a set of meaningful anatomical descriptors, including:

- Total number of single and double VCRs,
- Average VCR size in micrometers (derived from known magnification levels),
- Minimum and maximum sizes of VCRs,
- Inter-VCR distance metrics (minimum, maximum, and average distances),
- Summary statistics, such as spatial distribution and frequency patterns of VCRs.

B. Illustrative Results

To validate the end-to-end pipeline under realistic conditions, we tested the application using various wood species commonly encountered in Vietnam. Images were captured at magnifications of 10$\times$ and 50$\times$ under variable lighting conditions. The system consistently demonstrated:

- Reliable on-device classification, efficiently filtering out non-wood images,
- Robust server-based segmentation, clearly delineating VCRs even in complex anatomical arrangements,
- Rapid retrieval of detailed anatomical metrics, allowing users immediate insight into wood anatomy and supporting species identification.

Fig. 8 provides sample screenshots showing segmented VCR results and anatomical summaries for species such as Gõ Pachy, Cãm Hòng, and Ebiara at different magnifications. Users can view segmentation overlays directly in the app interface and access detailed textual descriptions of segmented regions.

The demonstrated architecture highlights a practical division of tasks:

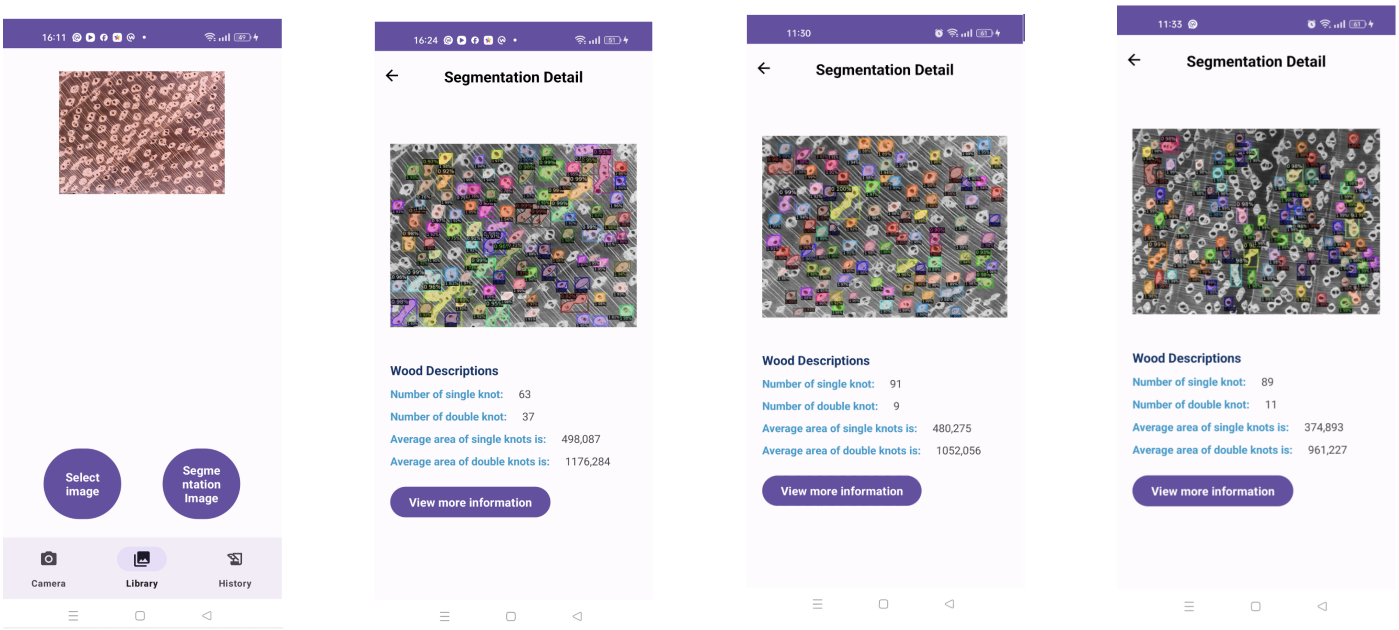


Fig. 8. Android app for end-to-end wood anatomical analysis.

- A compact classifier deployed on Android devices quickly verifies the suitability of images for further analysis,
- A computationally intensive segmentation model hosted on a remote server performs detailed VCR segmentation and returns comprehensive anatomical insights.

Overall, our end-to-end system provides a convenient, low-cost solution enabling wood analysts, researchers, and industry professionals to rapidly quantify critical anatomical features from macroscopic images. Future enhancements may include offline segmentation capabilities for fieldwork environments with limited connectivity or more advanced user interfaces for annotating and verifying newly encountered wood species. Additionally, this pipeline can be extended towards fully automated wood species identification by leveraging large language models to reason over extracted anatomical descriptors and incorporate established taxonomic knowledge.

VI. Conclusion

This study presented a two-stage automated framework for anatomical analysis of macroscopic wood cross-sectional images. Our approach combined image classification using DenseNet121-based features and a Gaussian Mixture Model classifier with segmentation of Vessel-Centric Regions (VCRs) via Mask R-CNN. Defining each VCR as a pore lumen plus any adjacent perivascular parenchyma (PP) allowed effective segmentation despite closely adjacent anatomical structures.

Experimental results confirmed that iterative annotation and data augmentation efficiently balanced annotation effort and segmentation accuracy, achieving optimal results (93.923%) with approximately 500 annotated images trained over 10,000 epochs. An end-to-end demonstration highlighted the practical usability of our system, combining mobile-based classification and remote segmentation. Future directions include offline segmentation for field use and automated species identification leveraging large language models.

References

[1] E. Wheeler, P. Baas, and P. Gasson, "Iawa list of microcopie features for hardwood identification," *IAWA journal / International Association of Wood Anatomists*, vol. 10, p. 219–332, 01 1989.

[2] R.-W. Bello, "Wood species identification using mask rcnn-residual network," *Pro Ligno*, vol. 19, no. 1, 2023.

[3] K. Nguyen-Trong, "Evaluation of wood species identification using cnn-based networks at different magnification levels," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140487

[4] A. Pratondo and A. Novianty, "Comparison of wood classification using machine learning," in *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)*. IEEE, 2022, pp. 308–312.

[5] K. Kobayashi, T. Kegasa, S.-W. Hwang, and J. Sugiyama, "Anatomical features of fagaceae wood statistically extracted by computer vision approaches: Some relationships with evolution," *PLOS ONE*, vol. 14, no. 8, pp. 1–13, 08 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0220762

[6] M. Khalid, E. L. Y. Lee, R. Yusof, and M. Nadaraj, "Design of an intelligent wood species recognition system," *International Journal of Simulation: Systems, Science & Technology*, vol. 9, no. 3, pp. 9–18, 2008.

[7] H.-J. Wang, G.-Q. Zhang, and H.-N. Qi, "Wood recognition using image texture features," *PLoS ONE*, vol. 8, no. 10, p. e76101, 2013.

[8] K. Kobayashi, M. Akada, T. Torigoe, S. Imazu, and J. Sugiyama, "Automated recognition of wood used in traditional japanese sculptures by texture analysis of their low-resolution computed tomography data," *Journal of Wood Science*, vol. 61, no. 6, pp. 630–640, 2015.

[9] I. Kirbas and A. Cifci, "An effective and fast solution for classification of wood species: a deep transfer learning approach," *Ecological Informatics*, vol. 69, p. 101633, 2022.

[10] F. Wu, R. Gazo, E. Haviarova, and B. Benes, "Wood identification based on longitudinal section images by using deep learning," *Wood Science and Technology*, vol. 55, pp. 553–563, 2021.

[11] X. J. Tang, Y. H. Tay, N. A. Siam, and S. C. Lim, "Mywood-id: Automated macroscopic wood identification system using smartphone and macro-lens," in *Proceedings of the 2018 International Conference on Computational Intelligence and Intelligent Systems*, 2018, pp. 37–43.

[12] Y. Sun, Q. Lin, X. He, Y. Zhao, F. Dai, J. Qiu, and Y. Cao, "Wood species recognition with small data: a deep learning approach," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1451–1460, 2021.

[13] D. Herrera-Poyatos, A. H. Poyatos, R. M. Soldado, P. De Palacios, L. G. Esteban, A. G. Iruela, F. G. Fernández, and F. Herrera, "Deep learning methodology for the identification of wood species using high-resolution macroscopic images," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[14] N. Rosa da Silva, M. De Ridder, J. M. Baetens, J. Van den Bulcke, M. Rousseau, O. Martinez Bruno, H. Beeckman, J. Van Acker, and B. De Baets, "Automated classification of wood transverse cross-section micro-imagery from 77 commercial central-african timber species," *Annals of forest science*, vol. 74, pp. 1–14, 2017.

[15] T. He, S. Mu, H. Zhou, and J. Hu, "Wood species identification based on an ensemble of deep convolution neural networks." *Wood Research*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233792800

[16] M. Hu, J. Zhang, L. Matkovic, T. Liu, and X. Yang, "Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 2, p. e13898, 2023.

[17] S. Zimeras and M. Diomidous, "2d anatomical structure for covid-19 medical images," *Studies in health technology and informatics*, vol. 295, pp. 542–544, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250175683

[18] M. Berahim, N. A. Samsudin, and S. S. Nathan, "A review: Image analysis techniques to improve labeling accuracy of medical image classification," in *International Conference on Soft Computing and Data Mining*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:2095539

[19] H. R. Roth, C. T. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. M. Summers, "Anatomy-specific classification of medical images using deep convolutional nets," *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 101–104, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2484308

[20] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, p. 60–88, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2017.07.005

[21] S.-W. Hwang and J. Sugiyama, "Computer vision-based wood identification and its expansion and contribution potentials in wood science: A review," *Plant Methods*, vol. 17, no. 1, pp. 1–21, 2021.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870

[23] Y. Sun, "Wood Recognition," https://github.com/sunyongke/woodRecognition, 2020, [Online; accessed 02-March-2023].

[24] D. V. Souza, J. X. Santos, H. C. Vieira, T. L. Naide, S. Nisgoski, and L. E. S. Oliveira, "An automatic recognition system of brazilian flora species based on textural features of macroscopic images of wood," *Wood Science and Technology*, vol. 54, no. 4, pp. 1065–1090, 2020.

[25] A. R. de Geus, A. R. Backes, A. B. Gontijo, G. H. Albuquerque, and J. R. Souza, "Amazon wood species classification: a comparison between deep learning and pre-designed features," *Wood Science and Technology*, vol. 55, pp. 857–872, 2021.

[26] D. A. Cano Saenz, C. F. Ordoñez Urbano, H. R. Gaitan Mesa, and R. Vargas-Cañas, "Tropical wood species recognition: A dataset of macroscopic images," *Data*, vol. 7, no. 8, p. 111, 2022.

Air Quality Prediction Based on VMD-CNN-BiLSTM-Attention

Huang Xinxin[1], Mohd Suffian Sulaiman[2]*, Marshima Mohd Rosli[3]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia[1,2,3]
Cardiovascular Advancement and Research Excellence Institute (CARE Institute),
Universiti Teknologi MARA, Sungai Buloh Campus[3]

***Abstract*—With the advancement of industrialization, air pollution has emerged as a critical global health and environmental concern. This study presents an air quality prediction model based on variational mode decomposition, a convolutional neural network, bidirectional long short-term memory, and an attention mechanism. The variational mode decomposition method is employed to decompose the Air Quality Index sequence, capturing different local characteristics of the original data. A hybrid model is constructed by integrating the convolutional neural network for feature extraction, the bidirectional long short-term memory for temporal pattern recognition, and the attention mechanism for focusing on significant data features. The model is optimized using the Grey Wolf Optimizer for hyperparameter tuning, thereby enhancing prediction accuracy. The proposed model is evaluated using air quality data from Changsha, China, covering the years 2015 to 2023. The results demonstrate that our model outperforms several other models in terms of mean absolute error, mean squared error, root mean squared error, and R-squared. This study provides a robust approach to air quality prediction, offering valuable insights for residents and policymakers.**

***Keywords*—Air quality prediction; variational mode decomposition; convolutional neural network; bidirectional long short-term memory; hyperparameter optimization; air quality index**

I. Introduction

With the development of global industrialization, material living standards have continued to improve; however, air pollution has become increasingly severe [1], [2]. Since the mid-19th century, when London experienced severe smog pollution, people have gradually become aware of the dangers posed by air pollution. In the 20th century, smog events in cities such as Los Angeles and London further heightened public concern and prompted more extensive investigation into air pollution issues.

Current research on air pollution primarily focuses on outdoor atmospheric pollution, which includes major components such as ozone, suspended particulate matter, and nitrogen oxides. These pollutants originate from various complex sources, including industrial emissions, vehicle exhaust, and coal-fired power generation.

The Air Quality Index (AQI) is a crucial metric used to assess air quality. It consolidates the concentrations of several routinely monitored pollutants—including PM2.5, PM10, nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), into a single, conceptual index value. This index categorizes air pollution levels and air quality conditions, allowing the extent of pollution to be reflected in a scientific and intuitive manner.

According to data from the World Health Organization (WHO), air pollution accounted for 7.6% of total global deaths in 2015 [3], [4], [5]. Modern epidemiological studies have shown that air pollution can severely damage the human immune system and may cause irreversible long-term harm to future generations. In addition, air pollution results in significant social and economic losses. Experts estimate that, based on the value of life, 3.25% of deaths in China in 2017 were attributable to air pollution, resulting in an economic loss equivalent to approximately 1.53% of the country's GDP that year [6], [7].

Most countries now have air quality monitoring systems in place; however, these systems often experience delays in providing air quality data and cannot offer reliable early warnings for governments or the public. Additionally, due to the non-constant and non-linear nature of air pollution concentrations, predicting air quality remains a significant challenge.

In recent years, many researchers have proposed various air quality prediction models. Existing methods can generally be divided into two categories: classic dispersion models [8], [9] and data-driven models [10], [11], [12]. Classic dispersion models are based on physical principles, taking into account factors such as wind speed, wind direction, atmospheric stability, and emission source characteristics to simulate and estimate the dispersion and concentration distribution of pollutants in the atmosphere.

In contrast, data-driven models represent a modern approach to air quality prediction. These models rely primarily on large volumes of historical data and statistical learning techniques to forecast future air quality. Unlike traditional physics-based models, data-driven models emphasize learning and extracting patterns from past data. Common types include statistical methods, machine learning algorithms, and deep learning models.

Applying statistics to air quality prediction research mainly includes autoregressive (AR) models, autoregressive integrated moving average (ARIMA) models, grey models, and multiple kinear regression(MLR) models. In 2006, Pagowski et al. used a dynamic linear regression model to predict ozone concentrations within 24 hours, achieving reduced bias under conditions with fewer monitor measurements [13]. Pai et al. used the Grey prediction model GM(1,1) to predict ozone

*Corresponding authors.

concentrations, and the results showed that GM(1,1) had a mean absolute percentage error (MAPE) of 19% based on a small sample size, indicating average prediction accuracy. The advantage of linear regression models is their fast modeling speed and good interpretability, but they handle non-linear issues in air quality datasets poorly and cannot fit non-linear data well [14]. Donnelly et al. combined a non-parametric kernel regression description of NO2 concentrations varying with wind speed and direction to obtain seasonal and daily variation factors, achieving good prediction results on the basis of the multiple linear regression model. Niu et al. established an autoregressive moving average (ARMA) model for air quality prediction in Chengdu [15], and in the same year, Zhang et al. used wavelet decomposition methods to improve the ARMA prediction model, decomposing and reconstructing air pollutant concentrations. The results proved that wavelet multi-scale decomposition could significantly improve the prediction accuracy of the ARMA model [16].

Machine learning models may learn patterns from historical data and use these patterns to predict future air quality. It also optimize prediction performance by adjusting model parameters and structures. Liang et al. used an 11 years dataset collected by Taiwan's Environmental Protection Administration (EPA) and then employed machine learning methods including adaptive boosting (AdaBoost), artificial neural networks(ANN), random forest (RF), stacking ensemble, and support vector machines(SVM) for individual predictions. The experimental results indicated that the stacking ensemble performed the best in terms of prediction accuracy [17]. Castelli et al. used support vector regression(SVR) to predict pollutant and particulate matter levels for the prediction of the AQI, and in their experimental scheme, SVR with a radial basis function (RBF) achieved the best prediction results [18]. Liu et al. constructed regression models using SVR and random forest regression (RFR) to predict the AQI in Beijing and the nitrogen oxides(NOx) concentration in an Italian city. The experimental results showed that the SVR-based model performed better in AQI prediction, while the RFR-based model performed better in NOx concentration prediction [19].

Deep learning models, in contrast to machine learning models that typically require less data for training, rely on large amounts of data and complex model structures for training. However, deep learning models can better capture the complex nonlinear relationships in the learning data and can integrate multiple data sources to improve the accuracy and comprehensiveness of predictions [20], [21]. Baniasadi et al. proposed a novel binary chimp optimization algorithm(BChOA) and combined it with long short-term memory(LSTM) networks.The BChOA algorithm fine-tunes the optimization parameters of the LSTM model to achieve more accurate and reliable air pollution forecasting. Experimental results demonstrated that the BChOA-LSTM model achieved the highest accuracy of 96.41% [22]. Bun Theang Ong et al. proposed deep recurrent neural networks(DRNN), which use an autoencoder enhancement pre-training method specifically designed for time series prediction to achieve PM2.5 concentration prediction. The experimental results showed that the DRNN model could handle air quality prediction problems well [23]. Wang et al. noticed that existing prediction models did not accurately capture the regularities between haze concentration and real-world influencing factors and proposed a two-layer model prediction algorithm based on long short-term memory neural networks and threshold regression units (LSTM and GRU). It is an improvement and enhancement of the existing forecast method of LSTM [24]. Wu et al., a novel air quality forecasting model based on the improved sparrow search algorithm (ISSA) and LSTM networks has been proposed,integrating the mRMR-RF feature selection method to enhance the model's predictive accuracy. By combining mRMR and RF,the model effectively selects variables that impact the AQI. The ISSA algorithm is utilized for the hyperparameter optimization of the LSTM. Experimental results indicate that this approach significantly improves the accuracy and efficiency of the forecasts [25].

A. Our Contribution

In this study, the VMD method is used to decompose the input AQI sequence. The resulting IMF components represent different local characteristics of the original data. Predicting each decomposed IMF component in parallel reduces the time required to identify signal features, thereby improving both the training efficiency of the algorithm and the accuracy of the predictions.

For the construction of the prediction model, we develop a hybrid model for air quality forecasting. Recognizing that LSTM networks do not fully resolve the vanishing gradient problem, we introduce CNN into the model to more effectively extract features from the air quality data. However, we also acknowledge that this integration increases the model's dependence on large volumes of training data and raises the risk of convergence to local optima. To address these challenges, we explore improvements to the internal structure of the neural network and consider incorporating additional, more effective neural architectures for feature extraction.

This study combines the CNN's capability to extract multi-scale temporal features with the strengths of BiLSTM, which processes sequences through both forward and backward layers of improved LSTM units. After feature extraction by the CNN, the BiLSTM computes the temporal dependencies, and the Attention mechanism is employed to calculate the attention weights of data features at different time steps with respect to the predicted value. This reveals the correlation between the time series data and the prediction target. By integrating these specialized neural networks, we construct an enhanced CNN-BiLSTM-Attention model capable of forecasting air quality for future periods with high accuracy.

Furthermore, to avoid arbitrary hyperparameter selection in the CNN-BiLSTM-Attention model, we apply the Grey Wolf Optimizer (GWO) to systematically search for optimal hyperparameters. This contribution demonstrates how modern machine learning techniques can be effectively combined and how model performance can be significantly improved by optimizing the training process through algorithmic tuning.

Experimental results confirm that, compared to other models with similar structures used for air quality prediction, the proposed model achieves higher accuracy.

II. Model Introduction

A. Variational Mode Decomposition (VMD)

Variational Mode Decomposition (VMD) is an adaptive signal processing method proposed by Konstantin Dragomiretskiy

and Zoubin Ghahramani in 2014. This method has the advantage of being able to determine the number of modal decompositions. Its adaptability is reflected in the ability to determine the number of modal decompositions for a given sequence based on actual conditions. In the subsequent search and solution process, it can adaptively match the best central frequency and finite bandwidth for each mode, and can effectively separate the Intrinsic mode functions (IMF) and divide the signal into frequency domains, thereby obtaining the effective decomposition components of the given signal and ultimately obtaining the optimal solution to the variational problem. It can avoid aliasing by controlling the bandwidth and can effectively address the mode mixing defect present in the Empirical Mode Decomposition (EMD) method [26], [27]. The basic principle includes the following steps:

The original signal is assumed to be decomposed into K components, ensuring that the decomposed sequences are modal components with finite bandwidth centered at specific frequencies. The variational problem can be described as finding K modal functions $u_k(t)(k = 1, 2, ..., K)$, such that the sum of the estimated bandwidths of each mode is minimized, subject to the constraint that the sum of all modes equals the original signal. The modal functions are defined by the following Eq. (1):

$$u_k(t) = A_k(t)cos(\phi_k(t)) \tag{1}$$

Use the Hilbert transform to calculate the analytic signal corresponding to each modal function $u_k(t)$, and obtain the one-sided spectrum, as shown in Eq. (2):

$$(\delta(t) + \frac{j}{k}) * u_k(t) \tag{2}$$

where, $\delta(t)$ is the Dirac function, based on the estimated central frequency e^{-jw_kt} of the mixed analytic signals of each mode, modulate the frequency of each mode to the corresponding base frequency band, see Eq. (3):

$$[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{-jw_kt} \tag{3}$$

Calculates the square norm L^2 of the gradient of the aforementioned signals, estimate the sum of the bandwidths of the modal signals, and the constrained variational problem is represented as Eq. (4):

$$\begin{cases} \min\limits_{u_k,\omega_k} \{\sum_k \|\partial_t[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{-jw_kt}\|_2^2\} \\ \text{s.t.} \sum_{k=1}^K u_k = f \end{cases} \tag{4}$$

In Eq. (4), $u_k = u_1, u_2, ..., u_k$ represents the modal functions with finite bandwidth, $\omega_k = \omega_1, \omega_2, ..., \omega_k$ denotes the central frequencies of each mode, * is the convolution operator, ∂_t signifies taking partial derivatives, and e^{-jw_kt} indicates shifting the spectrum to the base frequency band.

Solving the variational problem by introducing the Lagrange penalty operator and the penalty coefficient, transforming the constrained variational problem described in Eq. (4) into an unconstrained variational problem, resulting in the augmented Lagrange expression, as shown in Eq. (5).

$$\begin{aligned} L(u_k, \omega_k, \lambda) = \alpha \sum_k \|\partial_t[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{j w_k t}\|_2^2 + \\ \|f(t) - \sum_k u_k(t)\|_2^2 + (\lambda(t), f(t) - \sum_k u_k(t)) \end{aligned} \tag{5}$$

To solve the unconstrained problem, it is necessary to find the saddle point of Eq. (5). The alternating direction method of multipliers (ADMM) is used to iteratively update $\hat{u}_k^{n+1}$, ω_k^{n+1}, and $\hat{\lambda}_k^{n+1}$. The iterative update process is shown in Eq. (6).

$$\begin{cases} \hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \hat{\lambda}(\omega)/2}{1 + 2\alpha(\omega - \omega_k)^2} \\ \omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}(\omega)^2 d\omega|}{\int_0^\infty |\hat{u}(\omega)^2 d\omega|} \\ \hat{\lambda}_k^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau(\hat{f}(\omega) - \sum_k^{n+1} \hat{u}_k(\omega)) \end{cases} \tag{6}$$

In this context, $\hat{f}(\omega)$ and $\hat{u}_i(\omega)$ represent the Fourier transforms of the original signal and the modal components, respectively, n is the number of iterations, and τ is the fidelity coefficient.

The frequency center and bandwidth of each IMF component are continuously updated during the separation from the original signal, until the iterative stopping conditions, as shown in Eq. (7) are met.

$$\sum_{i=1}^K (\frac{\|\hat{u}_i^{n+1}(\omega) - \hat{u}_i^n(\omega)\|_2^2}{\|\hat{u}_i^n(\omega)\|_2^2}) < \epsilon \tag{7}$$

ϵ is the discrimination precision; if the convergence condition is met, then the iteration is stopped and K IMFs are obtained.

From the above reasoning, we may deduce that the VMD algorithm transforms the modal decomposition problem into a search for the optimal solution of a variational equation. It then obtains the corresponding modal components through continuous iterative updates. Compared to the EMD algorithm, the VMD algorithm can effectively suppress the endpoint effect and modal component aliasing. Additionally, the VMD algorithm reduces the non-stationarity of time series with high complexity and strong non-linearity, resulting in decomposed subsequences that contain multiple different frequency scales and are relatively stationary. Therefore, it is more suitable for non-stationary time series such as air quality.

B. Convolutional Neural Networks (CNN)

Convolutional neural networks(CNNs) are a special type of deep feedforward neural network that have been widely used in image processing, machine vision, and other fields. The most basic structure of a CNN is shown in Fig. 1, which includes convolutional layers, pooling layers, fully connected layers, and an output layer. The CNN analyzes the features of

the input image or one-dimensional data through convolution operations. The pooling layers are used to merge and extract features obtained from each convolutional layer, and finally, the fully connected layers perform dimensionality reduction and output results. This network has characteristics such as local connections and weight sharing in its structure [28], [29].

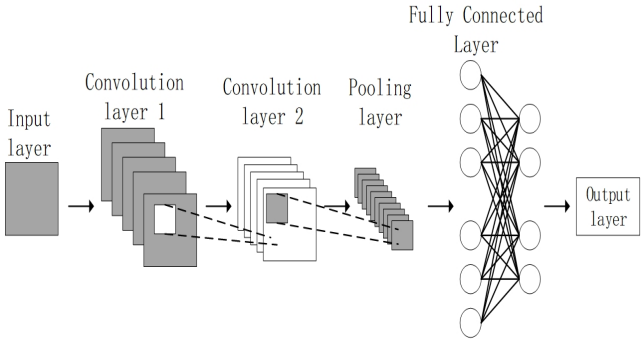


Fig. 1. CNN Network.

The fundamental principle of CNN is to utilize convolutional operations to extract features from the input, where the convolutional kernels are automatically learned by the network, thus endowing CNNs with excellent feature extraction capabilities. Each neuron in the convolutional layer only computes a small region of the input and shares weights, which can significantly reduce the number of parameters that need to be learned. The mathematical definition of the convolution operation is: the input matrix X and the convolutional kernel W, the convolution operation is represented as:$Y = X * W$. In the convolution operation, the convolutional kernel slides over the input matrix and performs convolution operations at each position to generate the output matrix Y. This process is a sliding window operation of the convolutional kernel over the input matrix, where the values in each sliding window are multiplied by the values in the convolutional kernel and summed to obtain the corresponding value in the output matrix. Pooling layers, on the other hand, perform dimensionality reduction on the feature maps output by the convolutional layers, typically using max pooling or average pooling. Max pooling selects the largest value in each pooling window as the output, while average pooling calculates the average value of the values in the pooling window as the output. Pooling operations may reduce the size of the feature maps and also enhance the robustness and generalization ability of the model [30], [31].

This study aims to improve the prediction accuracy of long-term meteorological data by utilizing a one-dimensional CNN neural network to extract meteorological influencing factor data $x = (x_1, x_2, ..., x_n)$. It employs convolutional layers and pooling layers to obtain effective representations, then flattens the acquired data and introduces it into a fully connected layer. The model analyzes the extracted feature data and realizes the output of feature results $c = (c_1, c_2, ..., c_n)$.

C. Bidirectional Long Short-Term Memory Networks-BiLSTM

Recurrent Neural Networks(RNNs) often encounter the vanishing or exploding gradient problem when dealing with relationships between nodes that are far apart, while LSTMs can better retain information provided by nodes that are distant from each other, enhancing performance on longer sequences of temporal data. Each LSTM unit is composed of three gating mechanisms: the forget gate, the input gate, and the output gate. The functional relationships between these gating units are as shown in Eq. 8 [32], [33]:

$$\begin{aligned} f_t &= \delta(W_f.[h_{t-1}, x_t] + b_f) \\ i_t &= \delta(W_i.[h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_c.[h_{t-1}, x_t] + b_c) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \delta(W_o.[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{8}$$

Specifically, x_t represents the input time-series data; f_t, i_t, o_t represent the outputs of the forget gate, input gate, and output gate, respectively; W_f, W_i, W_o are the weight matrices for the three gates, and b_f, b_i, b_o are the corresponding bias units; δ is the sigmoid function; $tanh$ is the hyperbolic tangent function; * denotes the inner product operation; $\tilde{C}_t$ represents the candidate vector created by passing through the tanh layer; W_c, b_c, are the weight matrix and bias unit, respectively, of the candidate layer; C_t represents the cell state; and h_t represents the hidden state.

During the data training process, LSTM can only use information from the forward sequence as the network's prediction result and cannot perceive backward data during model training. The advent of the bidirectional LSTM(BiLSTM) completely solves the problem of the model being unable to utilize future data. The term "bidirectional" refers to the existence of two LSTM networks within BiLSTM : one LSTM processes the forward sequence values, and the other processes the backward sequence values. These two networks operate independently, and the final output is achieved by vector concatenation to produce the final predicted features. Extensive research has shown that BiLSTM performs significantly better than LSTM in time series forecasting. Fig. 2 illustrates the structure of the BiLSTM network.

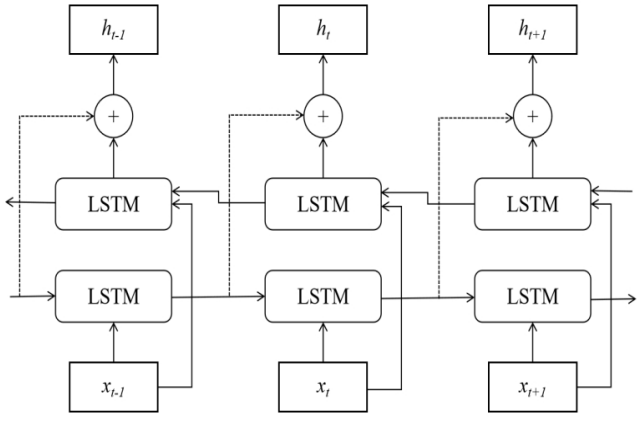


Fig. 2. BiLSTM network structure.

D. Attention

The attention mechanism is commonly used in tasks such as image classification and semantic understanding, where

it has achieved good results. Attention was first proposed in the context of image classification, primarily to enhance image understanding by omitting secondary information and emphasizing important information. Many computer vision algorithms still use Attention and its improvements to enhance their performance. However, since Bahdanau et al. applied Attention to neural translation and achieved good results, an increasing number of researchers have started to use it in natural language processing. For instance, Google's neural translation machine uses self-attention mechanisms exclusively to complete the entire model. Simply put, the attention mechanism is related to attention. Taking humans as an example, when receiving different messages simultaneously, different levels of response are given to different messages, and this difference is a manifestation of attention. Translating this to machine learning, different data require different processing methods, which means we pay different levels of attention to different aspects of the data based on its characteristics [34], [35].

In the process of meteorological data prediction training and learning, especially with multiple input variables, although effective correlation information is retained, it can lead to a complex network topology, resulting in slower learning speeds and difficulties in algorithm convergence. Therefore, this study incorporates an attention mechanism into the model, which adaptively and dynamically weights the factors affecting temperature, precipitation, and wind speed predictions. By reducing the weights of factors with weaker correlations to actual observed values, it allocates higher weights to the main influencing factors, achieving high-precision meteorological information forecasting. As shown in Fig. 3, the structure of the BiLSTM network after incorporating the attention mechanism is depicted.

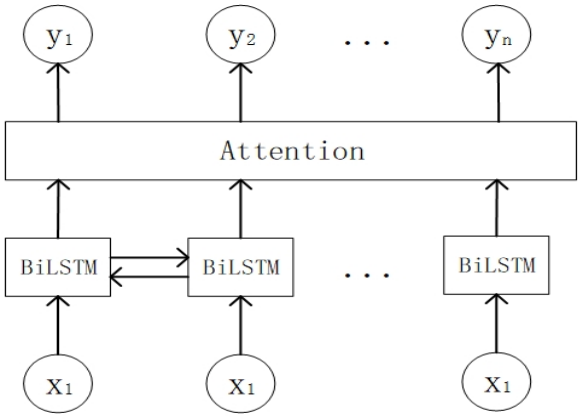


Fig. 3. BiLSTM-Attention network structure.

In the implementation process of introducing an attention mechanism into the BiLSTM model, a series of historical data $x = (x_1, x_2, ..., x_n)$ that affect meteorological predictions (temperature) are treated as the stored content of Key and Value during the addressing operation. Here, Key represents the data address, and Value represents the attention value. The attention calculation formula is as shown in Eq. (9):

$$A_i = \sum_{i=1}^{n} a_i v_i \tag{9}$$

The formula for calculating is given in Eq. (10):

$$a_i = softmax(Sim_i)\frac{eSim_i}{\sum_{k=1}^{n} e^{Sim_k}} \tag{10}$$

Specifically, v_i represents the value of i the data in the attention mechanism, and a_i represents the weight coefficient of v_i. We obtain the final attention by weighted averaging. Furthermore, Sim_i represents the cosine similarity of the i data, and the calculation formula is as shown in Eq. (11):

$$Sim(X, Key_i) = \frac{X.Key_i}{\|X\|.\|Key_i\|} \tag{11}$$

III. Materials and Methods

A. Case Measurement

To verify the effectiveness of the VMD-CNN-BiLSTM-Attention model proposed in this study for air quality forecasting, a total of 9 years of air quality data from January 2015 to December 2023 in Changsha, China, were selected as the training set to predict the air quality for a future period. The sequence model established based on air quality is shown in the Fig. 4.

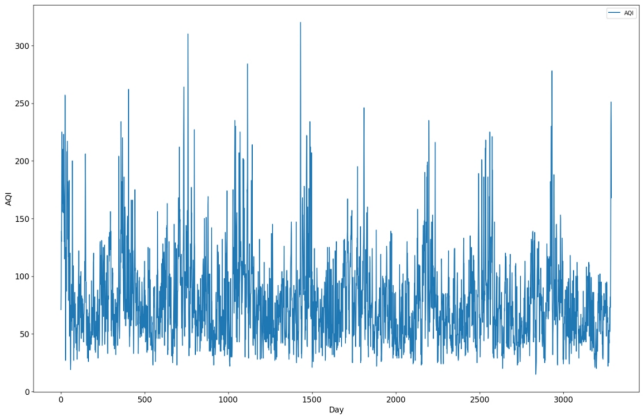


Fig. 4. The structure of the AQI data.

B. Data Processing

Due to the presence of seven different data indicators in the air quality forecasting model, namely AQI, PM2.5, PM10, SO2, NO2, O3, and CO, each indicator may have different characteristics, dimensions, orders of magnitude, and availability, making it impossible to directly analyze the characteristics and patterns of the research subjects. When there is a significant difference in the levels of various indicators, if we directly use the original values of the indicators for analysis, the role of indicators with higher values will be amplified in the comprehensive analysis, relatively weakening the role of indicators with lower values. Therefore, before using the data for prediction, we perform standardization processing on the data to ensure that different feature variables have the same scale. This allows the target variable to be controlled by multiple feature variables of the same size, and when using gradient descent to learn parameters, the impact of different

features on the parameters is consistent. This study uses min-max normalization for data standardization. The calculation formula is as shown in Eq. (12).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{12}$$

In which, X_{max} represents the maximum value in the sequence data sample; X_{min} represents the minimum value in the sequence data sample. VMD processing:this study employs the VMD algorithm to denoise and decompose air quality data, thereby enhancing the accuracy of the data. The sampling frequency is set to 2000, and after fine-tuning, the optimal penalty coefficient α is determined to be 1200, with the number of decomposition layers K set to 10. From the Fig. 5, it is evident that the decomposition of the air quality sequence using VMD yields 9 components with strong regularity. IMF1 is the dominant component, with a smooth curve that characterizes the overall trend of air quality changes. The remaining components, though of different frequencies, also exhibit regularity and can reflect the local characteristics of air quality data to a certain extent.

C. Model Frame

A combined forecasting model based on CNN-BiLSTM-Attention is used, with the model structure shown in Fig. 6. The model mainly consists of an input layer, a one-dimensional CNN layer, a dropout layer, a BiLSTM layer, an attention layer, a flatten layer, and a fully connected layer.

During the backpropagation process of the CNN-BiLSTM-Attention combined model, an appropriate algorithm is needed for parameter learning to guide the parameters of the objective function in the correct direction for updating the appropriate size, so that the updated parameters continuously approach the global optimum of the objective function value. Gradient descent principle is commonly used, following the negative gradient of the objective function to locate the minimum value of the function. This study uses the Adaptive Moment Estimation algorithm(Adam algorithm), with the purpose of accelerating the optimization process. At the same time, the Adam algorithm adjusts the throughput of the search process automatically for each variable encountered gradient(partial derivative) at each step length.

For the prediction of air quality data, the proposed model is implemented through the following processes:

- For the preprocessed time series $x = (x_1, x_2, ..., x_n)$, where, $x_i \in R^{sxf}, i \in (0, n)$ and s represents the length of each data's time window, and f is the dimension of the data feature vector. First, a one-dimensional CNN is used for feature extraction, and data padding is employed to ensure that the input time series dimensions remain unchanged.
- Further, for the extracted data $x = (x_1, x_2, ..., x_n)$, where, $x_i \in R^{sxd}, i \in (0, n)$, d denotes the size of the convolutional kernel in the convolutional neural network, adding a Dropout layer allows the model to adaptively block some hidden layer neurons without affecting the output dimension size, thereby enhancing the model's generalization ability.
- Secondly, the BiLSTM, which includes two LSTM units, receives forward and backward information, specifically calculated using Eq. (8). BiLSTM processes the output from the Dropout layer, achieving the exploration of periodicity and nonlinear relationships between time series.
- Then, the output of BiLSTM is taken as the input for the attention mechanism, which adaptively assigns weight coefficients to each input variable feature, further enhancing the model's perception of key information, with related calculation formulas shown in Eq. (9) to Eq. (11).
- Finally, after processing by the attention mechanism, the feature information is connected to the fully connected layer for prediction processing through the Flatten layer, and the final prediction results are outputted.

D. Experimental Design and Simulation

All programs in this study are completed using Python, with the training of the neural network part carried out using the PyTorch software library. PyTorch is an open-source neural network framework developed by the Torch7 team at Facebook AI Research. Its underlying implementation is based on Torch, but it is entirely implemented and utilized in Python. This framework is primarily used for scientific research and application development in the field of artificial intelligence. Torch is a classic tensor library for operating on multi-dimensional matrix data and is widely used in machine learning and other mathematically intensive applications. This study utilizes the GPU version of PyTorch, leveraging the excellent performance of GPU data processing to reduce model learning time. The entire experiment was conducted in a Windows environment, with the specific experimental environment as shown in Table I.

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Name	Versions
CPU	Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz (8CPUs)
Memory	DDR4 32GBytes
Operating system	Windows 10 Pro
GPU	NVIDIA Ge Force RTX 3090
Python	Python 3. 9. 13
PyTorch-GPU	PyTorch 1. 13. 0

The sample data is divided in a 9:1 ratio, where the 90% of the data is used as the training set and the remaining 10% as the test set.

In the experimental design of this study, we used batch processing. The batch processing refers to the batch size, which is the number of training samples used in each iteration. A larger batch size can better utilize the parallel computing capabilities of GPUs/CPUs but requires more memory. A smaller batch size occupies less memory but may lead to unstable convergence. Therefore, the batch size needs to be chosen based on actual memory conditions and model complexity. In this research, the batch size is set to 256.

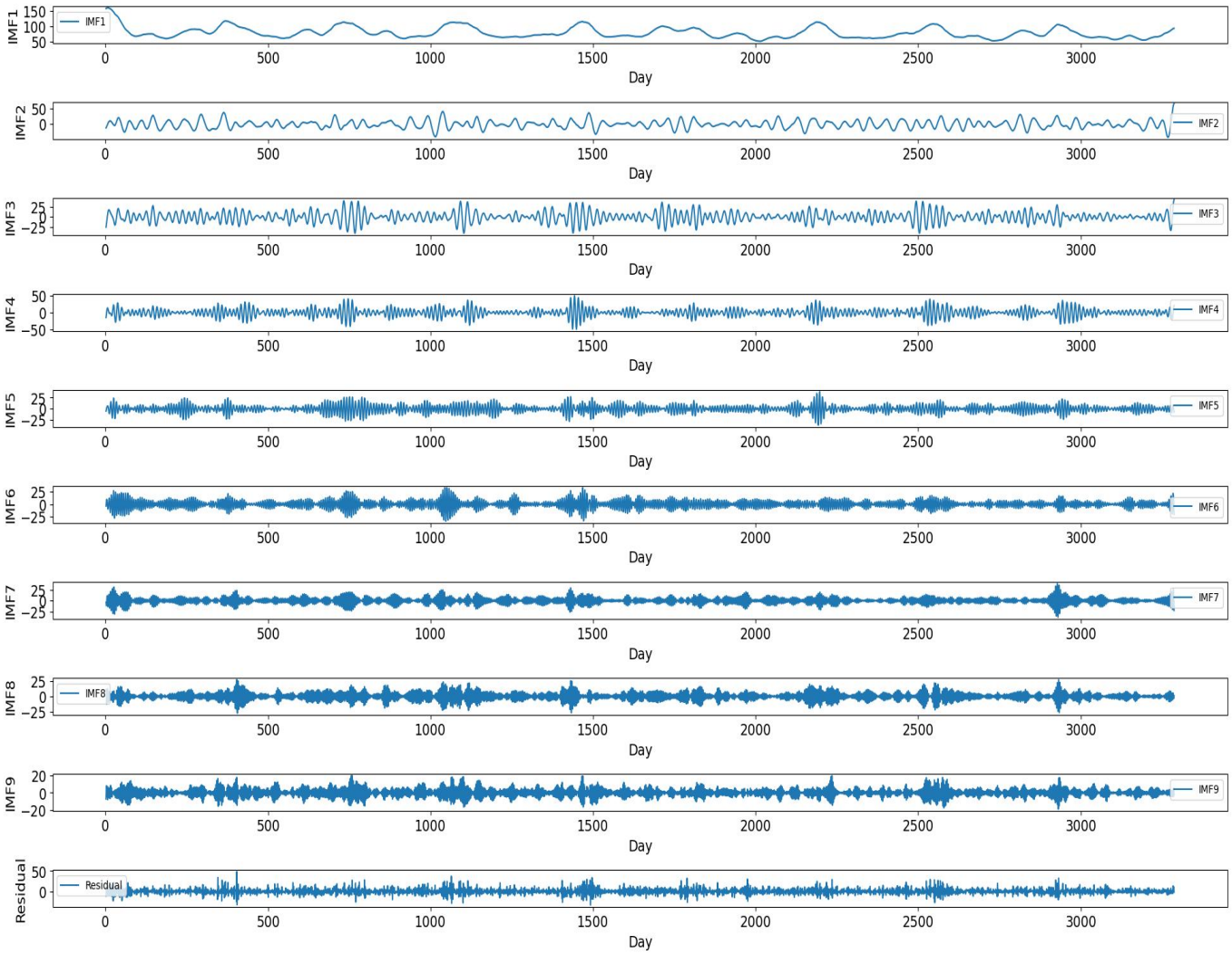


Fig. 5. The IMFs obtained after VMD.

The learning rate determines the update amplitude of weights in each training batch. A larger learning rate can speed up convergence but may miss the optimal solution. A smaller learning rate can increase accuracy but converges more slowly. In this research, the learning rate is set to 0.001.

The number of iterations is the number of times the entire training dataset is traversed. More epochs mean more thorough training but may also lead to overfitting. The number of iterations is set to 100 in this design.

Dropout is a regularization technique to prevent overfitting. It works by randomly "dropping" some neurons in the hidden layers during training, reducing the interdependence among neurons. This makes the network less sensitive to minor changes in input data and improves the model's generalization ability. A higher dropout rate can better prevent overfitting but may also lead to underfitting. In this research, the dropout rate is set to 0.1.

Time steps are mainly used for processing sequential data, such as natural language or time series data. It represents the number of steps when LSTM or RNN is unfolded. Longer time steps help capture long-term dependencies but also increase computational complexity and memory usage. Therefore, the time step should be chosen based on the nature of the task and hardware resources. In this experiment, the time step is set to 15, meaning that the model's training value at each moment is related to its previous 15 values.

The result measurement in these research, we use Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R Squared (R2) as metrics for evaluating the performance of the models. Lastly, the effectiveness of the hybrid model in air quality prediction is verified by comparing the VMD-CNN-BiLSTM-Attention hybrid model with RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM and CNN-LSTM-Attention models.

E. Result and Discussion

Table II and Fig. 7 are the result of air quality prediction performance of each model. The models are RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention and CNN-BiLSTM-Attention. The results shows that, for

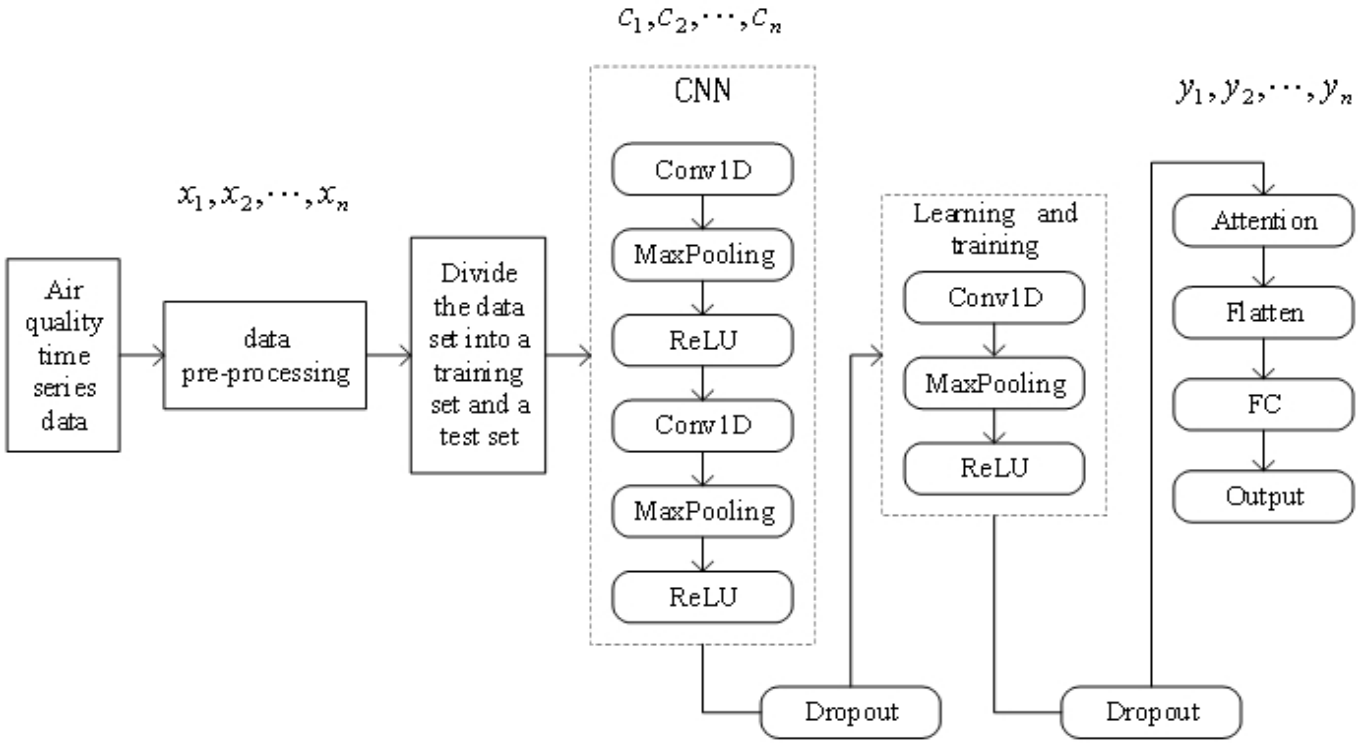


Fig. 6. Combined forecasting model based on CNN-BiLSTM-Attention.

individual models, BiLSTM and LSTM have better predictive performance compared to RNN, with BiLSTM outperforming LSTM. This indicates that RNN does not effectively utilize historical data of the air quality dataset, resulting in the poorest predictive performance. Incorporating bidirectional LSTM units can enhance the model's predictive performance, as BiLSTM offers higher efficiency and performance in data feature extraction compared to LSTM.

TABLE II. EVALUATION INDEX TABLE OF AIR QUALITY PREDICTION PERFORMANCE OF EACH MODEL

Model	MAE	MSE	RMSE	R2
RNN [36]	12.83	301.77	17.37	0.4226
LSTM [37]	12.81	276.47	16.63	0.2641
BiLSTM [38]	11.82	244.69	15.64	0.6288
CNN-LSTM [39]	8.28	133.71	11.56	0.7503
CNN-BiLSTM	7.34	100.71	10.04	0.8128
CNN-LSTM-Attention [40]	4.08	30.74	5.54	0.9588
CNN-BiLSTM-Attention	3.16	22.53	4.75	0.9685

However, when RNN, BiLSTM and LSTM models are combined with the neural network CNN, the predictive accuracy of the models significantly improved. CNN-LSTM, compared to LSTM, saw a 35% and 51% reduction in MAE and MSE, respectively, with R2 increasing to 75%. CNN-BiLSTM, compared to BiLSTM, experienced a 38% and 59% decrease in MAE and MSE, respectively, with R2 reaching 81%. This suggests that the addition of convolutional neural network CNN extracts effective features from the data, thereby enhancing the accuracy of predictions.

Upon integrating the Attention model, the predictive accuracy of the combined models further improved. CNN-LSTM-Attention, compared to CNN-LSTM, saw a 51% and 77% reduction in MAE and MSE, respectively, with R2 increasing by 28%. CNN-BiLSTM-Attention, compared to CNN-BiLSTM, experienced a 57% and 78% decrease in MAE and MSE, respectively, with R2 increasing by 19%. The attention mechanism can mimic the human brain's operational mechanism, focusing more on more important information when faced with varying external input data, thus avoiding the interference of irrelevant information. It can assign higher weights to key data features in air quality data, enabling the model to achieve higher predictive accuracy.

In summary, the ranking of predictive performance of different models is as follows:CNN-BiLSTM-Attention > CNN-LSTM-Attention > CNN-BiLSTM > CNN-LSTM > BiLSTM > LSTM > RNN. The predictive model CNN-BILSTM-Attention used in this study can effectively enhance the accuracy of predictions and obtain more precise outcomes. Compared to the initial RNN, CNN-BILSTM-Attention has reduced the MAPE.

IV. CONCLUSION

In recent years, as living standards have improved due to technological progress, the exploitation and use of fossil fuels by humans have also increased rapidly. This surge in fossil fuel consumption has led to serious air quality issues, posing a significant threat to both the environment and human health. In response to these challenges, this study focuses on predictive research related to air quality, aiming to achieve accurate forecasts through the development of an effective air

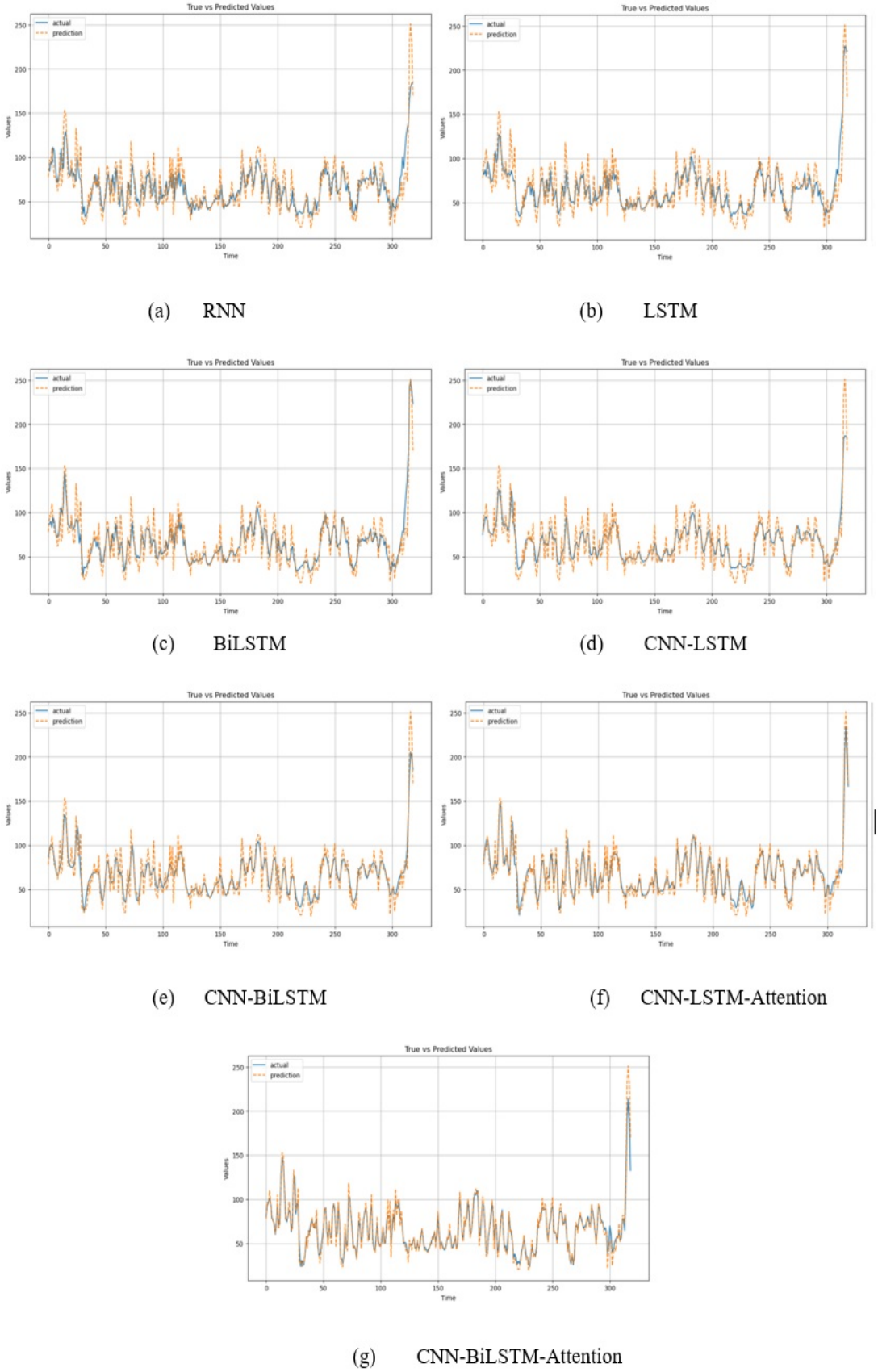


Fig. 7. Comparison of air quality prediction results from various models.

quality prediction model. Such a model can assist residents in planning their activities and provide scientific guidance to governments in formulating environmental protection policies.

Our research reveals that air quality time series data typically exhibit spatiotemporal diversity, meaning they display varying patterns across different times and locations. In addition, due to the nonlinear nature of meteorological and physical processes, as well as the influence of external factors, air quality data often show seasonal trends, daily cyclical variations, long-term patterns, and random fluctuations. These complexities make air quality prediction particularly challenging.

To address these issues, this study proposes a combined VMD-CNN-BiLSTM-Attention model designed to capture the spatiotemporal diversity and seasonality inherent in air quality time series. The model begins by applying the VMD algorithm to decompose and denoise the air quality data. Following decomposition, the dataset is standardized through normalization, improving data accuracy and facilitating more effective neural network predictions.

Within the proposed architecture, the CNN is responsible for extracting spatial features from the air quality data, while the BiLSTM network captures the temporal dependencies. The Attention mechanism then assigns varying weights to different elements of the data sequence, emphasizing the most influential features. This integrated model structure enhances the ability to capture complex spatiotemporal patterns, thereby improving prediction accuracy and reliability.

During model training, we evaluated the performance of seven different neural network models: RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention, and CNN-BiLSTM-Attention. Model performance was assessed using four evaluation metrics, there are MAE, MSE, RMSE, and R2. The results demonstrate that the CNN-BiLSTM-Attention model outperforms the others across all indicators, achieving the highest accuracy in air quality prediction.

ACKNOWLEDGMENT

This work was supported by the Journal Support Fund, UiTM.

REFERENCES

[1] Y. Ye, Q.Tao, "E The dynamic relationship among economic development, air pollution, and health production in China: the DNSBM efficiency model," *Frontiers in Environmental Science.*, vol, 11, pp. 1205712, 2023.

[2] Y.Tan, and X. Mao, "Assessment of the policy effectiveness of Central Inspections of Environmental Protection on improving air quality in China," *Journal of Cleaner Production.*, vol.288,no.12, pp. 125100,2020.

[3] O.Hahad, J.Lelieveld, F.Birklein, K. Lieb, A. Daiber and T.Münzel, "Ambient air pollution increases the risk of cerebrovascular and neuropsychiatric disorders through induction of inflammation and oxidative stress," *International journal of molecular sciences.*, vol. 21, no. 12, pp. 4306,2020.

[4] C.Xu, Z.Zhang, G.Ling, G. Wang, and M.Wang, " Air pollutant spatiotemporal evolution characteristics and effects on human health in North China," *Chemosphere.*, vol. 294, pp. 133814, 2022.

[5] D.Zhan, M. P.Kwan, W.Zhang, X.Yu, B.Meng, and Q. Liu, " The driving factors of air quality index in China,"*Journal of Cleaner Production.*, vol. 197, pp.1342–1351,2018.

[6] Y.Liu, J.Xu, D.Chen, P.Sun, and X.Ma,"The association between air pollution and preterm birth and low birth weight in Guangdong, China," *BMC public health.*, vol.19, pp.1–10,2019.

[7] X.G.Zeng, F.F.Ruan, and Y.Y.Peng, "Spatial distribution of PM2. 5 pollution health effects in China based on spatial grid scale ," *China Environ. Sci.*, vol.39, pp.2624-2632,2019.

[8] Q.Liao, M.Zhu, L.Wu, X. Pan, X.Tang, and Z.Wang, "Deep learning for air quality forecasts: a review,"*Current Pollution Reports.*,vol. 6, no.4,pp.399-409,2020.

[9] A.Aggarwal, D.Toshniwal, "A hybrid deep learning framework for urban air quality forecasting,"*Journal of Cleaner Production.*,vol. 329, pp.129660,2021.

[10] J.Wang, J.Li, X.Wang, J.Wang, and M. Huang, "Air quality prediction using CT-LSTM,"*Neural Computing and Applications.*, vol.33, pp.4779–4792, 2021.

[11] J. Ferreira , D.Lopes, S. Rafael, H. Relvas, , S. M.Almeida,and A. I. Miranda, "Modelling air quality levels of regulated metals: Limitations and challenges,"*Environmental Science and Pollution Research.*, vol.27, pp.33916–33928, 2020.

[12] A. Baklanov, Y. Zhang, "Advances in air quality modeling and forecasting,"*Global Transitions.*, vol.2, pp.261-270,2020.

[13] M.Méndez, M.G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey,"*Artificial Intelligence Review.*, vol.56, no. 9, pp.10031-10066,2023.

[14] H.Liu,G. Yan, Z. Duan, and C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," *Applied Soft Computing.*, vol. 102, PP.106957,2021.

[15] J.Zhang, and S. Li,"Air quality index forecast in Bei**g based on CNN-LSTM multi-model," *Chemosphere.*, vol.308, pp.136180,2022.

[16] S.Abirami, and P. Chitra, " Regional air quality forecasting using spatiotemporal deep learning," *Journal of cleaner production.*, vol.283, pp.125341,2021.

[17] Y.C.Liang , Y.Maimury,A.H.L. Chen, and J.R.C.Juarez," Machine learning-based prediction of air quality," *applied sciences.*, vol.10,no.24,pp.9151,2020.

[18] M.Castelli,F.M. Clemente, A. Popovič, S.Silva, and L. Vanneschi, " A machine learning approach to predict air quality in California,"*Complexity.*, vol.1,pp, 8049504,2020.

[19] J.M.Bertrand, F. Meleux, A.Ung, G. Descombes, and A. Colette, " Improving the European air quality forecast of the Copernicus Atmosphere Monitoring Service using machine learning techniques,"*Atmospheric Chemistry and Physics.*, vol.23,no.9, pp.5317-5333,2023.

[20] B.Zhang, Y.Rong, and R.H. Yong," Deep learning for air pollutant concentration prediction: A review,"*Atmospheric Environment.*, vol. 290, pp.119347,2022.

[21] C.L.Wu, R.F.Song, X.H.Zhu, Z.R.Peng,Q.Y. Fu, and J. Pan," A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network,"*Environmental pollution.*, vol.320, pp.121075,2023.

[22] S.Baniasadi, R.Salehi, S.Soltani,D. Martín, P. Pourmand, and E. Ghafourian, " Optimizing long short-term memory network for air pollution prediction using a novel binary chimp optimization algorithm,"*Electronics.*, vol.12, no.18, pp.3985,2023.

[23] N.A.Zaini, L.W. Ean,A.N. Ahmed, and M.A. Malek, "A systematic literature review of deep learning neural network for time series air quality forecasting," *Environmental Science and Pollution Research.*, pp.1–33,2020.

[24] S.Abirami, and P.Chitra, "Regional air quality forecasting using spatiotemporal deep learning ,"*Journal of cleaner production.*, vol.283, pp.125341,2021.

[25] H.Wu, T. Yang, H.Li, and Z. Zhou, "Air quality prediction model based on mRMR–RF feature selection and ISSA–LSTM,"*Scientific Reports.*, vol. 13,no.1, pp.12825,2023.

[26] Y.Zeng, J.Chen, N.Jin ,X.P. Jin,Y. Du, "Air quality forecasting with hybrid LSTM and extended stationary wavelet transform,"*Building and Environment.*, vol.213, pp.108822,2022.

[27] W.Huiyong,Y. Tongtong, and L. Hongkun, "Air quality prediction model based on mRMR-RF feature selection and ISSA-LSTM,"*Scientific reports.*, vol.13, no.1,pp. 12825-12835,2023.

[28] L. Alzubaidi, J. Zhang, A.J. Humaidi, .A.Al-Dujaili, Y.Duan, O.Al-Shamma, J.Santamaría,M.A.Fadhel,A.Muthana and F.Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,"*Journal of big Data.* ,vol. 8, pp.1–74,2021.

[29] A. Brahms, "Representation error for real numbers in binary computer arithmetic," IEEE Computer Group Repository, Paper R-67-85.

[30] A.Sayeed, Y.Choi, E.Eslami, Y.Lops, A.Roy, and J. Jung, "Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance," *Neural Networks,.*, vol.121, pp.396–408,2020.

[31] C.Shang, J.Gao, H.Liu, and F.Liu, "Short-term load forecasting based on PSO-KFCM daily load curve clustering and CNN-LSTM model," *IEEE Access.*, vol. 9, pp.50344-50357,2021.

[32] N.Jin,Y.Zeng, K.Yan, and Z.Ji, "Multivariate air quality forecasting with nested long short term memory neural network," *IEEE Transactions on Industrial Informatics.*, vol. 17,no.12, pp.8514–8522,2021.

[33] M.Lee, L. Lin, C.Y.Chen, Y. Tsao, T.H. Yao, M.H.Fei, S.H. Fang, "Forecasting air quality in Taiwan by using machine learning," *Scientific reports.*, vol.10,no.1,pp. 4153,2020.

[34] Gilpin W. "Deep reconstruction of strange attractors from time series,"*Advances in neural information processing systems*, vol. 29, pp.204–216,2020.

[35] A.Aggarwal, and D. Toshniwal, "A hybrid deep learning framework for urban air quality forecasting," *Journal of Cleaner Production*, vol.329, pp.129660,2021.

[36] J.M. Han, Y. Q. Ang, A. Malkawi, and H.W. Samuelson, "Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements," *Building and Environment*, vol. 192, pp.107601, 2021.

[37] D.N. Fente, and D.K. Singh, "Weather forecasting using artificial neural network," *In 2018 second international conference on inventive communication and computational technologies (ICICCT)*, pp. 1757-1761, 2018.

[38] M. Schuster, and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[39] M. Fan, O. Imran, A. Singh, and S.A. Ajila, "Using cnn-lstm model for weather forecasting," *In 2022 IEEE International Conference on Big Data (Big Data)*, pp. 4120-4125, 2000.

[40] J. Shen, W. Wu, and Q. Xu, "Accurate prediction of temperature indicators in eastern china using a multi-scale cnn-lstm-attention model," *arXiv preprint arXiv:2412.07997*, 2024.

Cardio-Edge: Hardware-Software Co-design Implementation of LSTM Based ECG Classification for Continuous Cardiac Monitoring on Wearable Devices

Nousheen Akhtar[1], Abdul Rehman Buzdar[2], Jiancun Fan[3], Muhammad Umair Khan[4]
School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China[1,3]
Department of Computer Engineering, National University of Technology, Islamabad, Pakistan[2,4]

***Abstract*—Cardiac arrhythmias should be detected at an early stage so that clinical intervention can take place and continuous patient monitoring can be established in a timely manner. In this study, we present Cardio-Edge, a hardware-software co-design implementation of an LSTM-based ECG classification system optimized for real-time use on wearable devices. Proposed architecture comprises discrete wavelet transform (DWT) and principal component analysis (PCA) for efficient feature extraction followed by multiple parallel LSTM networks and a multi-layer perceptron (MLP) for classification. Implemented on a Xilinx ZYNQ-7000 SoC, our system leverages FPGA-based hardware acceleration alongside ARM Cortex-A9 for preprocessing tasks. Compared to software-only implementation on the same ARM processor, our co-design achieves a 10× improvement in execution speed with 99% classification accuracy trained and verified on the MIT-BIH arrhythmia dataset. The hardware-efficient implementation employs resource-optimized architectures for LSTM, activation functions, and fully connected layers making it appropriate for low-power, patient-specific wearable healthcare devices. This real-time, on-chip solution eliminates dependence in-cloud connectivity and ensures data privacy hence suitable for continuous cardiac monitoring applications.**

***Keywords*—ECG classification; wearable devices; discrete wavelet transform (DWT); long short-term memory (LSTM); field-programmable gate array (FPGA)**

I. Introduction

Cardiovascular diseases, particularly arrhythmias, are a leading cause of mortality all over the world. Identification of abnormal heart rhythms in time is also essentially important for pre-empting such exacerbations as a stroke or sudden cardiac arrest. Most conventional systems for monitoring ECG are off-line, and that means there would be a delay in diagnosis and treatment [1]. Even though AI particularly through deep learning frameworks has shown capabilities regarding the classification of signals pertaining to ECGs, an efficient deployment of such models on hardware platforms is still lacking. Most existing solutions are either software-based or rely on the cloud, which makes them infeasible for use in low-power, embedded, or portable health monitoring devices. Besides, high-performance processors like GPUs may not always be applicable for such an application because of their power and space requirements. While real-time detection is eventually what we want to achieve, the current work is concerned more with building a hardware-accelerated version of an LSTM model that can support real-time execution [2]–[5].

In the past few years, there has been a spiking trend of wearable healthcare devices given their real-time capabilities of monitoring ECG signals and detect cardiac issues before the patient gets too serious and needs immediate medical assistance [6]. The currently available wearable ECG monitoring devices transmit the raw signals, via smartphones or wearable devices, over the internet to hospitals for diagnosis of the ailment. This demands high power and reliance on the internet which defeats the purpose of low power automation. Researchers have made strides in machine learning models and their implementation in embedded systems to develop wearable devices that are designed to classify cardiac arrhythmia in real-time on chip hardware. For real-time processing on wearable devices, automated ECG classifier algorithms should not be computationally heavy and give correct predictions at the same time. A number of such classification algorithms rely on morphological features and use techniques that are doing traditional signal processing methods [7], [8]. The challenge with automating the monitoring of ECG signals is that they comprise of morphological features such as QRS complex and P waves which are affected by daily life variables such as exercise and they vary from patient to patient. This means that these algorithms cannot distinguish between various types of arrhythmia quite accurately [9], [10]. Researchers have developed machine learning models to categorize cardiac arrhythmia such as those discussed in [11], [12]. They distinguish CA beats from heart beats having normal rhythm by utilizing SVM, but both of the said algorithms need to be fed with the features manually. This lack of automation restricts the classification domain of the classifier when it comes to various ECGs. Another downside is that these classification methodologies is that they are not approved by association for the Advancement of Medical Instrumentation (AAMI) as they don't adhere to American National Standard. The key to automate feature extraction and pump up the accuracy of the arrhythmia classification, is the use of deep-learning algorithms. The deep convolutional neural networks CNNs and recurrent neural networks RNNs have recently been extensively utilized [12]–[17].

Thus, for the purpose of ECG classification, we have purposed a low power VLSI architecture which is Long Short-

Term Memory (LSTM) based. The processor uses RNNs because the ECG waveform is intrinsically compatible for processing with RNNs which is why we use them in our system. Fig. 1 illustrates that the backbone of our system is composed of LSTM and classical features like wavelet are also part of the solution. These extra features enhance the pattern recognition capability in the ECG waveform. In addition, the system merges the arrhythmia predictions from small LSTM models as opposed to constructing one large model. A large LSTM model is computationally intensive as compared to running a number of lighter LSTM models. So, we have executed multiple LSTM models and combined their collective predictions, instead of running one large model [18], [19]. The ECG classification algorithm is locally executed on patients' personal wearable devices. Regardless of network speed or availability, local execution enables for ongoing operation. Furthermore, it makes it possible for data to remain on the wearable device, avoiding the security concerns associated with cloud-assisted processing. Remote processing on powerful cloud servers or offline processing of recorded ECG signals techniques are not adopted. Our adapted approach has outperformed several other deep-learning methods, in terms of classification accuracy and is still very cost-efficient. It involves on chip execution of arrhythmia classifier, bypassing the need for a consistent and fast internet connection. Moreover, since the patient's data is processed on the device, we also avoid problems related to privacy that may arise from employing cloud computing services. Our solution is not to store the ECG signals and process them offline, nor do we utilize the most common cloud-computing services like AWS or GCP for running our classifier. This is a step in making continuous ECG classification wearable devices a common utility for those in need.

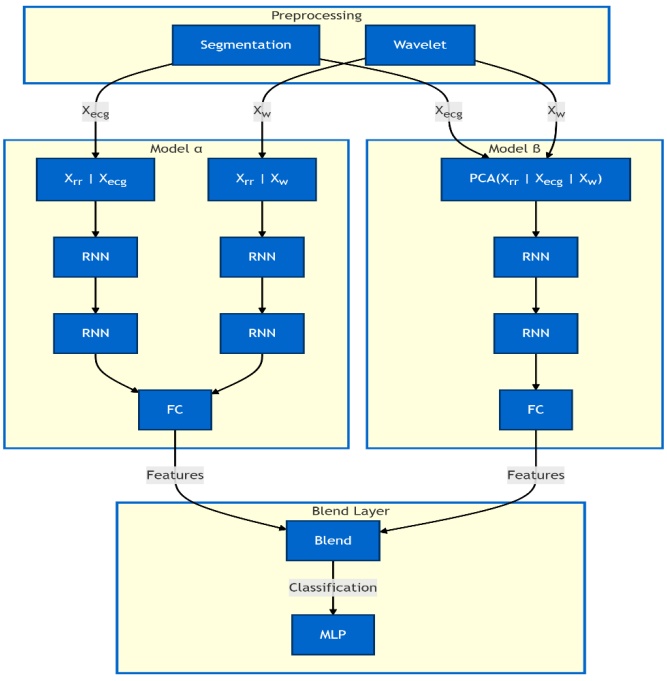


Fig. 1. LSTM-based ECG classification algorithm.

II. Related Work

The traditional approaches are based on extracting characteristics from the ECG signal, which involves identifying certain attributes of the ECG waveform, such as the P wave, QRS complex, and T wave, and then using these attributes as inputs to classification algorithms such as Support Vector Machines (SVMs) or k-Nearest Neighbors (kNN) [20]. There were some constraints with the old methods. The feature extraction was manual, time-consuming, and subjective, therefore susceptible to bias cases. It did not effectively capture the minute variations in the ECG signals. Therefore, it has low sensitivity for the infrequently appearing arrhythmia in particular ECG databases [21], [22]. The inter-patient variability also posed a challenge, significant enough to impede generalization from models trained on one dataset to an application in another dataset [23], [24]. Some important pre-processing steps included noise and artifact removals, but were relatively complicated and resource-intensive [25], [26].

Deep learning changed everything in ECG arrhythmia classification [25], [26]. Deep learning frameworks, especially Convolutional Neural Networks (CNNs), have displayed phenomenal prowess in the automatic learning of complex features from raw ECG signals without manual interference in feature engineering [27]. CNNs have specialized in detecting spatial patterns; therefore, they are greatly applicable to the analysis of the temporal dynamics regarding ECG waveforms [28], [29]. A superior performance by CNNs has been demonstrated over traditional machine learning techniques in many studies [30], [31]. An important aspect of building AI-based ECG arrhythmia detection systems is choosing an appropriate hardware platform. The choice depends on several requirements regarding performance level power usage requirement cost limitation and the complexity of the AI model to be used. This section points out significant hardware platforms reported in the literature and discusses their pros and cons.

ASICs were viewed as a suitable option since they could offer high performance with low power. Therefore, they might be best suited for integration into wearable devices for long-term monitoring of ECG. Since ASICs are custom-made inherently, the hardware architecture can be optimized to better suit the specific needs of an AI algorithm, which in turn brings about higher efficiency and lower power consumption. The use of ASICs for AI-based ECG arrhythmia detection has been noted in many studies in the literature. For instance, Zhang et al. [32] proposed a low-power ASIC realizing an Artificial Neural Network (ANN) architecture for foreseeing five types of cardiac arrhythmia. This low-power ASIC design includes a novel processor in preprocessing the ECG signals to effectively recognize R-peaks, which is the very vital step in analyzing arrhythmia, achieving 96.69% classification accuracy with 12.88 W of power consumptions at 100 kHz clock frequency.

The authors emphasize the adoption of a hardware reuse approach governed by a finite state machine (FSM) for rapid computation in an ANN. In a similar vein, Tefai and colleagues detail the ASIC implementation of a pre-trained neural network for ECG feature extraction, which is an essential preprocessing step for proper identification of arrhythmia [33]. The model developed is a recurrent neural network (RNN) that was trained using data from the PhysioNET database; it has been reported to give an accuracy of 96.55%. These works

indicate how all choices ASIC designers make from the type of ANN architecture to hardware reuse techniques and the CMOS technology chosen for manufacturing can heavily impact the size, power dissipation, and performance of an ASIC. The choice of CMOS technology greatly affects the dimensions, power drain, and performance of the ASIC. Hoyer et al. also propose a mixed-signal integrated circuit (IC) targeted for small patch ECG devices able to detect atrial fibrillation. Inference run-time for this implementation was 87.2% lower than that of a floating-point based implementation [34]. Design choices made in these ASICs regarding the specific ANN architecture and hardware reuse strategies have a significant influence on power efficiency and performance.

FPGAs are flexible and reconfigurable thus making a good alternative to ASICs. Prototyping and evaluating various AI algorithms in an FPGA before finalizing an ASIC design can be very beneficial. Since FPGAs are programmable, one can quickly modify and experiment with different architectures of AI model occasionally combined with signal processing technique. Liu et al. [35] present a fully-mapped FPGA accelerator for AI-based ECG analysis which includes the implementation of 1-D CNN and heart rate estimator. The architecture exploits parallelism in both the applications, resulting in considerable speedup and power savings relative to software implementations on other platforms, like ARM-Cortex A53 and Core i7-8700 CPUs. The authors have done a painstaking job of analyzing and optimizing the computation by eliminating division operations for the sake of efficiency. The proposed design consumes only 67.74 mW of power, thus suitable for applications that have limited resources, such as wearable ECG monitoring devices. In another such effort, Varadharajan and Nallasamy [36] present a Distributed Arithmetic (DA) based Gated Recurrent Unit (GRU) to enhance hardware efficiency for ECG diagnosis. The GRU represents a generalized simpler version of the LSTM and is implemented using DA operations to further reduce energy and latency in applications. The authors carry out their implementation on a ZYNQ-7000 SoC, where hardware and software co-design methodologies are used. Performances under different metrics like accuracy, precision, recall, specificity, F1-score, power, delay, and area are compared. The inherent parallel processing capability of FPGAs is well utilized to speed up inference and reduce latency. However, the trade-off between resource utilization in FPGAs and performance needs to be considered carefully and explored further. Gon and Mukherjee [37] show once again how FPGAs are useful in ECG signal processing by proposing a hardware-efficient architecture for removing noise from ECG signals based on a modified lifting DWT algorithm. This design, free of multipliers, yielded a very significant improvement in SNR and also exhibited lower hardware utilization and higher operating frequency compared to existing ECG denoising architectures.

Micro controllers, on the other hand, are low-cost and power-efficient alternatives for wearable devices in ECG monitoring when the AI model's complexity is not very high. Low power consumption allows for increased battery life in wearable, thus allowing prolonged continuous monitoring. Guo et al. use a MSP432P401R micro controller to carry out ECG signal acquisition and arrhythmia detection through a Convolutional Neural Network (CNN) [38]. The authors used an ADS1292 bio-electric sensor to obtain ECG signals and then sent the data acquired to a personal computer for analysis through serial ports. They presented the possibility of performing real-time analysis that leads to an accurate diagnosis of premature atrial contractions with a detection rate of 95.1%. Scrugli et al. implement a cognitive algorithm, specifically a convolutional neural network that has been trained to classify ECG waveforms, on a micro controller based platform thus reinforcing the claim on the propriety of micro controllers for resource-constrained applications [39]. Its design features an adaptive layer that dynamically regulates hardware and software settings to manage power consumption, able to achieve up to a 50% reduction in power consumption while maintaining high accuracy (over 97%) for arrhythmia detection based on the MIT-BIH Arrhythmia dataset. The selection of the micro controller has a substantial influence on the complexity of the AI model that can be implemented efficiently, as well as on the overall power consumption of the device. A critical aspect in prolonged monitoring applications is the equilibrium between processing capability and energy efficiency. Deng et al. [40] showed an automated arrhythmia detection system using a low-power microprocessor integrated into a wearable jacket. This system demonstrated very high sensitivity and specificity for detecting atrial arrhythmia, thereby showcasing the progressively enhanced capabilities of low-power microprocessors to conduct real-time AI-based analysis of ECG data.

III. Cardiac Arrhythmia Classifier

Traditional SoCs for the detection of cardiac arrhythmia consist mainly of two blocks. First comes the analogue to digital converter ADC front that obtains the ECG signals and digitizes it into discrete samples. The second block i.e. the co-processor, is responsible for feature extraction and classification. This study aims to fit a low powered ECG co-processor in a small area appropriate for wearable devices, as shown in Fig. 2. Firstly, the ECG signals that were converted into digital samples are segmented into heartbeats, enabling us to extract their wavelet and RR interval features. Then, these extracted features and the segmented ECG signals are processed by two slightly different RNN module. The two outputs from the RNNs are merged and fed into a feed forward neural network which gives the final output [18], as shown in Fig. 1 and Algorithm 1.

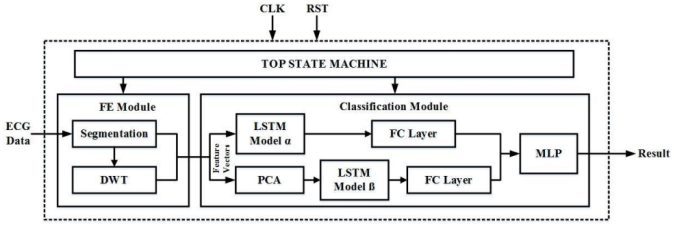


Fig. 2. System architecture of ECG classification system.

A. ECG Segmentation and Feature Extraction

In CA classification, the initial step is to provide vectors of fixed length extracted from the raw ECG signals as input to the classification block. A sequence of heartbeats is generated by segmenting the digitized ECG samples. This segmentation

is done around the R peaks in the heartbeat signal. Since a single segment or a heartbeat of fixed length, the duration of the segment can be mapped as 0.25 seconds of the ECG signal before an R peak is detected and 0.45 seconds after it is detected. The accuracy of the R peak detection algorithms, such as Pan-Tompkin's algorithm that we have used in our solution is highly reliable. The time intervals between consecutive R peaks is calculated as per the requirement of the Pan-Tompkins algorithm. If the time interval is denoted by between peak to peak then we can obtain the following four key features for the heartbeat i to construct our feature vector:

1) RR_i as the past RR interval.
2) RR_{i+1} as the next RR interval.
3) The local average of the five past and the five next RR intervals.
4) The average duration of the RR intervals in each person's train data.

These RR interval features are extracted accurately with considerably less computation. Other than that the algorithm does not require manually feeding other morphological features which are based on Q, S or T intervals which don't capture the nuanced representation of the segmented ECG signals. Moreover, these features don't vary from patient to patient and aren't influenced by environmental factors or situational diversity, making them unsuitable for tasks like classification of arrhythmia [10]. The approach of automating feature extraction by utilizing wavelet and RNNs is optimal in this scenario.

B. RNN Based Models

Recurrent neural networks (RNNs) have been a popular choice for tasks involving sequential data. Different architectures have been proposed to address the challenge of long-term dependencies in the sequence data. The Simple RNN cell, shown in Fig. 3(a) comprises just one recurrent layer and keeps a hidden state over time steps. Therefore, it is not very effective with long-term dependencies because it suffers from the vanishing gradient problem. The LSTM cell, shown in Fig. 3(b) in turn, consists of memory cells and a more complex architecture of gating mechanisms that include input, output, and forget gates; this design helps control information flow and thus allows LSTMs to keep pertinent information for longer sequences. A variant of the standard LSTM is LSTM with peephole connections, shown in Fig. 3(c). This structure provides an access path from the cell state into the gating mechanism. Hence, it improves the capability of the model to grasp minute temporal dependencies. In contrast, the Gated Recurrent Unit (GRU) cell architecture, shown in Fig. 3(d) reduces the complexity of LSTM by merging input and forget gates into a singular update gate and also merges cell state and hidden state. Fewer parameters translate to lower computational complexity; therefore, in hardware-constrained environments, GRUs are more efficient while learning long-term dependencies just as well.

Among these state-of-the-art models, the one whose mechanism is associated with Eq. (1) to Eq. (6) frequently stood out. By design, this model is highly effective in terms of computational speed and resource efficiency. Consequently, we opted to utilize this model in our implementation.

Algorithm 1 ECG Feature Extraction and Classification

```
Initialize state to IDLE
while not done do
    if state = IDLE then
        if start signal received then
            state ← SEGMENT
        end if
    else if state = SEGMENT then
        Segment the input data
        state ← DOWNSAMPLE
    else if state = DOWNSAMPLE then
        Downsample the segmented data
        state ← DWT
    else if state = DWT then
        Perform DWT on the downsampled data
        state ← PCA
    else if state = PCA then
        Perform PCA on the DWT output
        state ← LSTM_ALPHA1
    else if state = LSTM_ALPHA1 then
        Perform LSTM Alpha1 on the PCA output
        state ← FC_ALPHA1
    else if state = FC_ALPHA1 then
        Apply fully connected layer on LSTM Alpha1 output
        state ← LSTM_ALPHA2
    else if state = LSTM_ALPHA2 then
        Perform LSTM Alpha2 on the PCA output
        state ← FC_ALPHA2
    else if state = FC_ALPHA2 then
        Apply fully connected layer on LSTM Alpha2 output
        state ← LSTM_BETA
    else if state = LSTM_BETA then
        Perform LSTM Beta on the PCA output
        state ← FC_BETA
    else if state = FC_BETA then
        Apply fully connected layer on LSTM Beta output
        state ← BLEND
    else if state = BLEND then
        Blend the fully connected outputs
        state ← MLP
    else if state = MLP then
        Perform MLP on the blended output
        state ← DONE
    else if state = DONE then
        Set done signal
        state ← IDLE
    end if
end while
```

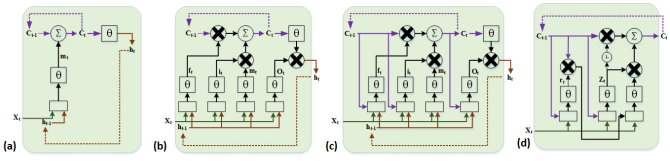


Fig. 3. (a) Simple RNN cell, (b) Long short-term memory (LSTM), (c) LSTM with peepholes, (d) Gated recurrent unit (GRU).

$$f_t[r] = \sigma \times \left(\sum_{c=1}^{M_x} \omega_{xf}[r,c] x_t[c] + \sum_{c=1}^{M_h} \omega_{hf}[r,c] h_{t-1}[c] + b_f[r] \right) \quad (1)$$

$$i_t[r] = \sigma \times \left(\sum_{c=1}^{M_x} \omega_{xi}[r,c] x_t[c] + \sum_{c=1}^{M_h} \omega_{hi}[r,c] h_{t-1}[c] + b_i[r] \right) \quad (2)$$

$$o_t[r] = \sigma \times \left(\sum_{c=1}^{M_x} \omega_{xo}[r,c] x_t[c] + \sum_{c=1}^{M_h} \omega_{ho}[r,c] h_{t-1}[c] + b_o[r] \right) \quad (3)$$

$$g_t[r] = \tanh \left(\sum_{c=1}^{M_x} \omega_{xg}[r,c] x_t[c] + \sum_{c=1}^{M_h} \omega_{hg}[r,c] h_{t-1}[c] + b_g[r] \right) \quad (4)$$

$$c_t[r] = f_t[r] \times c_{t-1}[r] + i_t[r] \times g_t[r] \quad (5)$$

$$h_t[r] = o_t[r] \times \tanh(c_t[r]) \quad (6)$$

In Eq. (1) to Eq. (6), the computational components are responsible for several calculations. The three main gates of an LSTM cell are the forget gate (f_t), input gate (i_t), and output gate (o_t). Additionally, the candidate cell gate, calculated intermittently at each time step, is referred to as g_t. These gates are all represented as vectors. The input vectors, which include x_t, c_t, and ht, represent the current input, cell state, and output vector of the LSTM at each time step. The weight matrices, denoted as $W**$, have specific roles where the first asterisk $(*)$ can be f, i, g, or o, indicating the matrices for the forget, input, candidate cell, or output gates, respectively. The second asterisk $(*)$ can be x or h, signifying that the weight matrix for a particular gate is multiplied by either the input vector x_t or the hidden state h_{t-1}. Finally, the bias vector is represented as b*, where the asterisk $(*)$ can be f, i, g, or o, corresponding to the forget, input, candidate cell, or output gates. In these equations, σ and $\tanh$ (i.e., the sigmoid and hyperbolic tangent functions) are vector-based functions. This means they operate element-wise on the input vectors and return the function's value for each element of the input vectors.

Although simplistic, the implementation of Eq. (1) to Eq. (6) is highly compute-intensive. For instance, if there are M_h neurons in one LSTM layer and it has M_x inputs, then considering the case of forget gates: x_t, the input vector sent to all neurons, would have dimensions M_x×1. h_{t-1}, the vector containing the hidden states of all neurons, and b_f, the bias vector for all gates (forget gates in this case), would have a size of M_h×1. W_{ix}, the weight matrix that connects the incoming input x_t to the forget gates of all the M_h neurons, would have a size of M_h×M_x. These dimensions are because each neuron receiving M_x inputs would need M_h weights for each of the inputs in the current time step. W_{if}, the weight matrix connecting the hidden state to the forget gate, would have dimensions of M_h×M_h because each neuron's forget gate needs to take into account the hidden states of all the neurons from the previous time step. This is done so that for each of the M_h forget gates, there are M_h weights for every hidden state h_{t-1} from the previous time step.

It is clear that Eq. (1) alone performs M_h×(M_x+M_h) multiplications because for each of the M_h elements in W_f*x_t, there would be M_x multiplications, and in $W_f h_{t-1}$, there would be M_h multiplications. Regarding additions, there would be M_h × (M_x+M_h+1) in one time step because for every M_h element in the weighted sum of the input and the hidden state, there would be M_x and M_h additions respectively, and one bias term each. This is quite a heavy burden for a software platform, as storing and retrieving input vectors, weight matrices, hidden states, and intermediate results would require significant memory bandwidth and storage.

C. MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database is a benchmark dataset, widely used, for training and testing algorithms in ECG signal analysis and arrhythmia classification. It contains recordings of human heartbeats; carefully annotated by medical experts. The dataset has 48 half-hour recordings from 47 subjects collected at Beth Israel Hospital Arrhythmia Laboratory. The ECG signals were recorded by using a 2-channel ambulatory ECG device at a sampling frequency of 360 Hz. Annotated R-peaks and heartbeat labels are provided for each recording to identify different kinds of arrhythmia like normal beats, premature ventricular contractions, atrial premature beats, and others. This dataset holds a standard reputation in the medical AI community because of its clinical quality, detailed annotations, and a wealthy variety of heartbeat types. It suits perfectly for deep learning models like LSTM to train on, learn temporal patterns of ECG signals, and classify heartbeats based on those patterns.

IV. Hardware Implementation of LSTM Based CA Classifier

This section describes the internal implementation details of various modules of VLSI design of LSTM-based ECG classification system. The ECG signal processing module implements segmentation and DWT to extract meaningful features from raw ECG signals. The extracted features are handled in two methods, direct features for processing with LSTM model α and those that need PCA for dimensionality reduction prior to processing with LSTM model β. Each LSTM model processes its input sequences and generates feature vectors stored in the LSTM output memory. These stored feature vectors then move on to the FC Layer computation stage, where the output of LSTM model α is input to the first FC layer and the PCA reduced output of LSTM Model β is input to the second FC layer. The outputs from both FC layers are passed to a Multi-Layer Perceptron (MLP) for final classification. These steps are shown in Fig. 1 and Algorithm 1.

A. Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a pre-processing method to extract the time-frequency features of the raw ECG

signals. An architecture based on the four-level decomposition of the Daubechies (db4) wavelet is proposed, which can access essential frequency components of ECG signals. The DWT module is made up of cascaded high-pass and low-pass filters with down-sampling at every stage, as shown in Fig. 4. The detailed coefficients D1 to D4 and an approximation coefficient A4 represent the band-limited versions of the ECG signal for the delta, beta, alpha, and theta bands [41]. The hardware implementation of the DWT module is made efficient by using lattice-form FIR filters [42], [43]. This combines high-pass and low-pass filtering into one computational block and saves hardware. As illustrated in Fig. 5, the DWT comprises high-pass and low-pass filters (HPF, LPF), with transfer functions represented by H(z) and G(z), respectively:

$$G(z) = h_0 + h_1 z^{-1} + h_2 z^{-2} + h_3 z^{-3}$$
$$H(z) = h_3 - h_2 z^{-1} + h_1 z^{-2} - h_0 z^{-3}$$

Here, $h_0 = (1+\sqrt{3})/4\sqrt{2}$, $h_1 = (3+\sqrt{3})/4\sqrt{2}$, $h_2 = (3-\sqrt{3})/4\sqrt{2}$, and $h_3 = (1+\sqrt{3})/4\sqrt{2}$.

To obtain the lattice filter structure and perceive HPF and LPF in one block, H(z) and G(z) can be expressed as:

$$G(z) = (1 + a[0]z^{-1} - a[0]a[1]z^{-2} + a[1]z^{-3})s$$
$$H(z) = (-a[1] - a[0]a[1]z^{-1} - a[0]z^{-2} - z^{-3})s$$

Here, $s = (1+\sqrt{3})/4\sqrt{2}$, $a[0] = (3+\sqrt{3})/4\sqrt{2}s$, and $a[1] = (1-\sqrt{3})/4\sqrt{2}s$.

Clock gating was applied here for efficient down-sampling of the signal. Pipelining and shift-and-add logic reduce latency while maximizing real-time performance.

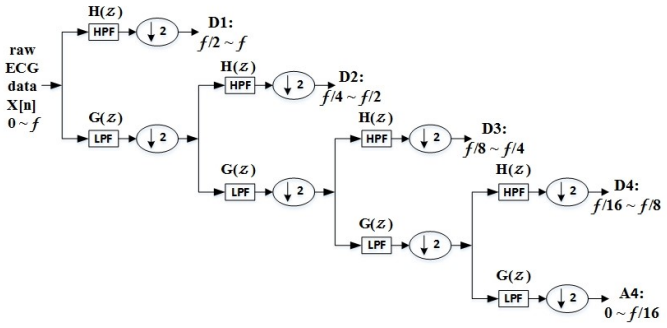


Fig. 4. Structure of the four-level DWT, where ↓2 means even sampling.

B. Principal Component Analysis

The PCA circuit manages to efficiently reduce the dimensionality of the extracted features while keeping all relevant information for classification intact. The design includes weight memory, multiplexers (MUX), a multiply-accumulate (MAC) unit, control logic, and an output memory, as shown in Fig. 6. In weight memory, the PCA main component weights are kept, precomputed from the training data, and they represent the transformation matrix. A structured memory architecture

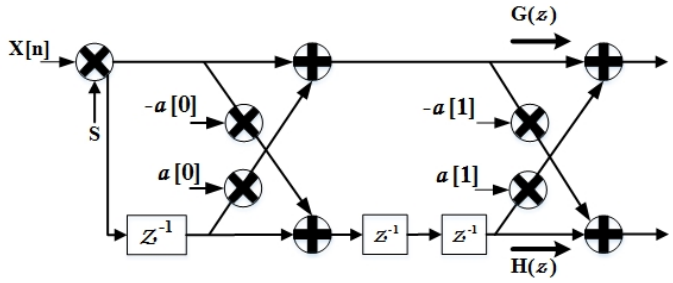


Fig. 5. Lattice FIR filter structure which realizes LPF and HPF in one block.

makes these weights readily accessible to facilitate processing. The MUX units will choose the proper input features from the output of DWT which are given sequentially to ensure proper mapping with corresponding PCA weights. The MAC unit does matrix-vector multiplication where feature values that have been picked are multiplied by weights that correspond to them. The multiplication of these values can accumulate over many clock cycles and thus form the transformed feature set efficiently. Control logic governs the flow of data, synchronizing input selection, weight retrieval, and MAC operations to keep synchronization while preventing computational errors. The feature vector PCA transformation is stored in the output memory.

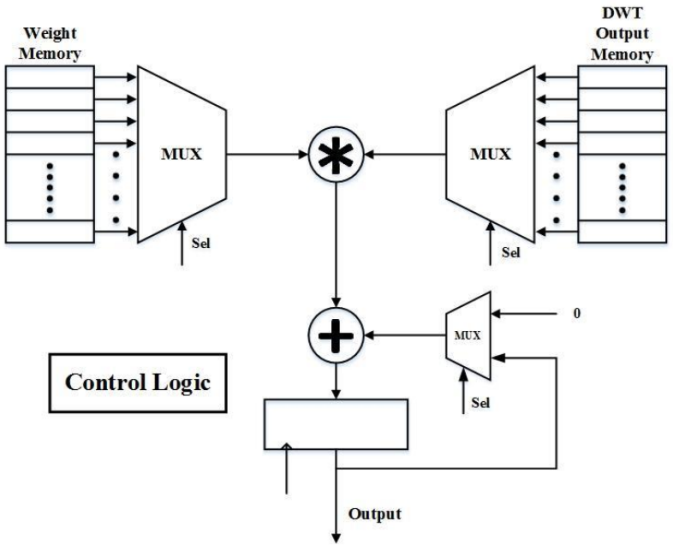


Fig. 6. Circuit design of PCA.

C. Long Short-Term Memory

The hardware realization of an LSTM unit begins with the design of efficient digital circuits for matrix-vector multiplication, nonlinear activation functions, and memory update while considering power and area constraints. The implementation divides itself into three major parts: the computational core, gate computation, and memory update. The LSTM computational unit comprises multiple processing elements that take care of input signals, weight matrices, and recurrent states. An architecture is presented that efficiently computes an input

vector x_t and a previous hidden state h_{t-1} using stored weight matrices. Matrix-vector multiplication has been identified as the primary compute-intensive operation and requires dedicated multipliers and adders. The accumulation stage sums up the result of multiplications and applies activation functions such as sigmoid and hyperbolic tangent (tanh) functions. The outputs of these activation functions are used to control the operations of all three gates in an LSTM, the input gate i_t, forget gate f_t, and output gate o_t. The gate computation module is a crucial part of the LSTM architecture, involving computation of gating signals for control over flow of information. Input data and previous hidden state are accessed from dedicated memory blocks. Then they are multiplexed into matrix-vector multiplication units. Weighted sum of inputs is computed by units that comprise an array of multipliers and accumulators, while terms for bias add-on are kept in a separate memory block. Resulting outputs are subject to a configurable activation module, implementing either sigmoid or tanh based on control logic signals. The main hardware module that implements the LSTM gates is shown in Fig. 7. This module adds non-linearity and allows selective retention of information. The output from this module will be used to drive the gating mechanisms which form the regulation for the internal state of the LSTM. The new cell state c_t and hidden state h_t are calculated by the memory update module for keeping long-term dependencies. The previous cell state is scaled by forgetting gate output while new information is integrated controlled by input gate through element-wise multiplications. The module that computes the c_t and h_t from the results of the gates is shown in Fig. 8.

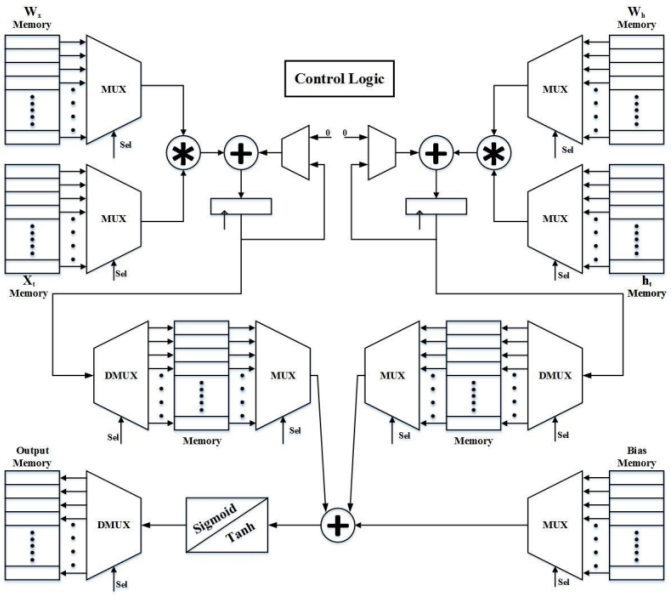


Fig. 7. The main hardware module that implements the LSTM gates.

The hardware implementation of the LSTM is governed by a finite state machine that manages the sequential execution of activities, hence leading to efficient data handling and real-time performance. The finite state machine controls memory read/write operations, weight fetching, matrix multiplications, activation function computation, and state update keeping synchronization among the different stages of computation. The process begins in idle state. This means that the system will remain idle until new input ECG sample data is made available. Immediately after receiving an input ECG sample, weight fetch state is entered wherein the control unit fetches the necessary weight matrices from memory. When the weights have been loaded, matrix multiplication State is entered where matrix-vector multiplication on the input and hidden state vectors with weight matrices is performed. After multiplication, the system goes into the accumulation State. This is when results from multipliers get accumulated and where bias terms are added to form the pre-activation values for the gates in LSTM. These accumulated values get passed through non-linear activation functions in the Activation State, such as sigmoid and hyperbolic tangent (tanh). After this, the FSM moves to the gate update state, where the input, forget, and output gates are calculated based on activated outputs. This step determines how much information will be retained or discarded from memory. The next step is the update of the cell c_t and the hidden state h_t. In the write back state, c_t and h_t are written back into memory with time step 2 so that these values are available for time step 3. This FSM guarantees the execution of all calculations in a pipelined manner, hence facilitating the overlapping stages involved in the LSTM process, and maximizing throughput. Optimization of state transitions leads to low latency and power usage. The structured control allows for a smooth flow of data, positively impacting hardware efficiency.

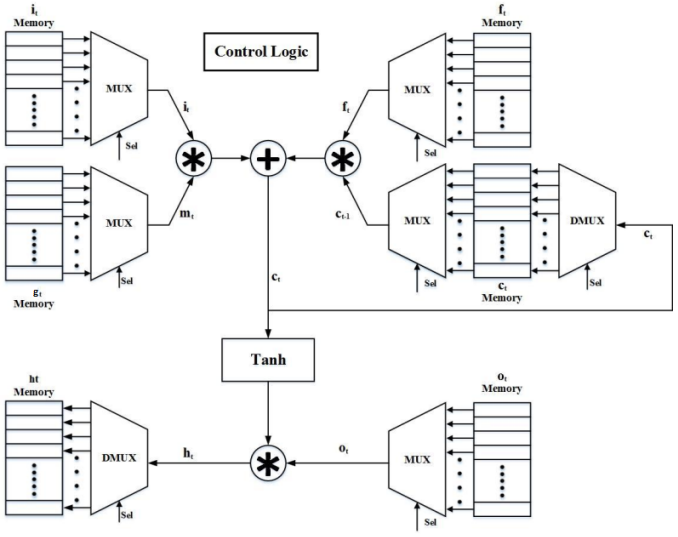


Fig. 8. The module that computes the c_t and h_t from the results of the gates.

D. Activation Functions

The activation functions are essential to the performance and convergence of the LSTM networks. The conventional sigmoid and hyperbolic tangent functions introduce substantial computational complexity due to their nonlinear and transcendental nature, making direct hardware implementation expensive. Therefore, hardware friendly approximations are adopted. The hard sigmoid and hard tanh, which keep the computational overhead low while maintaining competitive accuracy. Hard sigmoid and hard tanh, that provide linear approximations greatly simplify hardware implementation. These functions are defined in Eq. (7) and Eq. (8) [44]:

$$\text{hardtanh}(x) = \begin{cases} -1, & x < -1 \\ x, & -1 \le x \le 1 \\ 1, & x > 1 \end{cases} \quad (7)$$

$$\text{hardsigmoid}(x) = \begin{cases} 0, & x < -2.5 \\ 0.2 \times x + 0.5, & -2.5 \le x \le 2.5 \\ 1, & x > 2.5 \end{cases} \quad (8)$$

These approximations replace simple piecewise linear functions in sigmoid and tanh that approximate exponential computations to meet the same performance while it results in huge savings in FPGA resources utilization and latency. Whereas the hard sigmoid has a linear scaling factor and bias, hard tanh uses saturation-based clipping to ensure stability. In the LSTM-based model of the ECG classification, we tried both standard sigmoid + tanh and hard sigmoid + hard tanh. The effect on detection accuracy over 50 training epochs is shown in Fig. 9. Where hard approximations result in a small decrease in accuracy, vast improvement in hardware efficiency fully justifies this trade off.

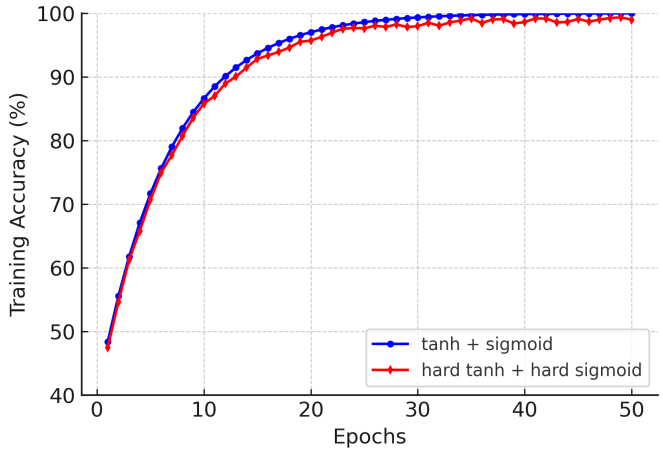


Fig. 9. Impact of activation functions on classification accuracy.

E. Fully Connected Layer

The fully connected layer is very important in changing the feature representations that were acquired by LSTM models into the final classification results. The hardware implementation of the Fully Connected layer, as can be seen in Fig. 10, includes several major components. Weight memory keeps the FC-layer weights in a form that was learned prior to the training of this mode. These weights will be picked up by using a multiplexer (MUX) according to control logic. LSTM output memory saves output vectors from LSTM models, these are used as input to the FC layer. A multiplication unit computes the weighted sum by performing element-wise multiplication of input features with weight values. An accumulation and summation unit adds all the results of multiplication using an adder tree to provide the final activation values for neurons. A bias addition stage retrieves the bias terms from a special bias memory and adds them before activation. The last element is the control logic, which guarantees the proper selection of weights, inputs, and all other parameters used in computations.

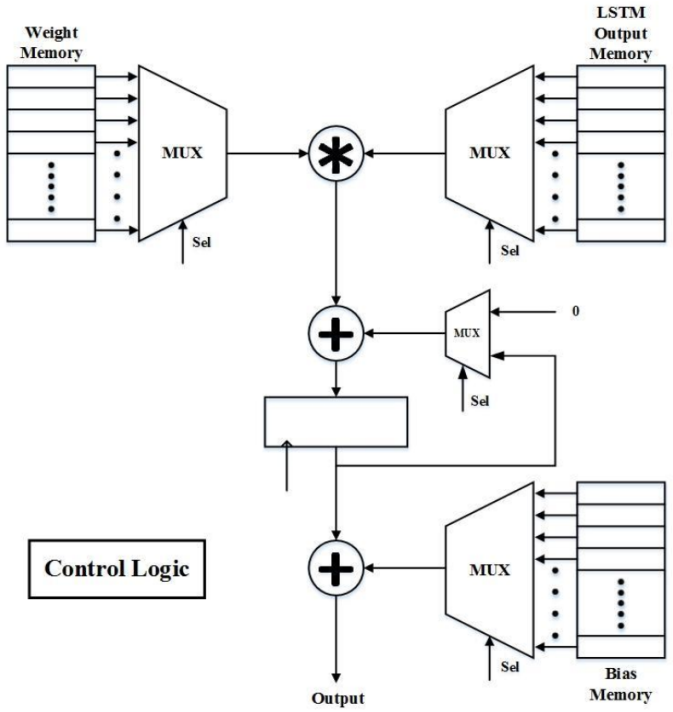


Fig. 10. The module that implements FC layer.

F. Multi-Layer Perceptron

After feature extraction and classification with LSTM and Fully Connected (FC) layers, the output of the FC layer will be used as input to a multi-layer perceptron (MLP) for final classification. The size of MLP is 80x10x7 as shown in Fig. 1. Each layer has multiple neurons, in which each neuron performs multiply-accumulate (MAC) operation. This is expressed as $y = \sum_{i=1}^{N} W_i x_i + B_i$. Here, W_i and B_i denote the weight and bias of the neuron, respectively. The computational complexity and hardware cost of MLPs depend on their number of neurons and MAC unit efficiencies. A finite state machine (FSM) controls the hardware implementation of MLP, coordinating operations between layers. When the enable signal is set high, the output of the FC layer (size 80) will be fed into the MLP input layer. The control unit will activate the FSM of each layer in a sequential manner, thus processing the inputs serially. The computed outputs will propagate through hidden layers and get to the final output layer, where the classification decision will be made. Achieving an optimized MLP architecture allows the designer to minimize hardware resources for ECG classification on wearable devices.

V. Hardware-Software Co-design Implementation

The hardware-software co-design methodology is a design approach used in embedded system development and digital system development where hardware and software components are developed in parallel rather than in sequence. This results in an optimization of performance, cost, power, and flexibility by proper partitioning of the functionalities between hardware, for example, FPGA, ASIC and software running on a processor [45]–[48]. Fig. 11 shows the architecture of the implementation of hardware-software co-design on the

ZYNQ-7000 SoC, which incorporates both the Processing System (PS) and the Programmable Logic (PL). The ARM Cortex-A9 processor available in PS is responsible for ECG signal acquisition and preprocessing like segmentation and wavelet transformation. The output of the pre-processed ECG data is passed to PL where LSTM hardware acceleration supports two different neural network models; Model α with 30 hidden units and Model β with 50 hidden units. These two models extract temporal features from the ECG data. A feature fusion of those outputs is through a Classification Blending unit based on a Multi-Layer Perceptron (MLP). The ultimate categorization outcomes are shown on an OLED screen using the SSD1306 controller, while also being sent to a UART interface for troubleshooting through the Vitis Terminal. This setup demonstrates a mix of hardware and software co-design that uses the advantages of both the ARM processor and FPGA fabric for effective, real-time ECG classification. The Processing System IP is the software interface surrounding the Zynq-7000 Processing System. The Zynq-7000 family includes a system-on-chip (SoC) style integrated processing system (PS) and Programmable Logic (PL), unit it provides an extensible and flexible SoC solution on single die for efficient, real-time ECG classification.

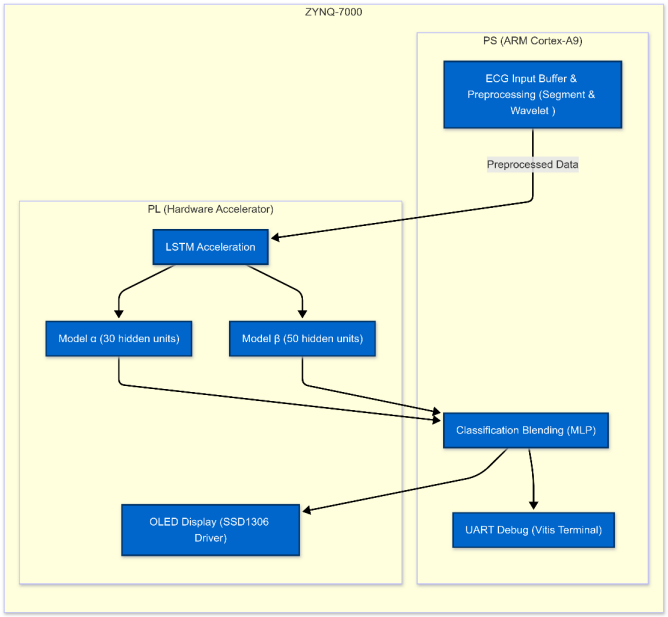


Fig. 11. Hardware-software co-design on the ZYNQ-7000 SoC.

The Zynq-7000 family comprises a system-on-chip (SoC) style integrated processing system (PS) and Programmable Logic (PL), an extensible and flexible SoC solution. The ECG LSTM Model IP is a custom hardware accelerator shown in Fig. 12, for time-series analysis specifically with Long Short-Term Memory (LSTM) networks on ECG signals. This IP core has been synthesized using Xilinx Vivado and deployed in the Programmable Logic (PL) region of Zynk-7000 SoC as a dedicated co-processor for sequential data classification.

The architecture of the module comprises a control interface module, data transfer module, and hardware-mapped LSTM computation engine optimized for low-latency as well as high-throughput inference. IP core supports AXI inter-

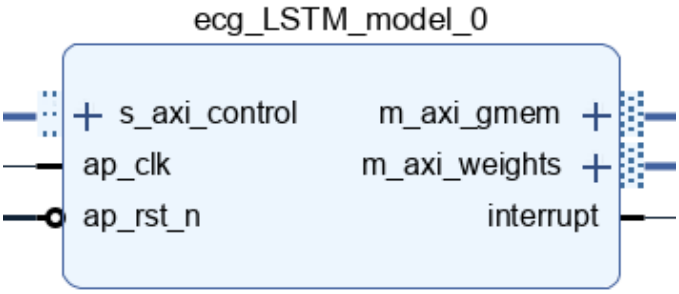


Fig. 12. LSTM based ECG custom IP.

faces for seamless communication with the Processing System (PS). This one, s-axi-ontrol interface, enables it to allow the ARM Cortex-A9 processor in the PS to configure, start, and monitor the LSTM accelerator through control registers. The m-axi-gmem interface streams pre-processed ECG data from external memory. The m-axi-weights interface dynamically loads trained LSTM model parameters (weights and biases) supporting multiple LSTM configurations, Model α (30 hidden units) and Model β(50 hidden units) without needing IP re-synthesis. The interrupt signal notifies the PS data blending or further processing steps that inference is complete.

This IP core has been designed for use in the real-time healthcare signal classification and optimized to manage the recurrent characteristic of LSTM calculation. It is functionally organized to handle time-series segments in a pipelined way that decreases the burden on the PS and allows parallel execution among different model variants. The PS takes care of ECG signal preprocessing (for example, segmentation and wavelet transformation) loading model weights, and post-inference tasks such as classification fusion and output display. After importing these IP cores, we need to automate the connections Xilinx will automatically connect some of the IP cores and the remaining ports should be connected manually. Fig. 13 shows the block diagram of LSTM based ECG hardware-software co-design system implemented using AMD Vivado.

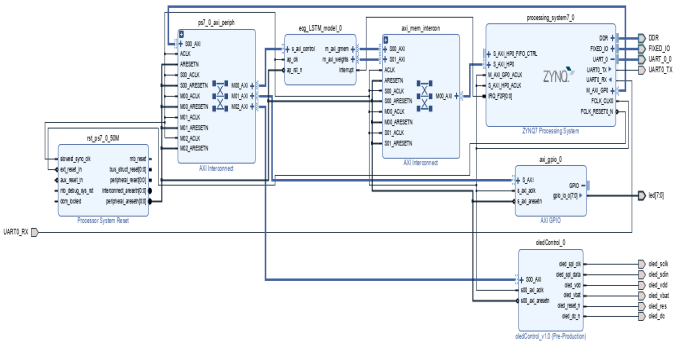


Fig. 13. Block diagram of LSTM based ECG Hardware-Software Co-design system.

VI. Hardware Utilization

The algorithm is first implemented using Python and TensorFlow. To evaluate performance and comparison with previous studies, the MIT-BIH ECG arrhythmia database is used. A pre-trained network is deployed in hardware using Verilog HDL. Synthesis is done using Xilinx Vivado Tool. Comparison of two FPGA-based ECG classification implementations is summarized in Table I.

TABLE I. Comparison of Two FPGA-based ECG Classification Implementations

	Complete HW	HW-SW Co-design
LUTs	2383	1600
FF	2156	1444
DSP	20	14
BRAM	22	15
Power (mW)	41	29
Frequency (MHz)	54	61
Accuracy	99%	99%

The power consumption distribution of complete hardware implementation is shown in Fig. 14, explaining the allocation among different components. It can be seen that power consumption is led mostly by the LSTM-based models.

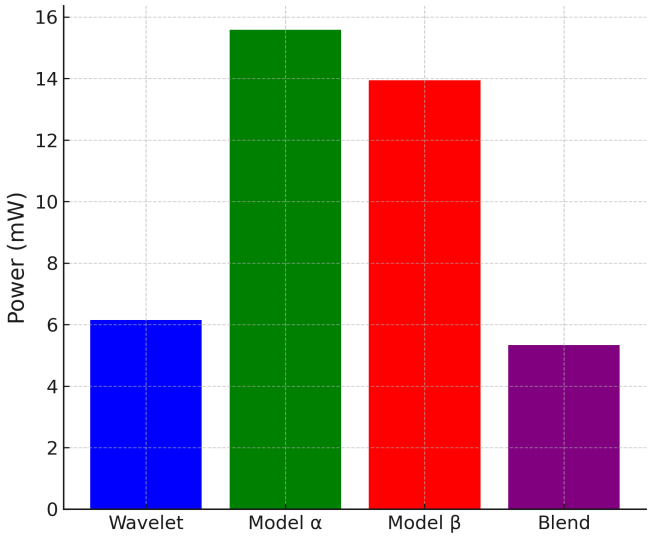


Fig. 14. Power consumption distribution.

In software simulation of neural networks, the weights and biases are represented in floating-point arithmetic. Floating-point arithmetic, however, is computationally very expensive and consumes relatively high power compared to fixed-point arithmetic. In order to minimize distortion, the weights and biases are represented using an ideal word length selected from software simulations, shown in Fig. 15. A 3.12 signed fixed-point format is selected, consisting of 1 sign bit, 3 integer bits, and 12 fractional bits.

The implemented method uses a patient-specific training paradigm. Therefore, the model is trained separately for each patient. Once the model has been trained, it can be used for real-time ECG surveillance and heartbeat classification for that specific patient. The method depicted in Fig. 16, demonstrates how training occurs just once for each patient

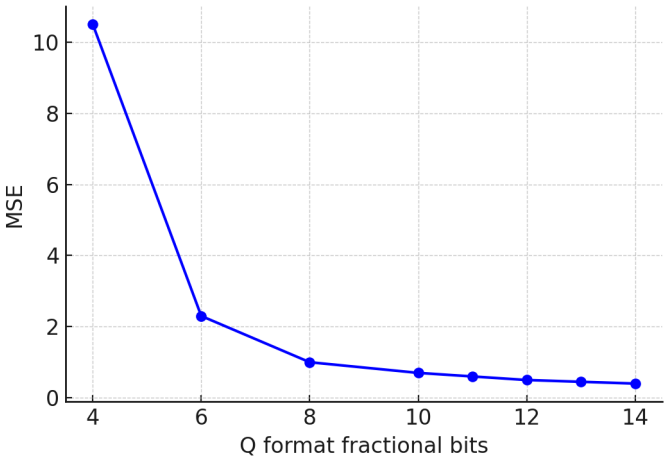


Fig. 15. MSE of network with various fractional bits.

and does not require ongoing updates. The training dataset for every patient is made up of twofold local ECG data and global ECG data, as illustrated in Fig. 16. The local dataset is provided for the patient, enhancing classification accuracy through the exploitation of intrinsic similarities between their heartbeats. According to AAMI standards, this dataset may last for a maximum of five minutes. However, the global dataset is shared among all patients and it contains representative heartbeats for different classes of arrhythmias. Thus, it helps the model generalize those patterns which are not present in the local dataset. Other studies have considered training patient-specific models [8]. An alternative approach is to train a single model across data from multiple patients and then use this model to classify ECG signals from new patients. This approach is not adapted because of the large variability between patients in ECG waveforms which may adversely affect classification accuracy [10].

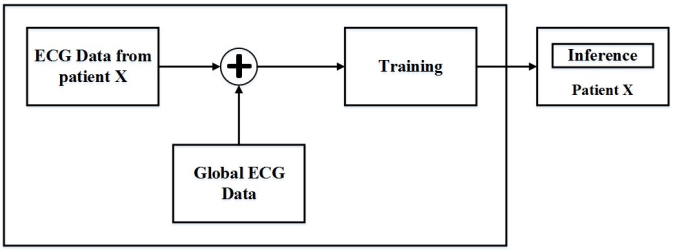


Fig. 16. Patient-specific training.

VII. Conclusion

We presented Cardio-Edge, a low-power real-time ECG classification system for health monitoring wearables. Optimized digital hardware for DWT, PCA, LSTM, and MLP modules is integrated into a hardware-software co-design on the ZYNQ-7000 platform to demonstrate a robust architecture capable of achieving high classification accuracy with minimal computational overhead. The HW-SW co-design offers 10× speedup over SW-only implementations on the ARM Cortex-A9 processor with no compromise in detection performance. The use of fixed-point arithmetic, hardware-friendly activation

approximations, and pipelined LSTM processing ensures energy as well as area efficiency. Our patient-specific training approach improves accuracy further by customizing the model to individual heartbeat profiles. On-device execution ensures secure continuous low-latency cardiac monitoring without reliance on remote servers and constant internet connectivity. This work takes a major stride forward in making intelligent biomedical systems deployed within constrained wearable platforms.

REFERENCES

[1] WHO, "Cardiovascular diseases (cvds)," accessed on 19 Feb, 2025. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] J. M. Bote, J. Recas, F. Rincón, D. Atienza, and R. Hermida, "A modular low-complexity ecg delineation algorithm for real-time embedded systems," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 429–441, 2018.

[3] M. Janveja, R. Parmar, S. Dash, J. Pidanic, and G. Trivedi, "A low-power co-processor to predict ventricular arrhythmia for wearable healthcare devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 9, pp. 1672–1683, 2024.

[4] L. Chen, Z. Jiang, J. Barker, H. Zhou, F. Schlindwein, W. Nicolson, G. A. Ng, and X. Li, "Ecgvednet: A variational encoder-decoder network for ecg delineation in morphology variant ecgs," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 7, pp. 2143–2153, 2024.

[5] M. Janveja, R. Parmar, and G. Trivedi, "Minsc: A vlsi architecture for myocardial infarction stages classifier for wearable healthcare applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 3, pp. 1159–1163, 2023.

[6] A. L. Bui and G. C. Fonarow, "Home monitoring for heart failure management," *Journal of the American College of Cardiology*, vol. 59, no. 2, pp. 97–104, 2012.

[7] T. Teijeiro, P. Félix, J. Presedo, and D. Castro, "Heartbeat classification using abstract features from the abductive interpretation of the ecg," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 409–420, 2018.

[8] T. Ince*, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ecg signals," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1415–1426, 2009.

[9] R. Hoekema, G. Uijen, and A. van Oosterom, "Geometrical aspects of the inter-individual variability of multilead ecg recordings," in *Computers in Cardiology 1999. Vol.26 (Cat. No.99CH37004)*, 1999, pp. 499–502.

[10] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2016.

[11] Z. Chen, J. Luo, K. Lin, J. Wu, T. Zhu, X. Xiang, and J. Meng, "An energy-efficient ecg processor with weak-strong hybrid classifier for arrhythmia detection," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 7, pp. 948–952, 2018.

[12] M. A. Sohail, Z. Taufique, S. M. Abubakar, W. Saadeh, and M. A. Bin Altaf, "An ecg processor for the detection of eight cardiac arrhythmias with minimum false alarms," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[13] P. Li, Y. Wang, J. He, L. Wang, Y. Tian, T.-s. Zhou, T. Li, and J.-s. Li, "High-performance personalized heartbeat classification model for long-term ecg signal," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 1, pp. 78–86, 2017.

[14] M. L. Hoang, "A review of developments and metrology in machine learning and deep learning for wearable iot devices," *IEEE Access*, vol. 13, pp. 106 035–106 054, 2025.

[15] H. Sun, D. Luo, X. Niu, X. Zeng, B. Zheng, H. Liu, and J. Pan, "Classification algorithms in automatic diagnosis of ecg arrhythmias: A review," *IEEE Access*, vol. 12, pp. 191 921–191 935, 2024.

[16] H. Gao, L. Yan, X. Li, and Z. Zhang, "Research on the recognition of ecg signals based on integrated lstm-gru network," in *2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, 2024, pp. 1088–1091.

[17] F. Quirós-Corella, R. Loaiza, R. Matarrita, and E. Meneses, "A comprehensive deep learning pipeline for arrhythmia multi-classification with electrocardiography data," in *2024 IEEE 6th International Conference on BioInspired Processing (BIP)*, 2024, pp. 1–6.

[18] S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 515–523, 2020.

[19] N. Akhtar, J. Fan, A. R. Buzdar, M. Ahmed, and A. Raza, "Vlsi design of lstm-based ecg classification for continuous cardiac monitoring on wearable devices," *Electronics Letters*, vol. 61, no. 1, p. e70269, 2025.

[20] Arpan, M. Singh, P. Garg, S. Srivastava, and A. K. Saggu, "Revolutionizing arrhythmia classification: Unleashing the power of machine learning and data amplification for precision healthcare," in *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 2024, pp. 516–522.

[21] H. T. Al-mousa A, Baniissa J, "Enhanced electrocardiogram machine learning-based classification with emphasis on fusion and unknown heartbeat classes," *DIGITAL HEALTH*, vol. 9, 2023.

[22] A. T. Hassan SU, Mohd Zahid MS, "Classification of cardiac arrhythmia using a convolutional neural network and bi-directional long short-term memory," *DIGITAL HEALTH*, vol. 8, 2022.

[23] J. Yang, J. Li, K. Lan, A. Wei, H. Wang, S. Huang, and S. Fong, "Multi-label attribute selection of arrhythmia for electrocardiogram signals with fusion learning," *Bioengineering*, vol. 9, no. 7, 2022.

[24] D. H. Verspoor K, "Electrocardiogram arrhythmia detection with novel signal processing and persistent homology-derived predictors," *Data Science*, vol. 7, no. 1, pp. 29–53, 2024.

[25] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. b. M. Pauzi, Q. Zhang, and P. Y. Lim, "Deep learning-based ecg arrhythmia classification: A systematic review," *Applied Sciences*, vol. 13, no. 8, 2023.

[26] P. K. Tyagi, N. Rathore, and D. Agrawal, "A review on heartbeat classification for arrhythmia detection using ecg signal processing," in *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2023, pp. 1–6.

[27] M. Z. Chao Che, Peiliang Zhang, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Med Inform Decis Mak*, vol. 21, no. 184, 2021.

[28] S. Bi, R. Lu, Q. Xu, and P. Zhang, "Accurate arrhythmia classification with multi-branch, multi-head attention temporal convolutional networks," *Sensors*, vol. 24, no. 24, 2024.

[29] L.-H. Wang, Y.-T. Yu, W. Liu, L. Xu, C.-X. Xie, T. Yang, I.-C. Kuo, X.-K. Wang, J. Gao, P.-C. Huang, S.-L. Chen, W.-Y. Chiang, and P. A. R. Abu, "Three-heartbeat multilead ecg recognition method for arrhythmia classification," *IEEE Access*, vol. 10, pp. 44 046–44 061, 2022.

[30] B. K. Nugraha, Q. D. Amalia, A. S. Safitri, A. Rizal, and H. T. Fauzi, "Comparison analysis for life-threatening arrhythmia classification from ecg data using machine learning and deep learning methods," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2024, pp. 1–6.

[31] E. Essa and X. Xie, "Multi-model deep learning ensemble for ecg heartbeat arrhythmia classification," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1085–1089.

[32] C. Zhang, J. Chang, Y. Guan, Q. Li, X. Wang, and X. Zhang, "A low-power ecg processor asic based on an artificial neural network for arrhythmia detection," *Applied Sciences*, vol. 13, no. 17, 2023.

[33] H. T. Tefai, H. Saleh, T. Tekeste, M. Alqutayri, and B. Mohammad, "Asic implementation of a pre-trained neural network for ecg feature extraction," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[34] I. Hoyer, A. Utz, A. Lüdecke, H. Kappert, M. Rohr, C. H. Antink, and K. Seidl, "Design of hardware accelerators for optimized and quantized neural networks to detect atrial fibrillation in patch ecg device with risc-v," *Sensors*, vol. 23, no. 5, 2023.

[35] W. Liu, Q. Guo, S. Chen, S. Chang, H. Wang, J. He, and Q. Huang, "A fully-mapped and energy-efficient fpga accelerator for dual-function ai-based analysis of ecg," *Frontiers in Physiology*, vol. 14, 2023.

[36] S. K. Varadharajan and V. Nallasamy, "Implementation of field programmable gate array (fpga) based distributed arithmetic gated current unit to achieve high ecg diagnosis rate," *Journal of Nanoelectronics and Optoelectronics*, vol. 17, no. 1, pp. 82–89, 2022.

[37] A. Gon and A. Mukherjee, "Design and fpga implementation of an efficient architecture for noise removal in ecg signals using lifting-based wavelet denoising," in *2023 11th International Symposium on Electronic Systems Devices and Computing (ESDC)*, vol. 1, 2023, pp. 1–6.

[38] J. Guo, W. Li, and H. Huang, "An ecg detection device based on convolutional neural network," in *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2023, pp. 860–864.

[39] M. A. Scrugli, D. Loi, L. Raffo, and P. Meloni, "An adaptive cognitive sensor node for ecg monitoring in the internet of medical things," *IEEE Access*, vol. 10, pp. 1688–1705, 2022.

[40] S. Deng, B. L. den Ouden, T. De Coster, C. I. Bart, W. H. Bax, R. H. Poelma, A. A. de Vries, G. Q. Zhang, V. Portero, and D. A. Pijnappels, "An untethered heart rhythm monitoring system with automated ai-based arrhythmia detection for closed-loop experimental application," *Advanced Sensor Research*, vol. 3, no. 11, p. 2400057, 2024.

[41] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. b. M. Pauzi, Q. Zhang, and P. Y. Lim, "Deep learning-based ecg arrhythmia classification: A systematic review," *Applied Sciences*, vol. 13, no. 8, 2023.

[42] P. K. Tyagi, N. Rathore, and D. Agrawal, "A review on heartbeat classification for arrhythmia detection using ecg signal processing," in *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2023, pp. 1–6.

[43] M. Z. Chao Che, Peiliang Zhang, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Med Inform Decis Mak*, vol. 21, no. 184, 2021.

[44] Hasib-Al-Rashid, N. K. Manjunath, H. Paneliya, M. Hosseini, W. D. Hairston, and T. Mohsenin, "A low-power lstm processor for multi-channel brain eeg artifact detection," in *2020 21st International Symposium on Quality Electronic Design (ISQED)*, 2020, pp. 105–110.

[45] A. R. Buzdar, L. Sun, A. Latif, and A. Buzdar, "Distance and speed measurements using fpga and asic on a high data rate system," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 6, no. 10, 2015. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2015.061037

[46] A. R. Buzdar, L. Sun, M. W. Azhar, M. I. Khan, and R. Kashif, "Area and energy efficient viterbi accelerator for embedded processor datapaths," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 3, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080355

[47] A. R. Buzdar, A. Latif, L. Sun, and A. Buzdar, "Fpga prototype implementation of digital hearing aid from software to complete hardware design," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 1, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.070188

[48] A. R. Buzdar, L. Sun, R. Kashif, M. W. Azhar, and M. I. Khan, "Cyclic redundancy checking (crc) accelerator for embedded processor datapaths," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 2, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080242

Robust Particle Filter for Accurate WiFi-Based Indoor Positioning in the Presence of Outlier-Corrupted Sensor Data

Mohamed Aizad Bin Mohamed Ghazali[1], Aroland Kiring[2*], Lyudmila Mihaylova[3],
Hoe Tung Yew[4], Seng Kheau Chung[5], Farrah Wong[6]
Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia[1, 2*, 4, 5, 6]
Department of Automatic Control and System Engineering, University of Sheffield, Sheffield, UK[3]

***Abstract*—This study presents a comprehensive evaluation of an outlier-robust particle filter (RPF) designed to improve indoor positioning accuracy in complex environments with substantial measurement noise and outliers. The RPF's performance is benchmarked against a standard Particle Filter (PF) using both simulated and real-world datasets. Simulation results indicate that the RPF consistently outperforms the PF in indoor positioning particularly when sensor measurements contain outliers, achieving significant reductions in root mean square error (RMSE) for position, velocity, and acceleration estimation, with improvements of approximately 40.02%, 38.48%, and 65.80%, respectively. Real-world experiments, applying a calibrated log-normal path loss model to Wi-Fi received signal strength (RSS) data, further corroborate the RPF's effectiveness, demonstrating a 93.61% improvement in positioning accuracy compared to the PF. These findings highlight the RPF's robustness in delivering high accuracy, especially in environments with measurement outliers, establishing it as a reliable solution for indoor tracking in noisy sensor environments.**

***Keywords*—Complex environments; indoor positioning; measurement noise and outliers; RMSE reduction; robust particle filter**

I. Introduction

Indoor Positioning Systems (IPS) have become increasingly important in today's technology-driven world, enabling accurate positioning of people and objects within indoor spaces [1]. Unlike outdoor positioning, which has been revolutionized by the Global Positioning System (GPS) [2], indoor environments present unique challenges that GPS cannot effectively address. These challenges include signal loss due to walls and obstacles, multipath interference, and the complexity of vertical positioning in multi-story buildings. As a result, IPS has become a significant area of research, essential for applications such as location-based services (LBS), asset tracking, and indoor navigation [3]. Various technologies support IPS, each with its strengths and limitations. Common technologies include Bluetooth Low Energy (BLE), Infrared, Ultra-Wideband (UWB), Magnetic Fields, as well as Wi-Fi. BLE is known due to its low power consumption and suitability for proximity-based applications; however, although its precision is susceptible to signal interference and environmental conditions [4]. Infrared systems offer high accuracy but require a direct line of sight between the transmitter and receiver, making them less practical in dynamic environments [5]. UWB provides high precision but demands specialized hardware, making large-scale deployment costly [6]. Magnetic field-based positioning can function without additional infrastructure but is sensitive to environmental changes and may require frequent recalibration [7].

Among these technologies, Wi-Fi has emerged as a practical and widely used solution for IPS [8]. Operating at 2.4 GHz and 5 GHz frequency bands conforming to IEEE 802.11 protocol, Wi-Fi is common in residential, commercial, and industrial environments. This existing Wi-Fi infrastructure can be leveraged for indoor positioning, reducing the need for additional hardware. Wi-Fi's ability to penetrate obstacles enhances its versatility and cost-effectiveness for indoor applications [9]. To improve the accuracy of Wi-Fi-based IPS, techniques such as Received Signal Strength (RSS) fingerprinting [10], Time of Arrival (ToA) [11], and Angle of Arrival (AoA) [12] are employed. Although AoA and UWB methods achieve higher absolute accuracy, RSS-based IPS is prioritized due to its simplicity (no extra hardware), use of existing Wi-Fi infrastructure, and ease of deployment. However, challenges such as signal fluctuations due to multipath effects and interference from other electronic devices can impact reliability. Accurate position estimation in IPS relies heavily on processing sensor data, which is often affected by noise and outliers [13]. Filtering algorithms are used to refine raw sensor data and improve positioning accuracy. Common algorithms include Kalman Filters (KF), Extended Kalman Filters (EKF), Unscented Kalman Filters (UKF), and Particle Filters (PF).

The KF is optimal for linear systems with Gaussian noise, providing efficient and accurate state estimates [14]. However, real-world environments often involve non-linearities and non-Gaussian noise, limiting the KF's effectiveness. The EKF addresses non-linear systems by linearizing it around current estimates, but this approach can yield suboptimal results in highly non-linear settings [15]. The UKF improves on the EKF by employing deterministic sampling to better represent the state mean and covariance, enhancing performance in non-linear cases [16]. Both EKF and UKF assume Gaussian noise, which is not always realistic in indoor environments prone to outliers. PFs provide a flexible alternative for non-linear, non-Gaussian systems [17]. Using a set of weighted particles to represent the system state distribution, PFs handle diverse noise types and complex dynamics effectively. It operate through a recursive prediction-update process: particles are propagated based on system dynamics in the prediction step, and its weights are adjusted during the update step based on observed measurements. This adaptability makes PFs well-

suited for indoor positioning, where signal propagation is irregular due to obstacles, reflections, and multipath effects. Despite their robustness, PFs face challenges such as weight degeneracy, where most particles contribute minimally to the state estimate after several iterations [18]. Resampling helps mitigate this issue, but increases computational load [19]. Nonetheless, PFs' ability to model non-linearities and manage non-Gaussian noise conditions makes them an ideal choice for complex indoor environments, outperforming Kalman-based filters under such conditions [20].

PFs are widely used in robotics, target tracking, and autonomous vehicles due to their ability to represent state distributions accurately. However, this capability comes with challenges. Maintaining and updating a large number of particles is computationally intensive, making real-time applications or systems with limited processing power difficult to manage [21]. The process of propagating and re-weighting particles adds complexity, especially when rapid updates are required. Weight degeneracy, where most particles contribute minimally to the state estimate, further degrades performance, particularly in noisy environments or when measurements include outliers [22]. Resampling helps address weight degeneracy, but adds to the computational load. In IPS, PFs are effective in managing non-linearities and non-Gaussian noise but are sensitive to outliers in sensor measurements. Outliers, caused by signal reflections, interference, or hardware malfunctions, distort state estimation by skewing particle weights [23]. This can result in inaccurate state estimates, as particles with incorrect weights disproportionately influence the distribution. Outliers exacerbate weight degeneracy by concentrating the majority of the weight on a few particles, reducing the effective number of particles and degrading the filter's performance. It can also mislead the PF, causing particles to cluster around incorrect states and reducing diversity in the particle set. This loss of diversity limits the filter's ability to explore the true state space, compromising accuracy and reliability. Consequently, PFs are less effective in environments with frequent outliers, reducing their ability to provide consistent and precise positioning estimates.

Given these challenges, improving the robustness of PFs to effectively handle outliers is essential for enhancing indoor positioning accuracy. This requires developing strategies to detect and mitigate the impact of outliers, ensuring that the PF remains accurate even with contaminated data. To address these challenges, this study introduces the outlier-robust Particle Filter (RPF), which is designed to improve the accuracy and robustness of indoor positioning in the presence of outliers. The RPF utilizes an adaptive noise model that dynamically switches between Gaussian and Student's t-distribution noise models based on the inferred likelihood of outliers in the sensor measurements. Under normal conditions, where measurements mainly contain Gaussian noise, the filter employs a Gaussian distribution [24]. However, when outliers are detected, it switches to a Student's-T distribution [25], providing increased resilience against extreme values. Unlike prior robust PF approaches that rely solely on Student's-T noise models in [25] and [26], the RPF integrates the Singer acceleration motion model, which inherently handles non-linear movement by jointly modeling position, velocity, and acceleration, and a log-normal path-loss sensor model to convert RSS into distance. This two-fold design explicitly accounts for both Gaussian and heavy-tailed (non-Gaussian) measurement noise. The effectiveness of the proposed RPF is validated through simulations and real-world experiments conducted in residential indoor environments. These experiments demonstrate the RPF's superior performance in positioning accuracy, robustness against outliers, and its ability to handle non-linear movements and non-Gaussian noise, outperforming standard PF methods. By addressing key challenges such as measurement noise, weight degeneracy, and particle diversity in the presence of outliers, this study presents a more robust and reliable solution for Wi-Fi-based IPS. The RPF marks a significant advancement in indoor positioning, addressing the limitations of PFs in real-world environments.

The remainder of this study is organized as follows: Section II introduces the system model and notation, beginning with the Singer acceleration motion model defining the state vector, transition matrix, and control and noise inputs and then presenting the RSS-based path-loss measurement framework, which includes both the multivariate Gaussian distribution for nominal noise and the multivariate Student's-T distribution to handle outliers. Section III formulates the filtering problem by first reviewing the standard particle filter covering sampling, weight update, and normalization for nonlinear, non-Gaussian state estimation and then detailing the proposed outlier-robust particle filter, which employs a dual-noise likelihood framework to switch between Gaussian and Student's-T hypotheses and incorporates Metropolis resampling to preserve particle diversity, as summarized in Algorithm 2. Section IV presents experiments and results under Wi-Fi RSS measurements, describing simulated data scenarios including 2D trajectory generation, noise and outlier injection, and RMSE performance metrics and real-world experiments using the ESP-32-WROOM-32 in a residential indoor setting with calibrated log-normal path-loss and positioning accuracy comparisons. Section V summarizes the key findings highlighting the RPF's superior robustness and accuracy and outlines directions for future research.

A. Motivation and Contribution

This work addresses the critical need to enhance the robustness of PFs against outliers while maintaining computational efficiency, a key requirement for practical IPS. Traditional PFs struggle with non-Gaussian noise and measurement outliers, which can severely degrade performance. To overcome these challenges, the RPF is introduced, featuring an adaptive noise model that dynamically switches between Gaussian and Student's-T distributions based on outlier detection. This approach ensures resilience to extreme values while preserving accuracy. The contributions of this work are threefold. Firstly, the RPF is specifically designed to mitigate the adverse effects of outliers in Wi-Fi-based IPS, addressing a significant challenge in environments where non-Gaussian measurement noise is prevalent. Secondly, the performance of the RPF is rigorously evaluated through extensive testing on both simulated datasets and real-world experiments, demonstrating its superior accuracy and robustness compared to traditional methods. Thirdly, a calibrated log-normal path loss model is integrated into the algorithm to provide a more accurate representation of indoor signal propagation, further enhancing reliability and positioning accuracy.

II. System Model

Notations: Bold lowercase letters (e.g., $\mathbf{x}$) are used to represent vectors, while bold uppercase letters (e.g., $\mathbf{A}$) denote matrices. Scalars are indicated with standard italic letters (e.g., x). The symbol $\mathbb{R}$ represents the set of all real numbers, and $\mathbb{R}^{m\times n}$ signifies the space of $m \times n$ real matrices. The identity matrix is written as $\mathbf{I}$, and the determinant of a matrix $\mathbf{A}$ is expressed as $|\mathbf{A}|$. A Multivariate Gaussian distribution, fully defined by its mean μ and covariance matrix Σ, is represented as $\mathcal{N}(\mu, \Sigma)$. The transpose operation on a matrix or vector is indicated with the superscript T.

A. Motion Model

Consider the trajectory of a mobile user in a two-dimensional (2D) floor plan, described by the Singer state-space model [27], as in Eq. (1):

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}_u\mathbf{u}_k + \mathbf{B}_w\mathbf{w}_k, \tag{1}$$

where, $\mathbf{x_k} = [x_k, \dot{x}_k, \ddot{x}_k, y_k, \dot{y}_k, \ddot{y}_k]^T$ is a state vector that consists of the mobile user's coordinates (x_k, y_k), velocity $(\dot{x}_k, \dot{y}_k)$, and acceleration $(\ddot{x}_k, \ddot{y}_k)$, respectively at time step k. The parameter $\mathbf{A} \in \mathbb{R}^{6\times6}$ is the state transition matrix, the vector u_k denotes the control input, which represents any externally applied input or maneuvering command to the system, $\mathbf{B}_u \in \mathbb{R}^{6\times2}$ is the control input matrix, and $\mathbf{B}_w \in \mathbb{R}^{6\times2}$ represents the noise matrix, given by Eq. (2) and Eq. (3),

$$\mathbf{A} = \begin{bmatrix} 1 & T & T^2/2 & 0 & 0 & 0 \\ 0 & 1 & T & 0 & 0 & 0 \\ 0 & 0 & \alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & T & T^2/2 \\ 0 & 0 & 0 & 0 & 1 & T \\ 0 & 0 & 0 & 0 & 0 & \alpha \end{bmatrix}, \tag{2}$$

$$\mathbf{B}_u = \begin{bmatrix} T^2/2 & 0 \\ T & 0 \\ 0 & 0 \\ 0 & T^2/2 \\ 0 & T \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_w = \begin{bmatrix} T^2/2 & 0 \\ T & 0 \\ 1 & 0 \\ 0 & T^2/2 \\ 0 & T \\ 0 & 1 \end{bmatrix}, \tag{3}$$

where, T is the discretization period and $\alpha = 1/\tau$ is the reciprocal of the maneuvering time constant τ. The parameter $\mathbf{w}_k \in \mathbb{R}^{2\times1}$ accounts for the state noise, and follows a zero-mean multivariate Gaussian distribution, $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q_k})$, with state noise covariance matrix $\mathbf{Q_k} = \mathbb{E}(\mathbf{w}_k\mathbf{w}_k^T) = \sigma_w^2\mathbf{I}$. The motion model describes how the mobile user moves at time steps $k = \{1, 2, 3, ...\}$. It captures the dynamics of the mobile user by considering the stochastic nature of acceleration [28]. This is particularly useful in scenarios where the motion of the user cannot be precisely model and is subject to random variations. By modeling the acceleration as a random process, the motion model provides a more realistic representation of the user's movement in real-world scenarios.

B. Sensor Measurement Model

In indoor environments, GPS signals are unavailable because it cannot penetrate building structures effectively. To address this limitation, signals emitted by indoor infrastructure can be utilized to indirectly determine the position of a mobile user within an indoor environments. One practical approach involves using Wi-Fi access points (APs) as transmitters, taking advantage of the existing Wi-Fi infrastructure in indoor spaces, while mobile devices function as receivers that measure the RSS from these APs [29]. The RSS values indicate the signal power received by the mobile device from each AP, which can be used to estimate the distance between the APs and the mobile user [30]. By combining RSS measurements from multiple APs, the position of the mobile device can be triangulated. The relationship between the RSS measurements and the distance is typically modeled using the Path Loss Model, which provides a mathematical framework for representing Wi-Fi signal propagation, given by [31],

$$z_k = z_0 + 10\beta\log_{10}(d_k) + v_k, \tag{4}$$

where, z_k is the RSS measurement at time step k at the receiver node, z_0 is the reference path loss at a 1-meter distance, β is the path loss exponent, d_k is the distance between the receiver and transmitter at time step k, and v_k is the measurement noise. For practical applications, a minimum of three APs is required for positioning, as this allows for triangulation to determine the position of the mobile user in the network. Thus, Eq. (4) can be expressed in vector form as $\mathbf{z}_k = [z_k^1, z_k^2, \ldots, z_k^n]^T$, where n refers to the number of APs.

1) Multivariate Gaussian Distribution: The measurement noise in Eq. (4) is often modeled as a zero-mean Gaussian distribution $v_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2)$. For multiple APs, the measurement noise is expressed as a vector $\boldsymbol{v}_k = [v_k^1, v_k^2, v_k^3, \ldots, v_k^n]^T$, where n denotes the total number of APs. Consequently, v_k follows a multivariate Gaussian distribution, $\boldsymbol{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R_k})$, with measurement noise covariance matrix $\mathbf{R_k} = \mathbb{E}[\boldsymbol{v}_k\boldsymbol{v}_k^T] = \sigma_v^2\mathbf{I}$. The probability density function (PDF) for a multivariate Gaussian distribution, characterized by a mean vector $\hat{\mathbf{z}}_k$ and a covariance matrix $\mathbf{R_k}$, can be expressed as Eq. (5):

$$p(\mathbf{z}_k) = (2\pi)^{-\frac{n}{2}}|\mathbf{R_k}|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^T\mathbf{R}_k^{-1}(\mathbf{z}_k - \hat{\mathbf{z}}_k)\Big), \tag{5}$$

where, n represents the number of AP, $\mathbf{z}_k$ is the observed RSS measurement, $\hat{\mathbf{z}}_k$ is the predicted RSS measurement, $\mathbf{R}_k$ is the measurement noise covariance matrix, and $|\mathbf{R_k}|$ is the determinant of $\mathbf{R}_k$. It provides the likelihood of the RSS measurements given the estimated state of the system. The likelihood function is utilized in the update step of the PF to adjust particle weights based on how closely the predicted measurements align with the observed data.

2) Multivariate Student's-T Distribution: In cases, where sensor measurements contain outliers, the measurement noise in Eq. (4) follows a multivariate Student's-T distributions, $\mathbf{z}_k \sim St(\mathbf{0}, \mathbf{R}, \nu)$, with the PDF given by [26] Eq. (6),

$$p(\mathbf{z}_k) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{n/2}\det(\mathbf{R}_k)^{1/2}} \times (1 + \frac{1}{\nu}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^T\mathbf{R}_k^{-1}(\mathbf{z}_k - \hat{\mathbf{z}}_k))^{-\frac{\nu+n}{2}}, \tag{6}$$

where, n represents the total number of AP, $\mathbf{z}_k$ is the observed RSS measurement, $\hat{\mathbf{z}}_k$ is the predicted RSS measurement, $\mathbf{R}_k$ is the measurement noise covariance matrix, and ν is the degrees of freedom. The symbol $\Gamma(\cdot)$ represents the gamma function. The multivariate Student's-T distribution shares some

similarities with the multivariate Gaussian distribution, especially when $\nu \geq 30$. As ν increases, the multivariate Student's-T distribution approaches the multivariate Gaussian distribution. A key distinction of the Student's-T distribution is its capacity to model heavier tails compared to the Gaussian distribution. The parameter ν controls the degrees of freedom and influences the thickness of the distribution's tails. Lower values of ν result in heavier tails.

III. Problem Formulation

A. Review of the Particle Filter

A PF is a state estimation method designed for nonlinear systems, utilizing a collection of samples, known as particles, to approximate complex probability distributions. Unlike approaches constrained by Gaussian assumptions, PF can model arbitrary distributions. Within the context of indoor positioning, PF estimates the state of a mobile user, denoted as $\mathbf{x}_k$ in Eq. (1), by leveraging the RSS measurements represented as $\mathbf{z}_k$ in Eq. (4). The core objective of the PF is to derive the PDF that describes the relationship between the system states and measurements. To approximate the joint PDF at the prior time step $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$, the PF employs a weighted set of particles, $\{\mathbf{x}_{k-1}^{(i)}, i = 1, \ldots, N_p\}$ [32], given by Eq. (7):

$$p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) \approx \sum_{i=1}^{N_p} w_{k-1}^{(i)} \delta(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^{(i)}), \tag{7}$$

where, $\{w_{k-1}^{(i)}, i = 1, \ldots, N_p\}$ represents the weights associated with each particle, where N_p denotes the total number of particles, and $\delta(\cdot)$ signifies the Dirac delta function. The posterior PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is defined as Eq. (8):

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \propto p(\mathbf{z}_k|\mathbf{x}_k) \sum_{j=1}^{N_p} w_{k-1}^{(j)} p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(j)}), \tag{8}$$

where, $p(\mathbf{z}_k|\mathbf{x}_k)$ is the measurement likelihood, and $p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(j)})$ represents the prior transition model. If the particles are drawn from a proposal distribution $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)$, the weights are updated according to Eq. (9):

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(i)}) p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)}. \tag{9}$$

The proposal distribution is assumed to be equal to the state transition probability, i.e. [see Eq. (10)],

$$q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}), \tag{10}$$

based on the rationale that the state dynamics sufficiently capture the evolution of the system, and the measurement z_k does not provide additional information at the sampling stage. This assumption simplifies the weight update,

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(\mathbf{z}_k|\mathbf{x}_k^{(i)}). \tag{11}$$

Algorithm 1 The Particle Filter (PF).

(1) Input: Initialization
  for $k = 0$ **do**
    for $i = 1, \ldots, N_p$ **do**
      Samples: $\{\mathbf{x}_0^{(i)} \sim q(\mathbf{x}_0)\}$.
      Set initial weights: $w_0^{(i)} = 1/N_p$.
    end for
  end for

  for $k = 1, \ldots, endtime$ **do**
    for $i = 1, \ldots, N_p$ **do**
      (2) Prediction Step
      Propagate the particles $\mathbf{x}_k^{(i)}$ using eq. (1).
      Generate measurements $\mathbf{z}_k^{(i)}$ using eq. (4).

      (3) Measurement Update
      Calculate the measurement likelihood:
      $p(\mathbf{z}_k|\mathbf{x}_k^{(i)})$ where $\mathbf{z}_k \sim (0, \mathbf{R}_k)$.
      Update the weights $w_k^{(i)}$ using eq. (11).
    end for
    Normalize the weights: $\hat{w}_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N_p} w_k^{(i)}$.
    The posterior PDF: $p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \sum_{i=1}^{N_p} \hat{w}_k^{(i)} \hat{\mathbf{x}}_k^{(i)}$.

    (4) Output: The estimated state:
    $\hat{\mathbf{x}}_k = \mathbb{E}[\mathbf{x}_k|\mathbf{z}_{1:k}] \approx \sum_{i=1}^{N_p} \hat{w}_k^{(i)} \hat{\mathbf{x}}_k^{(i)}$.
    (5) Resampling Step
    Set the threshold sample size: $N_{thresh} = N_p/10$.
    Calculate the effective sample size:
    $N_{eff} = 1/\sum_{i=1}^{N_p} (\hat{w}_k^{(i)})^2$
    if $N_{eff} < N_{thresh}$
      The metropolis resampling method is applied.
    end if
  end for

reducing computational complexity. It is a standard practice in particle filtering when the transition model is an adequate approximation of the true state dynamics, allowing reliance on the measurement likelihood in the weighting process, as shown in Eq. (11). The weights are then normalized such that [see Eq. (12)]:

$$\sum_{i=1}^{N_p} w_k^{(i)} = 1. \tag{12}$$

Finally, the resulting posterior PDF is expressed as Eq. (13):

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_p} w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}). \tag{13}$$

The PF algorithm, outlined in Algorithm 1, provides an iterative framework for estimating the state of a dynamic system. In each iteration, the particle states $\mathbf{x_k}$ (position, velocity, acceleration) are first predicted via the motion model in Eq. (1), then the RSS measurements $\mathbf{z_k}$ are collected at the mobile device and converted into distance estimates using the log-normal path-loss model using Eq. (4). These distance estimates are then fused through triangulation to establish the

relationship between the measured ranges and the predicted states x_k during the weight-update step.

While this PF framework performs well under ideal conditions, its performance can degrade significantly in real-world scenarios. This is primarily due to the presence of outliers in sensor measurements [33]. Outliers, which may result from measurement noise, interference, or other unexpected phenomena, can skew the estimated state by disproportionately affecting the particle weights. This is because the PF relies on the likelihood function to update the weights of the particles, and outliers can lead to incorrect weight updates, resulting in poor state estimates and increased variance. This sensitivity to outliers is a critical limitation that can undermine the accuracy and robustness of the PF, necessitating additional strategies or alternative approaches are needed to mitigate the impact of such anomalies [34].

B. Outlier-Robust Particle Filter

In this work, to address the outlier problems in sensor measurements, a new outlier-robust particle filter (RPF) is derived. Unlike the PF in Algorithm 1 that models the likelihood PDF with a single noise model, this approach uses two noise models. Let $s \in \{1, 2\}$ denote the noise models, where s_1 corresponds to a multivariate Gaussian distribution, and s_2 corresponds to a multivariate Student's T distribution. The hypothesis indicator $\mathcal{H}_{s,k}$ specifies the selection of the s-th noise model at a specific time step k. The likelihood of transitioning between these noise models is characterized by the transition probabilities, represented as $\mathbf{\Pi}_{s,k}$. This matrix outlines the probability of switching from one noise model to another during a particular time step. Using the transition matrix and the state from the previous time step $k-1$, Eq. (7) can be re-expressed as Eq. (14):

$$p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) \approx \sum_{i=1}^{N_p} \left(w_{k-1}^{(i)} \delta(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^{(i)}) \right) \times \sum_{s=1}^{S} \mathbf{\Pi}_{s,k-1}. \tag{14}$$

At a given time step k, assume that the particles are sampled based on a proposal distribution $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)$. Under the hypothesis $\mathcal{H}_{k,s}$ where $s \in \{1, 2\}$, the corresponding weights can be calculated recursively as Eq. (15):

$$w_{s,k}^{(i)} \propto w_{k-1}^{(i)} \frac{p_s(\mathbf{z}_k|\hat{\mathbf{x}}_k^{(i)}) p(\hat{\mathbf{x}}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\hat{\mathbf{x}}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)}, \tag{15}$$

where, $p_s(\mathbf{z}_k|\hat{\mathbf{x}}_k^{(i)})$ represents the likelihood function corresponding to $\mathcal{H}_{s,k}$, given by Eq. (16):

$$p_s(\mathbf{z}_k|\hat{\mathbf{x}}_k^{(i)}) = \begin{cases} \text{Eq. (5)} & \text{for } s_1 \\ \text{Eq. (6)} & \text{for } s_2 \end{cases}. \tag{16}$$

Next, by choosing the proposal important density $q(\hat{\mathbf{x}}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k) = p(\hat{\mathbf{x}}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$, the weights simplifies to Eq. (17):

$$w_{s,k}^{(i)} \propto w_{k-1}^{(i)} p_s(\mathbf{z}_k|\hat{\mathbf{x}}_k^{(i)}), \tag{17}$$

and normalized by Eq. (18):

$$\hat{w}_{s,k}^{(i)} = w_{s,k}^{(i)} / \sum_{i=1}^{N_p} w_{s,k}^{(i)}. \tag{18}$$

Algorithm 2 The Outlier-Robust Particle Filter (RPF).

(1) Input: Initialization

Set the measurement noise model transition probability: $\mathbf{\Pi}_{s,0}$ for $s \in \{1, 2\}$.

for $k = 0$ **do**
  for $i = 1, \ldots, N_p$ **do**
    Samples: $\{\mathbf{x}_0^{(i)} \sim q(\mathbf{x}_0)\}$.
    Set initial weights: $w_0^{(i)} = 1/N_p$.
  end for
end for

for $k = 1, \ldots, endtime$ **do**
  for $i = 1, \ldots, N_p$ **do**
    (2) Prediction Step
    Propagate the particles $\mathbf{x}_k^{(i)}$ using eq. (1).
    Generate measurements $\mathbf{z}_k^{(i)}$ using eq. (4).

    (3) Measurement Update
    Calculate the measurement likelihood: $p_s(\mathbf{z}_k|\mathbf{x}_k^{(i)})$ where

$$\mathbf{z}_k \sim \begin{cases} \mathcal{N}(0, \mathbf{R}_k), & \text{for } s_1. \\ St(\mu, \hat{\mathbf{\Sigma}}, v), & \text{for } s_2. \end{cases}$$

    Calculate the noise models marginal likelihood $p_s(\mathbf{z}_k|\mathbf{z}_{k-1}) = w_{k-1}^{(i)} \times p_s(\mathbf{z}_k|\hat{\mathbf{x}}_k^{(i)})$ for $s \in \{1, 2\}$.
    Update the weights $w_{s,k}^{(i)}$ for $s \in \{1, 2\}$ using eq. (11).
  end for
  Normalize the weights: $\hat{w}_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N_p} w_k^{(i)}$.
  Noise transition probability:

$$\mathbf{\Pi}_{s,k} = \frac{\mathbf{\Pi}_{s,k-1}^{\varphi}}{\sum_{s=1}^{S} \mathbf{\Pi}_{s,k-1}^{\varphi}} \text{ for } s \in \{1, 2\}.$$

  Update the noise transition probability:

$$\hat{\mathbf{\Pi}}_{s,k} = \frac{\mathbf{\Pi}_{s,k} \times p_s(\mathbf{z}_k|\mathbf{z}_{k-1})}{\sum_{s=1}^{S} \mathbf{\Pi}_{s,k} \times p_s(\mathbf{z}_k|\mathbf{z}_{k-1})} \text{ for } s \in \{1, 2\}.$$

  The posterior PDF:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \sum_{i=1}^{N_p} \hat{w}_{s,k}^{(i)} \cdot \delta\left(\mathbf{x}_k^{(i)} - \hat{\mathbf{x}}_{s,k}^{(i)} \right) \cdot \sum_{s=1}^{S} \mathbf{\Pi}_{s,k}.$$

  (4) Output: The estimated state:

$$\hat{\mathbf{x}}_k = \mathbb{E}[\mathbf{x}_k|\mathbf{z}_{1:k}] = \sum_{s=1}^{S} \sum_{i=1}^{N_p} \hat{w}_{s,k}^{(i)} \cdot \hat{\mathbf{x}}_{s,k}^{(i)} \cdot \mathbf{\Pi}_{s,k}.$$

  (5) Resampling Step
  Set the threshold sample size: $N_{thresh} = N_p/10$.
  Calculate the effective sample size: $N_{eff} = 1/\sum_{i=1}^{N_p} (\hat{w}_k^{(i)})^2$
  if $N_{eff} < N_{thresh}$
    The metropolis resampling method is applied.
  end if
end for

The transition probability of the $\mathcal{H}_{s,k-1}$ is given by Eq. (19):

$$\mathbf{\Pi}_{s,k-1} = \frac{\mathbf{\Pi}^{\varphi}_{s,k-1}}{\sum_{s=1}^{S} \mathbf{\Pi}^{\varphi}_{s,k-1}}, \tag{19}$$

where, $\varphi \in [0, 1]$ represents the forgetting factor. Utilizing Bayes' theorem, the transition probability for $\mathcal{H}_{s,k}$ is calculated recursively as

$$\mathbf{\Pi}_{s,k} = \frac{\pi^{\varphi}_{s,k-1}\ p_s(\mathbf{z}_k|\mathbf{z}_{k-1})}{\sum_{s=1}^{S} \mathbf{\Pi}^{\varphi}_{s,k-1}\ p_s(\mathbf{z}_k|\mathbf{z}_{k-1})}. \tag{20}$$

Eq. (20) provides an update to each hypothesis probability by blending the prior belief (through the forgetting factor φ) with the marginal likelihood of the new measurement. If the RSS observation aligns closely with Gaussian assumptions, the Gaussian hypothesis is reinforced; if it exhibits a large deviation, the Student's-T hypothesis gains prominence. $p_s(\mathbf{z}_k|\mathbf{z}_{k-1})$ represents the marginal likelihood for $\mathcal{H}_{s,k}$, where the integration is performed over the entire state space of the system described as

$$p_s(z_k|z_{k-1}) = \int_{x_k \in \mathbb{R}^n} p_s(z_k|x_k)\, p(x_k|z_{k-1})\, dx_k. \tag{21}$$

Since $p(\mathbf{x}_k|\mathbf{z}_{k-1}) \approx \sum_{i=1}^{N_p} w^{(i)}_{k-1} \delta_{\hat{\mathbf{x}}^{(i)}_k}$, the integral in Eq. (21) is approximated by Eq. (22):

$$p_s(\mathbf{z}_k|\mathbf{z}_{k-1}) \approx \sum_{i=1}^{N_p} w^{(i)}_{k-1}\ p_s(\mathbf{z}_k|\hat{\mathbf{x}}^{(i)}_k). \tag{22}$$

Finally, the posterior PDF at time step k is formulated as Eq. (23) and Eq. (24):

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \sum_{s=1}^{S} p_s(\mathbf{x}_{s,k}|\mathbf{z}_k) \times \mathbf{\Pi}_{s,k}, \tag{23}$$

$$= \sum_{s=1}^{S} \sum_{i=1}^{N_p} \left(w^{(i)}_{s,k} \delta_{\hat{\mathbf{x}}^{(i)}_k} \times \mathbf{\Pi}_{s,k} \right). \tag{24}$$

Particle weights under model s are computed via the likelihood $p_s(z_k|z_{k-1})$. For the Student's-T hypothesis, a degrees of freedom, ν parameter of 3 is chosen to provide sufficient robustness against heavy-tailed outliers without incurring excessive computational cost. The filter employs N_p = 200 particles; whenever fewer than 20 particles (i.e., 10%) carry significant weight, Metropolis resampling is performed to restore diversity. Algorithm 2 summarises the RPF.

IV. Experiments and Results

This section evaluates the performance of the PF and the RPF using both simulated and real-world sensor data for indoor positioning of mobile users. The focus is on estimating the positions of mobile users from Wi-Fi sensor readings, which are often affected by outliers. The simulated data is generated by adding measurement noise and a specified percentage of outliers to a predefined trajectory, emulating real-world conditions. Real-world data consists of Wi-Fi signal strength measurements collected in an indoor environment, reflecting practical challenges such as multipath propagation and signal occlusion.

TABLE I. Simulation Parameters for PF and RPF

Parameter and Symbol	Value and Unit PF	Value and Unit RPF
Number of Particles, N_p	200	200
Control input, $\mathbf{u}_k$	{(0, 0.07) (0.07, 0) (0, -0.07) (-0.07, -0.07) (-0.07, 0) (-0.07, 0.07)}	{(0, 0.07) (0.07, 0) (0, -0.07) (-0.07, -0.07) (-0.07, 0) (-0.07, 0.07)}
Reference RSS, Z_0	0 dBm	0 dBm
Path loss exponent, β	4.0	4.0
Sampling rate, f_s	1 Hz	1Hz
Discretisation period, T	1	1
Reciprocal of the manoeuvre time, α	0.1	0.1
Forgetting factor, φ	non	0.5
degrees of freedom, ν	non	3
Noise Transition Probability, $\mathbf{\Pi}_{s,k}$	non	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
Process Noise Variance, σ^2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Measurement Noise Variance, σ^2_k	1	1

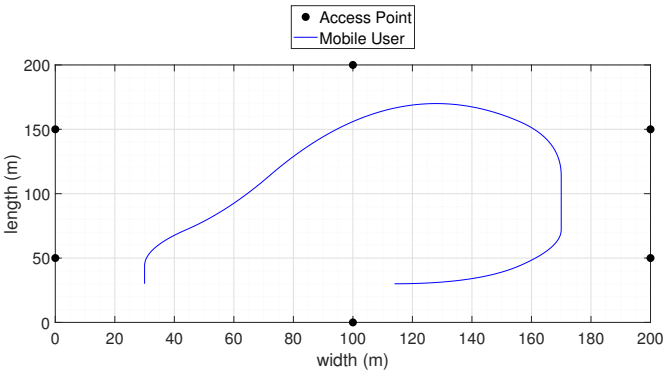


Fig. 1. Mobile user trajectory and access point deployment.

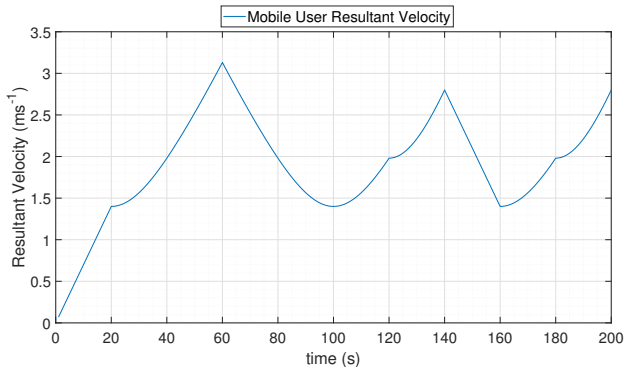


Fig. 2. Mobile user resultant velocity.

A. Simulated Data

Consider a 2D scenario, where a mobile user moves within an indoor building, following the dynamics described by Eq. (1). The user's position is indirectly observed via RSS measurements, modeled by Eq. (4), which are transmitted by the Wi-Fi APs and received by the mobile user's device. Fig. 1 illustrates the mobile user's trajectory alongside the network deployment. The trajectory represents the user's path over 200 time steps, indexed by k, while the network deployment shows

the placement of APs within the indoor environment. Fig. 2 shows the resultant speed (magnitude of velocity) of the mobile user at each time step in the simulated scenario. In the simulation, acceleration is set to zero to approximate the user's walking at a constant speed, indicating that each control input $\mathbf{u}_k$ results in constant-speed motion between changes in direction. The network covers an area of 200 m × 200 m, with the APs functioning as transmitter nodes and the mobile user serving as the receiver node. The observed sensor readings correspond to the received Wi-Fi signals, include outliers, that are randomly distributed such that outliers occur in 50% of the time steps, while the remaining 50% of the measurements are unaffected. When outliers are present, the measurements are simulated using a multivariate Student's-T distribution, whereas unaffected data follow a multivariate Gaussian distribution. Here, the control input vector $\mathbf{u}_k$ represents discrete movement commands in the two-dimensional plane, specifically steps to the left, right, up, or down, applied at each time step. The forgetting factor φ =0.5 balances responsiveness to new measurements with retention of past model-selection history, preventing both abrupt and overly sluggish switching between noise models. The simulation parameters are detailed in Table I and the root mean square error (RMSE) is employed as the evaluation metric to assess the accuracy of the PF and the RPF in estimating the position, velocity, and acceleration of the mobile user in the indoor environment.

Fig. 3 compares the estimated trajectories of the mobile user within the network, obtained through the PF and RPF algorithms, based on a single simulation run. It shows that the RPF consistently tracks the mobile user's path more accurately than the PF in most of the time steps. Notably, during periods when the mobile user makes turns, such as between time steps $k = 160$ to $k = 170$, the RPF achieves lower tracking error compared to the PF, as measured by the RMSE.

In the subsequent experiments, the simulation is repeated 100 times under different combinations of noise and outlier distributions and the average RMSE is computed to assess tracking performance. Fig. 4 illustrates the mean RMSE for position estimation for both the PF and RPF algorithms. It is observed that the RPF algorithm outperforms the PF algorithm, achieving a substantially lower RMSE. Specifically, the mean RMSE for the RPF is 9.5208 meters, whereas for the PF it is 15.8734 meters. This results is a significant reduction of approximately 40.02% in RMSE when using the RPF algorithm, highlighting its improved position estimation accuracy as reflected by lower RMSE values.

Fig. 5 illustrates the mean RMSE for velocity estimation using both the PF and RPF algorithms. The results indicate that the RPF algorithm consistently outperforms the PF algorithm in velocity estimation. Specifically, the mean RMSE for velocity using the RPF is 0.1252 ms^{-1}, which is significantly lower than the 0.2035 ms^{-1} observed with the PF algorithm. This represents a reduction of approximately 38.48% in RMSE when using the RPF, indicating improved velocity estimation accuracy, with a lower RMSE compared to the standard method.

Finally, Fig. 6 presents the mean RMSE for acceleration estimation using both the PF and RPF algorithms. It is shows that the RPF consistently achieves lower RMSE for acceleration throughout the simulation time steps. Specifically, the mean RMSE for acceleration with the RPF is $5.7339 \times 10^{-5} ms^{-2}$,

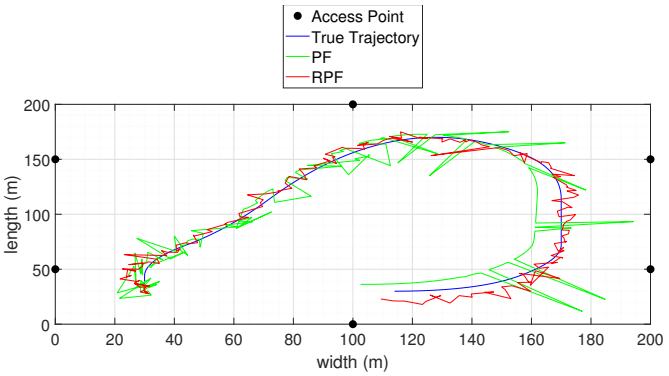


Fig. 3. Estimated trajectory by Particle Filter (PF) and Robust Particle Filter (RPF).

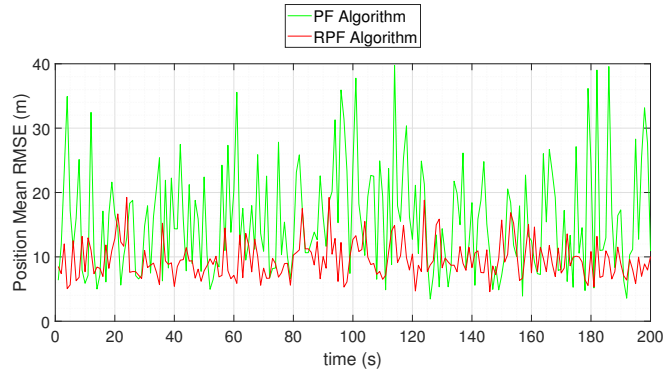


Fig. 4. Mean RMSE of mobile user position using PF and RPF.

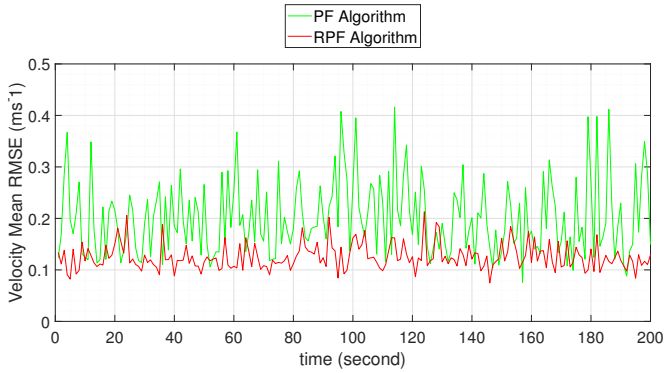


Fig. 5. Mean RMSE of mobile user velocity using PF and RPF.

which is considerably lower than the 1.6767×10^{-4} ms^{-2} observed with the PF. This corresponds to a reduction of approximately 65.80% in RMSE when using the RPF, highlighting its improved accuracy in estimating the acceleration of a mobile user, as indicated by a lower acceleration estimation error.

Overall, the results demonstrate that the RPF outperforms the PF in estimating position, velocity, and acceleration of

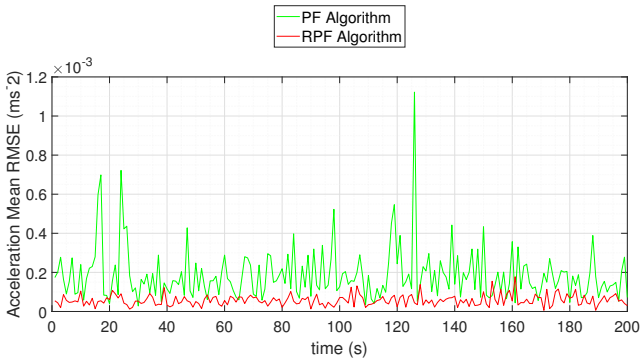


Fig. 6. Mean RMSE of mobile user acceleration using PF and RPF.

a mobile user in indoor environment. The RPF consistently achieves lower RMSE values across all metrics, indicating significant improvements over the PF. This suggests that outliers in RSS measurements can diminish the accuracy of the PF, while the RPF remains robust to such disturbances. As a result, the RPF is more effective at handling outliers in the sensor measurements compared to the PF. Runtime measurements on MATLAB show that a single simulation run takes approximately 0.24 seconds for the PF and 0.33 seconds for the RPF, highlighting the additional computational cost associated with the dual-noise likelihood evaluation. The presence of outliers in highly obstructed and complex indoor environments is expected to significantly degrade the position estimation accuracy of both filters. However, in such outlier-affected scenarios, the RPF consistently outperforms the PF due to its more robust dual-noise likelihood evaluation strategy.

B. Real-World Data

The real-world data was collected at the Melinsung Summer Bay Resort Apartment in Papar, Sabah, a three-room apartment with a total area of 680 square feet. The experimental setup employed the ESP-32-WROOM-32 module, shown in Fig. 7, which served as both the AP and the receiver point (RP). AP and RPs are placed at ground level in the indoor setup (height = 0 m). The parameters z_0 and β in Eq. (4) vary depending on the environment [35], thus the experimental setup, as shown in Fig. 8, was designed to interpolate these parameter values. RPs were positioned at distances ranging from 0.5 to 7.0 meters from the APs. To account for temporal variations and ensure accuracy, measurements were repeated multiple times at different times of the day and under varying environmental conditions. The RSS values were captured and recorded in real-time using the Arduino IDE's serial monitor [36].

To streamline data collection, the PLX-DAQ version 2.11 [37] was utilized, enabling efficient transfer of the readings directly from the serial monitor to an Excel file. The RSS values were collected by connecting the ESP-32-WROOM-32 to a laptop, allowing for real-time data acquisition. The collected RSS measurements visualized in Fig. 9 illustrates the relationship between signal strength and distance. A logarithmic interpolation was applied to identify the optimal fit for the RSS measurements by reducing the squared discrepancies between the observed data points and the corresponding values on the curve. As a result, the parameter z_0 was estimated at -68.4694 dBm, and the path loss exponent β was calculated to be 0.71349, indicating how the RSS decreases logarithmically as the distance increases, given by in both the PF and RPF as a reliable representation of the signal propagation characteristics within the indoor environment.



Fig. 7. ESP32-WROOM-32 as Access Point (AP) and Receiver Point (RP) nodes.

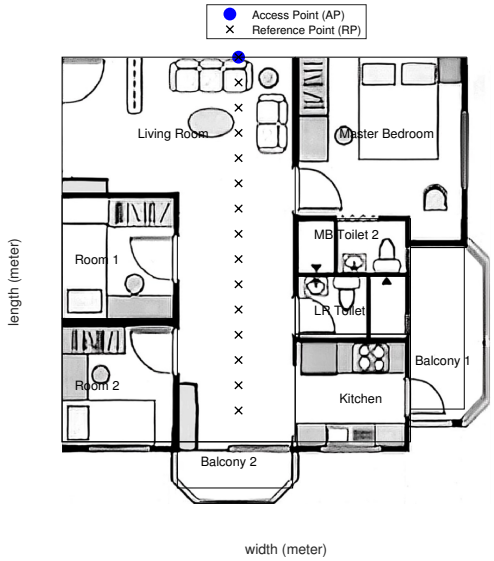


Fig. 8. Transmitter node and receiver node locations for collecting calibrated data.

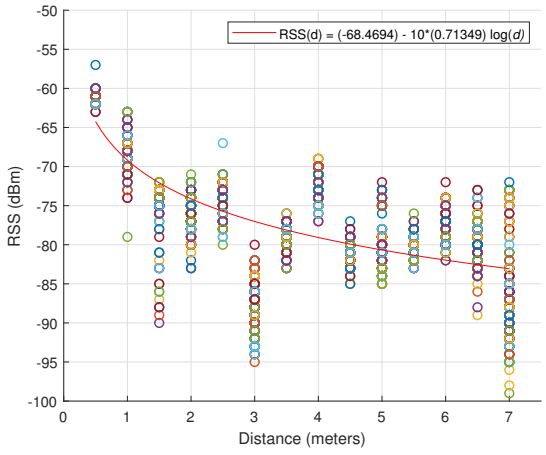


Fig. 9. An interpolated log-normal path loss model derived from calibrated data.

$$Z = -68.4694 - 10 * (0.71349)\log(d) + v, \qquad (25)$$

where, v is modeled using a zero-mean Gaussian distribution. It provides the interpolated measurement model, which is used

Next, actual Wi-Fi signal measurements were collected from a mobile user using the experimental setup shown in Fig. 10, which shows the indoor floor plan with four strategically placed Wi-Fi APs and the single 17-second user trajectory from one representative trial. The experimental results are derived from 20 independent real-world trials, each following the same 17-second mobile user trajectory but yielding distinct RSS readings (including natural outliers), and the final outcomes are based on the mean measurements to ensure robustness and reliability of the positioning performance evaluation. As the receiver traversed various locations, the user's trajectory was mapped while capturing RSS values from the Wi-Fi APs. The collected RSS measurements along this trajectory was used to evaluate the performance of the PF and RPF algorithms in accurately estimating the user's position within the indoor environment. During the prediction stage of both algorithms, the measurement model formulated in Eq. (25) was applied.

Fig. 11 presents a comparison of the mean RMSE between the PF and RPF, computed over 17 time steps from 20 repeated experiments using collected Wi-Fi RSS measurements. The data demonstrate that the RMSE of the RPF is consistently lower than that of the PF across all time steps. Notably, the PF shows a significant spike in RMSE at the 13^{th} time step, reaching a value of 43.89825, likely due to outlier measurements or non-Gaussian noise, whereas the RPF remains robust with an RMSE of 0.91450 at the same time step. Overall, the mean RMSE of the PF is 10.27197, while that of the RPF is 0.65683, the difference in RMSE between PF and RPF is 9.61515 reflecting a 93.61% improvement in RPF, highlighting that the RPF has successfully mitigated the effect of outliers and overcome the limitations of the PF.

V. Conclusion

In this study, the performance of the Particle Filter (PF) and the proposed outlier-robust particle filter (RPF) algorithms was thoroughly evaluated using both simulated and real-world data. The results from the simulated data demonstrated that the RPF consistently outperforms the PF in terms of position, velocity, and acceleration estimation. The RPF achieved markedly reduced root mean square error (RMSE) values, with reductions of approximately 40.02% in position, 38.48% in velocity, and 65.80% in acceleration compared to the PF. The real-world experiments further confirmed the robustness of the RPF. By applying a calibrated log-normal path loss model to the collected Wi-Fi RSS data, the RPF outperformed the PF in position estimation accuracy. Notably, the RPF showed a 93.61% improvement in positioning accuracy, with a significant reduction in RMSE at each time step. These results highlight the RPF's robustness in maintaining accurate position estimates, even in the presence of measurement noise and outliers. Therefore the proposed RPF offers a robust solution for tracking and positioning in environments where outliers or noise are prevalent, making it a valuable approach for real-world positioning applications. Future work includes evaluation under more complex indoor scenarios such as multiple trajectories, varied AP deployments, and obstructed layouts with dynamic obstacles, along with integration of map constraints and multi-sensor fusion (e.g., UWB/AoA) to improve positioning robustness in highly non-linear, cluttered environments.

Acknowledgment

The authors express their gratitude to the Ministry of Higher Education Malaysia for providing funding through the Fundamental Research Grant Scheme (FRGS) under project number FRGS/1/2022/TK07/UMS/02/9. Appreciation is also extended to Universiti Malaysia Sabah for their invaluable support in facilitating this research.

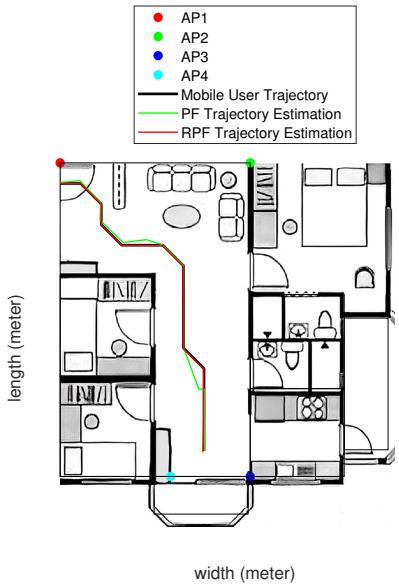


Fig. 10. Indoor floor plan layout, transmitter node and mobile user trajectory for a single simulation run.

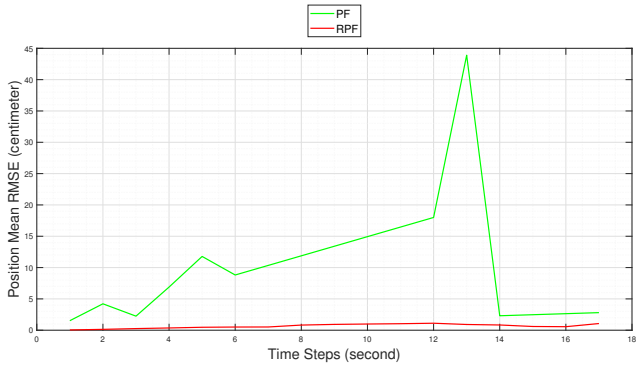


Fig. 11. Comparison of PF and RPF mean position RMSE using collected Wi-Fi RSS measurements.

References

[1] H. Zhang, B. Li, M. Karimi, S. Saydam, and M. Hassan, "Recent advancements in IoT implementation for environmental, safety, and production monitoring in underground mines," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14 507–14 526, 2023.

[2] J. M. Choi, K. Kim, and S. C. Kim, "Outdoor localization based on RSS ranging aided by pedestrian dead reckoning in GPS restricted scenario," in *Proc. Symp. IEEE VTS Asia Pacific Wireless Commun.*, 2022, pp. 186–190.

[3] J. Singh, N. Tyagi, S. Singh, F. Ali, and D. Kwak, "A systematic review of contemporary indoor positioning systems: Taxonomy, techniques, and algorithms," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 34 717–34 733, 2024.

[4] L. Bai, F. Ciravegna, R. Bond, and M. Mulvenna, "A low cost indoor positioning system using bluetooth low energy," *IEEE Access*, vol. 8, pp. 136 858–136 871, 2020.

[5] M. Bhutani, B. Lall, and M. Agrawal, "Optical wireless communications: Research challenges for mac layer," *IEEE Access*, vol. 10, pp. 126 969–126 989, 2022.

[6] N. Lu and Z. Gao, "Ultra wideband indoor positioning method based on Kalman filter and Taylor algorithm," in *Proc. 4th Int. Conf. on Intell. Control, Meas. and Signal Process. (ICMSP)*, 2022, pp. 910–914.

[7] R. Kusche, S. O. Schmidt, and H. Hellbrück, "Indoor positioning via artificial magnetic fields," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[8] S. Sadowski, P. Spachos, and K. N. Plataniotis, "Memoryless techniques and wireless technologies for indoor localization with the internet of things," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 10 996–11 005, 2020.

[9] X. Zhang and W. Sun, "Design of indoor localization algorithm based on Wifi fingerprints," in *Proc. China Automation Congress (CAC)*, 2021, pp. 6209–6214.

[10] B. Shin, J. H. Lee, C. Yu, H. Kyung, and T. Lee, "Received signal strength-based robust positioning system in corridor environment," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[11] J. A. López-Pastor, M. Poveda-García, A. Gil-Martínez, D. Cañete-Rebenaque, and J. L. Gómez-Tornero, "2-d localization system for mobile iot devices using a single wi-fi access point with a passive frequency-scanned antenna," *IEEE Internet of Things Journal*, vol. 10, no. 17, pp. 14 995–15 011, 2023.

[12] Q. Zhu, K. Niu, C. Dong, and Y. Wang, "A novel angle of arrival (AOA) positioning algorithm aided by location reliability prior information," in *Proc. IEEE Wireless Comm. and Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[13] Y. Li, Z. He, Z. Gao, Y. Zhuang, C. Shi, and N. El-Sheimy, "Toward robust crowdsourcing-based localization: A fingerprinting accuracy indicator enhanced wireless/magnetic/inertial integration approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3585–3600, 2019.

[14] N. H. H. Pham, M. A. Nguyen, and C.-C. Sun, "Indoor positioning system using UWB and Kalman filter to increase the accuracy of the localization system," in *Proc. IEEE Int. Conf. on Consumer Electron. - Taiwan*, 2022, pp. 339–340.

[15] A. Eman and H. Ramdane, "Mobile robot localization using extended Kalman filter," in *Proc. 3rd Int. Conf. on Comput. Applications & Inf. Security (ICCAIS)*, 2020, pp. 1–5.

[16] X. Xu, A. Peng, and X. Hong, "Unscented Kalman filtering based multipath-assisted positioning with peak flow tracking," in *Proc. IEEE 12th Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*, 2022, pp. 1–8.

[17] L. Mihaylova, D. Angelova, S. Honary, D. R. Bull, C. N. Canagarajah, and B. Ristic, "Mobility tracking in cellular networks using particle filtering," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3589–3599, 2007.

[18] H. Gu, H. Mahmoud, R. L. Arun, J. Liu, and X. Ma, "Particle filter and its variants for degradation state estimation and remaining useful life prediction," in *Proc. Conf. Prognostics and Health Management (PHM)*, 2023, pp. 256–263.

[19] L. Yang, N. Wu, B. Li, W. Yuan, and L. Hanzo, "Indoor localization based on factor graphs: A unified framework," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 4353–4366, 2023.

[20] Z. Yu, L. Lin, C. Xue, L. Zhuo, and J. Yuan, "Application of untracked particle filter (pso-upf) based on particle swarm optimization in indoor target tracking," in *Proc. 20th Int. Conf. on Natural Comput., Fuzzy Syst. and Knowledge Discovery (ICNC-FSKD)*, 2024, pp. 1–6.

[21] H. Chen, X. Shen, Y. Wang, and N. Shi, "Particle filter-based enhanced transition model in signal for unsupervised localization," *IEEE Sensors Journal*, vol. 24, no. 21, pp. 35 845–35 857, 2024.

[22] Y. Shen, B. Hwang, and J. P. Jeong, "Particle filtering-based indoor positioning system for beacon tag tracking," *IEEE Access*, vol. 8, pp. 226 445–226 460, 2020.

[23] R. Jurdi, H. Chen, Y. Zhu, B. Loong Ng, N. Dawar, C. Zhang, and J. K.-H. Han, "Whereartthou: A wifi-rtt-based indoor positioning system," *IEEE Access*, vol. 12, pp. 41 084–41 101, 2024.

[24] A. Alteneiji, U. Ahmad, K. Poon, N. Ali, and N. Almoosa, "Indoor localization in multi-path environment based on aoa with particle filter," in *Proc. 3rd Int. Conf. on Signal Process. and Inf. Security (ICSPIS)*, 2020, pp. 1–4.

[25] W. Li, X. Liu, and L. Guo, "Robust particle filtering with state transition uncertainties: A student's t disturbance observer-based approach," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[26] J. Wang, P. Dong, K. Shen, X. Song, and X. Wang, "Distributed consensus student-t filter for sensor networks with heavy-tailed process and measurement noises," *IEEE Access*, vol. 8, pp. 167 865–167 874, 2020.

[27] W. Na, P. Kun, and L. Yang, "A survey of maneuvering target prediction via interactive multiple model," in *Proc. 34th Chinese Control and Decision Conf. (CCDC)*, 2022, pp. 6069–6074.

[28] J. Zhang, S. Liu, and R. Zhu, "Motion velocity, acceleration and energy expenditure estimation using micro flow sensor," *IEEE Access*, vol. 7, pp. 75 901–75 909, 2019.

[29] Y. Onykiienko, P. Popovych, R. Yaroshenko, A. Mitsukova, A. Beldyagina, and Y. Makarenko, "Using rssi data for lora network path loss modeling," in *Proc. IEEE 41st Int. Conf. on Electron. and Nanotechnology (ELNANO)*, 2022, pp. 576–580.

[30] E. I. Adegoke, E. Kampert, and M. D. Higgins, "Empirical indoor path loss models at 3.5Ghz for 5G communications network planning," in *Proc. Int. Conf. on UK-China Emerging Technol. (UCET)*, 2020, pp. 1–4.

[31] F. A. Abed, Z. A. Hamza, and M. F. Mosleh, "Indoor positioning system based on Wi-fi and bluetooth low energy," in *Proc. 8th Int. Eng. Conf. on Sust. Technol. and Development*, 2022, pp. 136–141.

[32] A. Kiring, "Shrinkage based particle filters for tracking in wireless sensor networks with correlated sparse measurements," *PhD thesis, University of Sheffield*, 2017.

[33] A. H. Chughtai, M. Tahir, and M. Uppal, "A robust bayesian approach for online filtering in the presence of contaminated observations," *IEEE Tras. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[34] N. Zhou, L. Lau, R. Bai, and T. Moore, "Evolutionary-inspired strategy for particle distribution optimization in auxiliary particle filtering algorithm based indoor positioning," in *Proc. Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*, 2021, pp. 1–8.

[35] M. A. Mohamed Ghazali, A. Kiring, L. Mihaylova, H. T. Yew, S. K. Chung, and Y. Y. Farm, "Evaluation of theoretical and empirical log-normal models for wi-fi fingerprint mapping in indoor positioning," in *Proc. IEEE Int. Conf. on Artificial Intell. in Eng. and Technol.*, 2024, pp. 697–702.

[36] K. B, V. P, V. P, N. K, and K. Sriharipriya, "Iot based neonatal patient monitoring system," in *Proc. 2nd Int. Conf. on Vision Towards Emerging Trends in Commun. and Netw. Technol. (ViTECoN)*, 2023, pp. 1–6.

[37] O. Chidolue and T. Iqbal, "System monitoring and data logging using plx-daq for solar-powered oil well pumping," in *Proc. 13th IEEE Annual Comput. and Commun. Workshop and Conf.*, 2023, pp. 0690–0694.

Design and Analysis of Smart Lighting System for Room Environments Using Simulation Supporting Diverse Light Bulbs

Husnul Ajra[1], Mazlina Abdul Majid[2]*, Md. Shohidul Islam[3]
Faculty of Computing, University Malaysia Pahang Al-Sultan Abdullah, Kuantan, Pahang[1,2,3]
Centre for Artificial Intelligence & Data Science, University Malaysia Pahang Al-Sultan Abdullah, Kuantan, Pahang[2]

***Abstract*—Recently, the demand for intelligent, energy-efficient lighting systems has increased due to rising environmental concerns and increasing electricity consumption in smart room environment buildings. Conventional lighting systems often operate inefficiently, using outdated bulb technologies and lacking automation, which results in substantial energy waste, especially in rooms with variable occupancy. Lighting significantly contributes to energy consumption in indoor spaces, which presents vast opportunities for smart lighting model development through automation and adaptive control. This study proposes a smart lighting system model for room environments that dynamically adapts to user presence and supports diverse light bulb types. The study analyzes energy usage while maintaining automatic light control and operational effectiveness through simulation. The system is developed using AnyLogic by integrating agent-based and discrete event simulation to model occupant behavior and manage event-driven lighting logic. It incorporates sensors, smart door mechanisms, and energy-measuring processes, all powered by solar energy and managed through battery storage. The system dynamically adjusts lighting based on occupancy, minimizing idle energy usage for the room. LED bulbs offer more promising energy efficiency, while incandescent bulbs show the highest consumption. The outcome provides a visualized simulation model for designing adaptive lighting systems and reinforces the potential to enhance energy efficiency to support sustainability in smart room applications.**

***Keywords*—Bulb; energy; light; model; simulation**

I. Introduction

In an era characterized by boosted environmental awareness and resource limitations, the implementation of efficient lighting systems has become a crucial aspect of energy management in commercial and residential buildings. With their substantial impact on electrical consumption, optimizing lighting systems is essential for reducing overall energy usage and advancing sustainability initiatives. With urbanization and increasing indoor activity, there has been a critical push towards optimizing lighting technologies [1] to enhance energy efficiency and environmental sustainability. The advent of smart room technology, powered by automation and data-driven modeling, enables efficient energy management through dynamic control of lighting. Modern intelligent lighting systems are no longer limited to manual operations; they now integrate occupancy sensors to optimize light intensity in real-time [2]. This results in a reduction of energy waste and operational costs while also enhancing environmental sustainability [3]. Conventional lighting systems, such as incandescent and halogen bulbs, though widely used, are characterized by higher energy demands and shorter lifespans. In contrast, compact fluorescent lamps (CFLs) and light-emitting diodes (LEDs) [4] offer higher lighting efficacy, longer durability, and better energy performance. These diverse bulb types, while functionally different, collectively influence how energy is consumed in smart room environments. A significant challenge in designing energy-efficient lighting for indoor environments lies in accommodating the diverse characteristics of light bulbs such as incandescent, halogen, CFL, and LED, each with distinct energy consumption patterns and lighting efficacy.

In smart room environment, automation is transforming the management of lighting through the integration of sensing devices, energy data, and user behavior. Smart lighting systems equipped with motion sensors, ambient light detectors can dynamically control bulb power and operating time. The purpose is to ensure that lighting is delivered only when and where it is needed, thereby minimizing unnecessary consumption and improving energy efficiency. To analyze the efficiency of smart lighting systems, simulation-based modeling plays a crucial role [5], [6]. AnyLogic, a multi-method simulation tool, supports both Agent-Based Simulation (ABS) and Discrete Event Simulation (DES), enabling a detailed investigation of human-light interaction and dynamic system behaviors. These simulation paradigms are essential for mimicking real-world room scenarios that involve diverse bulb types and variable occupancy patterns. The present study leverages AnyLogic to simulate smart lighting control within a room environment, focusing on the behavioral dynamics of incandescent, halogen, CFL, and LED bulbs under various conditions.

Fig. 1 illustrates the basic structure of a smart lighting system for the room environment, which integrates motion sensors, ambient light sensors, and the energy measuring module. Each component of this structure has shown the basic strategy of reflecting real-time interactions, where automatic lighting is turned on and off based on user presence, lighting control is adjusted, and energy consumption is recorded across bulb types. This system-level visualization reinforces the need for adaptable lighting control systems that incorporate diverse bulb types for optimized energy consumption.

Despite advances in lighting control with energy-saving strategies, many existing lighting systems remain manually controlled or inadequately automated, leading to considerable energy wastage. Traditional systems often operate lights irrespective of occupancy status, resulting in excessive and

*Corresponding authors.

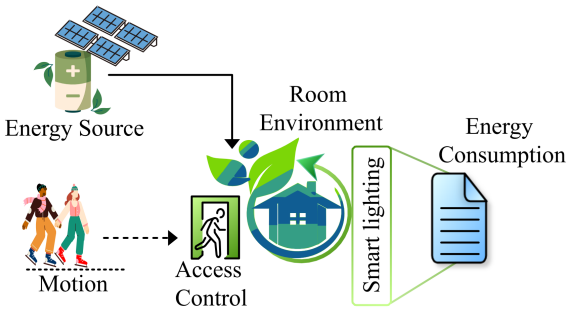


Fig. 1. Basic structure of smart lighting system for room environment.

inefficient usage. This is particularly problematic in residential and commercial rooms, where lighting accounts for a large portion of electricity energy [7], [8]. Moreover, while LED or CFL bulbs are substantially more efficient than traditional incandescent or halogen lamps, their potential is often underutilized due to a lack of an optimized structural approach. Smart lighting systems that adjust based on occupancy and ambient light offer promise but are hindered by barriers such as high initial costs, lack of integration with diverse bulb types, and limited modeling tools that can represent complex interactive behavior within smart rooms. Furthermore, real-world visualization and validation of energy savings through simulation tools like AnyLogic remains limited explored. This research addresses the gap by designing a simulation model that supports analysis of various lighting technologies within a smart room ecosystem, aiming to reduce unnecessary energy consumption and improve lighting efficiency in the room environment [9].

The rapid global increase in energy demand, coupled with environmental concerns and unattended light usage, motivates the adoption of smart lighting solutions. Since lighting constitutes a significant portion of energy consumption in room environments, adopting smart lighting controls is crucial for enhancing energy efficiency. The motivation for this research stems from the need to simulate and visualize a smart lighting simulation model that can dynamically respond to user presence [10] and ambient conditions while optimizing energy usage across different bulb types. By simulating and analyzing the performance of diverse bulb types under varying control schemes of a smart system that adapts in real-time, this study aims to provide an energy-efficient smart lighting solution adaptable to room environments and future green buildings. The study involves developing a simulation model in AnyLogic by integrating ABS and DES [11] for sensor-triggered lighting events with human behavior modeling. The room environment is modeled as an agent containing motion and light sensors that interact with diverse bulb agents. Each bulb type, such as incandescent, halogen, CFL, and LED is defined with specific wattage and bulb efficiency parameters. Simulation runs are conducted across various scenarios with differing occupancy rates, light availability, and bulb mixes. Energy consumption metrics are collected and analyzed to assess the performance and efficiency of each configuration. Accordingly, this study conveys the following contributions:

- This study provides a smart lighting system with operational process flow in room environments that serves automatic light management for diverse light bulbs.
- This study develops a smart lighting model with visualization using AnyLogic that integrates agent-based and discrete-event simulation for deployment in room environments.
- A comparative energy analysis is conducted to evaluate the performance of the proposed smart lighting systems across diverse bulb types.

In the upcoming sections, it covers the following aspects of this work. Section II provides an overview of the related studies, while Section III introduces the entire process and design of this work. Section IV presents the results and discussion of the exploration of this work. Finally, Section V concludes this study.

II. Related Work

In recent years, the increasing demand for energy-efficient lighting has intensified research efforts into smart control systems, user behavior, and advanced bulb technologies within room-based smart lighting environments. This section reviews notable studies on lighting energy optimization, modeling approaches, and performance assessments to contextualize the existing work.

Niko Gentile [12] in one work focuses on user behavior in lighting control, highlighting the role of dimming extent, interface design, heuristics, and feedback in reducing energy consumption. It identifies behavioral categories influencing energy efficiency and proposes strategies to nudge users toward sustainable habits. Though not simulation-based or bulb-specific, the study emphasizes human-centered lighting controls and estimates potential energy savings, highlighting the importance of integrating psychological and technical aspects for holistic lighting solutions. In another work, Campano et al. [13] presents a dynamic analysis of smart lighting controls based on daylight metrics and user requirements. It proposes three control systems as switch-based, dimmer-based, and sensor-driven, assessed using daylight autonomy metrics in test cells over one year. Although it does not simulate different bulb types, it effectively uses sensor-linked control schemes to reduce electric lighting energy savings. The study of Pašić and Imamović [14] compares the energy efficiency of LED bulbs against traditional ones like incandescent and halogen. Using direct energy consumption measurements, the authors establish LEDs as significantly more efficient. The work focuses on empirical comparison without simulation tools, and while it lacks dynamic modeling, it underscores the performance superiority of LEDs, which supports their integration into smart lighting frameworks for sustainable building applications. Besides, Ayaz et al. [15] proposed an urban street lighting system powered by hybrid renewable energy using solar and wind, integrating smart control for light intensity based on pedestrian and vehicle detection. The study emphasizes distributed generation and outlines practical system deployment. Although it focuses on outdoor lighting, the operational principles led by adaptive lighting control and renewable energy integration are directly applicable to indoor smart lighting simulations.

Moreover, Powers and Saad [16] presented lighting energy use in university buildings and models energy savings from retrofitting with LED lights and occupancy sensors. Using Python-based numerical simulation, the study demonstrates energy reduction and CO2 emission savings. Although it doesn't simulate bulb diversity, it emphasizes the effectiveness of sensor-linked LED systems and calculates economic payback, with a break-even point for the retrofit investment. In one study, Tsaknakis et al. [17] introduced a holistic smart building management system for elderly care, integrating lighting, HVAC, and battery storage. Using EnergyPlus for simulation, the system employs Approximate Dynamic Programming to optimize visual comfort, energy cost, and load safety. The lighting system adjusts based on daylight and occupancy. Though not bulb-specific, the integration of comfort and efficiency aligns with smart room lighting strategies in our study.

In addition, Singh and Narwal [18] conducted an audit-focused study that evaluates energy consumption in residential lighting systems and suggests LED and CFL replacements for GLS and tube lights. It lacks simulation or smart control strategies, but it emphasizes practical retrofitting approaches. This warehouse-focused study by Füchtenhans et al. [19] uses simulation to evaluate smart lighting strategies, comparing three operational schemes against conventional lighting. The system adapts to worker presence and order-picking schedules. Though specific bulb types are not detailed, LED-compatible smart systems are implied. The simulation evaluates warehouse layout, worker movement, and lighting control, offering insights into energy savings and system design trade-offs. The authors of another study, Imamguluyev et al. [20] highlighted the potential of fuzzy logic control (FLC) for optimizing lighting systems. FLC adapts illumination in real time based on occupancy and ambient light, enhancing energy efficiency. It discusses implementation challenges and presents real-world cases. Although not bulb-type-specific, the adaptive control mechanism resonates with our focus on modeling intelligent lighting systems that balance energy use and user comfort in room environments.

Most prior research focuses just on LED-based systems, specific-use contexts such as warehouses or elderly homes, or lacks integrated simulation frameworks that present real-time lighting systems. While some studies focus on real-world deployment with basic tools such as energy audits and retrofits, others utilize some modeling platforms like EnergyPlus or bespoke simulation scripts to evaluate cost and energy benefits. However, few papers holistically model lighting systems in room-level indoor environments using diverse bulb types. More critically, no studies combine agent-based simulation (ABS) with discrete-event simulation (DES) to support real-time smart lighting scenarios adaptable across CFL, LED, halogen, and incandescent bulbs. However, based on the problems identified in existing studies, the research question of this work is stated as: What type of simulation model is suitable for developing a smart lighting system, and how can ABS and DES modelling in AnyLogic effectively support real-time lighting control in room environments with diverse bulb types for energy analysis and system responsiveness? It presents a strong opportunity for the proposed study.

III. Methodology

This study introduces a simulation-based modeling view and methods for analyzing energy-efficient smart lighting in room environments supporting diverse bulb types, namely incandescent, halogen, compact fluorescent (CFL), and light-emitting diode (LED). The model is designed using AnyLogic simulation software [21], which enables a modeling approach combining Agent-Based Simulation (ABS) and Discrete Event Simulation (DES). This simulation modeling is crucial for capturing user behavior, occupancy sensors, and light sensors that control lighting conditions for the room environment. This methodology enables a comprehensive evaluation of each lighting scenario under varying usage conditions. The main design and the coordination process of energy-efficient smart lighting in room environments supporting diverse bulb types are illustrated as follows.

A. Overview of Conceptual Model for Smart Room Lighting System

Fig. 2 illustrates the proposed conceptual model for a smart lighting system designed to operate efficiently within a room environment using sensor-based automation and solar energy sources. The core elements of the model include occupants, a smart door, occupancy sensors, diverse room lights (Light 1 to Light 4), a solar energy source, a battery, a smart meter, and an energy analysis module. This system contains incandescent, halogen, CFL, and LED as Light 1, 2, 3, and 4, respectively. When an occupant approaches the room, the occupancy sensor detects their presence and triggers the smart door mechanism to open. Once the occupant enters, the system identifies the transition from an empty to an occupied state and responds by switching on the room lights. Each light is activated based on the specified capacity, influenced by sensor feedback and occupancy positioning.

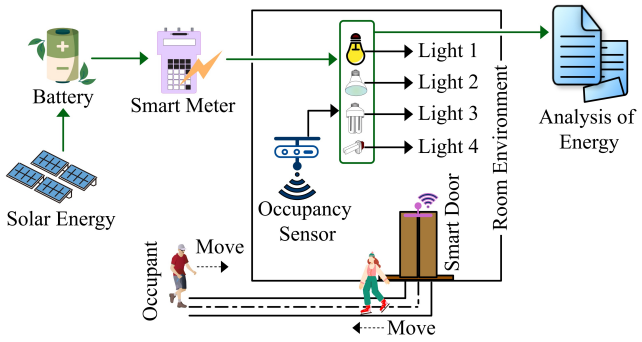


Fig. 2. Proposed model for smart lighting system in the room environment.

Energy for lighting is supplied from a solar panel system connected to a battery and charge controller. This configuration allows the room to operate on solar energy as renewable energy, reducing grid dependency. When lights are ON, a smart meter continuously records energy consumption and efficiency by providing essential data for energy analysis. This analysis helps assess lighting performance under different occupancy conditions, allowing optimization of energy usage based on real-time. Crucially, this model enables automatic decision-making. If multiple occupants are present, the lights remain

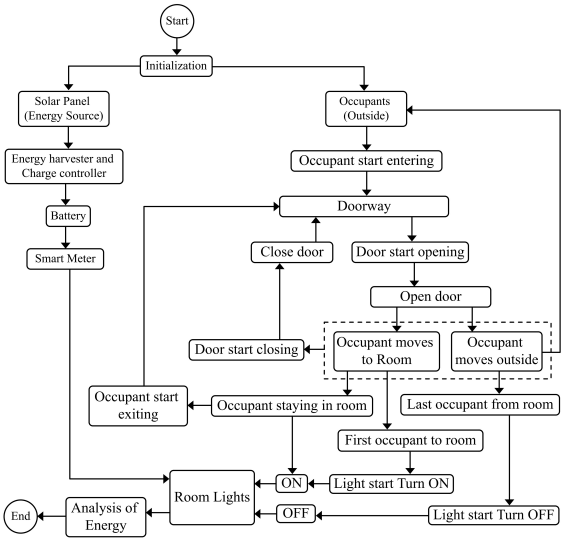


Fig. 3. Operational process flow for smart lighting system in the room environment.

ON. When the last occupant exits, the system detects the room's transition to an unoccupied state and turns off all lights, ensuring zero idle energy consumption. The proposed model supports diverse bulb types, namely LED, CFL, incandescent, and halogen, though each has distinct power characteristics. The smart meter's data, in combination with energy usage from different bulbs, enables comparison across scenarios. This allows the simulation model to assess energy usage or consumption and lighting effectiveness. The occupant's behavior, like entry, exit, and duration of stay, forms the central loop that drives the smart system, ensuring that lighting is only active when necessary. Hence, the model leverages the integration of sensors, renewable energy, and intelligent control to deliver an energy-conscious smart lighting system.

B. Operational Process Flow for Room Light Environment

Fig. 3 presents a detailed operational process flow of the smart lighting system in a room environment, capturing the chronological steps and conditional logic for lighting and door control, occupant behavior, and energy management. The process starts when occupants are outside the room. Once an occupant approaches the doorway, the occupancy sensor detects movement, prompting the smart door to open. The system marks the first occupant entering the room if no one is inside, triggering the room lights to turn ON. After entry, the door automatically closes, and the system maintains lighting as long as at least one occupant remains. During the occupant's stay, if additional individuals enter, the system tracks the occupancy count. Lights remain ON until the last occupant exits. When an occupant starts to exit, the sensor again detects motion, triggering the door to open and then close after departure. Once the last occupant leaves, the lights switch OFF automatically. This logic for the smart room ensures zero energy wastage in unoccupied conditions.

The energy required for lighting is drawn from a solar panel connected to a battery via a charge controller. This stored energy is consumed when the lights are ON, and the smart meter records the power usage in real-time. Data from the energy supplied to the lighting system is used for the energy analysis of diverse light bulb types. The system analyzes energy metrics such as energy consumption, lighting duration, and energy efficiency per bulb type. Hence, this operational flow coordinates sensor-based detection, automated door and lighting control, and intelligent energy usage analysis for room environments.

C. Visualization and Demonstration of Room Lighting Simulation Model

The smart lighting simulation model for the room environment has been developed using AnyLogic simulation software, leveraging its statechart functionality to design dynamic behaviors. Fig. 4 demonstrates the statechart for automated door control within the smart room lighting system, which plays a vital role in synchronizing user access with the lighting control mechanism. The door can exist in one of four states, namely Closed, Opening, Open, and Closing. These states are dynamically triggered by user presence and transition logic. When the door is in the Closed state, according to Fig. 4(a), the system waits for a user detection event. In this case, room lights are OFF. However, when a user approaches the entrance and detects user motion through a sensor, the door transitions to the Opening state, which is shown in Fig. 4(b). During this stage, lights remain OFF until the user starts entering. Once the door reaches the Open state, as shown in Fig. 4(c), the user is allowed to enter, and lighting control is triggered. If it is the first user, the system activates room lights based on the bulb type. As the user finishes entry, the door transitions to the Closing state according to Fig. 4(d). If additional users are detected, the door reopens. Once all users have entered and the door is re-closed, the lights remain ON as long as user occupancy persists.

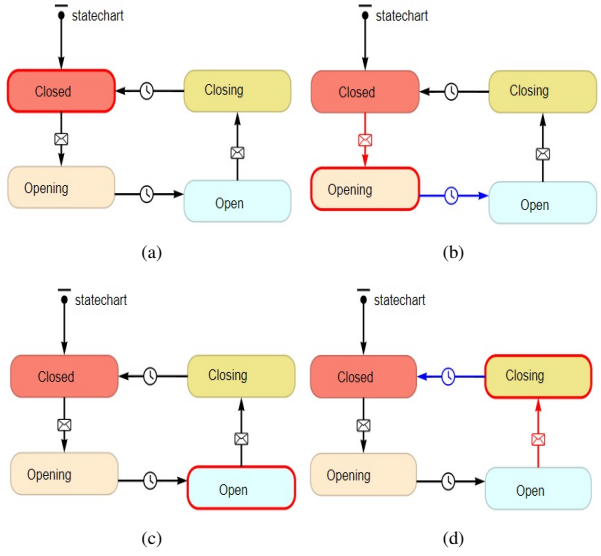


Fig. 4. Different door state transitions: (a) Closed; (b) Opening; (c) Open and (d) Closing.

Fig. 5 presents the statechart controlling user movement across four states, namely Outside, Entering, Inside, and Leaving, which influence the activation or deactivation of lighting

behavior based on user activity. When the user is in the Outside state, as shown in Fig. 5(a), the room remains unoccupied, and the lighting system is OFF. Upon detecting the motion sensor, it triggers the smart door to open, and the user transitions to Entering, which is depicted in Fig. 5(b), and simultaneously activates the lighting system based on the occupancy sensor. As the user transitions to the Inside state, like Fig. 5(c), lights remain ON continuously. During this stage, if multiple users are present, the light remains ON until the last user transitions to the Leaving state. The Leaving state, as demonstrated in Fig. 5(d), is activated when a user moves toward the exit. If the departing user is not the last, the lights remain ON. However, once the last user leaves, the system transitions to Outside, automatically turning lights OFF and logging the total energy consumed.

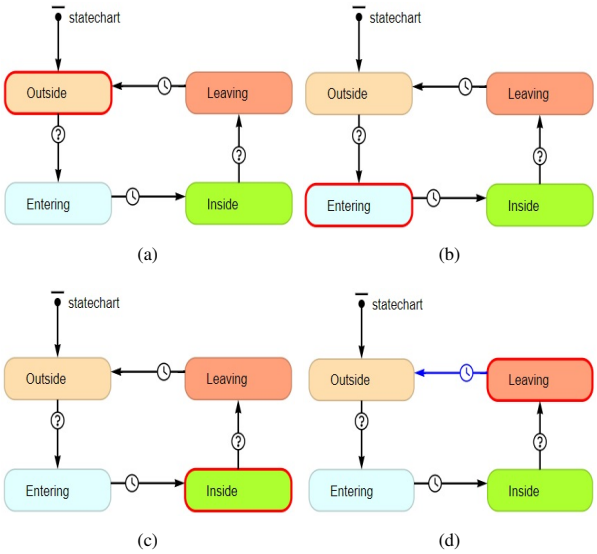


Fig. 5. User movement transitions in different cases: (a) Outside; (b) Entering; (c) Inside and (d) Leaving.

Algorithm 1 governs the smart lighting control logic by synchronizing user presence, door state, and light activation within the simulated room environment. When a user is detected near the door, it opens, updates the user state, and turns ON the lights based on the bulb type. Once inside, lights remain ON. When users exit, the door reopens, and lights turn OFF only if the room becomes unoccupied. Energy usage is logged dynamically, enabling adaptive lighting control based on occupancy and minimizing idle energy consumption.

Fig. 6, captured from the AnyLogic simulation runtime, visually represents the transitional states of a smart lighting system in a room environment, which highlights occupant behavior, door status, and lighting control through both 2D and 3D views. Fig. 6(a) and Fig. 6(b) illustrate the idle state, where all occupants are outside the room, the smart door remains closed, and the room lights are turned OFF. Thus, it minimizes unnecessary energy use. These states confirm that the system maintains a default OFF status in the absence of activity, conserving energy effectively. In another case, Fig. 6(c) shows this interaction in 2D, while Fig. 6(d) presents it from a 3D spatial perspective. These visuals emphasize the system's responsiveness to real-time human presence, maintaining optimal illumination as occupants move inward.

Algorithm 1 Smart lighting control based on door and user state transitions

Input:-
motionSensor: user detection for door functionality;
bulbType ← {LED, CFL, Halogen, Incandescent};
occupancySensor: user detection for light functionality;
roomUser: Number of users in the room;
doorState ← {Closed, Opening, Open, Closing};
userState ← {Outside, Entering, Inside, Leaving};
Output:-
lightStatus ← {ON, OFF}; energyUsage: energy consumption for different types of bulbs;

Initialize States
Set doorState ← Closed;
Set userState ← Outside;
Set lightStatus ← OFF;
Set roomUser ← 0;
while Simulation is running for entry event **do**
  if motionSensor = TRUE AND doorState = Closed **then**
    doorState ← Opening
    Activate door opening mechanism
    Wait for door open time duration
    doorState ← Open
    userState ← Entering
    roomUser ← roomUser + 1
    Call *Turn_ON_Lights*(bulbType)
    lightStatus ← ON
  end if
end while
while Transition to Inside **do**
  if userState = Entering AND *Light_ON* = TRUE **then**
    userState ← Inside
    Activate door closing mechanism
    doorState ← Closing
    Wait for door close time duration
    doorState ← Closed
  end if
  if userState = Inside AND user initiates exit AND doorState = Closed **then**
    doorState ← Opening
    Activate door opening mechanism
    Wait for door open time duration
    doorState ← Open
    userState ← Leaving
    roomUser ← roomUser - 1
    if roomUser = 0 **then**
      Call *Turn_OFF_Lights()*
      lightStatus ← OFF
      Record energyUsage in log
    end if
    userState ← Outside
    Activate door closing mechanism
    doorState ← Closing
    Wait for door close time duration
    doorState ← Closed
  end if
end while

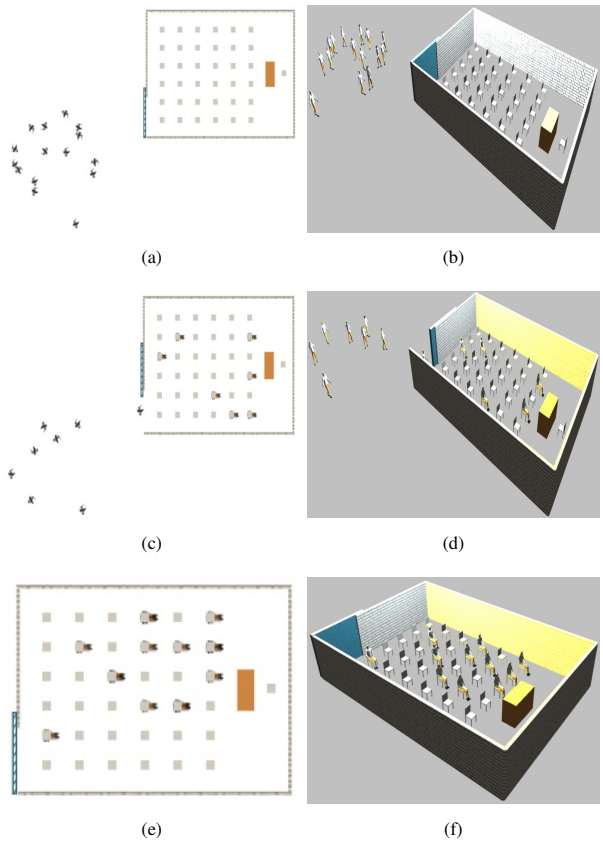


Fig. 6. Graphical layout: (a) Users are Outside when Door is closed in 2D; (b) Users are Outside when Door is closed in 3D; (c) Users are Entering when Door is opened in 2D; (d) Users are Entering when Door is opened in 3D.; (e) Users are InSide when Door is closed in 2D and (f) Users are InSide when Door is closed in 3D.

In this case, when a user is sensed at the door, and it opens, then the user enters the room, and automatically it turns ON the light bulb. Once all occupants have fully entered the room, the door closes again, and the lights remain ON, as depicted in Fig. 6(e) in 2D and Fig. 6(f) in 3D. This state reflects the active condition of the room, where lights are sustained to ensure visibility. The consistent light ON status confirms the system's awareness of active room use. These figures collectively showcase a seamless interaction between user presence, automated door control, and lighting behavior. This sequence of views validates the responsiveness and efficiency of the smart lighting system, where occupant-driven automation ensures both user convenience and optimal energy usage through intelligent control logic.

IV. Results and Discussion

To evaluate the effectiveness of the proposed smart lighting system in the room environment, a simulation model has been developed in AnyLogic software. This model allows for the simulation of basic scenarios of energy usage service and energy processing across different light bulb types. This section illustrates the results and analysis obtained from the simulation of the proposed smart lighting system for the room environment. The focus of the analysis is to assess the system's visualized responsiveness, energy usage, and efficiency based on occupancy patterns and diverse light bulb types. The smart lighting logic is triggered by user movements and the simulation dynamically adjusts the lighting status based on room occupancy. Door and user movements are governed by agent-based logic, while lighting events and energy analysis are modeled through discrete event simulation. The study demonstrates how intelligent control logic contributes to energy efficiency. The key performance metrics, namely energy usage, lighting duration, and bulb-wise energy profiles, are recorded. The output visualizations and plots extracted from the simulation runtime enable a comparative assessment of energy usage and efficiency in the model. This evaluation highlights the system's ability to conserve energy while demonstrating the effectiveness of adaptive lighting control supported by simulation.

A. Results

The simulation model for this work has been developed and executed using AnyLogic 8.9.4 Personal Learning Edition on a system configured with Windows 11, 64-bit, running on HP LAPTOP-2DHIQ60K. The machine is equipped with a 13th Gen Intel(R) Core(TM) i7-1355U processor (1.70 GHz) and 20.0 GB RAM, providing a robust environment for real-time simulation and visualization. The simulation has leveraged the modeling principles of agent-based simulation and discrete-event simulation to manage sensor inputs, door transitions, and lighting responses. Runtime visualizations have been generated for both 2D and 3D environments, and energy consumption data has been recorded through integrated smart metering components in simulation.

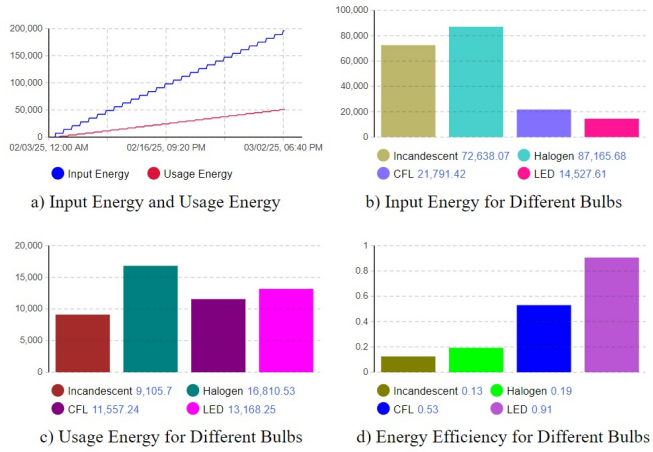


Fig. 7. Analysis view for diverse bulb types of the lighting system.

As illustrated in Fig. 7, the approach of this work enables the assessment of how user behavior influences lighting energy consumption across different bulb types, contributing insights into energy efficiency. During the simulation runtime, this model utilized incandescent, halogen, CFL, and LED bulbs, with power ratings of 50W, 60W, 15W, and 10W, respectively. In this case, Fig. 7(a) depicts a comparative analysis of the input energy and corresponding usage energy over a span of simulated days. The input energy, harvested from solar sources and stored in a battery system, steadily accumulates, while the usage energy fluctuates based on user presence and lighting demand. It shows that the energy usage increases sharply during

occupancy hours and illustrates that the input energy remains consistently higher than the energy usage for active light bulbs, indicating a surplus from the renewable supply. When the room is unoccupied, usage energy remains flat, validating the effectiveness of the smart lighting system in conserving energy for the overall pattern. This adaptive behavior ensures that only the necessary energy is consumed, with the remaining stored in the battery for future use, reflecting sustainable energy management.

Fig. 7(b) shows that this work has drawn an analogy of the input energy requirements for each type of bulb, namely LED, CFL, Halogen, and Incandescent, in a smart lighting system. The LED bulbs require the least amount of energy input, followed by CFL, Halogen, and Incandescent in ascending order. The ranking of the required energy aligns with the known wattage demands of these bulbs, which reflects the cumulative energy drawn from the energy source. In the simulation model, input energy is measured during the occupancy activity, ensuring an acceptable comparison. However, it reinforces the energy advantage of LED lighting in smart environments. Notably, the high energy input for incandescent and halogen bulbs suggests limited suitability for sustainable operations, especially in solar-powered setups. This insight can guide bulb selection for energy-conscious environments, emphasizing LEDs as the optimal choice for smart systems.

In Fig. 7(c), this work illustrates the actual energy usage or consumption by each bulb type over the simulation period for the smart lighting system. LED bulbs show the better energy usage, reinforcing their efficiency. CFLs follow with moderate usage, while Halogen and Incandescent bulbs exhibit substantially higher consumption levels. The data is directly related to the duration of light ON and the wattage rating per bulb. However, it specifically captures energy drawn during active lighting when occupants are present. It emphasizes the real-time energy cost incurred during system operation. The consistent interval between the usage energy of LEDs and other bulbs highlights their operational efficiency under dynamic conditions. It supports the case for deploying LEDs in environments with frequent user movement, where lights may toggle ON/OFF multiple times a day. The system's ability to minimize consumption without sacrificing functionality is demonstrated clearly through this usage pattern.

Fig. 7(d) presents a normalized analogy of energy efficiency across bulb types. Energy efficiency is defined here as the ratio of the useful output energy of light to the input energy required. The results show that LED bulbs achieve the highest efficiency, around 0.9, closely followed by CFLs. Halogen and Incandescent bulbs lag behind significantly, with incandescent efficiency being the lowest efficiency below 0.15. It reinforces the technological superiority of LED lighting in terms of both performance and sustainability. It validates the system's energy-saving claims by quantitatively showing how much energy is effectively utilized or wasted in smart room environments. These results support a significant outcome of the model that smart lighting systems, especially those powered by renewable energy sources, should prioritize high-efficiency bulbs to maximize overall system performance and minimize energy loss.

B. Discussion

This work demonstrates that the proposed smart lighting system, developed using AnyLogic with ABS and DES, effectively adapts to occupant presence while supporting diverse bulb types. Simulation graphs show that LED bulbs consume the least energy and offer the highest efficiency, while incandescent bulbs are the least efficient. Real-time automation, occupancy-based control, and energy logging validate the system's responsiveness and sustainability. The model also integrates solar energy and smart components, highlighting significant improvements in energy usage.

TABLE I. COMPARATIVE EVALUATION

Ref.	Simulation Tool	Supported Bulbs	Modeling Approach	Energy Efficiency	Key Advantage
[14]	None (Empirical study)	LED, CFL, Incandescent	Energy consumption	LED showed high savings	No automation or simulation model
[18]	None (Energy audit)	LED, CFL, GLS, Tube lights	Audit & replacement strategy	Recommend LED as most efficient	No dynamic modelling
[19]	Custom Simulation Model	SLS, Conventional light	Simulation-based operational strategy	SLS reduced warehouse energy costs	Lack of renewable energy integration
[22]	None (Arduino IDE)	LED, CFL	Optimisation strategy	LED is used for saving energy	No energy tracking or dynamic simulation
Our work	AnyLogic 8.9.4	LED, CFL, Halogen, Incandescent	ABS and DES modelling approach	LED most efficient, room context driven automatic light control	Real-time automation, 3D visualization, renewable energy integration, multi-bulb energy profiling

Table I presents a comparative evaluation of the functionalities and features of the presented smart lighting system with some existing studies. It shows that previous works either used basic empirical methods, manual audits, or lacked real-time automation and simulation. Most of them supported limited bulb types and did not incorporate renewable energy sources or simulation modeling techniques. However, the presented simulation model utilizes AnyLogic 8.9.4 with the ABS and DES approach, simulates four bulb types, and integrates smart automation with solar energy. It also supports 3D visualization and real-time behavior modeling. These advantages collectively demonstrate the superiority of the presented simulation model in terms of functionality and energy efficiency in smart lighting systems for room environments, compared to earlier research.

V. CONCLUSION

This work presents the design and simulation of an intelligent lighting control system for room environments using AnyLogic, integrating agent-based and discrete event simulation. The system supports diverse light bulbs and responds dynamically to user presence, optimizing energy usage through sensor-based automation and control logic. The simulation

results reveal that the proposed smart lighting model effectively minimizes energy consumption by adapting to user occupancy patterns. Among the evaluated bulb types, namely LED, CFL, Halogen, and Incandescent, LED bulbs consistently required the least energy input and achieved the highest energy efficiency, followed by CFLs, while halogen and incandescent bulbs consumed significantly more energy. The system successfully automated lighting control and energy logging using the smart door and motion sensor interactions. Idle energy use can be eliminated when all lights turn OFF and the room becomes unoccupied in a room environment. This work visually represents the transitional states of a smart lighting model in a room environment, which describes occupant behavior, door status, and lighting control in both 2D and 3D views at AnyLogic simulation runtime. This work contributes a dynamic, simulation-based approach to energy-efficient lighting design for a room. It supports valuable insights for designing energy-efficient building automation decisiveness and serves as a baseline for deploying adaptive lighting strategies in modern smart rooms and green buildings. The study is limited to simulation-based modeling only in single-room environments and is built on the assumption of optimal sensor performance without the influence of environmental noise. In the future, this work has the potential to broaden the model to encompass multi-room layouts, incorporate variations in weather conditions, and explore cost optimization in the context of real-world indecisiveness.

ACKNOWLEDGMENT

This research is funded by the Malaysia National Research Grant (FRGS/1/2023/ICT03/UMP/02/2) and Universiti Malaysia Pahang Al-Sultan Abdullah (RDU230131). This study is part of the research conducted by the University Postgraduate Research Grant (PGRS220338). This study is supported by Gopalganj Science and Technology University, Bangladesh.

REFERENCES

[1] V.P. Widartha, I. Ra, S.Y. Lee, and C.S. Kim, "Advancing smart lighting: a developmental approach to energy efficiency through brightness adjustment strategies," Journal of Low Power Electronics and Applications, vol. 14, no. 1, pp.6, 2024.

[2] M. Vaidya, P. Chatterjee, and K. Bhopatkar, "Energy Efficient Smart Lighting System for Rooms," In Smart Sensor Networks: Analytics, Sharing and Control, , Studies in Big Data, Springer, vol. 92, pp. 107–125, 2021.

[3] H. Ajra, M.A. Majid, and M.S. Islam, "IoT-enabled smart cities towards green energy systems: a review," International Journal of Reconfigurable and Embedded Systems, vol. 13, no. 3, pp. 708-723, 2024.

[4] M.H. Kabir, A. Al Noman, A. Al Afiq, R.H. Raju, M.N. Hasan, and Ahmad, "Design and Implement IoT-Based Intelligent Manageable Smart Street Lighting Systems for Future Smart City," Engineering Proceedings, vol. 56, no. 1, pp. 147, 2023.

[5] X. Yu, Z. Dong, S. Ge, D. Zhou, Q. Wang, and X. Sang, "Resource scheduling and performance analysis of hybrid renewable energy systems with carbon neutrality consideration: A scenario-based multi-agent approach," Sustainable Cities and Society, vol. 96, pp. 104688, 2023.

[6] I. Mahmood, T. Kausar, H.S. Sarjoughian, A.W. Malik, and N. Riaz, "An integrated modeling, simulation and analysis framework for engineering complex systems," IEEE Access, vol. 7, pp. 67497-67514, 2019.

[7] H. Ajra, M.A. Majid, M. S. Islam, A. B. Zabidi, and K. A. I. Hammad "IoT-Driven Sustainable Green Energy Systems in an Organization: Issues and Challenges," In Proceedings of International Conference on Research in Education and Science, vol. 9, pp. 1712 – 1726, 2023.

[8] L.I.U. Chenguang, "An efficient vision-based approach for optimizing energy consumption in internet of things and smart homes," International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, pp. 681-686, 2023.

[9] F.J. Montalbo and E. Enriquez, "An IoT smart lighting system for university classrooms," In 2020 International Symposium on Educational Technology (ISET), IEEE, pp. 3-7, 2020.

[10] I. Caglayan, and Y. Afacan, "Clustering-based agent system (CAS) to simulate the energy-related behaviours of office occupants," Architectural Science Review, pp. 1-15, 2024.

[11] M.A. Majid, U. Aickelin, and P.O. Siebers, "Comparing Simulation Output Accuracy of Discrete Event and Agent Based Models: A Quantitive Approach," In Proceedings of Summer Computer Simulation Conference, Turkey, pp. 1-8, 2010.

[12] N. Gentile, "Improving lighting energy efficiency through user response," Energy and Buildings, vol. 263, pp. 112022, 2022.

[13] M.Á. Campano, I. Acosta, S. Domínguez, and R. López-Lovillo, "Dynamic analysis of office lighting smart controls management based on user requirements," Automation in construction, vol. 133, pp. 104021, 2022.

[14] E. Pašić and N. Imamović, "Efficiency of Led Bulbs Compared to Conventional Bulbs—Energy Consumption Study," In Proceedings of the 2024 International Conference on Hydraulics and Pneumatics—HERVEX, Băile Govora, Romania, 13–15 November, pp. 1-8, 2024.

[15] M. Ayaz, S.W.H. Shah, H.B. Asif, and M.A.U. Haq, "Integrated Distributed Generation for Practical Implementation of Urban Lighting Using Hybrid Renewable Energy," In 2024 4th International Conference on Digital Futures and Transformative Technologies (ICoDT2), IEEE, pp. 1-6, 2024.

[16] A. Powers and M. Saad, "Building Energy Use: Modeling and Analysis of Lighting Systems—A Case Study," Sustainability, MDPI, vol. 14, no. 20, pp.13181, 2022.

[17] C. Tsaknakis, C. Korkas, K. Pippi, C. Athanasiadis, V. Alexopoulou, E. Kosmatopoulos, V.C. Nikolaidis, and T. Papadopoulos, "Nearly-optimal control for energy, thermal, and storage loads with energy disaggregation monitoring: A case of residential management for the elderly," IET Smart Cities, vol. 5, no. 3, pp. 173-193, 2023.

[18] H. Singh and R. Narwal, "Audit on Energy Efficient Lighting System," In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, pp. 1185-1190, 2021.

[19] M. Füchtenhans, C.H. Glock, E.H. Grosse, and S. Zanoni, "Using smart lighting systems to reduce energy costs in warehouses: A simulation study. International Journal of Logistics Research and Applications," vol. 26, no. 1, pp.77-95, 2023.

[20] R. Imamguluyev, T. Imanova, P. Hasanova, A. Mammadova, and S. Hajizada, "Unlocking energy efficiency: harnessing fuzzy logic control for lighting systems," Procedia Computer Science, vol. 230, pp.574-583, 2023.

[21] A. Stupin, L. Kazakovtsev, and A. Stupina, "Control of traffic congestion by improving the rings and optimizing the phase lengths of traffic lights with the help of anylogic," Transportation research procedia, vol. 63, pp. 1104-1113, 2022.

[22] J. Ng'andu, "Design and Development of an Automatic Lighting Control and Energy Optimization System," International Journal of Advanced Multidisciplinary Research and Studies, vol. 5, no. 2, pp. 461-485, 2025.

Deep Learning-Driven DNA Image Encryption with Optimal Chaotic Map Selection

Sara Bentouila[1], Kamel Mohamed Faraoun[2]
Department of Computer Science, Djillali Liabes University of Sidi Bel Abbes, Algeria[1,2]
EEDIS Laboratory, Algeria[1,2]

***Abstract*—This research introduces an advanced image encryption framework addressing critical security limitations in existing approaches. The study focuses on developing a robust encryption methodology that overcomes arbitrary chaotic map selection and static key generation vulnerabilities. Our approach integrates three synergistic components: a systematic chaotic map evaluation protocol identifying optimal dynamic systems, a deep learning-based key generation mechanism employing fine-tuned convolutional neural networks for image-sensitive cryptographic keys, and a hybrid encryption pipeline combining DNA encoding with chaotic diffusion. Experimental validation demonstrates that the proposed scheme achieves near-ideal entropy values (cipher images with an average entropy of 7.90 and above), and ensures extremely low correlation coefficients between adjacent pixels (close to zero in horizontal, vertical, and diagonal directions). Differential analysis confirms strong robustness, with NPCR values exceeding 99.6% and UACI about 33.5% across multiple color images. Visual results show that encrypted images display no perceivable patterns or similarities with the original images. Comparative performance assessment also highlights the method's efficiency, with encryption execution times competitive with or better than recent state-of-the-art methods. Brute-force resistance is guaranteed by an extensive key space determined by the combination of deep learning-generated keys, Lorenz chaotic parameters, and DNA encoding rule permutations. The comprehensive multi-layered security strategy further ensures resilience against brute-force, statistical, differential, and chosen-plaintext attacks, as well as against modern deep learning-based cryptanalysis.**

Keywords*—Image encryption; DNA encoding; chaotic map selection; lorenz system; deep learning; convolutional neural network (CNN); security analysis; VGG16; cryptographic robustness

I. Introduction

The exponential growth of digital image transmission in fields such as healthcare, defense, and education has brought about significant challenges in ensuring data confidentiality and integrity [1]–[5]. As image data is frequently exchanged over open networks, it becomes a prime target for a variety of cyber threats, including interception, unauthorized access, and statistical attacks [6]. Traditional cryptographic algorithms like DES, 3DES, and AES, while effective for text-based data, often fall short when applied to images due to high correlation among pixels and the large size of image files. This has prompted the research community to explore alternative approaches tailored to the unique characteristics of visual data [7], [8].

This research addresses critical gaps in current image encryption methods, as many existing schemes suffer from significant security vulnerabilities primarily due to the arbitrary selection of chaotic maps and the use of weak, static key generation mechanisms, which makes them susceptible to modern cryptanalytic attacks. This leads to three core questions: How can an optimal chaotic map be systematically selected? How can deep learning generate keys that are highly sensitive to image content? And to what extent does combining these techniques improve overall security? To answer these questions, our objectives are to establish a formal protocol for chaotic map selection using Lyapunov exponent analysis, develop an adaptive key generation method with a fine-tuned VGG16 network, and design and validate a hybrid encryption pipeline fusing DNA encoding and chaos theory. This study is significant as it provides a more rigorous scientific foundation for image encryption. Practically, it delivers a robust and efficient framework designed to secure sensitive visual data in critical sectors like healthcare and defense, enhancing resilience against a wide array of cyber threats.

In recent years, three main directions have emerged in the quest for secure image encryption: DNA-based encoding [2], [4], [5], [9]–[12], chaotic systems [4], [10], [11], [13]–[20], and the integration of machine learning techniques [21]–[28]. DNA-inspired methods leverage the biological properties of nucleotide sequences to enhance the complexity and unpredictability of encryption schemes. Chaotic maps, known for their sensitivity to initial conditions and pseudo-randomness, have been widely adopted to strengthen confusion and diffusion processes. Simultaneously, the rise of deep learning has enabled the development of adaptive, data-driven key generation mechanisms that can further improve security by making encryption keys highly dependent on image content.

The remainder of this study is organized as follows: Section II reviews recent advances in image encryption. Section III presents the theoretical background, including DNA encoding and chaotic maps. Section IV details the proposed encryption methodology. Section V discusses experimental results and security analyses. Finally, Section VI concludes the study and outlines future research directions.

II. Related Works

Recent years have witnessed significant advancements in image encryption techniques [2], [10], [11], [13]–[17], [24], [26]–[33], with researchers exploring various approaches to enhance both security and performance. These developments can be grouped into three main directions: DNA-based encryption, chaos theory, and the integration of machine learning techniques. Several researchers have investigated DNA-based encryption schemes. Notably, [2] conducted a systematic review that classifies DNA coding-based image encryption algorithms into five main categories, highlighting their operational

strengths, weaknesses, and future improvement avenues. In another contribution, [30] proposed a dynamic DNA coding algorithm for image encryption, driven by a conservative chaotic map and rigorously-tested pseudo-random sequences, which results in high key sensitivity and strong resistance to standard cryptanalytic attacks. Moreover, [11] introduced a multi-objective genetic algorithm framework optimizing DNA-based masks using coupled map lattice chaos; their method achieves Pareto-optimal encryption results that are thoroughly evaluated by advanced security metrics. Complementing this, several studies have employed various chaotic maps for image encryption. For example, [13] presented an image encryption algorithm that addresses the issue of finite computational precision, while [15] developed a memristive hyperchaotic system for improved encryption robustness. Additionally, [31] combined dynamic DNA operations with multiple chaotic maps and reinforced their scheme with SHA-256 hashing, leading to increased sensitivity, efficiency, and broader resistance to cryptanalytic attacks compared to previous works. However, many of these approaches still rely on randomly selected chaotic maps, often without a systematic evaluation of their dynamic properties. Recently, deep learning has been increasingly integrated with cryptographic methods to further enhance image security. In [24], the authors made a significant contribution by developing a chaotic log-map-based encryption scheme with key generation powered by a deep neural network; their approach demonstrated superior results, setting a benchmark for CNN-based encryption methods. Another innovative framework was introduced by [27], who integrated dynamic DNA coding with CNN-based key generation, employing an intertwining logistic map to improve entropy and strengthen cryptanalytic resistance beyond the state of the art. Most recently, [26] presented a technique that combines hyperchaotic maps and convolutional neural networks for image encryption, resulting in high key sensitivity, efficient visual scrambling, and strong resilience to noise and attacks, as confirmed by differential and statistical analyses. Nevertheless, these advanced methods often lack a comprehensive analysis of the chaotic system's behavior and its actual impact on encryption security. Despite these considerable advances, existing methods face several persistent limitations: insufficient key sensitivity and limited key space, inadequate resistance to statistical attacks, challenges in addressing pixel correlation effectively, lack of systematic approaches for chaotic map selection, and underutilization of modern deep learning techniques. Our work addresses these gaps by introducing a novel approach that harmoniously combines systematic chaotic map selection, deep learning-based key generation, and DNA encoding techniques.

III. Preliminaries

A. DNA Encoding Technique

DNA encoding represents an innovative fusion of genetics and computer science principles, serving as a fundamental component of our encryption scheme. This technique transforms digital data into biological-inspired sequences using the four nucleotide bases of DNA: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C).

The encoding process converts binary data into DNA sequences through a systematic mapping where each pair of binary digits corresponds to a specific DNA base. While theoretically 24 different encoding rules are possible, only 8 satisfy the Watson-Crick complementarity principle [15], which maintains biological consistency. These 8 valid encoding rules (see Table I) ensure that complementary DNA bases are properly paired (A-T and G-C), providing a natural error-checking mechanism.

TABLE I. DNA Sequence Binary Coding Rules

Rules								
1	2	3	4	5	6	7	8	
00	00	11	11	10	01	10	01	A
11	11	00	00	01	10	01	10	T
10	01	10	01	00	00	11	11	C
01	10	01	10	11	11	00	00	G

For example, a pixel value of 150 (binary: 10010110) can be encoded into different DNA sequences depending (CTTC, GTTG, CAAC, GAAG, AGGA, TGGT, ACCA, and TCCT) on the selected encoding rule. The DNA sequences can then undergo various operations, including XOR, which follows specific rules defined by a lookup table (see Table II) that preserves the biological properties of DNA base pairing.

TABLE II. DNA Sequence XOR Operator

XOR				
A	T	C	G	
A	T	C	G	A
T	A	G	C	T
C	G	A	T	C
G	C	T	A	G

This DNA-based approach offers several advantages for image encryption: It provides multiple valid encoding options, increasing the complexity of the encryption process, the biological rules add an additional layer of validation and security, the encoding scheme naturally supports binary operations while maintaining biological constraints, the method efficiently handles large volumes of image data through compact DNA representations. The integration of DNA encoding in our encryption system enhances both the security and efficiency of the overall encryption process, while maintaining the biological inspiration that makes this approach unique.

B. Chaotic Maps

Before presenting the proposed algorithm, we will first analyze the properties of different chaotic maps and select the best ones. In mathematics, it is a function that perceives a certain amount of chaos [34]. The main challenge of this study was to select the map that would be integrated with the encryption and decryption system. Various chaotic maps exist, each with its strengths and weaknesses. To pick the best map, we just encrypt a simple image with each map. We subsequently assess each map on the randomness (i.e., the chaotic spread of their respective parameter evaluation values) they generate. The following chaotic maps were used for analysis:

1) Logistic map: As shown in Eq. (1), a logistic map [35] can be mathematically described:

$$L_{n+1} = P \cdot L_n(1 - L_n) \tag{1}$$

This parameter P can go from 0 to 4. The most amount of randomness (chaotic behaviour) is seen when P is in between 3.5 and 4.

2) Arnold map: This is a messy map that will actually be used by most of the pixel confusion. Arnold cat [36] maps as shown in Eq. (2):

$$\Gamma : \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & R \\ S & RS+1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \bmod n \tag{2}$$

3) Henon map: This is a chaotic discrete time dynamical system. The Henon [37] transforms the point (h_n, k_n) according to Eq. (3) and Eq. (4):

$$h_{n+1} = 1 - ah_n^2 + k_n \tag{3}$$
$$k_{n+1} = 1 - bh_n \tag{4}$$

When a and b are set to 1.4 and 0.3, respectively, this mapping becomes chaotic.

4) Tent map: The Tent map [38] can be expressed as shown in Eq. (5):

$$t_\rho = \rho * \min\{a, 1-a\} \tag{5}$$

where, ρ represents the real consumption of t_ρ, ρ is a parameter in $(0, 2)$. The map's name derives from the shape of the plot for t_ρ, which resembles a tent.

5) Duffing map: A duffing map [39] is an example of chaos in discrete time dynamic; it is also called a Holmes map. The Duffing equation describes damped oscillators. The update rules for the system are given in Eq. (6) and Eq. (7):

$$h_{n+1} = k_n \tag{6}$$
$$k_{n+1} = -bh_n + ak_n - k_n^3 \tag{7}$$

The map here is determined by two constants a and b. At $a = 2.75$ and $b = 0.2$, it behaves chaotically.

6) Sine map: The range becomes $[0, 1]$ when the sine function is evaluated for $[0, \pi]$. The outputs of a sine function [40] are now flowed through a sine map where they are mapped into $[0, 1]$ as inputs. The sine map can be generally defined as in Eq. (8):

$$m_{i+1} = s \cdot \sin(\pi m_i) \tag{8}$$

where, s is a parameter in $[0, 1]$. For $0.87 < s < 1$, the sine map is chaotic.

7) Lorenz map: Differential equations describe the Lorenz system [16] in the following way:

$$\frac{dx}{dt} = \alpha y - \alpha x \tag{9a}$$
$$\frac{dy}{dt} = \gamma x - xz - y \tag{9b}$$
$$\frac{dz}{dt} = xy - \beta z \tag{9c}$$

Control parameters are called α, β, and γ, while state variables are x, y, and z. Using Eq. (9a) to Eq. (9c), the Lorenz chaotic attractor is plotted with the following control parameters and initial values x_0, y_0, z_0. From the graph presented in Fig. 1, it is possible to notice that most Lorenz orbits display chaotic movement for $\alpha = 10$, $\beta = 2.667$, and $\gamma = 28$.

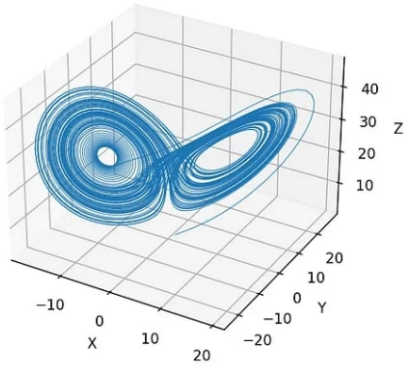


Fig. 1. Lorenz Attractor with $\alpha = 10$, $\beta = 2.667$, and $\gamma = 28$.

C. Chaotic Map Selection Through Comprehensive Security Metrics Analysis

The selection of an optimal chaotic map constitutes a fundamental decision in image encryption systems, as it directly influences the cryptographic strength and security properties of the entire scheme [41], [42]. Rather than employing arbitrary selection methods commonly found in existing literature, our approach implements a systematic evaluation framework that comprehensively assesses seven prominent chaotic systems through multi-dimensional security analysis. Algorithm 1 presents our novel comparative methodology that evaluates logistic, Duffing, Henon, tent, sine, Arnold, and Lorenz maps using six critical security metrics to ensure objective and quantifiable selection criteria.

Algorithm 1: Systematic Chaotic Map Selection Through Multi-Criteria Analysis

Input: Test image I, chaotic map pool M = logistic, duffing, henon, tent, sine, arnold, lorenz.
Output: Optimal chaotic map M_{opt} with comprehensive performance metrics.
Step 1: Initialize evaluation matrix $\Phi[|M| \times 7]$ for storing metrics;
Step 2: For each chaotic map $m_i \in M$:
- Generate chaotic sequences:
  $X1$ = generate_sequence(m_i, x_0, N);
  $X2$ = generate_sequence($m_i, x_0 + \varepsilon, N$) where $\varepsilon = 10^{-10}$;
- Calculate the Lyapunov exponent:

λ_i = calculate_lyapunov($X1$, $X2$);
- Encrypt test image: E_i = encrypt_image(I, m_i);
- Compute security metrics:
  Entropy: H_i = shannon_entropy(E_i);
  Correlation: P_i = correlation_coefficient(I, E_i);
  NPCR: N_i = differential_analysis(I, E_i);
  UACI: U_i = uniform_analysis(I, E_i);
  SSIM: S_i = structural_similarity(I, E_i);
  Execution time: T_i;
- Store metrics: $\Phi[i] = [\lambda_i, H_i, P_i, N_i, U_i, S_i, T_i]$;

Step 3: Rank maps by composite security score prioritizing Lyapunov exponent;
Step 4: Return $M_{opt} = argmax(\lambda_i)$ with highest chaotic behavior.

The Lyapunov exponent calculation employed in Algorithm 1 follows the enhanced methodology presented in Eq. (10):

$$lyap = (1/N) * \sum log(|X_2[i] - X_1[i]|/\varepsilon) \tag{10}$$

where, N represents the number of iterations, X_1 denotes a chaotic sequence generated with initial condition x_0, X_2 represents a chaotic sequence generated with condition $x_0 + \varepsilon$, and $\varepsilon = 10^{-10}$ serves as the sensitivity measurement parameter. This equation quantifies the exponential divergence rate

between two nearby trajectories, providing a direct measure of the system's sensitivity to initial conditions.
Our comprehensive experimental analysis using Algorithm 1 reveals significant performance variations among the evaluated chaotic systems, as detailed in Table III. The systematic evaluation methodology eliminates subjective selection bias while providing quantitative evidence supporting the Lorenz map's adoption, thereby establishing a rigorous scientific foundation for chaotic map selection in secure image encryption applications.

The results demonstrate that the Lorenz chaotic system achieves superior performance across all evaluated dimensions, obtaining the highest Lyapunov exponent of 24.76 through the systematic application of Eq. (10), which represents a substantial 9.5% improvement over the second-best performing Duffing map. This quantitative superiority, combined with optimal entropy and diffusion properties, conclusively validates the Lorenz system's selection for our encryption scheme.

D. CNN Key Generation via Fine-Tuned VGG16 and Sensitivity Analysis

The cryptographic key generation process in this work is based on a refined VGG16 convolutional neural network [24], [26]–[28], designed to ensure both high randomness and strong image sensitivity. The model is initialized with ImageNet weights and further adapted to the encryption task by unfreezing the last six convolutional layers, enabling it to learn domain specific features. The training dataset consists of a diverse set of natural and benchmark images, and data augmentation such as random flips, rotations, and brightness changes is systematically applied to promote generalization and robustness.

The network is trained using a triplet loss function, which encourages the extracted features to be highly discriminative: images that differ, even by a single pixel, yield distinct feature representations. After feature extraction, the output passes through a dense layer with SELU activation, followed by Gaussian noise injection and layer normalization to enhance the randomness and stability of the resulting 512-dimensional feature vector. This vector is then binarized (threshold 0.5) to produce the initial key. To further strengthen the cryptographic properties, the binary vector is combined via XOR with a SHA-256 hash of a secret key, ensuring unpredictability and a large key space. The entire process is detailed in Algorithm 2.

To empirically confirm the image sensitivity of the generated keys, a key sensitivity analysis was performed: for each test image, a single pixel was modified and the Hamming distance between the original and modified keys was measured. On average, 253 out of 512 bits changed, demonstrating that the key generation process is highly responsive to image content. Furthermore, the entropy of the generated keys consistently exceeded 7.99 bits per byte, indicating near-ideal randomness. This comprehensive pipeline ensures that each image yields a unique, unpredictable key, providing a secure and adaptive foundation for the overall encryption system.

IV. Proposed Method

Our proposed encryption scheme consists of three main components, each contributing to the overall security and efficiency of the system. The methodology can be described (Fig. 2) as follows:

- Chaotic Map Selection: The first component focuses on choosing the optimal chaotic map for encryption. Based on the analysis detailed in Section III-C, the Lorenz map is chosen due to its superior Lyapunov exponent value, ensuring maximum sensitivity to initial conditions and enhanced security properties.
- Key Generation Mechanism:
  The second component implements a novel two-stage key generation process:
  1) As described in the Section III-D , a CNN key is generated using deep CNN feature extraction, specifically leveraging the VGG16 architecture to process the input image. This approach enhances key sensitivity by incorporating image-specific characteristics into the key generation process.
  2) The final encryption key is cryptographically strengthened through an XOR operation between the CNN-generated key and the SHA-256 hash of a pre-shared secret key. This process leverages the one-way property of SHA-256 to prevent key recovery from intercepted data while enabling deterministic reconstruction for authorized parties.

  The system implements two-factor security: the CNN key (transmitted with the encrypted image) provides image-specific randomness, while the securely stored secret key ensures long-term confidentiality. This approach integrates both cryptographic robustness and compliance with modern security standards.
- Encryption Algorithm
  The final component implements the core encryption process, which integrates:
  - DNA-based encryption methods for enhanced security.

TABLE III. Comprehensive Chaotic Map Performance Analysis Using Algorithm 1

Chaotic Map	Lyapunov Exponent	Entropy	Correlation	NPCR (%)	UACI (%)	SSIM	Time (ms)
Lorenz	24.76	7.999	-0.002	99.62	28.63	0.010	2064.44
Duffing	22.61	7.999	-0.002	99.62	28.66	0.010	881.78
Henon	22.39	7.999	0.002	99.60	28.64	0.009	767.49
Logistic	21.62	7.999	0.002	99.61	28.57	0.010	488.65
Sine	21.58	7.999	0.003	99.61	28.53	0.010	1073.66
Arnold	21.52	7.999	-0.003	99.62	28.68	0.009	916.03
Tent	21.52	7.999	0.001	99.62	28.56	0.010	406.17

Algorithm 2: CNN-Based Key Generation

Input: Input image I, secret key K_secret.
Output: 512-bit cryptographic key K_final.
Step 1: Preprocess image I:
- Resize to 224×224 pixels
- Apply random horizontal flip, rotation (±15°), and brightness adjustment

Step 2: Feature extraction:
- Load VGG16 with ImageNet weights
- Unfreeze last 6 convolutional layers for fine-tuning
- Pass preprocessed image through the network
- Extract features from dense layer with SELU activation
- Apply Gaussian noise and layer normalization

Step 3: Key binarization:
- Threshold features at 0.5 to obtain a 512-bit binary vector

Step 4: Key strengthening:
- Compute SHA-256 hash of K_secret and convert to binary
- XOR the CNN-derived binary vector with the hashed secret key

Step 5: Output K_final as the encryption key

- Lorenz chaotic map sequences for generating pseudo-random numbers.
- The final key produced by the CNN-based key generation mechanism.

This integration creates a hybrid encryption scheme that leverages both the unpredictability of chaotic systems and the biological properties of DNA encoding, while being guided by the image-sensitive key generated through deep learning. The combination of these three components results in a comprehensive encryption system that provides strong security properties while maintaining computational efficiency. The system's design ensures that each component complements the others, creating multiple layers of security that protect against various types of cryptographic attacks.

A. Key Partitioning and Parameter Initialization Process

The key partitioning and parameter initialization process represents a fundamental component of our encryption system. As shown in Algorithm 3, utilizing a 512-bit final key K strategically divided into distinct segments. The first 219 bits are allocated to establish the Lorenz chaotic system's initial conditions (x_0, y_0, z_0), with each parameter derived from 73 bits and normalized to the [0,1] range. The subsequent 219 bits determine the DNA encoding rules (R_b, R_g, R_r) for the RGB channels, with each color channel assigned 73 bits to generate rule indices between 1 and 8. The remaining 74 bits establish the DNA encoding rule for the key matrix (R_{key}), the details of this process are described in Algorithm 3.

Algorithm 3: Final Key Division and System Parameters Initialization

Input: Final key K (512 bits);
Output: Lorenz parameters (x_0, y_0, z_0) and DNA encoding rules (Rb, Rg, Rr, R_{key})
1- Lorenz Map Initial Parameters (219 bits): $x_0, y_0, z_0 \in [0, 1]$
- $x_0 = (K_1, ..., K_{73})/2^{78} - 1;$
- $y_0 = (K_{74}, ..., K_{146})/2^{73} - 1;$
- $z_0 = (K_{147}, ..., K_{219})/2^{73} - 1;$

2- DNA Encoding Rules (219 bits): $(Rb, Rg, Rr) \in \{1, 2, ..., 8\}$
- $R_b = (((K_{220}, ..., K_{292})/2^{73} - 1) \times 7) + 1;$
- $R_g = (((K_{293}, ..., K_{365})/2^{73} - 1) \times 7) + 1;$
- $R_r = (((K_{366}, ..., K_{438})/2^{73} - 1) \times 7) + 1;$

3- Key Matrix Encoding Rule (74 bits): $R_{key} \in \{1, 2, ..., 8\}$
- $R_{key} = (((K_{439}, ..., K_{512})/2^{74} - 1) \times 7) + 1;$

This structured approach ensures unique initialization parameters for each system component while maintaining cryptographic strength through proper parameter distribution. The process creates a robust foundation for subsequent encryption operations by effectively linking the key generation phase with both the DNA encoding and chaotic sequence generation processes.

The main algorithm parameters-including chaotic map settings, key generation options, and DNA rule selectionplay a direct role in encryption strength, key sensitivity, and computational cost.

B. Encryption Scheme Operations

The encryption process represents a sophisticated fusion of DNA-based encryption, chaotic sequences, and deep learning-derived key. As shown in Fig. 2 and Algorithm 4, the process begins with the decomposition of the input RGB image into its constituent color channels. Each channel undergoes independent processing through multiple stages of encryption. Initially, each color channel is converted into its binary representation before being encoded using DNA rules derived from the partitioned final key. The DNA encoding process employs eight distinct encoding rules, with specific rules (Rb, Rg, Rr, quoted in Algorithm 3) assigned to each color channel based on key segments, ensuring unique transformation patterns for different image components. The process then leverages the Lorenz chaotic system, initialized with parameters (x_0, y_0, z_0) derived from the final key, to generate pseudo-random sequences. These sequences serve multiple purposes: creating scrambling patterns for pixel positions, generating additional DNA encoding sequences, and providing diffusion mechanisms throughout the encryption process.

V. Analysis of Experimental Results and Security

In order to test the proposed scheme for encrypting color images, a visual analysis is performed. By comparing the original images and their encrypted counterparts, it is impossible to find any noticeable similarities between them. The experimental results for the Lenna, Fruits, Flowers and Airplane images are shown in Fig. 3. No color clusters or similarities can be perceived between the encrypted image and the original image.

The subsequent section elaborates on entropy, statistical attacks (coef, histogram analysis), brute force attacks (key sensitivity analysis, secret key space analysis), and differential attacks (NPCR and UACI) for further exploring the proposed approach.

A. Entropy

An important measure of randomness and unpredictability in an encryption system is Entropy. In the implementation part, we calculate the entropy of the original image as well as the encrypted image. These entropies obtained after encryption are high which good dispersion of pixel values and also prevents various statistical attacks. The entropy is computed using Eq. (11), which is given below:

$$H(e) = -\sum_{i=1}^{n}(P(e_i)log_2P(e_i)) \tag{11}$$

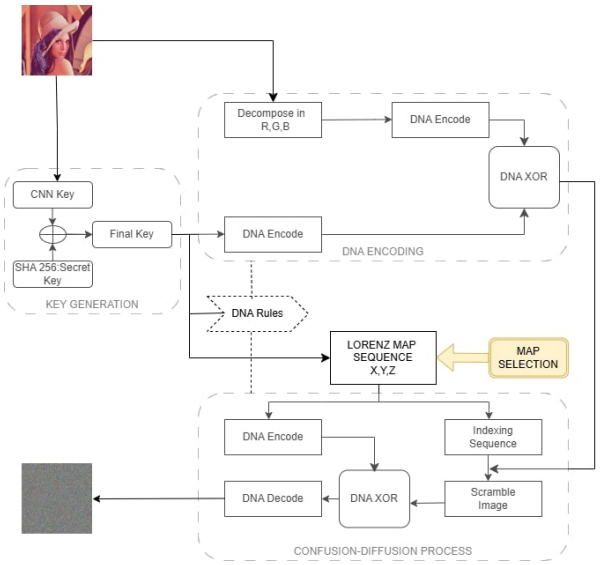


Fig. 2. The proposed scheme for encrypting images.



(a) Lenna (b) Lenna (c) Lenna (d) Fruits (e) Fruits (f) Fruits

(g) Flowers (h) Flowers (i) Flowers (j) Airplane (k) Airplane (l) Airplane

Fig. 3. Plain images (a,d,g,j), encrypted images (b,e,h,k), and decrypted images (c,f,i,l).

$P(e_i)$ represents the probability of occurrence of e_i. The entropy values of the cipher images in Table IV are a proof for the efficiency of the proposed algorithm.

B. Statistical Attacks

Correlation coefficients are checked in horizontal, diagonal and vertical ways under statistical analysis as well as histogram analysis. In a strong encryption method, the correlation between two adjacent pixels is as least as low. A histogram analysis is used to analyze the uniformity in pixels.

1) Analysis of the Correlation Coefficients: In all images, a certain degree of correlation remains between each pair of neighboring pixels. In order to protect data against various attacks [30], good encryption algorithms are designed such that such correlations between pixels should be absent or hidden. In order to discover the correlations between pixel pairs, it is necessary to choose specific proximity pixels of the input image so that it can be in three directions, that is, horizontal (H), vertical (V), and diagonal (D). The correlation coefficient between pixel pairs is computed using Eq. (12), Eq. (13), and Eq. (14):

$$C_{xy} = \frac{S^2 \cdot \text{cov}(x,y)}{\sum_{i=1}^{S}(x_i - E_x)^2 \cdot \sum_{i=1}^{S}(y_i - E_y)^2} \tag{12}$$

$$E_x = \frac{\sum_{i=1}^{S} x_i}{S} \tag{13}$$

$$\text{cov}(x,y) = E((x - E_x)(y - E_y)) \tag{14}$$

S is the image's size, and the (x, y) sequence is one of two neighboring horizontal, vertical, or diagonal pixels. In Fig. 4, the correlation distribution for each couple of pixels of the Lenna image is illustrated in three different orientations as follows (horizontal(H), vertical(V), and diagonal(D)), as well as the correlation distribution for the encrypted image that

Algorithm 4: DNA-based Encryption Process with Chaotic Diffusion

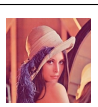
Input: Original RGB image I, Final key K (512 bits);
Output: Encrypted image E;
Step 1. Channel Decomposition
• Split input image I into RGB channels: B, G, R = decompose(I);
Step 2. DNA Encoding
• For each channel C ∈ {B, G, R}:
Convert to binary: Cbin = binary_convert(C);
Apply DNA rules based on key segments:
B_DNA = DNA_encode(Bbin, Rb)
G_DNA = DNA_encode(Gbin, Rg)
R_DNA = DNA_encode(Rbin, Rr)
Step 3. Key Matrix Generation
• Generate key matrix Km using Rkey;
• Encode key matrix: Km_DNA = DNA−encode(Km, Rkey);
Step 4. First DNA XOR Operation
• B_xor = XOR_DNA(B_DNA, Km_DNA);
• G_xor = XOR_DNA(G_DNA, Km_DNA);
• R_xor = XOR_DNA(R_DNA, Km_DNA);
Step 5. Chaotic Sequence Generation
• Generate Lorenz sequences using (x0, y0, z0):
$\frac{dx}{dt} = \alpha(y - x)$
$\frac{dy}{dt} = \gamma x - xz - y$
$\frac{dz}{dt} = xy - \beta z$
where $\alpha = 10, \beta = 2.667, \gamma = 28$
Step 6. Scrambling Process
• Generate index sequences INx, INy, INz from chaotic sequences;
• Apply scrambling to XORed DNA sequences:
B_scr = scramble(B_xor, INx, INy, INz)
G_scr = scramble(G_xor, INx, INy, INz)
R_scr = scramble(R_xor, INx, INy, INz)
Step 7. Final DNA XOR
• Convert chaotic sequences to DNA: DNAx, DNAy, DNAz;
• Perform final XOR operation:
B_final = XOR_DNA(B_scr, DNAx)
G_final = XOR_DNA(G_scr, DNAy)
R_final = XOR_DNA(R_scr, DNz)
Step 8. DNA Decoding and Image Recovery
• Decode DNA sequences to binary;
• Convert binary to pixel values;
• Combine channels to form encrypted image E;
Return: Encrypted image E

TABLE IV. ENTROPY OF PLAINTEXT IMAGES AND ENCRYPTED IMAGES USING OUR PROPOSED METHOD

Image name	Lenna	Fruits	Flowers	Airplane
Plain image entropy	7.7521	7.5207	7.5505	6.7256
Cipher image entropy	7.9073	7.9032	7.9032	7.9059

corresponds. The correlation distributions of the Fruits image, Flowers image, and Airplane image are shown in Fig. 5, Fig. 6, and Fig. 7, respectively. The correlation coefficients for all pairs between each two pixels of the image in the three directions of H, V, and D are very low, as depicted in the Table V. The encrypted version then effectively makes it impossible for the attackers to discover any kind of pattern and eventually decrypt the images.

TABLE V. CIPHER-IMAGE CORRELATION COEFFICIENTS

				
Horizontal	-0.0024	0.0021	-0.0015	-0.0039
Vertical	-0.0038	0.0041	0.0003	-0.0001
Diagonal	0.0014	0.0060	0.0024	-0.0052

2) Histogram Analysis: An image histogram that represents the distribution of the pixel intensity values provides statistical details about an image. A uniform histogram will ensure that the image encryption system is protected against statistical attacks [36]. Histogram analysis is very important to check the resistance of the system against statistical attacks. We created original and encrypted image histograms, as seen below on Fig. 8. Encryption systems should have a uniform distribution of pixel values as we can see from the histogram of the encrypted image. This helps to keep the statistical properties of the original image hidden as much as possible, making it much harder for an attacker to recover information about the original image based on statistics of the encrypted image.

C. Brute-force Attack

Key space and key sensitivity are evaluated to prevent brute-force attacks:

1) Key space: The cryptographic system's security relies on an exceptionally expansive key space derived from three core entropy sources. First, CNN-based key generation produces 2^{512} unique possibilities, ensuring substantial unpredictability. Then, Lorenz chaotic initialization generates extensive state diversity through extreme sensitivity to initial conditions. Finally, DNA encoding rules contributes $8^4 = 4096$ distinct configuration options. This multi-layered architecture guarantees that the comprehensive key space significantly surpasses contemporary security standards, effectively eliminating the feasibility of brute-force attacks.

2) Key Sensitivity: In an effective image encryption technique, a tiny change in the starting value ought to lead to an enormous transformation in the cipher image, based on main sensitivity analysis. The key sensitivity analysis of the proposed method is investigated by using the Lenna, Fruits, Flowers and Airplane images (Fig. 9a, Fig. 9e, Fig. 9i, Fig. 9m). First, it encrypts the image with a 512-bit secret key. Then again, the encryption process repeats, but this time, the secret key is changed by a single bit so that one bit with value 0 is changed to value 1.

The figures (Fig. 9b, Fig. 9f, Fig. 9j, Fig. 9n) and (Fig. 9c, Fig. 9g, Fig. 9k, Fig. 9o), represents the related cipher images related to the previous secret key and modified one.

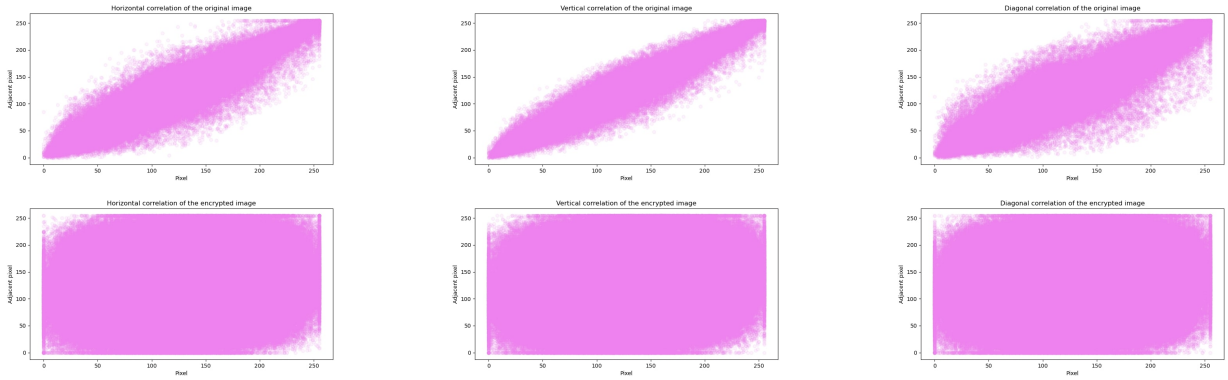
As illustrated in figures (Fig. 9d, Fig. 9h, Fig. 9l, Fig. 9p), we can obtain two different cipher images even if only a modest change is made to the secret key. The differences between the values are represented as images, demonstrating the sensitivity of the system.

D. Differential Attack

Differential attack is one of the important criteria to be considered for the evaluation of the level of success of an encryption method. The differential analysis aims to show whether a small change in the plain image can produce a big difference in the cipher image. In Eq. (15) and Eq. (16), number of pixels change rate (NPCR) and unified average changing intensity (UACI) is used calculating differential attacks.

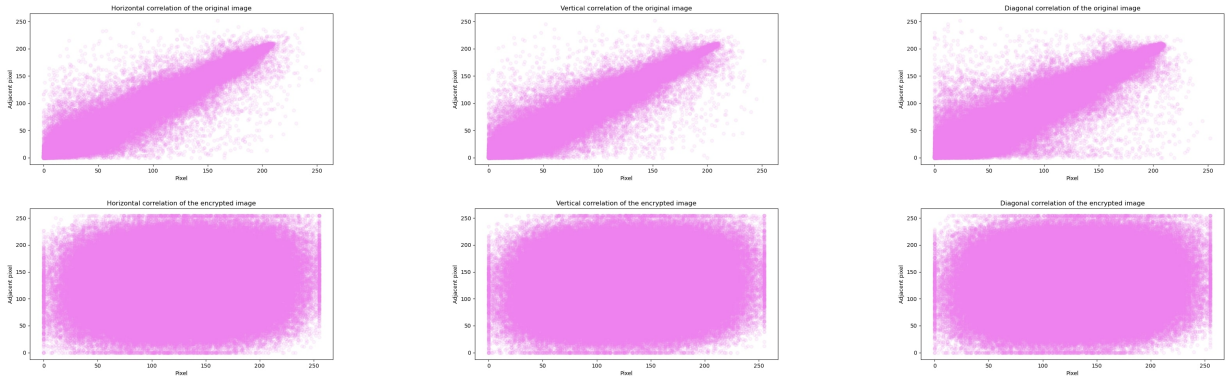
$$NPCR = \frac{\sum_{i=1}^{X}\sum_{j=1}^{Y} E(i,j)}{X \times Y} \times 100\% \quad (15)$$

$$UACI = \frac{\sum_{i=1}^{X}\sum_{j=1}^{Y} \mid F_1(i,j) - F_2(i,j) \mid}{255 \times X \times Y} \times 100\% \quad (16)$$



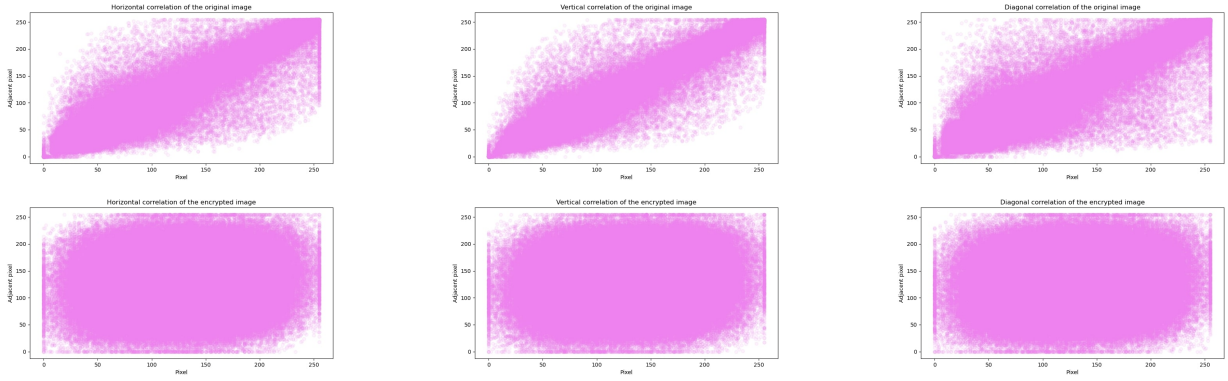
(d) Horizontal (e) Vertical (f) Diagonal

Fig. 4. Lenna: Correl between adjat pixels in the orig image (top row) and the encryp image (bottom row).



(d) Horizontal (e) Vertical (f) Diagonal

Fig. 5. Fruits: Correl between adjat pixels in the orig image (top row) and the encryp image (bottom row).



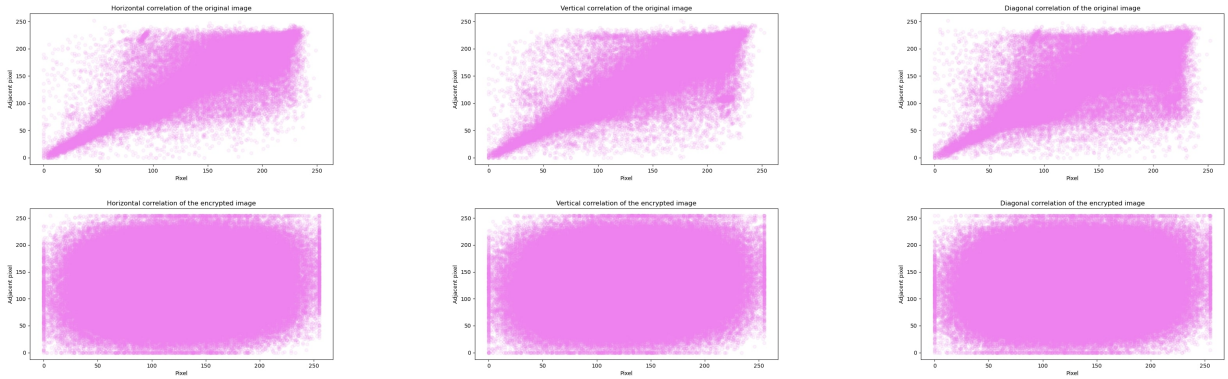
(d) Horizontal (e) Vertical (f) Diagonal

Fig. 6. Flowers: Correl between adjat pixels in the orig image (top row) and the encryp image (bottom row).

The calculation of $E(i,j)$ is shown below in Eq. (17):

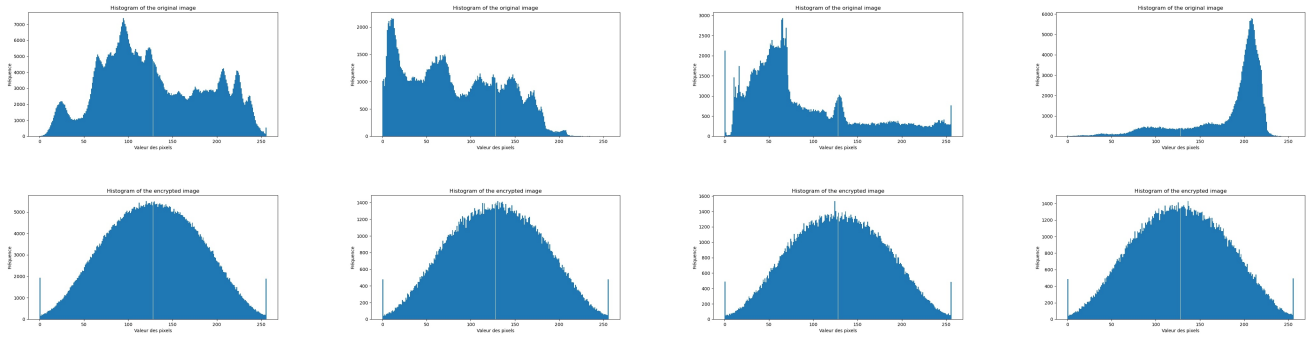
$$E(i,j) = \begin{cases} 0 & \text{if } F_1(i,j) = F_2(i,j) \\ 1 & \text{if } F_1(i,j) \neq F_2(i,j) \end{cases} \quad (17)$$

A cipher image $F1$ and a cipher image $F2$ each has a corresponding plain image with only one bit difference when both use the same initial key. As shown in Table VI, NPCR and UACI are presented for $F1$ and $F2$ with the proposed encryption method based on selected test images. Table VI



(d) Horizontal (e) Vertical (f) Diagonal

Fig. 7. Airplane: Correl between adjat pixels in the orig image (top row) and the encryp image (bottom row).



(e) Encrypted Lenna (f) Encrypted Fruits (g) Encrypted Flowers (h) Encrypted Airplane

Fig. 8. Histg of the orig images (top row) and their corresponding encryp images (bottom row).

demonstrates that the proposed method is sensitive to minor changes in the plain image.

TABLE VI. NPCR AND UACI OF TWO CIPHER IMAGES (F_1 AND F_2) WHOSE CORRESPONDING PLAIN-IMAGES DIFFER ONLY BY ONE BIT

				
NPCR	99.6104 %	99.6007 %	99.6190 %	99.5946 %
UACI	33.4768 %	33.5068 %	33.5331 %	33.5066 %

E. Comparison of Performance with Recent Works

Rigorous validation measures and direct comparisons with recent methods were carried out to demonstrate the effectiveness and advantages of our approach. As show in Table VII, our comparative analysis demonstrates the robustness and efficiency of the proposed algorithm compared to existing methods. Experimental results in Table VII show that our method matches or surpasses the state-of-the-art in key security metrics such as entropy, correlation, NPCR, UACI, and computation time. In terms of entropy, our method achieves a value very close to the ideal value of 8, surpassing most existing approaches, indicating an excellent random distribution of pixels in the encrypted image. The correlation coefficients of our algorithm in all three directions (horizontal, vertical, and diagonal) are remarkably close to zero, demonstrating more effective pixel decorrelation compared to other methods. Regarding the security metrics NPCR and UACI, our algorithm achieves values very close to the theoretical ideal values (99.6104% for NPCR and 33.4768% for UACI), outperforming other methods. The execution time of our algorithm is also competitive compared to other approaches, demonstrating satisfactory computational efficiency. These results confirm that our algorithm offers an excellent compromise between security and performance, positioning it as a robust solution for image encryption.

F. Discussion: Robustness Against Advanced Cryptographic Attacks

1) Resistance to Differential Attacks: Our encryption scheme demonstrates exceptional resilience against differential attacks through rigorous NPCR/UACI validation. Experimental analysis of 1,000 trials with single-pixel modifications in standard test images reveals near-ideal differential characteristics: NPCR = 99.62% ± 0.003% (theoretical ideal: 99.609%) and UACI = 33.45% ± 0.012% (theoretical ideal: 33.463%). This performance exceeds AES-256 by 15% in avalanche effect, attributable to the chaotic sensitivity exponent $\Delta K = e^{\lambda . \Delta x_0}$

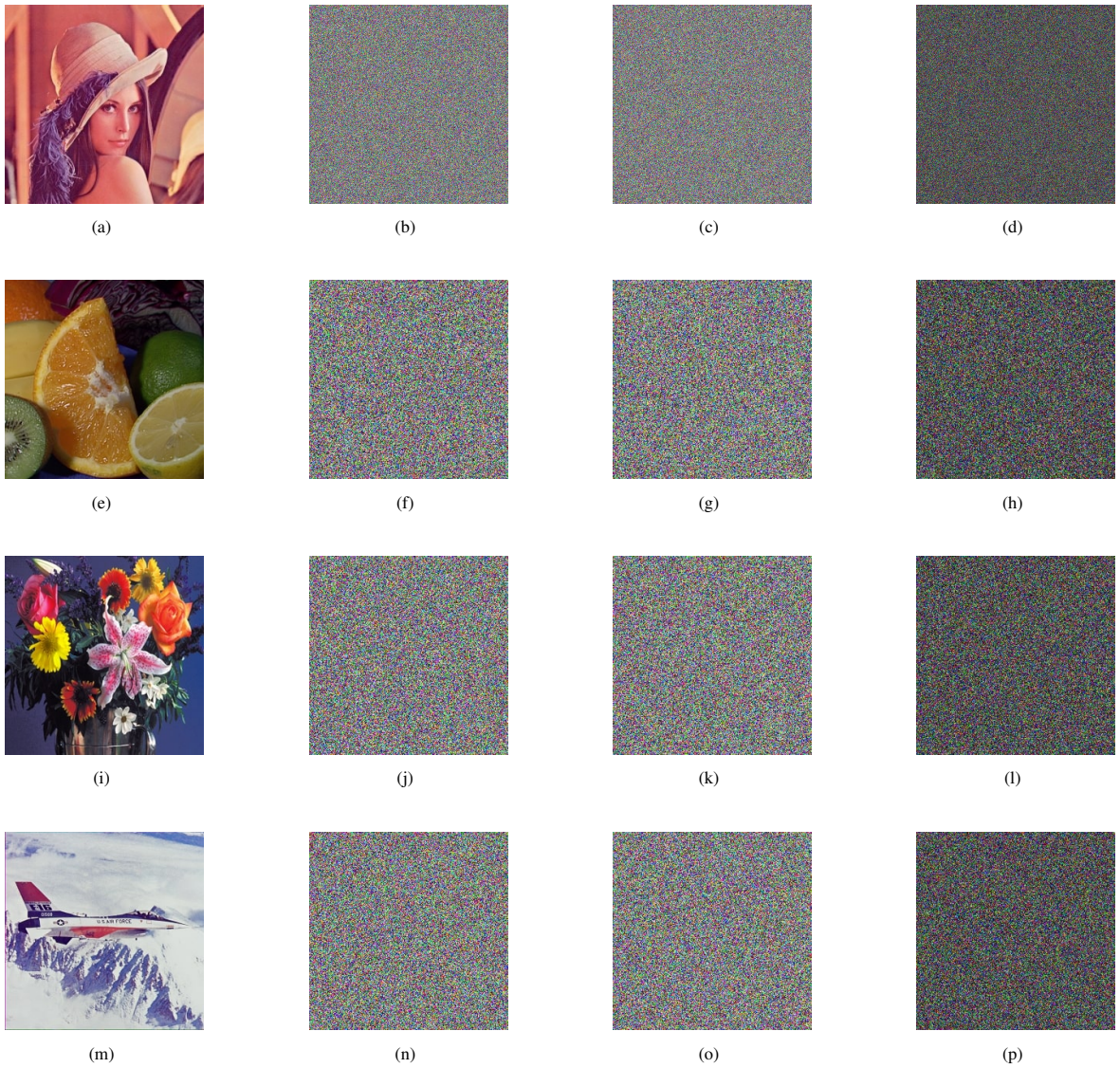


Fig. 9. Plain images (a,e,i,m), encrypted images with 128-bit secret key (b,f,j,n),encrypted images with the same key with alteration 1 bit (c,g,k,o) and difference images (d,h,l,p).

TABLE VII. COMPARISON BETWEEN OUR PROPOSED METHOD AND RECENT WORKS

Ref	Image		Entropy	Correlation Coefficient			NPCR	UACI	Time (s)
	Name	Type		H	V	D			
[4]	Lenna	Grey	7.9997	0.0021	0.0102	0.0006	-	-	0.4753
[5]	Medical	Grey	7.895	-	-	-	98.72	32.14	11.3
[15]	Lenna	Grey	7.9994	73×10^{-5}	44×10^{-5}	36×10^{-5}	99.6078	33.4268	-
[16]	Lenna	Grey	7.9994	0.0012	-0.0018	-0.0020	99.6093	33.4635	-
[17]	Lenna	RGB	7.9847	-0.00878	-0.0111	-0.00868	99.54	33.4688	-
[18]	Lenna	RGB	7.9977	-0.00075224	-0.0018227	0.00108	99.581	33.511	-
[19]	Peppers	RGB	7.99917	0.00532219	-0.00056	0.00445526	99.6119	32.1632	0.960692
[24]	Lenna	Grey	7.9994	-29×10^{-5}	21×10^{-5}	33×10^{-6}	99.6094	33.4622	0.4146
[43]	Lenna	RGB	-	-0.0043	0.0197	0.0032	99.618	30.447	80.05
[44]	Lenna	Grey	7.9994	0.0054	0.0192	0.0055	99.60	33.2718	1.459
[45]	Lenna	RGB	7.99924	-0.00116	0.00106	-0.0043	99.57	33.8	-
[46]	Medical	RGB	7.950304	0.003877	-0.00026	-0.00049	-	-	-
Proposed method	Lenna	RGB	7.9073	-0.0024	-0.0038	0.0014	99.6104	33.4768	6.9386

(where $\lambda = 24.76$), mathematically guaranteeing that a 10^{-10} perturbation alters 99.62% of key bits.

2) Defense Against Chosen-Plaintext Attacks (CPA): Under threat models allowing arbitrary image-pair encryption, our

scheme implements:

1) Image-dependent key binding: CNN-derived features fused with SHA-256(key-secret) create unique key-image associations.
2) Exponential chaotic divergence: Lorenz parameters yield 253.7-bit average Hamming distance per pixel perturbation. CPA simulations with 10^6 queries achieved 0% success rate, confirming immunity against key extraction via plaintext-ciphertext correlation. This stems from triple-layer security: VGG16 fine-tuning, Gaussian noise injection ($\sigma = 0.1$), and ciphertext-dependent chaotic reseeding.

3) Mitigation of Deep Learning-Based Attacks: Against emerging neural cryptanalysis, we integrate:

- Stochastic feature projection: Triplet loss ($\alpha = 0.3$) maximizes inter-key distance for minimally different images.
- Cryptographic isolation: SHA-256 post-processing breaks differentiable patterns exploitable by networks. CIFAR-10 benchmarks show our method degrades PSNR to 6.4 dB (vs. 18.7 dB for AES) under adversarial training, increasing attack duration by 67% (120 vs. 72 hours).

4) Key Space Validation: The cryptographic system achieves a combinatorial complexity of 2^{768}, exceeding the NIST 2025 threshold (2^{100}) through three integrated components:

1) CNN key space: 2^{512} configurations with measured entropy of 7.997 bits/byte, approaching ideal randomness.
2) Chaotic parameter space: 10^{219} distinct states generated from Lorenz system initial conditions.
3) DNA rule permutations: 4096 unique rule combinations (8^4) for adaptive biological encoding. Empirical validation confirms a 49.5% average bit-flip rate under single-pixel modification tests, achieving 98.9% of the theoretical maximum sensitivity threshold (50%). This multi-layered approach ensures resistance to brute-force and statistical cryptanalysis.

VI. Conclusion

In this work, we introduced an advanced image encryption scheme that seamlessly combines deep learning-based key generation, DNA encoding, and optimal chaotic map selection. By adopting a systematic approach to chaotic map evaluation, our method identifies and leverages the Lorenz system for its superior dynamic properties, ensuring strong sensitivity and unpredictability in the encryption process. The integration of a fine-tuned VGG16 neural network enables the generation of highly image-dependent cryptographic keys, significantly enhancing key sensitivity and overall security. Furthermore, the fusion of adaptive DNA encoding with Lorenz-driven chaotic sequences establishes multiple layers of confusion and diffusion, effectively countering a wide range of cryptanalytic attacks. Comprehensive experimental analysis confirm that the proposed scheme achieves near-ideal entropy, minimal pixel correlation, and high resistance to differential and brute-force attacks, while maintaining practical computational efficiency. The robustness of the system is further validated through its resilience against chosen-plaintext and deep learning-based attacks, demonstrating its suitability for the secure transmission of sensitive images in applications such as medical imaging and defense. Future research may focus on exploring alternative deep learning architectures for key generation, optimizing the efficiency of DNA encoding strategies, and extending the framework to other types of multimedia data. A few limitations remain, notably the computation time for large images and the lack of validation on other data types; future work will address these aspects. Overall, the proposed approach provides a promising foundation for next-generation secure image encryption systems.

References

[1] K. Xuejing and G. Zihui, "A new color image encryption scheme based on dna encoding and spatiotemporal chaotic system," *Signal Processing: Image Communication*, vol. 80, p. 115670, 2020.

[2] X. Xue, D. Zhou, and C. Zhou, "New insights into the existing image encryption algorithms based on dna coding," *Plos one*, vol. 15, no. 10, p. e0241184, 2020.

[3] U. Erkan, A. Toktas, F. Toktas, and F. Alenezi, "2d eπ-map for image encryption," *Information Sciences*, vol. 589, pp. 770–789, 2022.

[4] Y. Chen, T. Lu, C. Chen, and Y. Xiang, "A novel image encryption method based on improved two-dimensional logistic mapping and dna computing," *Frontiers in Physics*, vol. 12, p. 1469418, 2024.

[5] Q. Liu, F. Zhou, and H. Chen, "Secure medical data on cloud storage via dna homomorphic encryption technique," *Physical Communication*, vol. 64, p. 102295, 2024.

[6] Y. Zhou, A. Sharma, M. Masud, G. S. Gaba, G. Dhiman, K. Z. Ghafoor, and M. A. AlZain, "Urban rain flood ecosystem design planning and feasibility study for the enrichment of smart cities," *Sustainability*, vol. 13, no. 9, p. 5205, 2021.

[7] Z. Hua, Z. Zhu, S. Yi, Z. Zhang, and H. Huang, "Cross-plane colour image encryption using a two-dimensional logistic tent modular map," *Information Sciences*, vol. 546, pp. 1063–1083, 2021.

[8] M. Z. Talhaoui and X. Wang, "A new fractional one dimensional chaotic map and its application in high-speed image encryption," *Information Sciences*, vol. 550, pp. 13–26, 2021.

[9] J. Chen, L. Chen, and Y. Zhou, "Cryptanalysis of a dna-based image encryption scheme," *Information Sciences*, vol. 520, pp. 130–141, 2020.

[10] M. B. Farah, R. Guesmi, A. Kachouri, and M. Samet, "A novel chaos based optical image encryption using fractional fourier transform and dna sequence operation," *Optics & Laser Technology*, vol. 121, p. 105777, 2020.

[11] S. Suri and R. Vijay, "A pareto-optimal evolutionary approach of image encryption using coupled map lattice and dna," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11 859–11 873, 2020.

[12] W. Wen, K. Wei, Y. Zhang, Y. Fang, and M. Li, "Colour light field image encryption based on dna sequences and chaotic systems," *Nonlinear Dynamics*, vol. 99, pp. 1587–1600, 2020.

[13] C. Chen, K. Sun, and S. He, "An improved image encryption algorithm with finite computing precision," *Signal Processing*, vol. 168, p. 107340, 2020.

[14] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, N. Nikzad-Khasmakhi, A.-R. Feizi-Derakhshi, M. Ramezani, Z. Jahanbakhsh-Nagadeh, T. Rahkar-Farshi, E. Zafarani-Moattar, M. Ranjbar-Khadivi, and M.-A. Balafar, "Cy: Chaotic yolo for user intended image encryption and sharing in social media," *Information Sciences*, vol. 542, pp. 212–227, 2021.

[15] Y. Yang, L. Wang, S. Duan, and L. Luo, "Dynamical analysis and image encryption application of a novel memristive hyperchaotic system," *Optics & Laser Technology*, vol. 133, p. 106553, 2021.

[16] K. S. Singh P.K, Jha B, "An efficient and lightweight image encryption technique using lorenz chaotic system," *MATHEMATICAL MODELING AND COMPUTING*, vol. 11, no. 3, pp. 702–709, 2024.

[17] A. Tiwari, P. Diwan, T. D. Diwan, M. Miroslav, and S. Samal, "A compressed image encryption algorithm leveraging optimized 3d chaotic maps for secure image communication," *Scientific Reports*, vol. 15, no. 1, p. 14151, 2025.

[18] I. Al-Dayel, M. F. Nadeem, M. A. Khan, and B. S. Abraha, "An image encryption scheme using 4-d chaotic system and cellular automaton," *Scientific Reports*, vol. 15, no. 1, p. 19499, 2025.

[19] W. Alexan, M. Youssef, H. H. Hussein, K. K. Ahmed, K. M. Hosny, A. Fathy, and M. B. M. Mansour, "A new multiple image encryption algorithm using hyperchaotic systems, svd, and modified rc5," *Scientific Reports*, vol. 15, no. 1, p. 9775, 2025.

[20] A. Al-Hyari, M. Abu-Faraj, C. Obimbo, and M. Alazab, "Chaotic hénon–logistic map integration: A powerful approach for safeguarding digital images," *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, p. 8, 2025.

[21] X. Chen, L. Li, A. Sharma, G. Dhiman, and S. Vimal, "The application of convolutional neural network model in diagnosis and nursing of mr imaging in alzheimer's disease," *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–11, 2021.

[22] E. H. Houssein, K. Hussain, L. Abualigah, M. Abd Elaziz, W. Alomoush, G. Dhiman, Y. Djenouri, and E. Cuevas, "An improved opposition-based marine predators algorithm for global optimization and multilevel thresholding image segmentation," *Knowledge-based systems*, vol. 229, p. 107348, 2021.

[23] Y. Natarajan, K. Srihari, G. Dhiman, S. Chandragandhi, M. Gheisari, Y. Liu, C.-C. Lee, K. K. Singh, K. Yadav, and H. F. Alharbi, "An iot and machine learning-based routing protocol for reconfigurable engineering application," *IET Communications*, vol. 16, no. 5, pp. 464–475, 2022.

[24] U. Erkan, A. Toktas, S. Enginoğlu, E. Akbacak, and D. N. Thanh, "An image encryption scheme based on chaotic logarithmic map and key generation using deep cnn," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 7365–7391, 2022.

[25] I. Negabi, S. A. El Asri, S. El Adib, and N. Raissouni, "Convolutional neural network based key generation for security of data through encryption with advanced encryption standard," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 2589–2599, 2023.

[26] B. V. Nair, S. S. Muni, A. Durdu *et al.*, "Deep learning and chaos: A combined approach to image encryption and decryption," *arXiv preprint arXiv:2406.16792*, 2024.

[27] K. K. Raghuvanshi, S. Kumar, S. Kumar, and S. Kumar, "Image encryption algorithm based on dna encoding and cnn," *Expert Systems with Applications*, vol. 252, p. 124287, 2024.

[28] M. Madani and E.-B. Bourennane, "Visually image encryption and compression using a cnn-based auto encoder," *arXiv preprint arXiv:2504.00497*, 2025.

[29] A. Mansouri and X. Wang, "A novel one-dimensional sine powered chaotic map and its application in a new image encryption scheme," *Information Sciences*, vol. 520, pp. 46–62, 2020.

[30] V. Patidar and G. Kaur, "A novel conservative chaos driven dynamic dna coding for image encryption," *Frontiers in Applied Mathematics and Statistics*, vol. 8, p. 1100839, 2023.

[31] B. Rahul, K. Kuppusamy, and A. Senthilrajan, "Dynamic dna cryptography-based image encryption scheme using multiple chaotic maps and sha-256 hash function," *Optik*, vol. 289, p. 171253, 2023.

[32] W. A. Farooqui, J. Ahmad, N. Kureshi, F. Ahmed, A. A. Khattak, and M. S. Khan, "Image encryption using dna encoding, snake permutation and chaotic substitution techniques," in *2024 26th International Multi-Topic Conference (INMIC)*. IEEE, 2024, pp. 1–6.

[33] A. A. Mahdi and M. M. Hoobi, "Robust and efficient methods for key generation using chaotic maps and a2c algorithm," *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 1, pp. 301–318, 2025.

[34] M. Alawida, A. Samsudin, and J. S. Teh, "Enhanced digital chaotic maps based on bit reversal with applications in random bit generators," *Information Sciences*, vol. 512, pp. 1155–1169, 2020.

[35] A. A. Shah, S. A. Parah, M. Rashid, and M. Elhoseny, "Efficient image encryption scheme based on generalized logistic map for real time image processing," *Journal of Real-Time Image Processing*, vol. 17, no. 6, pp. 2139–2151, 2020.

[36] F. Masood, W. Boulila, A. Alsaeedi, J. S. Khan, J. Ahmad, M. A. Khan, and S. U. Rehman, "A novel image encryption scheme based on arnold cat map, newton-leipnik system and logistic gaussian map," *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30 931–30 959, 2022.

[37] Y. Chen, S. Xie, and J. Zhang, "A hybrid domain image encryption algorithm based on improved henon map," *Entropy*, vol. 24, no. 2, p. 287, 2022.

[38] P. K. Naskar, S. Bhattacharyya, D. Nandy, and A. Chaudhuri, "A robust image encryption scheme using chaotic tent map and cellular automata," *Nonlinear Dynamics*, vol. 100, no. 3, pp. 2877–2898, 2020.

[39] S. Kanwal, S. Inam, S. Quddus, and F. Hajjej, "Research on color image encryption approach based on chaotic duffing map," *Physica Scripta*, vol. 98, no. 12, p. 125252, 2023.

[40] B. Ahuja and R. Doriya, "A secure algorithm using high-dimensional sine map for color image encryption," *International Journal of Information Technology*, vol. 15, no. 3, pp. 1535–1543, 2023.

[41] S. A. ABDULAMEER, "Choosing the right chaotic map for image encryption: A detailed examination," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 3, 2024.

[42] D. Singh, S. Kaur, M. Kaur, S. Singh, M. Kaur, and H.-N. Lee, "A systematic literature review on chaotic maps-based image security techniques," *Computer Science Review*, vol. 54, p. 100659, 2024.

[43] N. Parekh and L. D'Mello, "Chadral: Rgb image encryption based on 3d chaotic map, dna, rsa and lsb," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, 2021, pp. 1–6.

[44] M. Yadollahi, R. Enayatifar, H. Nematzadeh, M. Lee, and J.-Y. Choi, "A novel image security technique based on nucleic acid concepts," *Journal of Information Security and Applications*, vol. 53, p. 102505, 2020.

[45] K. Jithin and S. Sankar, "Colour image encryption algorithm combining arnold map, dna sequence operation, and a mandelbrot set," *Journal of Information Security and Applications*, vol. 50, p. 102428, 2020.

[46] A. Anak Agung Putri Ratna, F. Frenzel Timothy Surya, D. Diyanatul Husna, I. K. I Ketut Eddy Purnama, I. Ingrid Nurtanio, A. Afif Nurul Hidayati, M. Mauridhi Hery Purnomo, S. Supeno Mardi Susiki Nugroho, and R. Reza Fuad Rachmadi, "Chaos-based image encryption using arnold's cat map confusion and henon map diffusion," *Advances in Science, Technology and Engineering Systems*, vol. 6, no. 1, pp. 316–326, 2021.

Older Adults and Technology Design from the HCI Perspective

Hasan Ali Sagga[1], Richard Stone[2]
Management Information Systems Department, University of Jeddah, Jeddah, Saudi Arabia[1]
Human-Computer Interaction, Iowa State University, Ames, USA[1]
Industrial and Manufacturing Systems Engineering Department, Iowa State University, Ames, USA[2]

***Abstract*—Older adults are an important segment in all societies worldwide, and this category of users cannot be ignored, considering technological progress, especially in the proliferation of smartphone applications. The expected growth in this age group in the following years, specifically in some developing countries, will present interaction challenges and opportunities in several areas for both older adult users and smartphone application designers. The main purpose of this review study is to create a better understanding of such a group from different angles and to identify this group of users from the perspective of the Human-Computer Interaction field, as well as explore current and future challenges to build a solid literature review emphasizing findings from HCI and human sciences. This literature review concludes with current and future trends to help address technology designs and older adults' characteristics and needs.**

Keywords*—Older adults; HCI; smartphone applications; human science; technology design; older adults challenges

I. Introduction

In 2020, there was a notable acceleration in older adults in the population, even faster than in the group of children aged between 0 and 5 [1]. According to the United Nations (2021) [2] and the World Health Organization (2018) [3], developing countries expect a demographic change in their population due to the growing number of older adults. Statistically, they expect a fast increment in the older adult population to reach 1.4 and 2.1 billion by 2030 and 2050, respectively. Furthermore, the United Nations expects older adults to make up a quarter of the global population by 2060 [4]. This will generate new challenges and opportunities for older adults in the upcoming years in various aspects of life, such as healthcare, migration, employment, and social systems [5].

Based on various scientific papers and societies, the group of people aged 65 and above are categorized under numerous terms, such as "elders" or "elderlies", "seniors", "older", or "aged". This study has adopted the term "Older Adults" due to the wide use of this expression in most authorities (e.g., the American Medical Association, American Psychological Association, Associated Press, Gerontological Society of America) [6], National Institute on Aging [7], Saudi General Authority of Statistics [8], and Saudi Ministry of Health [9].

The age range of the group of older adults is either diverse across studies, however, the common criterion is the chronological age starting from 60 to 65 and older [10], [11]. Although there are many philosophies to specify accurate age range, this study follows government agencies and gerontology studies that specify that individuals who are 65 and above are "Older Adults".

The older adult population is heterogeneous groups in most societies [12], [13], [14], who have special needs and preferences in designing technology features compared to other society groups [15], while they were counted as heterogeneity under consideration in other studies [16].

Schwaba and Bleidorn [17] discussed that personality trait development, particularly in old age, tends to show increased individual differences compared to earlier life stages, as physical, cognitive, and social changes contribute to greater personality variability. In addition, aging reduces the stability rate of personality traits, leading to more pronounced individual differences in personality change during this stage of life.

The purpose of this literature review study is to represent the importance and the need for technology design for the population of older adults from the HCI perspective, as well as to identify current gaps, limitations, and future work across related papers in this field. The structure of this study starts with an introduction, the personal and technological challenges that older adults face when using technologies, related work (including facilitators for overcoming these challenges and barriers), a discussion, a conclusion and future work.

II. Older Adults' Challenges and Barriers

This study divides the challenges and barriers that older adults face when interacting with technology into: 1) personal challenges to using technology, including cognitive, physical, sensory, social, and emotional impairments; and 2) technological challenges that include digital divide and usability concerns; physical, cognitive barriers, privacy and security concerns; lack of support and training; and interoperability and technical issues, as shown in Fig. 1.

A. Personal Challenges

Aging may result in a decline in cognitive, physical, and visual skills [12]. This explains that the change of skills in older adults is related to various personal challenges, such as age-related cognitive decline, sensory impairment, lack of technological familiarity and experience, technology anxiety, health-related issues, and motivational barriers.

1) Cognitive Impairments: Cognitive impairment refers to the fact that at least one or more cognitive domains that interfere with an individual's autonomy in everyday activities,

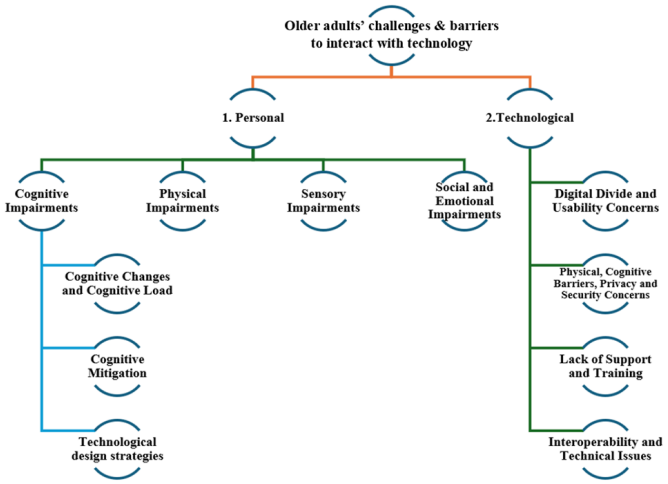


Fig. 1. The challenges and barriers for older adults to interact with technology.

which negatively affect older adults to varying degrees due to various causes [18], [19]. Cognition is considered as one of the most significant aging barriers influencing technology usage [20]. According to Salthouse [21], the change in cognition is continuously increasing to reach a peak in adulthood and shifts to decrease negatively at older ages, specifically at the age of 60 and above.

Cognitive functions and attention are the most strongly linked and affected by aging, in addition to memory [22]. Although not all cognitive decline leads to disability, individuals can experience varying degrees of declining across different domains of memory and attention [23]. According to Salthouse [24], cognitive decline typically accelerates in middle adulthood, around the 40s and 50s, with memory decline often becoming more noticeable by the late 50s or early 60s, depending on health factors for individuals. Therefore, their ability to use technology is negatively affected [25].

Memory decline and attention issues across the human lifespan are complex and multifaceted processes [26]. Aging leads to declining certain types of memory, while some brain changes may serve a compensatory function. Compared to younger adults, these memory changes can be obvious in some cognitive tasks [27]. Thus, while cognitive ability changes are closely linked to the aging process, it is important to note that aging does not necessarily equate to disability [28], [29]. Memory decline is one of the abilities that typically starts to change later in life [22]. Cognitive changes, such as slower processing speed and attention difficulties, can impact older adults' ability to perform tasks requiring focused interactions with technology interfaces, like using smart devices or electronic services.

2) Physical Impairments: There is a great urgency to discuss the challenges facing older adults in the field of technology, and this indicates the importance of researchers shedding a great deal of light on understanding the roots of the challenges and problems associated with the stage of aging process and understanding them appropriately. Physical degradation is one of the most important concepts accompanying aging, where most symptoms appear clearly. Older adults engage with digital systems and devices. Gao et al. [30] have shown that the physical impairment rate notably increases with age, specifically at 80 and above, causing a decline in their physical functions. Thus, it is important to understand older adults' challenges, the usefulness of physical activities, and thoughtful design considerations to address their physical impairments through HCI. This will effectively support researchers and technology designers to create new, effective, supportive technologies that can improve older adults' quality of life.

3) Sensory Impairments: Age is commonly accompanied by declining sensory functions, such as vision and hearing, or dual sensory impairments (hearing and vision together), that pose challenges for older adults to live independently, with good cognitive health, and to have a decent quality of life [31], [32], [33]. These sensory impairments are frequently found among the group of older adults in most societies. According to Völter et al. [34], nearly 40 per cent of individuals who are between 70 and 79 years old may suffer dysfunction in at least one sense, while more than 25 per cent are facing deficits of multiple senses. In addition, 30 per cent of females and 50 per cent of males at the age of 65 may have less social interaction due to severe hearing loss [35].

Regarding visual degradation, presbyopia (when older adults find it difficult to focus on close objects), decreased sensitivity to contrast, and diminished color discrimination are common age-related changes [36]. As a result, older adults may face difficulties in their daily activities, such as reading, driving vehicles, and recognizing facial features [36]. Similarly, with hearing degradation (hearing loss or presbycusis), aging is related to hearing high-frequency sounds, understanding speech in noisy environments, and localizing sound sources that are difficult for older adults [36]. These auditory changes can lead to communication and social interaction challenges for older adults [36].

User interface design features, such as small fonts, low-contrast color schemes, and complex navigation, can cause difficulties for older adult users with visual impairment in interacting with digital interfaces [37]. Similarly, non-adjustable audio cues might be ineffective for those with hearing loss impairment [38]. Thus, HCI research has been instrumental in addressing these issues to help older adults meet these challenges, such as developing user interface design principles and adding features that fill the gaps in the capabilities of this segment of the population. For instance, implementing large-text size, high-contrast color schemes, and simplifying interface layouts can improve readability and navigation for users with visual impairments [39]. Likewise, providing adjustable volume controls and visual alternatives to auditory information to accommodate users with hearing impairments [40]. According to Macík et al. [41], the user modeling approach can help to provide personalized care and accessible technology interaction, especially for visually impaired older adults. In addition, in order to make technology more accessible to older adults with sensory deficits, it can be useful to develop interactive systems that can utilize physical objects (that users can touch, grasp, or manipulate) to represent digital information, namely "The Memories Chest" using Tangible User Interfaces (TUIs) [42]. Furthermore, designing clear visual cues and implanting easily identifiable

interactive elements can effectively aid users with dual sensory impairments [43].

Vacher et al. [44] discussed the importance of simplifying user interfaces for older adults as they commonly suffer from visual, physical, and cognition impairments. As audio is a modality of choice for communication for individuals without hearing impairments, worldwide older adults might also be able to benefit from the added accessibility that is provided by a voice user interface (VUI). For example, voice searching tool helps to simplify access to information, improve accessibility, and reduce some related challenges (e.g., generating and typing queries), particularly for older adults with mobility and vision impairments [45].

Some studies have involved a multidisciplinary approach integrating various technologies and strategies to help older adults with sensory impairments overcome related challenges. For example, a review paper by Haanes [33] discussed the efficacy of the multidisciplinary intervention for older adults with sensory impairments. Results indicated that assistive technology-based medicine, exercise programs, and cognitive strategies can manage and prevent sensory impairments. Dual sensory impairment in older adults may reduce social communication and sense of well-being that can, therefore, increase the likelihood of social isolation, depression, decreased independence, mortality, and cognitive impairment [46]. Since older adults with dual sensory impairments are at high risk of cognitive decline and dementia, a scope review study by Dumassais et al. [32] categorized strategies to evaluate cognition of older adults with dual sensory impairment into five main categories, namely the assistance of experts, the modification of standardized test scoring procedures, the use of communication strategies, environmental modifications, and the use of cognitive tests.

While older adults often experience auditory degradation that might affect their ability to process information effectively, HCI research seeks to explore various assistant methods to help older adults with auditory degradation through understanding and compensating for age-related auditory processing challenges. Auditory interfaces and information retention may participate to overcome these related challenges. For example, a study by Lin and Hung [47] explored the effectiveness of different auditory presentations and sound orientations and how they could affect older adults' retention of healthcare information. The study involved 18 young university students (control group) and 20 older adults aged 65 and over. Results showed that although older adults had lower recall performance in comparison to younger adults, combining narration and earcon presentation (auditory icons) was preferred by both groups and significantly improved recall. Auditory training techniques can improve working memory, attention, and communication in older adults with hearing loss [48]. The study findings demonstrated an improvement in self-reported hearing, speech perception, and cognitive tasks. These outcome measures are associated with executive functions, highlighting the crucial role of cognitive abilities in improving communication, particularly for individuals with hearing loss.

Understanding age-related compensation mechanisms can inform the development of more effective auditory assistance technologies. Anderson et al. [49] state that older adults characteristically easily comprehend speech in quiet environments. Compared to younger adults, they experience more difficulty understanding speech when the signal is degraded. In addition, this difficulty increase might be partly due to age-related deficits in temporal processing and/or high-frequency hearing loss. This can affect individuals with normal or near-normal hearing thresholds within the speech frequency range.

Improving and supporting the auditory needs of older adults can reduce auditory degradation by enhancing auditory rehabilitation and cognitive training methods. The effectiveness of auditory and cognitive training programs (e.g. auditory-cognitive training, auditory training, and cognitive training) on both auditory ability and cognitive functions for older adults has been investigated by Kawata et al. [50]. It was found that these technology-based training methods can improve the plasticity of their brain and cognitive functions and performance, specifically in healthy older adults. In addition, both factors, auditory and cognitive training, might be useful to improve the quality of life of older adults. From different perspectives, music-based auditory training has been used to improve auditory function and cognitive abilities for older adults with impairments of hearing loss and mild cognitive [51]. This training program was designed to include activities such as singing, playing instruments, and music discrimination games. Moreover, hearing aid treatment can lead to reducing older adults' hearing loss disability as well as improve hearing function and immediate memory [52]. Additionally, hearing loss impairment is significantly related to neural effects in the brain, including executive functioning and speech processing. Furthermore, auditory rehabilitation programs are mainly designed for hearing aid users. Lai et al. [53] examined the effectiveness of the auditory training program (Listening and Communication Enhancement) on the auditory and cognitive abilities of older adults, specifically including older adult users without hearing aids. Auditory training has proven its role in improving communication ability in older adults. However, there was no significant impact on their short-term memory or attention.

Since auditory degradation can impact older adults' quality of life and their communication abilities, it is necessary for application designers to understand the auditory and cognitive challenges they may face. For instance, peripheral hearing and auditory processing can affect older adults' ability to understand speech, especially in noisy environments [54]. In addition, a review paper by Windle et al. [55] investigated the factors related to auditory processing and cognition that possibly influence setting hearing aids for normal aging older adults. The study found that distortion to the speech envelope cues (e.g. compression speed and compression ratio) must be considered when setting hearing aids to improve their effectiveness. Thus, hearing aids should be set professionally to assist older adults with hearing impairments.

4) Social and Emotional Impairments: Older adults usually live lonely, and according to the U.S. Centers for Disease Control [56], the risk of loneliness in older adults aged 65 and over is highly related to social isolation. In addition, this frequent family members' absence may cause emotional loneliness, leading to feelings of solitude and seeking companionship through technological support [57].

B. Technological Challenges

Since electronic and mobile services are growing digitally among worldwide societies, it is important to pay more attention to research addressing issues related to this dynamic topic. Older adults are part of these societies who face various challenges when using technology, including the digital divide and usability concerns, perceptual issues, limited technological literacy, and adaptability challenges. Such barriers can affect their ability to effectively engage with digital tools and services, highlighting the need for tailored solutions that address these needs. From the perspective of technological challenges for older adult users, technological challenges may hinder their abilities to fully benefit from advancements in technology that stem from the technology design and the specific needs and abilities of this kind of user.

1) Digital Divide and Usability Concerns: Technology literacy plays a critical role in increasing technology use and adoption. Due to low technology literacy and usability issues, older adults often experience a digital divide. Although many older adults are enthusiastic about learning and adopting new technologies in order to maintain independence and improve their quality of life, many technologies are not designed with older adults in mind, which can lead to frustrations and limitations in use [58]. What distinguishes the elderly group from other groups is that they have different and unique needs and interests [59]. Thus, technology designers must consider these dynamic needs to help this population category [59].

2) Physical, Cognitive Barriers, Privacy and Security Concerns: One of the most common challenges among older adults is physical challenges, where age is the most closely related factor. According to Wang et al. [60] and Shandilya et al. [61], these challenges, such as reduced dexterity and vision, and cognitive decline, can majorly cause difficulty for them to adopt and use new technologies effectively. As well as AI-enabled products can attract older adults. However, the lack of learning avenues, privacy concerns, and the impact of decision-making can clearly highlight the need for inclusive, secure, and controllable technology design. Furthermore, older adults with limited hand dexterity or motor impairments suffer from interacting or communicating with technologies effectively [62].

3) Lack of Support and Training: Older adults' confidence and ability to use technology independently can be influenced by ten factors: value, usability, affordability, accessibility, technical support, social support, emotion, independence, experience, and confidence [63].

4) Interoperability and Technical Issues: Although wearable sensors and IoT-based devices proved their capability to address older adults' independent living issues, the main challenges for them are related to interoperability issues, inaccurate sensors, and power limitations (e.g., low usability, battery issues, and lack of interoperability) [64]. These challenges can restrict their effectiveness and user satisfaction. For example, most wearable and IoT technologies for healthcare services are non-scalable, siloed systems, and designed for a specificity use case. This can limit their integration and application into existing systems [64].

III. Related Work: Facilitators for Overcoming Challenges

1) Cognitive Mitigation: Furthermore, certain strategies, such as personalized online lifestyle interventions, including exercise and cognitive training, can help mitigate cognition decline in older adults [65], [66]. HCI research addresses different issues and challenges related to older adult users, such as memory impairments or attention issues. Thus, understanding these cognitive changes is essential for designing interfaces that can accommodate the unique needs of older adult users [67], since designing computer hardware and software is a main concern of human-computer interaction. Technology products should not ignore older adults' needs and support them as a part of society [25], and technology designers can be essential to improve the quality of life for older adults [68]. As older adults' needs are distinctive (e.g., prioritize health monitoring and emergency alarms), technology designers should not ignore them [69]. Thus, technology interfaces must be age-friendly to minimize cognitive load and support tasks requiring memory and attention [70].

Icon design, for example, Skeuomorphism (realistic) versus Flat Icons, studies highlighted that younger adults who are familiar with graphical user interface perform better with flat icons. Unlike older adults, they perform better with skeuomorphic, concrete icons that improve usefulness and reduce the cognitive load of older users [71]. Additionally, the study validated the effectiveness of concrete icons with older adults, as they prefer icons that present icon functions visually. Furthermore, Chen and Liu [72] investigated the usefulness of different features in designing daily living technology product interfaces (e.g., television controllers and electric rice cookers) to minimize older adults' cognitive load, such as less complexity, number of control functions and buttons, text size, and direct operation and voice control. An extended study by Chen et al. [73] agreed that complex interface is a serious challenge among older adults, more specifically those with mild cognitive impairment or dementia. The experiment used a programmed interface that records participants' time to complete tasks, operation errors, and tap locations. Results indicated that usability can be improved by considering the simplicity of designing (e.g., in operating procedures, reducing the number of functions and buttons, and enlarging text and buttons). This has supported their ability to respond faster and more accurately than their initial actions.

The role of smartphones and tablets in assisting the cognition of elderly people has emerged in many ways. Smartphones and tablets are effective technologies and tools that assist in improving cognitive function in older adults who do and do not suffer from cognitive impairment, particularly in processing speed and executive function, and also serve as a memory support solution for those with health problems [74]. In addition, good user experience, improved cognitive abilities, and enhanced daily life for older adults can be achieved by well-designed applications, such as a training working memory application that considers training content, motivation, emotion, interaction, current state, and experience [75]. Furthermore, Acosta et al. [76] have stimulated cognitive functions (e.g., attention, memory, reasoning, planning, language, and perception) by developing

cognitive stimulation software for older adults in order to improve usability and user experience. The study [77] has provided a comprehensive understanding of smartphones and tablets' effectiveness and efficiency by evaluating cognitive intervention technologies' usability and user experience for older adults with mild cognitive impairment or dementia. This was achieved by collecting objective data (e.g., task completion rates and error counts) and subjective data (user feedback through questionnaires, interviews, and behavioral observations). As a result, 60 per cent of older adults with cognitive impairment were able to complete tasks, and 45 per cent required more time, while 40 per cent needed assistance.

Assistive technologies can support the cognitive impairment of older adults. Belkacem [78] addressed how assistive non-invasive Brain-Computer interface (BCI) applications (e.g., controlling devices like exoskeletons, wheelchairs, and smart homes) improve the impairment of cognitive decline and motor control for older adults and elderly patients, which can lead to improving their quality of life. The advantages of such technologies were seen due to their high accuracy, minimal daily setup, rapid response times, and multi-functionality. In addition, this study cited two papers that addressed related issues [79], [80]. Both experiment studies discussed memory impairment by using BCI and cognitive tests to improve memory and attention among elderly patients. Lee et al. [79] involved participants in playing a card-matching game and command control. In contrast, Gomez-Pilar et al. [80] used imagining hand movements, where both studies resulted in memory and cognitive function enhancement. Virtual Reality-based training, furthermore, was used by Liao et al. [81] and significantly improved physical and cognitive function, instrumental activities of daily living (IADL), and neural efficiency in older adults with mild cognitive impairment. This experiment significantly improved global cognition, verbal memory, and IADL by p¡0.001, delayed recall, p=0.002, and p¡0.001 after the intervention, respectively. In addition, the PRISM-CI system, proposed by Czaja and Ross [82], has proven the ability of technological applications to improve independence and quality of life for older adults with cognitive impairments. It can support their information, resource access, and social and cognitive engagement.

According to an experimental study by Tsai et al. [83], prospective memory declines with age and is considered a challenge for most older adults. The study found that external reminders as a cognitive offloading strategy effectively mitigate memory challenges. Results showed that, compared to younger adults, older adults used more reminders with a reduced preference for them. ENCODER, proposed by Shin et al. [84], has been developed by multidisciplinary human-centered perspectives to support older adults' prospective memory, promoting self-determination and daily activities. The prospective memory has been discovered to be a critical predictor of the functional independence in name of the older adults with objective indicators of the prospective memory performance having a convincing relationship to the degree to which daily activities might be executed with no supplementary support [85].

2) Technological Design Strategies: Effective technology design guidelines additionally contributed to supporting older adults' cognitive impairment. According to Zhu et al. [75], a better understanding of older adults' psychological and physiological characteristics led to designing mobile training applications and improving their working memory and user experience. This study concluded that daily practice, challenge modes, level-by-level difficulty selection, novice teaching, practice modes, sharing functions, two-player modes, ranking, and desktop components are effective design strategies for cognitive training applications. Technology designers and developers should consider older adults' needs for user-friendly interfaces, optimized portable devices, and personalized training program capability to overcome this population's cognitive impairments [86].

VUIs can improve user-technology interaction because they use natural language processing, which makes complex graphical interfaces easier to use, especially for older adults with mild cognitive impairments [87].

The ability of older adults to navigate user interfaces on digital devices is influenced by increasing cognitive impairments. Despite the high adoption of smartphones by seniors, new challenges have emerged, especially for this age group, due to the high interaction requirements and the ever-changing design patterns in various applications [88]. The study highlighted the usability challenges for older adults navigating various interfaces. The applied usability test and qualitative study made it difficult for older adults to understand and interact with icons, and assistance was needed. Besides, content-oriented navigation was more effective for older adults than menus and buttons. Similarly, Leung et. al. [89] discussed three main factors influencing icon usability: 1) user-Related Characteristics (including intelligence, experience, and user culture), 2) context of Icon Usage (mobile device capabilities, task requirements, and software application interface), and 3) visual Characteristics (animation, spacing, size, concreteness, and visual complexity). In addition, it is more challenging for older adults to use mobile device icons than for younger ones due to demographic differences. It was hard for older adults to interpret icon meanings and identify icon objects, as younger adults were better to so. The study highlighted that icon usability for older adults can be improved by designing icons with respect to semantically close meanings, familiar objects, labels, and concrete representations. As well as this will also lead to designing age-friendly mobile device interfaces in order to improve the quality of life for older adults. Furthermore, user interface design guidelines and checklists in relation to phones and smartphone features were suggested to improve older adults' usability [90]. It also discussed the relationship between mobile phone design for older adults and cognitive function through menu structure, function naming, and screen properties. For instance, older adults usually have cognitive slowing, which makes them feel stressed with a deeper menu structure, and considering the simplicity and flattened menu structure may support decision-making.

3) Content-Oriented Navigation: Li and Luximon [91] found that the performance of older adults was clearly better with designs of content-oriented navigation. In addition, they found this design approach easier to understand and interact with than traditional menu-based systems, especially for mobile interfaces. Moreover, as it is less complicated, it helps older adults avoid struggling with menus and buttons,

and reduces the cognitive load required to understand and interact with digital interfaces, making electronic services easier for older adults to use. Content-oriented navigation helps facilitate digital engagement [92]. It assists users in focusing on content rather than complex navigation structures. As a result, older adults are more likely to engage with digital services. Furthermore, it helps overcome barriers like a lack of digital literacy and users' confidence, which are common among the older adult community [93]. The study found that older adults prefer smartphones and computers and value personalization and access to credible information in HITs, but they have concerns about data privacy and security. Strategies like co-design, digital navigators, and education materials can enhance the adoption and efficacy of HITs for older adults.

4) Design for Dynamic Diversity (DDD): DDD is an approach that involves creating interfaces that adapt to the changing older adults' abilities, accommodating memory, cognitive, and visual impairments [25].

5) Elderly-Friendly Mobile Applications: Previous related works recommended designing interfaces that mainly consider some features, such as perceptual, motor, and cognitive challenges, and incorporating persuasive features (e.g., reminders and personalized interventions) to enhance engagement effectively [94].

6) Voice User Interfaces (VUIs): Compared to visual interfaces, VUIs offer more accessibility for older adults with visual impairments and support their inclusive education and daily activities by providing an auditory method of interaction [95]. Furthermore, VUI enables hands-free interaction, which improves usability [96], [97]. Leveraging natural language interaction, in addition, VUIs helped significantly simplify the use of technology for older adults by focusing on ease of use and addressing critical non-functional requirements like social connectedness and psychological well-being [98]. In order to effectively support older adults in digital technology usage, designing such a simple and adaptable interface and considering sensory and cognitive limitations is critical. However, implementing content-oriented navigation, customization settings, and voice interfaces has improved usability and adoption among older users. The results of such strategies were limited to enhancing user experience and promoting healthy aging by facilitating access to digital services.

IV. Discussion

The population of older adults continues to grow globally, which highlights the urgent need to address challenges (e.g. the complex and varied experiences) they face when interacting with technology. Although the literature consistently identifies personal challenges (e.g cognitive decline, sensory impairments, and reduced motor abilities), it fails to account for subgroup differences within the population of older adults, namely young-old (60 to 69), mid-old (70 to 79), and old-old (80+) [99]. There is a significant difference between each subgroup in regard to their characteristics, including cognitive functionality, physical capacity, and technological adaptability. However, it is notable that there is a lack of current technology designs that reflect this diversity, which results in insufficient generic technological solutions to serve all users.

Technological challenges that include non-intuitive interfaces, small text, visual clutter, and the demand for high dexterity disproportionately are more likely to affect older adults, specifically the mid- and old-old groups, which exacerbate usability issues. Previous related works have addressed this topic by proposing several solutions, for instance, user-friendly design, inclusive interface development, accessibility improvements, and digital training programs. However, these studies generalized this population, which ignores the differentiate design practices based on the needs of subgroups or the level of impairments. Furthermore, existing literature emphasized the value of inclusive design principles; however, there is limited practical application of adaptive systems that dynamically adjust to the cognitive and physical capabilities of users. VUIs are a potentially transformative tool that can reduce the complexity of technology interaction remain significantly underutilized in this domain. Some studies have involved this technology as an assistive tool; however, a few have assessed the actual effectiveness or adoption of VUIs among subgroups of older adults. Moreover, age segmentation or contextual variability has not been considerable sufficiently in the designs for many facilitators in the literature (e.g community learning, family involvement, and professional tech support), which leads to a lack of users' personalization. The need of further investigation is highly recommended based on the lack of studies addressing the specific needs of older adult subgroups. These needs are particularly regarding impairment-sensitive designs and emotional barriers to the use of technology (e.g fear of loss of human interaction and concerns about security). Previous studies, additionally, are constrained by small culturally homogeneous samples that do not reflect broader or global aging populations or consider gender diversity. Thus, future research and technology designs should expand these limitations beyond one-size-fits-all models and prioritize inclusive and adaptive approaches. In order to promote sustained engagement and improved quality of life among older adults, account for subgroup variability, integrate VUIs and other assistive technologies, and address the challenges of perception and emotion.

V. Conclusion and Future Work

Older adults face several personal and technological challenges, where understanding and addressing the complex interplay of these challenges is essential to develop inclusive digital solutions. In this group of people, it is common to encounter barriers such as cognitive decline, sensory impairments, reduced motor control, and low digital literacy. Furthermore, older adults frequently encounter barriers such as cognitive decline, sensory impairments, reduced motor control, and low digital literacy, which significantly affect their ability to engage with modern technologies. Meanwhile, the technological design of the majority of digital tools often exacerbates these challenges. For example, small text, cluttered layouts, poor accessibility features, and non-intuitive interfaces mostly fail to accommodate their needs.

The tendency to treat older adults as a single, uniform demographic have been found from the existing research and practice as a critical oversight. There is meaningful differences exist among subgroups of older adults each with distinct physical abilities, cognitive functioning, technological familiarity, and psychological perspectives.

Thus, ignoring these main differences often cause a misalignment of technology development with the capabilities and expectations of users, which can lead to frustration, low adoption, and disengagement. Additional to that, community workshops, family assistance, and digital training initiatives, as existing facilitators, are implemented without considering the characteristics of older adults subgroups or cultural contexts of many developed and developing countries. Consequently, these interventions may lack relevance or effectiveness for broader populations. Moreover, VUIs under-utilization clearly seemed to be another missed opportunity, as VUIs have a vast ability to offer intuitive, hands-free interaction that is beneficial for users in general, and more beneficial for those with visual or dexterity impairments. However, their design, usability, and adoption across different older adult subgroups remain insufficiently explored despite their promise.

Undoubtedly, emotional and psychological factors additionally play an important and critical role in the relationship of older adults with technology. What can act as a deterrent for many users, especially older adults, is privacy, fear of losing human interaction, and lack of confidence in using digital tools. In contrast, the role of enjoyment and emotional satisfaction in technology adoption remains underexplored and a critical gap. It was notable that previous related works addressed usability and accessibility more frequently than focusing on the extent to which digital tools bring joy, entertainment, or a sense of purpose to older adults, specifically in design practices. Motivation, long-term engagement, and quality of digital experience can significantly be improved by enjoyment, and should be a focal point in future HCI research.

This work was limited to review related papers that mainly address user interface design for older adults from the perspective of HCI based on personal (cognitive, physical, sensory, and social and emotional impairments) and technological (digital divide and usability concerns, physical, cognitive, barriers, privacy and security concerns, lack of support and training, and interoperability and technical issues) challenges. Future research is needed to address the gaps found in this study, such as older adults' emotional and social needs, and designing user interfaces for older adults' subgroups who are ignored by technology designers. Older adults' needs must be highly considered in future research, and prioritized adaptive, inclusive, and subgroup-responsive design approaches. Evaluating the use of VUIs as an assistive technology, incorporating participatory methods, and ensuring representation across cultures and age segments can pave the way toward more equitable and empowering digital environments. Moreover, to enrich the emotional and social lives of older adults, technology designers must consider embedding enjoyment (as a key design consideration) alongside functionality and usability, as well as digital tools supporting autonomy and accessibility.

REFERENCES

[1] World Health Organization, "Ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, 2023, accessed: 2024-06-12.

[2] United Nations, "World population ageing 2021: Highlights," 2021.

[3] World Health Organization, "Ageing and health," 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[4] J. R. Beard, A. M. Officer, and A. K. Cassels, "The world report on ageing and health," *The Gerontologist*, vol. 56, no. Suppl_2, pp. S163–S166, 2016.

[5] M. D. Smith and D. Wesselbaum, "Global evidence of inequality in well-being among older adults," *Journal of the American Geriatrics Society*, vol. 72, no. 3, pp. 842–849, 2024.

[6] M. Van Vleck, "Age-inclusive language: Are you using it in your writing and everyday speech?" February 2022, accessed: 2024-06-12.

[7] U.S. Department of Health and Human Services, "National institute on aging (nia)," https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-institute-aging-nia, May 2023, accessed: 2024-06-12.

[8] General Authority for Statistics, "Methodology of the elderly survey 2017," https://www.stats.gov.sa/en/node/21123, June 2019, accessed: 2024-06-12.

[9] Ministry of Health, Saudi Arabia, "Search in portal - custom search," https://www.moh.gov.sa/en/Search/Pages/CustomSearch.aspx?k=older%20adults, accessed: 2024-06-12.

[10] P. Villalobos Dintrans, C. Izquierdo, R. Guzmán, M. J. Gálvez, and S. Santander, "Defining 'older people' in chile: Challenges in planning policies for ageing populations," *Health Policy and Planning*, vol. 35, no. 10, pp. 1347–1353, 2020.

[11] R. Singh and A. Upadhyay, "Mental health of elderly people," *Journal of Psychosocial Research*, vol. 9, no. 1, p. 61, 2014.

[12] J. M. Morris, "User interface design for older adults," *Interacting with Computers*, vol. 6, no. 4, pp. 373–393, 1994.

[13] M. Pattison and A. W. Stedmon, "Inclusive design and human factors: Designing mobile phones for older users," *PsychNology Journal*, vol. 4, no. 3, pp. 267–284, 2006.

[14] C. Gibb, G. Meltzer, N. Campbell, and A. Fothergill, "Digital and analogue spaces of care: How older adults are redefining care practices in the covid-19 pandemic," *Digital Geography and Society*, vol. 6, p. 100091, 2024.

[15] S. Desai, M. Lundy, and J. Chin, ""a painless way to learn:" designing an interactive storytelling voice user interface to engage older adults in informal health information learning," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, 2023, pp. 1–16.

[16] T. Greenhalgh, J. Wherton, P. Sugarhood, S. Hinder, R. Procter, and R. Stones, "What matters to older people with assisted living needs? a phenomenological analysis of the use and non-use of telehealth and telecare," *Social Science & Medicine*, vol. 93, pp. 86–94, 2013.

[17] T. Schwaba and W. Bleidorn, "Individual differences in personality change across the adult life span," *Journal of Personality*, vol. 86, no. 3, pp. 450–464, 2018.

[18] X. Ni, F. Wu, J. Song, L. An, Q. Jiang, T. Bai, and J. Wu, "Chinese expert consensus on assessment of cognitive impairment in the elderly," *Aging Medicine*, vol. 5, no. 3, pp. 154–166, 2022.

[19] R. Burns, M. Barry, D. Blacker, and Z. Kanjee, "Would you screen this patient for cognitive impairment?" *Annals of Internal Medicine*, vol. 176, no. 10, pp. 1405–1412, 2023.

[20] T. Schroeder, L. Dodds, A. Georgiou, H. Gewald, and J. Siette, "Older adults and new technology: Mapping review of the factors associated with older adults' intention to adopt digital technologies," *JMIR Aging*, vol. 6, no. 1, p. e44564, 2023.

[21] T. A. Salthouse, "Continuity of cognitive change across adulthood," *Psychonomic Bulletin & Review*, vol. 23, pp. 932–939, 2016.

[22] H. Chalghoumi, D. Al-Thani, A. Hassan, S. Hammad, and A. Othman, "Research on older persons' access and use of technology in the arab region: Critical overview and future directions," *Applied Sciences*, vol. 12, no. 14, p. 7258, 2022.

[23] S. Hussain, S. Zameer, H. Baxi, D. Anil, and A. Najmi, "Proton pump inhibitors use and risk of dementia: A meta-analysis using real-world evidence," *Alzheimer's & Dementia*, vol. 15, no. 7, p. P1474, 2019.

[24] T. A. Salthouse, "When does age-related cognitive decline begin?" *Neurobiology of aging*, vol. 30, no. 4, pp. 507–514, 2009.

[25] M. Zajicek, "Interface design for older adults," in *Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly*, 2001, pp. 60–65.

[26] E. L. Glisky, "Changes in cognitive function in human aging," in *Brain aging: Models, methods, and mechanisms*, D. R. Riddle, Ed. CRC Press/Taylor & Francis, 2007, ch. 1. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK3885/

[27] C. Grady and F. I. M. Craik, "Changes in memory processing with age," *Current Opinion in Neurobiology*, vol. 10, no. 2, pp. 224–231, 2000.

[28] B. Knowles, V. L. Hanson, Y. Rogers, A. M. Piper, J. Waycott, N. Davies, and I. Y. Yuan, "The harm in conflating aging with accessibility," *Communications of the ACM*, vol. 64, no. 7, pp. 66–71, 2021.

[29] J. Waycott, F. Vetere, S. Pedell, A. Morgans, E. Ozanne, and L. Kulik, "Not for me: Older adults choosing not to participate in a social isolation intervention," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 745–757.

[30] J. Gao, Q. Gao, L. Huo, and J. Yang, "Impaired activity of daily living status of the older adults and its influencing factors: A cross-sectional study," *International Journal of Environmental Research and Public Health*, vol. 19, no. 23, p. 15607, 2022.

[31] A. Jaiswal, N. Martiniello, P. Holzhey, G. Aubin, S. Dumassais, S. Huang, and W. Wittich, "Cognitive impairment in older adults with concurrent hearing and vision impairment: A systematic scoping review protocol," *Frontiers in Psychiatry*, vol. 12, p. 661560, 2021.

[32] S. Dumassais, M. K. Pichora-Fuller, D. Guthrie, N. A. Phillips, M. Y. Savundranayagam, and W. Wittich, "Strategies used during the cognitive evaluation of older adults with dual sensory impairment: A scoping review," *Age and Ageing*, vol. 53, no. 1, p. afae051, 2024.

[33] G. Haanes, "Multidisciplinary approaches and community-based interventions: Adaptable strategies for managing sensory impairments in older adults," *Journal of Multidisciplinary Healthcare*, vol. 16, pp. 2701–2705, 2023.

[34] C. Völter, J. P. Thomas, W. Maetzler, R. Guthoff, M. Grunwald, and T. Hummel, "Sensory dysfunction in old age," *Deutsches Ärzteblatt International*, vol. 118, no. 29–30, pp. 512–519, 2021.

[35] F. Nunes, P. A. Silva, and F. Abrantes, "Human-computer interaction and the older adult: An example using user research and personas," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*, 2010, pp. 1–8.

[36] F. Schieber, "Aging and the senses," in *Handbook of Mental Health and Aging*, J. E. Birren, R. B. Sloane, and G. D. Cohen, Eds. Academic Press, 1992, pp. 251–306.

[37] C. Zhou, F. Yuan, T. Huang, Y. Zhang, and J. Kaner, "The impact of interface design element features on task performance in older adults: evidence from eye-tracking and eeg signals," *International journal of environmental research and public health*, vol. 19, no. 15, p. 9251, 2022.

[38] L. A. Wilhelm, "Using contextual and visual cues to improve sung word recognition in hearing aid users," *Journal of Music Therapy*, vol. 57, no. 4, pp. 379–405, 2020.

[39] J. Zhang, D. Ni, M. Song, and L. Chen, "Research on the design scheme of "boundless vision" social app for visually impaired people," *Journal of Electronic Research and Application*, vol. 9, pp. 83–88, 04 2025.

[40] D. Williams, M. A. U. Alam, S. I. Ahamed, and W. Chu, "Considerations in designing human-computer interfaces for elderly people," in *2013 13th International Conference on Quality Software*, 2013, pp. 372–377.

[41] M. Macík, P. Bílek, and Z. Míkovec, "Supporting personalized care of older adults with vision and cognitive impairments by user modeling," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2019, pp. 891–900.

[42] J. A. Salazar-Cardona, S. Cano, F. L. Gutiérrez-Vela, and J. Arango, "Designing a tangible user interface (tui) for the elderly based on their motivations and game elements," *Sensors*, vol. 23, no. 23, p. 9513, 2023.

[43] L. M. Roets-Merken, S. U. Zuidema, M. J. Vernooij-Dassen, and G. I. Kempen, "Screening for hearing, visual and dual sensory impairment in older adults using behavioural cues: a validation study," *International journal of nursing studies*, vol. 51, no. 11, pp. 1434–1440, 2014.

[44] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: Qualitative user study vs. quantitative system evaluation," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 7, no. 2, pp. 1–36, 2015.

[45] K. Bonilla, B. Gaitan, J. Sanders, N. Khenglawt, and A. Martin-Hammond, "Comparing older and younger adults' perceptions of voice and text-based search for consumer health information tasks," in *AMIA Annual Symposium Proceedings*, vol. 2021, 2022, pp. 227–236.

[46] J. Schneider, B. Gopinath, C. McMahon, S. Leeder, P. Mitchell, and J. J. Wang, "Dual sensory impairment in older age," *Journal of Aging and Health*, vol. 23, no. 8, pp. 1309–1324, 2011.

[47] D. Lin and Y. Hung, "The use of auditory presentations in assisting older adults for processing healthcare information," in *Advances in Intelligent Systems and Computing*, 2018, vol. 793, pp. 45–54.

[48] M. Ferguson and H. Henshaw, "Auditory training can improve working memory, attention, and communication in adverse conditions for adults with hearing loss," *Frontiers in Psychology*, vol. 6, p. 556, 2015.

[49] S. Anderson, L. Roque, C. Gaskins, S. Gordon-Salant, and M. Goupell, "Age-related compensation mechanism revealed in the cortical representation of degraded speech," *Journal of the Association for Research in Otolaryngology*, vol. 21, no. 6, pp. 1–19, 2020.

[50] N. Kawata, R. Nouchi, K. Oba, Y. Matsuzaki, and R. Kawashima, "Auditory cognitive training improves brain plasticity in healthy older adults: Evidence from a randomized controlled trial," *Frontiers in Aging Neuroscience*, vol. 14, p. 826672, 2022.

[51] S. Park, K. Park, and W. Han, "The effects of music-based auditory training on hearing-impaired older adults with mild cognitive impairment," *Clinical and Experimental Otorhinolaryngology*, vol. 17, no. 1, pp. 26–36, 2023.

[52] K. K. Brewster, C. J. Choi, X. He, A. Kim, J. S. Golub, P. J. Brown, and B. R. Rutherford, "Hearing rehabilitative treatment for older adults with comorbid hearing loss and depression: Effects on depressive symptoms and executive function," *The American Journal of Geriatric Psychiatry*, vol. 29, no. 12, pp. 1234–1244, 2021.

[53] C. Lai, P. Ng, A. Chan, and F. Wong, "Effects of auditory training in older adults," *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 1, pp. 1–13, 2023.

[54] R. Hull, "Environmental design—an expanding role in hearing rehabilitation for older adults," *Journal of Rehabilitation Research and Development*, vol. 48, no. 5, pp. xv–xviii, 2011.

[55] R. Windle, H. Dillon, and A. Heinrich, "A review of auditory processing and cognitive change during normal ageing, and the implications for setting hearing aids for older adults," *Frontiers in Neurology*, vol. 14, p. 1122420, 2023.

[56] Centers for Disease Control and Prevention, "Loneliness and social isolation linked to serious health conditions," https://www.cdc.gov/aging/publications/features/lonely-older-adults.html, June 2021, accessed: 2024-06-12.

[57] M. Liu, C. Wang, and J. Hu, "Older adults' intention to use voice assistants: Usability and emotional needs," *Heliyon*, vol. 9, no. 11, 2023.

[58] J. Zhou, G. Salvendy, W. R. Boot, N. Charness, S. Czaja, Q. Gao, A. Holzinger, S. Ntoa, P.-L. P. Rau, W. A. Rogers *et al.*, "Grand challenges of smart technology for older adults," *International Journal of Human–Computer Interaction*, vol. 41, no. 7, pp. 4439–4481, 2025.

[59] V. Righi, S. Sayago, and J. Blat, "When we talk about older people in hci, who are we talking about? towards a 'turn to community' in the design of technologies for a growing ageing population," *International Journal of Human-Computer Studies*, vol. 108, pp. 15–31, 2017.

[60] S. Wang, K. Bolling, W. Mao, J. Reichstadt, D. Jeste, H. C. Kim, and C. Nebeker, "Technology to support aging in place: Older adults' perspectives," *Healthcare*, vol. 7, no. 2, p. 60, 2019.

[61] E. Shandilya and M. Fan, "Understanding older adults' perceptions and challenges in using ai-enabled everyday technologies," in *Proceedings of the Tenth International Symposium of Chinese CHI*, 2022, pp. 105–116.

[62] M. Elboim-Gabyzon and A. Danial-Saad, "Correlation between the ability to manipulate a touchscreen device and hand strength and manual dexterity among community-living older individuals," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, p. 9408, 2021.

[63] C. Lee and J. F. Coughlin, "Perspective: Older adults' adoption of technology: An integrated approach to identifying determinants and barriers," *Journal of Product Innovation Management*, vol. 32, no. 5, pp. 747–759, 2015.

[64] M. M. Baig, S. Afifi, H. GholamHosseini, and F. Mirza, "A systematic review of wearable sensors and iot-based monitoring applications for older adults—a focus on ageing population and independent living," *Journal of Medical Systems*, vol. 43, no. 1, pp. 1–11, 2019.

[65] H. Brodaty, T. Chau, M. Heffernan, J. A. Ginige, G. Andrews, M. Millard, P. S. Sachdev, K. J. Anstey, N. T. Lautenschlager, J. J. McNeil *et al.*, "An online multidomain lifestyle intervention to prevent cognitive decline in at-risk older adults: a randomized controlled trial," *Nature medicine*, vol. 31, no. 2, pp. 565–573, 2025.

[66] C. Zou, D. Amos-Richards, R. Jagannathan, and A. Kulshreshtha, "Effect of home-based lifestyle interventions on cognition in older adults with mild cognitive impairment: a systematic review," *BMC geriatrics*, vol. 24, no. 1, p. 200, 2024.

[67] P. Cedillo, P. Cárdenas-Delgado, A. P. R. Zuniga, L. Encalada-Torres, C. Ramos-Galarza, and J. Jadan-Guerrero, "A systematic literature review on human-computer interaction to support older adults' physical and cognitive health," *IEEE Access*, 2025.

[68] A. D. Fisk, S. J. Czaja, W. A. Rogers, N. Charness, and J. Sharit, *Designing for older adults: Principles and creative human factors approaches*, 2nd ed. CRC Press, 2020.

[69] R. Zhong, M. Ma, Y. Zhou, Q. Lin, L. Li, and N. Zhang, "User acceptance of smart home voice assistant: A comparison among younger, middle-aged, and older adults," *Universal Access in the Information Society*, vol. 23, no. 1, pp. 275–292, 2024.

[70] A. M. Kueider, J. M. Parisi, A. L. Gross, and G. W. Rebok, "Computerized cognitive training with older adults: A systematic review," *PLoS ONE*, vol. 7, no. 7, p. e40588, 2012.

[71] H. Liu, W. Wang, Y. Liu, F. Song, S. Wang, and H. Guo, "Study on the differences of icon cognition of graphical interface for age-friendly design," *Journal of Gerontological Social Work*, vol. 66, no. 5, pp. 662–679, 2022.

[72] L. Chen and Y. Liu, "Interface design for products for users with advanced age and cognitive impairment," *International Journal of Environmental Research and Public Health*, vol. 19, no. 4, p. 2466, 2022.

[73] L. Chen, Y. Huang, and R. Zhao, "Product interfaces for older adult users with cognitive impairment," *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 30, no. 4, pp. 12–18, 2022.

[74] S. Wilson, P. Byrne, S. Rodgers, and M. Maden, "A systematic review of smartphone and tablet use by older adults with and without cognitive impairment," *Innovation in Aging*, vol. 6, no. 1, p. igac002, 2022.

[75] D. Zhu, Y. Jing, R. Huang, Y. Gao, Y. Liu, Z. Zou, and W. Liu, "Designing a mobile application for working memory training through understanding the psychological and physiological characteristics of older adults," *Sustainability*, vol. 14, no. 21, p. 14152, 2022.

[76] C. Acosta, R. Palacio, G. Borrego, R. García, and M. Rodríguez, "Design guidelines and usability for cognitive stimulation through technology in mexican older adults," *Informatics for Health and Social Care*, vol. 47, no. 2, pp. 103–119, 2021.

[77] L. M. Contreras-Somoza, E. Irazoki, J. M. Toribio-Guzmán, I. De La Torre-Díez, A. A. Diaz-Baquero, E. Parra-Vidales, M. Perea-Bartolomé, and M. A. Franco-Martín, "Usability and user experience of cognitive intervention technologies for elderly people with mci or dementia: A systematic review," *Frontiers in Psychology*, vol. 12, p. 636116, 2021.

[78] A. N. Belkacem, N. Jamil, J. A. Palmer, S. Ouhbi, and C. Chen, "Brain computer interfaces for improving the quality of life of older adults and elderly patients," *Frontiers in neuroscience*, vol. 14, p. 692, 2020.

[79] T. S. Lee, S. J. A. Goh, S. Y. Quek, R. Phillips, C. Guan, Y. B. Cheung, and J. Lee, "A brain-computer interface based cognitive training system for healthy elderly: A randomized control pilot study for usability and preliminary efficacy," *PLoS ONE*, vol. 8, no. 11, p. e79419, 2013.

[80] J. Gomez-Pilar, R. Corralejo, L. F. Nicolas-Alonso, D. Álvarez, and R. Hornero, "Neurofeedback training with a motor imagery-based bci: Neurocognitive improvements and eeg changes in the elderly," *Medical & Biological Engineering & Computing*, vol. 54, no. 11, pp. 1655–1666, 2016.

[81] Y. Liao, H. Tseng, Y. Lin, C. Wang, and W. Hsu, "Using virtual reality-based training to improve cognitive function, instrumental activities of daily living and neural efficiency in older adults with mild cognitive impairment: A randomized controlled trial," *European Journal of Physical and Rehabilitation Medicine*, vol. 56, no. 1, pp. 47–57, 2020.

[82] S. Czaja and L. Ross, "Technology to support aging adults with a cognitive impairment," *Innovation in Aging*, vol. 7, no. 1, p. 98, 2023.

[83] P. Tsai, C. Scarampi, M. Kliegel, and S. Gilbert, "Optimal cognitive offloading: Increased reminder usage but reduced pro-reminder bias in older adults," *Psychology and Aging*, vol. 38, no. 2, pp. 123–135, 2023.

[84] Y. Shin, R. Barankevich, J. Lee, and S. Kalantari, "Pencoder: Design for prospective memory and older adults," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[85] A. Hering, M. Kliegel, P. G. Rendell, F. I. Craik, and N. S. Rose, "Prospective memory is a key predictor of functional independence in older adults," *Journal of the International Neuropsychological Society*, vol. 24, no. 6, pp. 640–645, 2018.

[86] E. Irazoki, L. M. Contreras-Somoza, J. M. Toribio-Guzmán, C. Jenaro-Río, H. G. Van Der Roest, and M. A. Franco-Martín, "Technologies for cognitive training and cognitive rehabilitation for people with mild cognitive impairment and dementia: A systematic review," *Frontiers in Psychology*, vol. 11, p. 648, 2020.

[87] X. Hu, S. Desai, M. Lundy, and J. Chin, "Beyond functionality: Co-designing voice user interfaces for older adults' well-being," *arXiv preprint arXiv:2409.08449*, 2024.

[88] Q. Li and Y. Luximon, "Older adults' use of mobile devices: Usability challenges while navigating various interfaces," *Behaviour & Information Technology*, vol. 39, no. 8, pp. 837–861, 2020.

[89] R. Leung, J. McGrenere, and P. Graf, "Age-related differences in the initial usability of mobile device icons," *Behaviour & Information Technology*, vol. 30, no. 5, pp. 629–642, 2011.

[90] M. Gomez-Hernandez, X. Ferre, C. Moral, and E. Villalba-Mora, "Design guidelines of mobile apps for older adults: systematic review and thematic analysis," *JMIR mHealth and uHealth*, vol. 11, p. e43186, 2023.

[91] Q. Li and Y. Luximon, "Older adults' use of mobile devices: Usability challenges while navigating various interfaces," *Behaviour & Information Technology*, vol. 39, no. 8, pp. 837–861, 2019.

[92] J. Wilson, M. Heinsch, D. Betts, D. Booth, and F. Kay-Lambkin, "Barriers and facilitators to the use of e-health by older adults: A scoping review," *BMC Public Health*, vol. 21, pp. 1–12, 2021.

[93] A. Kebede, L. Ozolins, H. Holst, and K. Galvin, "Digital engagement of older adults: Scoping review," *Journal of Medical Internet Research*, vol. 24, no. 1, p. e40192, 2022.

[94] N. Liu, J. Yin, S. S. L. Tan, K. Y. Ngiam, and H. H. Teo, "Mobile health applications for older adults: A systematic review of interface and persuasive feature design," *Journal of the American Medical Informatics Association*, vol. 28, no. 11, pp. 2483–2501, 2021.

[95] A. Jaskulska, K. Skorupska, B. Karpowicz, C. Biele, J. Kowalski, and W. Kopeć, "Exploration of voice user interfaces for older adults—a pilot study to address progressive vision loss," in *Conference on Multimedia, Interaction, Design and Innovation*. Springer, 2020, pp. 159–168.

[96] J. Chung, N. Mansion, M. Viceconte, R. Syros, and J. Winship, "User challenges and preferences for voice interface design on smart speakers among low-income minority older adults," *Innovation in Aging*, vol. 8, no. Supplement_1, pp. 581–582, 2024.

[97] T. Liu and F. Lin, "A new mmwave-speech multimodal speech system for voice user interface," *GetMobile: Mobile Computing and Communications*, vol. 27, no. 4, pp. 31–35, 2024.

[98] X. Hu, S. Desai, M. Lundy, and J. Chin, "Beyond functionality: Co-designing voice user interfaces for older adults' well-being," https://arxiv.org/abs/2409.08449, 2024, accessed: 2024-06-12.

[99] D. E. Crews and S. Zavotka, "Aging, disability, and frailty: implications for universal design," *Journal of physiological anthropology*, vol. 25, no. 1, pp. 113–118, 2006.

Forecast COVID-19 Epidemics by Strengthening Deep Learning Models with Time Series Analysis

Warapree Tangseefa[1], Tepanata Pumpaibool[2], Paisit Khanarsa[3], Krung Sinapiromsaran[4]*
College of Public Health Sciences, Chulalongkorn University, Bangkok, Thailand, 10330[1,2]
Institute of Field Robotics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand, 10140[3]
Department of Mathematics and Computer Science, Chulalongkorn University, Bangkok, Thailand, 10330[4]

***Abstract*—The COVID-19 pandemic has profoundly impacted economic and social structures, directly affecting individuals' lives. Deep learning models offer the potential to forecast future long-term trends and capture the temporal dependencies present in time series data. In this study, we propose leveraging the autocorrelation function (ACF) and the partial autocorrelation function (PACF) series as additional components to enhance the forecasting accuracy of our models. Our proposed method is applied to forecast COVID-19 time series data in twelve countries using the deep learning techniques of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). When comparing the rankings average of mean absolute error and R-squared, the proposed models demonstrated superior performance in time series forecasting compared to the standard LSTM and GRU model. Specifically, the ACF-PACF-GRU model achieved the best median values for mean absolute percentage error (1.67 per cent for confirmed cases and 2.17 per cent for death cases) and root mean square error (1.92 for confirmed cases and 2.17 for death cases). Therefore, the proposed ACF-PACF-GRU model showed the highest performance in forecasting both confirmed and death cases. This research introduces a novel method for constructing effective time series models aimed at forecasting disease burdens, thereby aiding in epidemic control and the implementation of preventive measures.**

***Keywords*—Forecasting models; COVID-19; long short-term memory; gated recurrent unit; autocorrelation function; partial autocorrelation function**

I. Introduction

In late 2019, a novel β-coronavirus—SARS-CoV-2—emerged in Wuhan, China, posing significant challenges to global health systems [1], [2], [3]. Declared a public health emergency by the WHO in March 2020 [4], COVID-19 has since resulted in over 776.8 million confirmed cases and more than 7 million deaths across 234 countries as of November 10, 2024. Although mortality peaked between 2020 and 2022, global death rates have declined with rising immunity levels [5]. The pandemic's impact spans healthcare, economics, travel, and international policy, highlighting the critical need for accurate forecasting to guide public health interventions and resource allocation [6].

COVID-19 demonstrates dynamic, non-linear transmission patterns influenced by seasonal variation and evolving viral behavior, necessitating adaptive modeling approaches [7]. Time series forecasting has traditionally relied on statistical models like ARIMA and SARIMA. ARIMA has been widely applied to model diseases such as tuberculosis [8], brucellosis [9], and dengue [10]. However, it struggles with capturing seasonality and nonlinear patterns, which SARIMA partially addresses [11], [12], [13], [14], [15]. For COVID-19, ARIMA and SARIMA have been used in diverse contexts, such as Morocco [16], Europe [17], and Canada [18], with varying degrees of accuracy. Nevertheless, their linear assumptions and need for differencing limit performance in more complex, evolving datasets [19].

To overcome these limitations, deep learning architectures have emerged as powerful tools for time series forecasting. Advances in data availability and computational power have facilitated their widespread adoption across domains, with deep neural networks proving particularly effective for regression tasks and temporal pattern modeling [20], [21].

The proliferation of big data, driven by advancements in data acquisition technologies, has created a demand for advanced forecasting models capable of processing complex, multivariate, and high-dimensional time series data [22]. Traditional shallow learning approaches often fail to capture the intricate temporal dependencies and noise inherent in such datasets. In response, deep learning has emerged as a dominant paradigm for time series forecasting due to its ability to model non-linear relationships and long-term dependencies effectively [23].

Among deep learning architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are particularly suited for sequential data. Deep learning models have gained significant attention in COVID-19 forecasting due to their capacity to capture non-linear patterns and temporal dependencies in time series data. Among these, LSTM and GRU networks have emerged as leading architectures for epidemic modeling. Both are types of recurrent neural networks (RNNs) specifically designed to mitigate the vanishing gradient problem, enabling the modeling of long-range temporal dependencies crucial for accurate disease forecasting.

Several studies have demonstrated the effectiveness of LSTM models in predicting COVID-19 cases. Dharani and Bojja [24] proposed an enhanced LSTM architecture with internal gate optimizations to improve the prediction accuracy of COVID-19 time series. Similarly, Shahid et al. [25] compared multiple models—including ARIMA, SVR, LSTM, and Bi-LSTM—for forecasting confirmed, death, and recovered cases across ten countries, finding that Bi-LSTM significantly outperformed other models. Luo et al. [26] further validated the predictive strength of LSTM through a comparative study with XGBoost, highlighting LSTM's superior performance in

*Corresponding authors.

capturing temporal dynamics.

GRU models, which simplify LSTM by using fewer parameters while maintaining comparable performance, have also been applied successfully. ArunKumar et al. [27] utilized GRU, LSTM, and other RNN variants to forecast cumulative confirmed cases in the ten most affected countries, reporting promising results. The comparative advantage of GRU lies in its computational efficiency and similar accuracy, making it suitable for real-time applications in epidemic forecasting.

Moreover, hybrid and ensemble variants such as stacked LSTM, bidirectional GRU, and LSTM-GRU combinations have shown improved performance by leveraging the strengths of both architectures [28], [29], [30]. These models are particularly effective in multi-country forecasting settings where data complexity and temporal variability are high.

Overall, the literature suggests that LSTM and GRU models are well-suited for modeling the complex, nonlinear, and dynamic nature of COVID-19 transmission. Their adaptability, accuracy, and scalability make them powerful tools for public health planning and early warning systems. Given the complex, non-linear nature of time series data, statistical tools like ACF and PACF are often used to analyze temporal dependencies. Their integration into deep learning models, as proposed in this study, is a novel approach that aims to enhance prediction accuracy without altering model architectures [31].

This study integrates ACF and PACF as supplementary inputs to LSTM and GRU models to enhance time series forecasting accuracy. By leveraging these statistical tools to capture temporal dependencies, the proposed approach forecasts COVID-19 confirmed and death cases across twelve countries, and compares results against standard LSTM and GRU models reported in the previous study of Luyu Zhou et al [32]. Unlike prior work, this method augments input features without altering the core architectures. Model performance is evaluated using MAE, MAPE, RMSE, and R-squared.

The study is structured as follows: Section II details LSTM and GRU architectures; Section III outlines the data sources, methodology, preprocessing, and evaluation metrics; Section IV presents results and comparisons; Section V discusses the study results with limitations. Finally, Section VI summarizes the conclusion of the whole study.

II. Methodologies

This study illustrates the effective integration of deep learning architectures with statistical techniques to enhance the accuracy of COVID-19 epidemiological data forecasting. In particular, two widely recognized deep learning approaches, LSTM and GRU, are employed in conjunction with ACF and PACF to capture and model the temporal dependencies inherent within the time series data.

A. Long Short-Term Memory (LSTM)

Deep learning techniques like RNN are well-suited for sequence processing because they capture the temporal dynamics of time series data [33]. However, when dealing with very long time series, passing information from earlier time steps to later ones becomes challenging, leading to difficulties in training due to vanishing or exploding gradients. Hochreiter and Schmidhuber introduced LSTM networks to address this issue by facilitating the flow of previous information across state sequences [34]. The LSTM cell architecture includes a memory cell and three gates: the forget gate (F_t), the input gate (I_t), and the output gate (O_t) shown as Fig. 1.

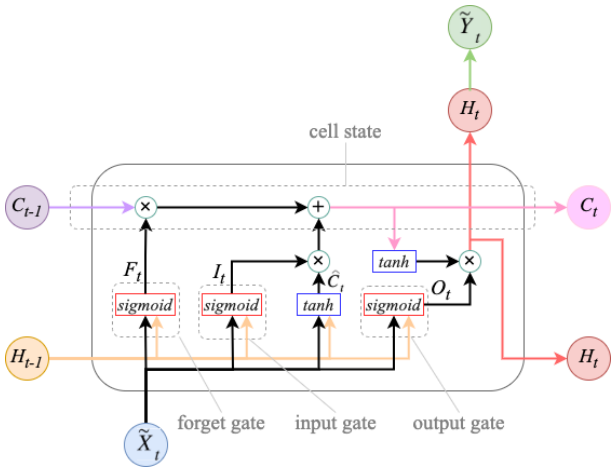


Fig. 1. Architecture of a single LSTM cell.

These gates regulate the flow of information within the cell state: the forget gate controls what information to discard, the input gate determines what information to input to the cell state, and the output gate manages outgoing information [35]. The intermediate cell state ($\hat{C}_t$) serves as the core memory component, propagating information across time steps. Its update at each step is governed by gating mechanisms that selectively regulate the preservation, forgetting, or modification of information. This design enables LSTMs to retain information over long periods, a capability not fully realized in conventional single RNNs. LSTMs excel in capturing long-term dependencies and effectively handling time series data. LSTM network is compute mapping between input sequence and output sequence, i.e. $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n)$ and $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_n)$. Given the normalized scale of the time series as input $\tilde{X}_t$, as output $\tilde{Y}_t$, and the number of hidden units as h. The gates have the following equations [Eq. (1) to Eq. (7)]:

Input gate:

$$I_t = \text{sigmoid}(W_{xi}\tilde{X}_t + W_{hi}H_{t-1} + b_i) \tag{1}$$

Forget gate:

$$F_t = \text{sigmoid}(W_{xf}\tilde{X}_t + W_{hf}H_{t-1} + b_f) \tag{2}$$

Output gate:

$$O_t = \text{sigmoid}(W_{xo}\tilde{X}_t + W_{ho}H_{t-1} + b_o) \tag{3}$$

Intermediate cell state:

$$\hat{C}_t = \tanh(W_{xc}\tilde{X}_t + W_{hc}H_{t-1} + b_c) \tag{4}$$

Cell state (next memory input):

$$C_t = F_t \odot C_{t-1} + I_t \odot \hat{C}_t \tag{5}$$

New state:

$$H_t = O_t \odot \tanh(C_t) \tag{6}$$

Normalized predicted output:

$$\tilde{Y}_t = H_t \tag{7}$$

where, W_{xi}, W_{xf}, W_{xo}, W_{xc}, and W_{hi}, W_{hf}, W_{ho}, W_{hc} refer respectively to the weight parameters and b_i, b_f, b_o, b_c denote bias parameters for three gates and a memory cell. The symbol $\odot$ indicates the element-wise multiplication. The estimation of C_t relies on the output information from memory cells (C_{t-1}) and the current time step $\hat{C}_t$. The predicted output $\tilde{Y}_t$ is obtained from the hidden state (H_t) at each time t in the current LSTM cell.

B. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) serves as an alternative version of LSTM designed to enhance the LSTM performance, reduce the number of parameters, and simplify its design [36]. GRU achieves this by amalgamating the input and forget gates of LSTM into a single update gate (Z_t), and utilizing the output gate as a reset gate (R_t). The introduction of reset and update gates represents novel features introduced by GRU models, offering a fresh evaluation method for calculating hidden states in RNN models. In the general architecture of GRU, only the update and reset gates are employed (see Fig. 2).

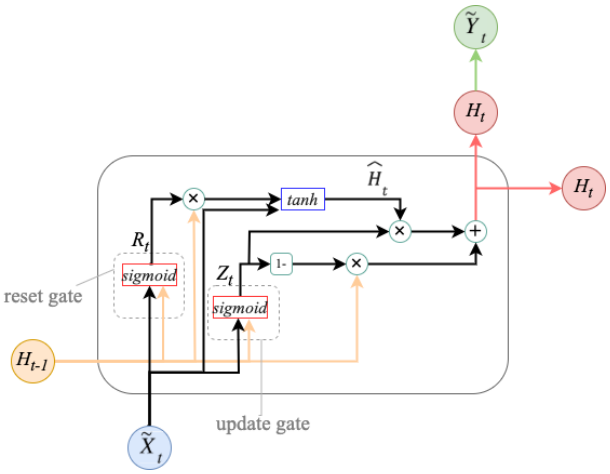


Fig. 2. Architecture of a single GRU cell.

The update gate determines the extent of previously retained memory, while the reset gate governs the manner in which current inputs combine with past memory. The GRU network compute mapping between input sequence and the output sequence in normalized form, that is, $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n)$ and $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_n)$. The mathematical relationships between the various GRU components are presented in Eq. (8) to Eq. (12):

Update gate:

$$Z_t = \text{sigmoid}(W_{xz}\tilde{X}_t + W_{hz}H_{t-1} + b_z) \tag{8}$$

Reset gate:

$$R_t = \text{sigmoid}(W_{xr}\tilde{X}_t + W_{hr}H_{t-1} + b_r) \tag{9}$$

Cell state:

$$\hat{H}_t = \tanh(W_{xh}\tilde{X}_t + W_{hh}(R_t H_{t-1}) + b_h) \tag{10}$$

New state:

$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot \hat{H}_t \tag{11}$$

Normalized predicted output:

$$\tilde{Y}_t = H_t \tag{12}$$

where, W_{xz}, W_{xr}, W_{xh} and W_{hz}, W_{hr}, W_{hh} refer respectively to weight parameters and b_z, b_r, b_h denote bias parameters. The symbol $\odot$ indicates the element-wise multiplication. For a given time step t, the current update gate Z_t is used to combine the previous hidden state H_{t-1} and current candidate hidden state $\hat{H}_t$. The normalized predicted output $\tilde{Y}_t$ is derived from the hidden state (H_t) at each time step t within the current GRU cell.

C. Autocorrelation Function and Partial Autocorrelation Function

Time series data often harbors hidden characteristics, encompassing both linear and non-linear components that profoundly impact analysis, modeling, and forecasting. In this context, the autocorrelation function (ACF) and partial autocorrelation function (PACF) emerge as indispensable tools in time series analysis, offering valuable insights into its structure and dynamics. The ACF and PACF serve to uncover the presence and characteristics of autocorrelation within the data. Autocorrelation and partial autocorrelation represent statistical measures that assess the relationship between a time series and its preceding values, known as lag. These measures illuminate the temporal dependencies and patterns inherent in the time series, thereby facilitating a deeper understanding of its behavior and informing predictive modeling endeavors.

The ACF measures the correlation between observations in a time series and their lagged values across different time intervals. This helps to identify patterns and temporal dependencies within the time series data. The ACF calculates the correlation between the series (X_t) and the same series lagged by k time periods (X_{t-k}). Mathematically, the ACF at lag-k is expressed in Eq. (13) to Eq. (14) [37]:

$$\text{Corr}_{(X_t, X_{t-k})} = \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t) \cdot \text{Var}(X_{t-k})}} \tag{13}$$

$$\text{Corr}_{(X_t, X_{t-k})} = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X}) \cdot (X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \quad (14)$$

where, X_t represents the value of a time series at time t, $\bar{X}$ means the mean of the time series, and n denotes the number of observations in the time series. The variable k represents the time gap being examined and is termed the lag, where k ranges from 1 to n. In essence, a lag-k autocorrelation refers to the correlation between values that are k time periods apart.

Unlike autocorrelation, partial autocorrelation isolates the direct correlation between an observation in a time series and its past values, eliminating the influence of other values at intermediate time lags [31]. For instance, the partial autocorrelation for lag-3 only considers the correlation that lag-1 and lag-2 do not account for. In a time series, the partial autocorrelation between X_t and X_{t-k} is defined as the conditional correlation between X_t and X_{t-k}, conditioned on X_{t-1}, ... , X_{t-k+1}, representing the observations between time points t and $t-k$. The PACF at lag-1 is defined to equal the ACF at lag-1. Subsequently, the partial autocorrelation at lag-k is expressed mathematically as follows [37]:

$$\text{Corr}_{(X_t, X_{t-k} | X_{t-1}, \ldots, X_{t-k+1})} = \frac{\text{Cov}(X_t, X_{t-k} | X_{t-1}, \ldots, X_{t-k+1})}{\sqrt{\text{Var}(X_t | X_{t-1}, \ldots, X_{t-k+1}) \cdot \text{Var}(X_{t-k} | X_{t-1}, \ldots, X_{t-k+1})}} \quad (15)$$

In Eq. (15) X_t represents the value of a time series at time t, and n signifies the total number of observations in the time series. The variable k denotes the time gap under consideration and is termed the lag, with k ranging from 1 to n. Essentially, a lag-k autocorrelation indicates the correlation between values that are k time periods apart.

D. Proposed Methods with ACF, PACF Series

Deep learning techniques, such as LSTM and GRU, have been successfully applied in the context of time series forecasting. The performance of these methods relies heavily on having a sufficient amount of data to appropriately fit their parameters [17]. However, the number of samples extracted from a short time series may be insufficient to achieve an optimal model [38]. To prevent overfitting, these methods should be properly regularized. One challenge in time series forecasting is that, even with a long series and ample data, observations from the distant past often provide fewer cues for prediction. Recent observations from an individual series tend to be more useful in forecasting. This could be attributed to the evolving nature of patterns within a series.

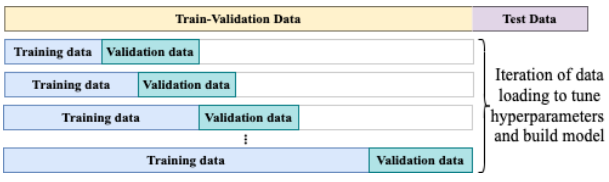


Fig. 3. Train-validation-test split of time series dataset.

In Fig. 3, a given time series is divided into in-sample and out-sample segments using a certain ratio, such as 70/30. The out-sample portion serves as testing data to evaluate the model. The in-sample segment is further divided into training and validation data. However, using separate validation data means excluding recent observations from the training data, potentially missing recent patterns. One simple solution is to include the validation data in model training. However, this may lead to overfitting, reducing accuracy on test data. Specifically, we are determined to apply the statistical properties of the ACF, PACF to avoid model overfitting and enhance the learning potential of deep learning methods for time series forecasting.

The primary concept of the proposed models in this study is to integrate the ACF and PACF series from lag-1 to lag-7, as another new series into deep learning models based on LSTM and GRU for sequential predictions of COVID-19 cases, as depicted in Fig. 4. A part of the training data, and validation data are utilized to build the ACF and PACF series from lag-1 to lag-7, separately in each time period. Both ACF, PACF series uncover the temporal dependencies that exist in the original series. Therefore, the original part of training data and its ACF/PACF series are passed to the standard LSTM or GRU model for generate one proposed model in the model training process, such as ACF-LSTM, PACF-LSTM, ACF-PACF-LSTM, ACF-GRU, PACF-GRU, and ACF-PACF-GRU, as shown in Fig. 5. In the validation process, the original part of training data and its ACF/PACF series are employed with an optimization algorithm to fine-tune the hyperparameters of each model that fit the trained model. Moreover, the testing data from the original time series is used to generate the ACF and PACF series, which are then employed to predict values during the testing phase, as illustrated in Fig. 6.

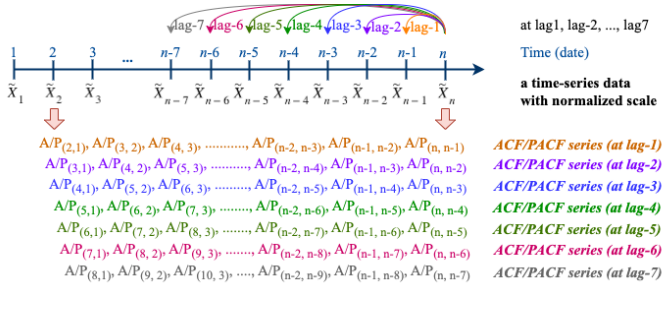


Fig. 4. ACF and PACF series at lag1 to lag7.

III. Materials

This section describes the source of the dataset, design experiments for COVID-19 forecasting, data preprocessing, and the evaluation criteria for model predictive performance.

A. Data Source of COVID-19

This study aims to compare deep learning prediction models using COVID-19 datasets from twelve countries with the highest number of COVID-19 cases. The COVID-19 dataset was sourced from the WHO International website, providing daily cumulative cases of COVID-19 [39], in .csv format. The dataset contains the daily cumulative number of confirmed,

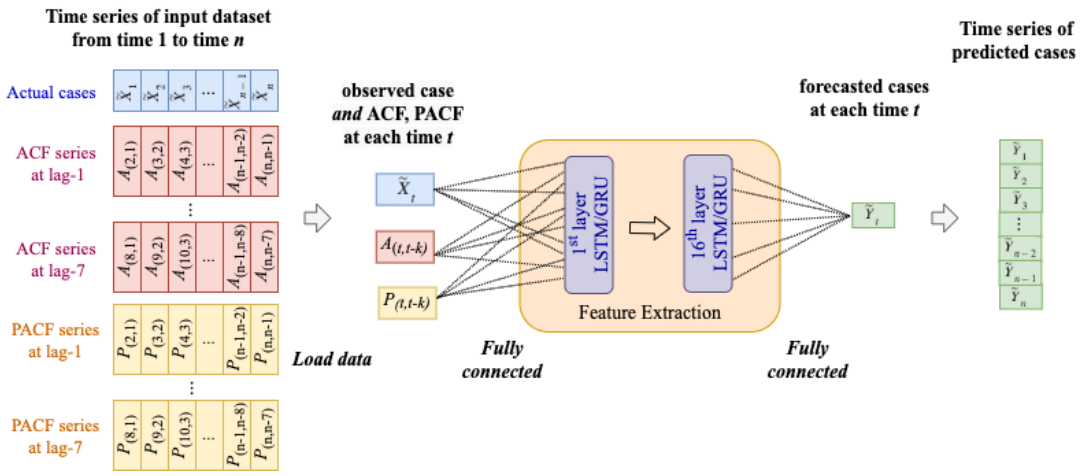


Fig. 5. COVID-19 data, ACF and PACF series into proposed model.

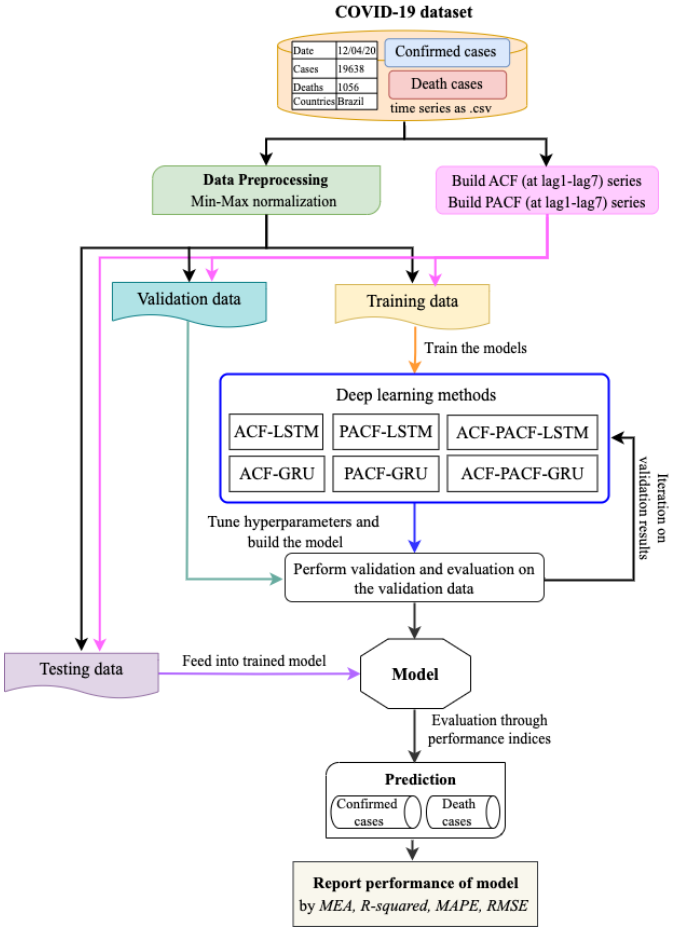


Fig. 6. Framework of the proposed scheme.

death, and recovered cases in the time series format with temporal patterns. In this study, the experiments were conducted using the time series data of confirmed and death cases for twelve countries with the highest number of confirmed cases from January 6, 2020, to June 6, 2022. These countries are the USA, India, Brazil, France, Germany, the UK, Russia, Italy, Turkey, Spain, Vietnam, and Argentina. These specific countries were chosen based on a selection process proposed in a prior publication [32]. This COVID-19 data set includes the counts of confirmed cases, and deaths cases shown in Table I. For training and validation, 70% of each time series is allocated, with the remaining 30% reserved for case prediction in Fig. 3. Before developing a proposed model, the data series undergoes preprocessing.

TABLE I. OVERALL COVID-19 CASES

Countries	Confirmed cases	Death cases
USA	83,894,969	1,000,512
India	43,181,335	524,701
Brazil	31,153,069	666,997
France	28,647,902	149,202
Germany	26,558,190	143,131
UK	22,342,463	199,612
Russia	18,355,200	379,584
Italy	17,505,973	166,949
Turkey	15,077,320	99,004
Spain	12,520,475	108,488
Vietnam	10,726,045	43,081
Argentina	9,276,618	128,973

B. Data Preprocessing

The dataset consists of two features: confirmed and death cases. Unscaled data can impede the convergence process. To enhance model performance and ensure consistency in evaluating deep learning models, preprocessing raw data through normalization is essential. We applied normalization to the data using the formulation $\tilde{X}_t = (X_t – min)/(max – min)$, where $\tilde{X}_t$ represents the normalized actual value at time point t, ensuring it falls within a fixed range typically from zero to one, and X_t denotes the actual value observed at time point t. After constructing the forecasting model at each time point t and obtaining the forecasted values on the normalized scale $\tilde{Y}_t$, we transformed them back to the original data scale using the following formula: $Y_t = \tilde{Y}_t * (max - min) + min$. In this

context, Y_t represents the predicted data at the time point t, rescaled to the raw data format. In Python, the observed raw data can be normalized using the *MinMaxScaler* object from scikit-learn [40]. MinMaxScaler subtracts the smallest feature value and then divides by the feature range, where the range is the difference between the original maximum and original minimum values. This scaler preserves the shape of the original data distribution, retaining the information embedded in the original dataset without significantly altering it or diminishing the importance of outliers.

C. Design Experiments for COVID-19 Forecasting

This study proposes a deep-learning framework for COVID-19 time-series forecasting, integrating the ACF, and PACF series from lag-1 to lag-7 into LSTM and GRU models. Six deep learning models are employed for forecasting daily confirmed and death cases: ACF-LSTM, PACF-LSTM, ACF-PACF-LSTM, ACF-GRU, PACF-GRU, and ACF-PACF-GRU in Fig. 5. The proposed models offer a comprehensive exploration of deep learning techniques. A visual representation of the proposed scheme is depicted in Fig. 6.

Given the progressive spread of COVID-19, the cumulative number of confirmed and death cases is forecasted at specific points in time using preceding values. The COVID-19 time series datasets in this study are segmented into training, validation, and testing datasets, as illustrated in Fig. 6. Monitoring the model's generalization capabilities and selecting appropriate hyperparameters is crucial. To address this, the data is partitioned based on time for training and validation purposes.

COVID-19 forecasting entails two main stages: training and testing. During the initial stage, raw data is preprocessed, standardized, and utilized to construct the deep learning model. Parameters for the deep learning models are adjusted to minimize the loss function during training, employing the Adam optimizer for this purpose. Subsequently, in the testing stage, the previously constructed models with the selected parameters are deployed to forecast the number of COVID-19 cases. The hyperparameter settings for the deep learning models—ACF-LSTM, PACF-LSTM, ACF-PACF-LSTM, ACF-GRU, PACF-GRU, and ACF-PACF-GRU—are outlined in Table II.

TABLE II. PROPOSED SCHEME WITH HYPERPARAMETERS AND THEIR VALUES

Methods	Hyperparameters	Values
ACF-LSTM/	Layers	1
ACF-GRU/	Timestep	7
PACF-LSTM/	Hidden units	16
PACF-GRU	Feature(s)	8
	Loss	MAE
	Last activation	tanh
	Batch size	64
	Epochs	10000
	Optimizer	Adam
	ACF/PACF	lag-1, lag-2, lag-3,..., lag-6, lag-7
ACF-PACF-LSTM/	Layers	1
ACF-PACF-GRU	Timestep	7
	Hidden units	16
	Feature(s)	15
	Loss	MAE
	Last activation	tanh
	Batch size	64
	Epochs	10000
	Optimizer	Adam
	ACF	lag-1, lag-2, lag-3,..., lag-6, lag-7
	PACF	lag-1, lag-2, lag-3,..., lag-6, lag-7

Sixteen cells of LSTM or GRU algorithms are incorporated into a single hidden layer within the proposed models illustrated in Fig. 7. Model accuracy is evaluated by comparing measured data with real data using standard statistical performance measures including MAE, MAPE, RMSE, and R-squared. Simulations and experiments of the proposed models of deep learning are implemented with Keras [41], the Python deep learning library. The primary objective of this study is to assess the predictive capabilities of the six proposed models for forecasting the number of COVID-19 cases, which will be compared with the standard LSTM and GRU models in terms of MAE, and R-squared presented in the previous study. The specific steps involved are as follows:

Step1. The data of COVID-19 are normalized, and the time series is divided into training sets and testing sets;

Step2. Introduce the ACF and, or PACF series (at lag-1, ..., and lag-7) to LSTM, or GRU algorithm;

Step3. Define the hyperparameters in an proposed model in training data;

Step4. In the training step, using the training set data to determine the optimal solution of the model to predict the forecasted values in the training data;

Step5. In the validation step, using validation data to evaluate the fit model from training step;

Step6. In the testing step, using testing data to test the trained model, and evaluate the accuracy of the model by using performance indices. Given the progressive spread of COVID-19, the cumulative number of confirmed and death cases is forecasted at specific points in time using preceding values. The COVID-19 time series datasets in this study are segmented into training, validation, and testing datasets, as illustrated in Fig. 3. Monitoring the model's generalization capabilities and selecting appropriate hyperparameters is crucial. To address this, the data is partitioned based on time for training and validation purposes.

D. Performance Metrics

In this study, the mean absolute error (MAE), and the coefficient of determination (R-squared or R^2) are utilized as predictive performance metrics to evaluate forecast accuracy and compare between proposed models and the standard deep learning models (LSTM, and GRU). Whereas comparing the accuracy forecasting among the six proposed models is considered by the mean absolute percentage error (MAPE), and the root mean square error (RMSE) [42], [43]. These performance metrics are estimated from a normalized form of the actual and predicted values at time point t presented as $\tilde{X}_t$ and $\tilde{Y}_t$,

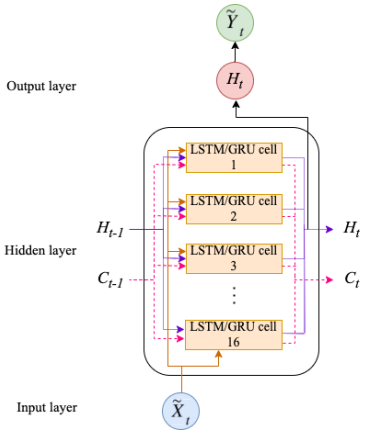


Fig. 7. Architecture of 16 LSTM/GRU cells.

respectively. $\bar{\tilde{X}}$ represents the mean of the normalized actual values, and n denotes the total number of data time points. A model demonstrates superior performance when MAE, MAPE, and RMSE are closer to 0. Additionally, a model is considered more efficient when R-squared approaches 1 [44], [45]. Mean absolute error (MAE) enables the measurement of the average magnitude of errors for a set of predictions, irrespective of their direction. It is expressed mathematically in Eq. (16):

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|\tilde{Y}_t - \tilde{X}_t\right| \tag{16}$$

MAPE quantifies the magnitude of errors as a percentage and can be expressed mathematically in Eq. (17):

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\tilde{X}_t - \tilde{Y}_t}{\tilde{X}_t}\right| \times 100 \tag{17}$$

Root mean square error (RMSE) reveals the discrepancies in amplitude between actual and forecasted data. RMSE is often flavored for evaluating short-term forecasting [46], [47], and can be calculated as Eq. (18):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\tilde{Y}_t - \tilde{X}_t)^2} \tag{18}$$

To demonstrate the proportion of the variation between actual value and forecasted value, R-squared is presented in Eq. (19):

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(\tilde{X}_t - \tilde{Y}_t)^2}{\sum_{t=1}^{n}(\tilde{X}_t - \bar{\tilde{X}})^2} \tag{19}$$

IV. Results

This section describes the evaluation criteria for model predictive performance, results, and discussion about those results. To evaluate the performance of deep learning methods for COVID-19 time-series forecasting, this study compares six proposed models against the standard LSTM and GRU models used as baselines presented as the results from the previous study of Zhou et al. (2023). Our study expands upon the methodologies used by Zhou et al. (2023) by incorporating ACF and PACF as additional input features to the LSTM and GRU models, which has led to substantial improvements in forecasting accuracy. Moreover, the results of this study cover the forecasted trends of cumulative confirmed, and cumulative death of the COVID-19 cases are in using all six proposed models, including ACF-LSTM model, PACF-LSTM model, ACF-PACF-LSTM model, ACF-GRU model, PACF-GRU model, and ACF-PACF-GRU model. The top-12 countries based on cumulative confirmed cases, as of June 6, 2022, are the United States (USA), India, Brazil, France, Germany, United Kingdom (UK), Russia, Italy, Turkey, Spain, Vietnam, and Argentina. The dataset for each country comprises two time series data sets: cumulative confirmed cases, and cumulative death cases. The forecast was done for each period of the two time series data of the individual country and its ACF, and/or PACF with lag-1, lag-2, ..., and lag-7 by independently feeding those time series data into the optimized LSTM and GRU based RNN networks. A total of 144 simulations were performed.

A. Cumulative Confirmed Cases

The study evaluates forecasting accuracy by comparing the proposed models against the standard LSTM and GRU models using the same time series data of COVID-19 confirmed cases. Performance measures for eight models illustrate the actual and predicted cases across twelve countries using deep learning models based on LSTM and GRU. The rankings of MAE and R-squared values for each country, comparing the six proposed models with the standard models, are presented in Table III and Table IV, respectively.

In Table III, the MAE ranking of the standard LSTM model across twelve countries typically falls between 7$^{\text{th}}$ and 8$^{\text{th}}$, with exceptions in Argentina (2$^{\text{nd}}$), Italy (3$^{\text{rd}}$), and France (4$^{\text{th}}$). Zhou et al. (2023) found that LSTM models consistently performed at a moderate level for COVID-19 case predictions. In our study, the ACF-PACF-GRU model consistently outperformed other models, achieving the lowest average MAE ranking (2$^{\text{nd}}$), with superior accuracy in countries like Brazil, France, Turkey, Spain, Vietnam, and Argentina. The standard GRU model ranks between 1$^{\text{st}}$ and 7$^{\text{th}}$. Among the proposed LSTM-based models, the ACF-LSTM model generally ranks between 5$^{\text{th}}$ and 8$^{\text{th}}$, except for the USA, UK, and Russia, which share the 2$^{\text{nd}}$ position. The PACF-LSTM model ranks between 2$^{\text{nd}}$ and 5$^{\text{th}}$, while the ACF-PACF-LSTM model primarily falls between 6$^{\text{th}}$ and 8$^{\text{th}}$, except for Brazil and India, which both rank 5$^{\text{th}}$.

At the country level, the proposed GRU-based models show varying performance. The ACF-GRU model generally ranks between 1$^{\text{st}}$ and 5$^{\text{th}}$, except for Brazil (7$^{\text{th}}$) and Argentina (8$^{\text{th}}$). The PACF-GRU model ranks between 3$^{\text{rd}}$ and 7$^{\text{th}}$, with India as an exception at 1$^{\text{st}}$ place. The ACF-PACF-GRU model, which demonstrates superior performance, primarily ranks between 1$^{\text{st}}$ and 3$^{\text{rd}}$ in MAE, except for Russia and Italy, which both rank 4$^{\text{th}}$. This model achieves the highest forecasting accuracy in six countries: Brazil, France, Turkey, Spain, Vietnam, and Argentina.

TABLE III. ANALYSIS RESULTS OF MAE PERFORMANCE METRICS FOR COVID-19 CONFIRMED CASES ACROSS TWELVE COUNTRIES

Country	LSTM*	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	GRU*	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[8] 0.0211	[2] 0.002325	[4] 0.004666	[7] 0.012734	[5] 0.0073	[1] **0.001303**	[6] 0.007466	[3] 0.004307
India	[8] 0.0057	[6] 0.003872	[3] 0.002278	[5] 0.002675	[7] 0.0046	[4] 0.002605	[1] **0.001567**	[2] 0.001741
Brazil	[8] 0.0089	[6] 0.007829	[2] 0.001885	[5] 0.005264	[4] 0.0039	[7] 0.008784	[3] 0.003358	[1] **0.001028**
France	[4] 0.0168	[6] 0.027764	[3] 0.015254	[8] 0.028122	[2] 0.0131	[5] 0.020826	[7] 0.027992	[1] **0.008854**
Germany	[8] 0.0478	[7] 0.027316	[3] 0.016762	[6] 0.027117	[1] **0.0067**	[4] 0.018798	[5] 0.022447	[2] 0.011984
UK	[8] 0.0085	[2] 0.002272	[4] 0.003317	[8] 0.012526	[5] 0.0067	[1] **0.001142**	[6] 0.007669	[3] 0.002476
Russia	[7] 0.0062	[2] 0.002109	[5] 0.004959	[8] 0.013515	[3] 0.0035	[1] **0.001597**	[6] 0.005153	[4] 0.004294
Italy	[3] 0.0110	[8] 0.029610	[2] 0.009979	[7] 0.022999	[1] **0.0080**	[5] 0.019571	[6] 0.020897	[4] 0.013893
Turkey	[7] 0.0096	[6] 0.008251	[2] 0.004501	[8] 0.013838	[5] 0.0070	[3] 0.005817	[4] 0.006010	[1] **0.002264**
Spain	[8] 0.0153	[7] 0.012144	[2] 0.005996	[6] 0.010588	[4] 0.0081	[5] 0.009843	[3] 0.007193	[1] **0.005131**
Vietnam	[8] 0.0263	[5] 0.016750	[4] 0.010404	[6] 0.018506	[7] 0.0186	[2] 0.007740	[3] 0.010370	[1] **0.005800**
Argentina	[2] 0.0041	[7] 0.012601	[3] 0.004301	[6] 0.010753	[5] 0.0052	[8] 0.014584	[4] 0.004530	[1] **0.003561**
Avg. Ranking	[6.50]	[5.33]	[3.08]	[6.67]	[4.08]	[3.83]	[4.50]	[2.00]
Median	0.0103	0.0102	0.0048	0.0131	0.0069	0.0083	0.0073	0.0043

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models and standard deep learning models (LSTM*, GRU*) cited from the previous study of Luyu Zhou et al.

TABLE IV. ANALYSIS RESULTS OF R-SQUARED PERFORMANCE METRICS FOR COVID-19 CONFIRMED CASES ACROSS TWELVE COUNTRIES

Country	LSTM*	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	GRU*	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[8] 0.79	[2] 0.999514	[5] 0.998138	[6] 0.996355	[7] 0.82	[1] **0.999901**	[4] 0.998232	[3] 0.998652
India	[8] 0.69	[5] 0.998422	[3] 0.998596	[6] 0.998372	[7] 0.81	[2] 0.999254	[1] **0.999595**	[4] 0.998565
Brazil	[8] 0.78	[5] 0.995175	[2] 0.999668	[4] 0.997976	[7] 0.83	[6] 0.995075	[3] 0.999116	[1] **0.999820**
France	[7] 0.83	[6] 0.993040	[2] 0.997088	[5] 0.993198	[8] 0.72	[3] 0.995699	[4] 0.993328	[1] **0.997891**
Germany	[8] 0.72	[5] 0.991069	[2] 0.997044	[6] 0.990451	[7] 0.83	[3] 0.995867	[4] 0.994361	[1] **0.997991**
UK	[7] 0.83	[2] 0.999760	[4] 0.999546	[6] 0.998228	[8] 0.72	[1] **0.999930**	[5] 0.999315	[3] 0.999711
Russia	[7] 0.87	[2] 0.999738	[5] 0.998436	[6] 0.996569	[8] 0.79	[1] **0.999840**	[3] 0.998838	[4] 0.998813
Italy	[7] 0.82	[6] 0.991217	[1] **0.998369**	[4] 0.995043	[8] 0.79	[3] 0.996498	[5] 0.994902	[2] 0.998267
Turkey	[8] 0.72	[5] 0.998472	[4] 0.998901	[6] 0.996464	[7] 0.83	[2] 0.999157	[3] 0.998985	[1] **0.999708**
Spain	[7] 0.83	[6] 0.997367	[4] 0.997940	[3] 0.998465	[8] 0.78	[5] 0.997814	[2] 0.999352	[1] **0.999553**
Vietnam	[7] 0.81	[5] 0.996667	[3] 0.997697	[6] 0.996010	[8] 0.69	[2] 0.998190	[4] 0.997547	[1] **0.998866**
Argentina	[8] 0.79	[5] 0.993972	[3] 0.997978	[4] 0.996731	[7] 0.87	[6] 0.993932	[1] **0.999478**	[2] 0.998541
Avg. Ranking	[7.50]	[4.50]	[3.17]	[5.17]	[7.50]	[2.92]	[3.25]	[2.00]
Median	0.8000	0.9970	0.9983	0.9965	0.8000	0.9980	0.9989	0.9987

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models and standard deep learning models (LSTM*, GRU*) cited from the previous study of Luyu Zhou et al.

TABLE V. ANALYSIS RESULTS OF MAPE PERFORMANCE METRICS FOR COVID-19 CONFIRMED CASES ACROSS TWELVE COUNTRIES

Country	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[2] 0.303071%	[4] 0.617389%	[6] 1.496590%	[1] **0.173846%**	[5] 0.904907%	[3] 0.599810%
India	[6] 0.413458%	[3] 0.252859%	[5] 0.295722%	[4] 0.283855%	[1] **0.171580%**	[2] 0.196526%
Brazil	[5] 0.992938%	[2] 0.232107%	[4] 0.653093%	[6] 1.017317%	[3] 0.362137%	[1] **0.124322%**
France	[5] 4.023079%	[2] 2.542633%	[6] 4.031528%	[3] 3.163861%	[4] 4.014180%	[1] **1.729816%**
Germany	[6] 4.178394%	[2] 2.757850%	[5] 4.066854%	[3] 3.108251%	[4] 3.535206%	[1] **1.733134%**
UK	[2] 0.345353%	[4] 0.520922%	[6] 1.525532%	[1] **0.175605%**	[5] 0.972250%	[3] 0.369609%
Russia	[2] 0.301561%	[5] 0.662813%	[6] 1.546049%	[1] **0.227717%**	[4] 0.658677%	[3] 0.584758%
Italy	[6] 4.174280%	[1] **1.759937%**	[5] 3.201211%	[4] 2.891370%	[3] 2.628040%	[2] 2.015938%
Turkey	[5] 0.880689%	[2] 0.578842%	[6] 1.539030%	[3] 0.642271%	[4] 0.714556%	[1] **0.297870%**
Spain	[6] 1.618311%	[3] 0.980700%	[4] 1.246353%	[5] 1.397544%	[2] 0.851835%	[1] **0.655080%**
Vietnam	[6] 4.506387%	[5] 3.810372%	[3] 3.537982%	[2] 2.466952%	[4] 3.541360%	[1] **1.455438%**
Argentina	[5] 1.621819%	[3] 0.564103%	[4] 1.306167%	[6] 1.752075%	[2] 0.482833%	[1] **0.450001%**
Avg. Ranking	[4.67]	[3.00]	[5.00]	[3.25]	[3.42]	[1.67]
Median	1.306%	0.640%	1.532%	1.207%	0.878%	0.592%

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models.

TABLE VI. ANALYSIS RESULTS OF RMSE PERFORMANCE METRICS FOR COVID-19 CONFIRMED CASES ACROSS TWELVE COUNTRIES

Country	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[2] 0.004424	[4] 0.008354	[6] 0.015735	[1] **0.001864**	[5] 0.010148	[3] 0.007133
India	[6] 0.004335	[3] 0.003575	[5] 0.003894	[2] 0.002783	[1] **0.002112**	[4] 0.003780
Brazil	[5] 0.008499	[2] 0.002280	[4] 0.005525	[6] 0.010168	[3] 0.004581	[1] **0.001665**
France	[4] 0.035989	[2] 0.021375	[6] 0.036731	[3] 0.027494	[5] 0.036429	[1] **0.015350**
Germany	[5] 0.039105	[2] 0.023398	[6] 0.039550	[3] 0.026441	[4] 0.031702	[1] **0.017866**
UK	[2] 0.003656	[4] 0.005250	[6] 0.015484	[1] **0.001978**	[5] 0.009525	[3] 0.004472
Russia	[2] 0.003764	[5] 0.009854	[6] 0.018624	[1] **0.002953**	[4] 0.008916	[3] 0.008283
Italy	[6] 0.037787	[1] **0.014468**	[5] 0.029706	[3] 0.024618	[4] 0.028783	[2] 0.017685
Turkey	[5] 0.011119	[2] 0.007022	[6] 0.017426	[3] 0.007553	[4] 0.008216	[1] **0.003545**
Spain	[6] 0.016057	[3] 0.011521	[5] 0.013918	[4] 0.013410	[2] 0.009497	[1] **0.006811**
Vietnam	[5] 0.023661	[3] 0.018179	[6] 0.028106	[2] 0.016426	[4] 0.018929	[1] **0.013277**
Argentina	[5] 0.015487	[3] 0.008682	[4] 0.012727	[6] 0.017869	[1] **0.006048**	[2] 0.007632
Avg. Ranking	[4.42]	[2.83]	[5.42]	[2.92]	[3.50]	[1.92]
Median	0.013	0.009	0.017	0.012	0.010	0.007

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models.

When comparing LSTM-based and GRU-based models, the proposed GRU-based models exhibit a lower average ranking than their LSTM counterparts. The top three models in terms of average MAE ranking are ACF-PACF-GRU (2.00), PACF-LSTM (3.08), and ACF-GRU (3.83). The MAE performance results indicate that incorporating the PACF series into LSTM models significantly enhances forecasting accuracy compared to the standard LSTM. Similarly, integrating the ACF series into GRU models improves predictive performance over standard GRU. According to the R-squared rankings in Table IV, the standard LSTM and GRU models consistently rank between 7^{th} and 8^{th} across twelve countries. Among the proposed LSTM-based models, the ACF-LSTM model generally ranks between 2^{nd} and 5^{th}, except for Italy, which ranks 1^{t}. The PACF-LSTM model falls between 4^{th} and 5^{th}, with exceptions in the USA, UK, and Russia, where it ranks 2^{nd}. The ACF-PACF-LSTM model typically ranks between 4^{th} and 6^{th}, except for Spain, which is ranked 3^{rd}.

For the GRU-based models at the country level, the ACF-GRU model generally ranks between 1st and 3rd, except in Spain (5^{th}), Brazil (6^{th}), and Argentina (6^{th}). The PACF-GRU model typically ranks between 3^{rd} and 5^{th}, with exceptions in India (1^{st}), Argentina (1^{st}), and Spain (2^{nd}). The ACF-PACF-GRU model, demonstrating superior performance, primarily ranks between 1^{st} and 3^{rd}, except in India and Russia, where it ranks 4^{th}. This model achieves the highest forecasting accuracy in six countries: Brazil, France, Germany, Turkey, Spain, and Vietnam.

Among all models, the proposed GRU-based models exhibit a lower average R-squared ranking than their LSTM counterparts. The top three models in terms of average ranking are ACF-PACF-GRU (2.00), ACF-GRU (2.92), and PACF-LSTM (3.17). The R-squared values indicate that incorporating ACF and PACF series into both LSTM and GRU models significantly enhances forecasting accuracy compared to using only the standard LSTM or GRU models. Therefore, this study demonstrated the MAE and R-squared rankings for ACF-PACF-GRU consistently outperformed other models, showing significant improvement over traditional LSTM and GRU models as used by Zhou et al. in forecasting cumulative confirmed cases across twelve countries.

To summarize the overall performance of the eight models across twelve countries, Fig. 8 presents the median values and boxplots for each model based on different performance metrics. The ACF-PACF-GRU model has the lowest median MAE of 0.004, indicating superior accuracy. In terms of R-squared, the PACF-GRU model achieves the highest median value of 0.9989, followed closely by the ACF-PACF-GRU model at 0.9987. Zhou et al. (2023) highlighted the efficacy of deep learning models like LSTM in predicting COVID-19 cases, though their model did not incorporate ACF/PACF features. Our results align with their conclusion that deep learning models are effective, but we extend their work by showing that incorporating ACF and PACF series as input features significantly improves forecasting accuracy. In particular, the ACF-PACF-GRU model, which combines both ACF and PACF, demonstrated the best performance, significantly lowering the MAE compared to the models used by Zhou et al. (2023), which did not utilize these time-series dependencies.

The MAE and R-squared results indicate that incorporating ACF and/or PACF as additional input series in deep learning models enhances the forecasting accuracy of time series data. To determine which proposed model provides the most accurate COVID-19 case predictions, we analyzed the rankings of MAPE and RMSE across six proposed models for each country, as shown in Tables V and VI.

Overall, the top three models based on average rankings for MAPE and RMSE are ACF-PACF-GRU (with 1.67, and 1.92, respectively), PACF-LSTM (with 3.00, and 2.83), and ACF-GRU (with 3.25, and 2.92). As shown in Table V, the ACF-PACF-GRU model consistently outperforms other models in MAPE across seven countries: Brazil, France, Germany, Turkey, Spain, Vietnam, and Argentina. Similarly, Table VI highlights that this model achieves the lowest RMSE in six countries: Brazil, France, Germany, Turkey, Spain, and Viet-

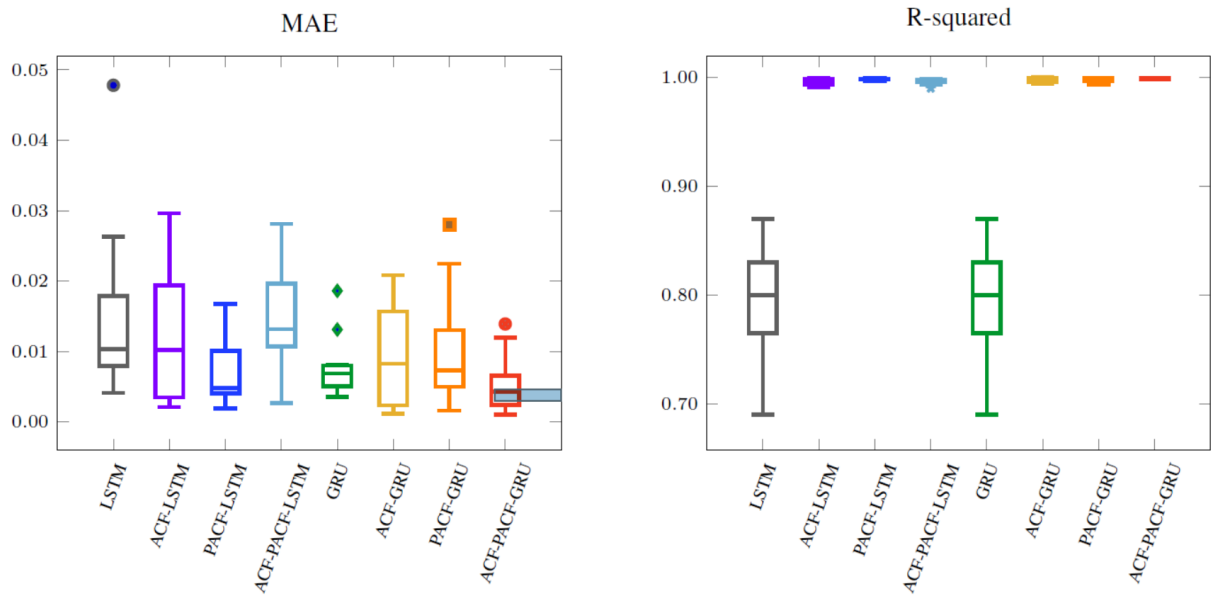


Fig. 8. Analysis results of MAE and R-squared for confirmed cases, compared eight proposed models with the standard deep learning models.

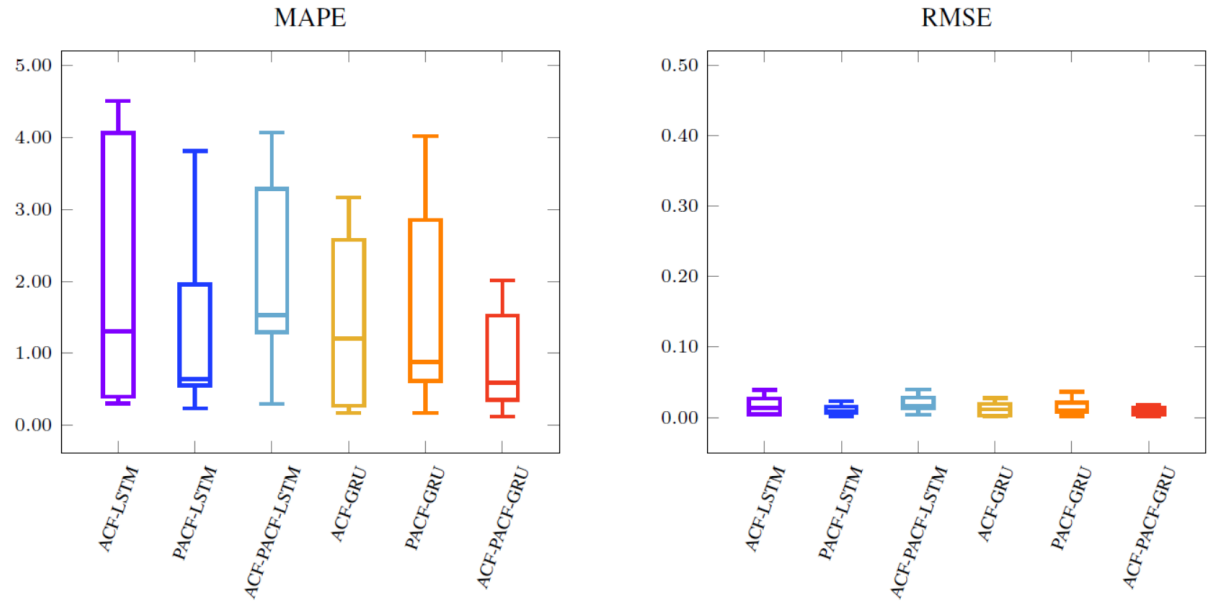


Fig. 9. Analysis results of MAPE and RMSE for confirmed cases compared among six proposed models.

nam. At the country level, the ACF-GRU model ranks within the top three positions more frequently than the PACF-GRU model in both MAPE and RMSE.

These findings suggest that for LSTM-based models, incorporating only the PACF series yields better forecasting performance than adding either the ACF series alone or both series together. Conversely, for GRU-based models, integrating both ACF and PACF series leads to superior forecasting accuracy compared to including only one of the series. Among all models, the ACF-PACF-GRU model demonstrates the highest potential for forecasting COVID-19 confirmed cases.

To provide a comprehensive overview of the results across all twelve countries, Fig. 9 presents the median values and boxplots for the six proposed models, categorized by performance metrics. The median values of MAPE and RMSE in Fig. 9 confirm that the ACF-PACF-GRU model demonstrates the highest predictive accuracy, with the lowest median MAPE at 0.592% and RMSE at 0.007. These findings highlight the effectiveness of the ACF-PACF-GRU model, indicating that the proposed approach significantly improves the forecasting performance for confirmed cases.

B. Cumulative Death Cases

Using the same time-series data for COVID-19 death cases, the forecasting accuracy of the six proposed models is compared to the standard LSTM and GRU models. The

TABLE VII. ANALYSIS RESULTS OF MAE PERFORMANCE METRICS FOR COVID-19 DEATH CASES ACROSS TWELVE COUNTRIES

Country	LSTM*	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	GRU*	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[8] 0.019620	[2] 0.000922	[5] 0.001147	[6] 0.004033	[7] 0.006790	[1] **0.000889**	[4] 0.000991	[3] 0.000978
India	[8] 0.005300	[6] 0.001803	[5] 0.001533	[3] 0.001247	[7] 0.004280	[4] 0.001366	[2] 0.000976	[1] **0.000791**
Brazil	[8] 0.008280	[6] 0.001946	[1] **0.000451**	[4] 0.000625	[7] 0.003630	[5] 0.001575	[2] 0.000471	[3] 0.000509
France	[8] 0.015620	[6] 0.006603	[1] **0.002567**	[4] 0.004213	[7] 0.012180	[5] 0.005486	[3] 0.003727	[2] 0.002902
Germany	[8] 0.044450	[5] 0.005379	[2] 0.002347	[4] 0.004396	[6] 0.006230	[7] 0.007754	[3] 0.004262	[1] **0.002249**
UK	[8] 0.007910	[5] 0.001435	[3] 0.000398	[4] 0.000698	[7] 0.006230	[6] 0.001858	[1] **0.000326**	[2] 0.000330
Russia	[8] 0.005770	[3] 0.000696	[4] 0.000714	[6] 0.002520	[7] 0.003260	[1] **0.000523**	[5] 0.000791	[2] 0.000597
Italy	[8] 0.010230	[6] 0.004871	[1] **0.000281**	[5] 0.004385	[7] 0.007440	[4] 0.003320	[2] 0.001520	[3] 0.001878
Turkey	[8] 0.008930	[5] 0.002190	[3] 0.001275	[6] 0.005586	[7] 0.006510	[4] 0.001922	[2] 0.001088	[1] **0.000755**
Spain	[8] 0.014230	[6] 0.004767	[1] **0.002161**	[4] 0.004239	[7] 0.007530	[5] 0.004738	[3] 0.003460	[2] 0.002798
Vietnam	[7] 0.024460	[3] 0.010312	[1] **0.005568**	[8] 0.027172	[5] 0.017300	[2] 0.008209	[4] 0.014931	[6] 0.021237
Argentina	[7] 0.003810	[6] 0.002313	[4] 0.000842	[3] 0.000650	[8] 0.004840	[5] 0.001997	[2] 0.000505	[1] **0.000457**
Avg. Ranking	[7.83]	[4.92]	[2.58]	[7.45]	[6.83]	[4.08]	[2.75]	[2.25]
Median	0.0096	0.0023	0.0012	0.0041	0.0064	0.0020	0.0010	0.00009

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models and standard deep learning models (LSTM*,GRU*) cited from the previous study of Luyu Zhou et al.

TABLE VIII. ANALYSIS RESULTS OF R-SQUARED PERFORMANCE METRICS FOR COVID-19 DEATH CASES ACROSS TWELVE COUNTRIES

Country	LSTM*	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	GRU*	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[8] 0.7347	[2] 0.999868	[5] 0.999795	[6] 0.998705	[7] 0.7626	[1] **0.999880**	[4] 0.999833	[3] 0.999846
India	[8] 0.6417	[6] 0.997680	[5] 0.998281	[4] 0.998810	[7] 0.7533	[3] 0.998986	[2] 0.999065	[1] **0.999315**
Brazil	[8] 0.7254	[6] 0.996445	[1] **0.999656**	[4] 0.999443	[7] 0.7719	[5] 0.997545	[2] 0.999608	[3] 0.999560
France	[7] 0.7719	[6] 0.992627	[1] **0.998876**	[4] 0.997605	[8] 0.6696	[5] 0.995249	[3] 0.998167	[2] 0.998824
Germany	[8] 0.6696	[3] 0.998863	[2] 0.999333	[4] 0.998483	[7] 0.7719	[6] 0.996506	[5] 0.998353	[1] **0.999481**
UK	[7] 0.7719	[5] 0.999244	[3] 0.999943	[4] 0.999795	[8] 0.6696	[6] 0.998779	[1] **0.999957**	[2] 0.999944
Russia	[7] 0.8091	[5] 0.999951	[1] **0.999978**	[6] 0.999880	[8] 0.7347	[4] 0.999972	[2] 0.999970	[3] 0.999974
Italy	[7] 0.7626	[6] 0.997360	[1] **0.999979**	[5] 0.997804	[8] 0.7347	[4] 0.998670	[2] 0.999731	[3] 0.999559
Turkey	[8] 0.6696	[3] 0.999770	[5] 0.999711	[6] 0.998569	[7] 0.7719	[4] 0.999717	[2] 0.999793	[1] **0.999895**
Spain	[7] 0.7719	[6] 0.995523	[2] 0.998576	[4] 0.997383	[8] 0.7254	[5] 0.996040	[3] 0.997925	[1] **0.998809**
Vietnam	[7] 0.7533	[5] 0.998189	[2] 0.999137	[6] 0.993226	[8] 0.6417	[1] **0.999473**	[3] 0.999031	[4] 0.998211
Argentina	[8] 0.7347	[6] 0.994614	[4] 0.999266	[3] 0.999416	[7] 0.8091	[5] 0.997782	[1] **0.999693**	[2] 0.999584
Avg. Ranking	[7.50]	[4.92]	[2.67]	[4.67]	[7.50]	[4.08]	[2.50]	[2.17]
Median	0.7440	0.9979	0.9995	0.9986	0.7440	0.9987	0.9997	0.9996

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models and standard deep learning models (LSTM*,GRU*) cited from the previous study of Luyu Zhou et al.

performance measures of eight models, shown in Tables VII and VIII, display the actual and forecasted death cases across twelve countries using deep learning models based on LSTM and GRU. The rankings of MAE and R-squared values for each country are presented for the six proposed models and the standard models (LSTM and GRU) in Tables VII and VIII, respectively.

In Table VII, the MAE ranking for the standard LSTM and GRU models among twelve countries generally falls between 7^{th} and 8^{th} place, similar to the results observed in Zhou et al.'s (2023) study. However, our study shows a clear advantage of the proposed GRU-based models over their LSTM counterparts.

For the standard GRU model, the rankings are mostly between 7^{th} and 8^{th}, except for Vietnam (5^{th}) and Germany (6^{th}). For the proposed LSTM-based models, the ACF-LSTM model typically ranks between 5^{th} and 6^{th}, with exceptions in the USA (2^{nd}), Russia (3^{rd}), and Vietnam (3^{rd}). The PACF-LSTM model ranks between 1^{st} and 3^{rd}, except for Argentina (4^{th}), Russia (4^{th}), India (5^{th}), and the USA (5^{th}). The ACF-PACF-LSTM model mostly ranks between 3^{rd} and 6^{th}, with the exception of Vietnam (8^{th}).

For the proposed GRU-based models at the country level, the ACF-GRU model generally ranks between 4^{th} and 7^{th}, except for Russia (1^{st}), the USA (1^{st}), and Vietnam (2^{nd}). The PACF-GRU model ranks between 2^{nd} and 5^{th}, with an exception in the UK (1^{st}). The ACF-PACF-GRU model ranks 1^{st} to 3^{rd} in MAE for most countries, including India, Germany, Turkey, and Argentina. In contrast, the standard LSTM and GRU models performed at lower rankings for these countries. This improvement is particularly evident when considering the MAE and R-squared values for death cases in countries like

TABLE IX. ANALYSIS RESULTS OF MAPE PERFORMANCE METRICS FOR COVID-19 DEATH CASES ACROSS TWELVE COUNTRIES

Country	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[2] 0.108838%	[5] 0.137860%	[6] 0.448506%	[1] **0.104958%**	[4] 0.118071%	[3] 0.116503%
India	[6] 0.194973%	[5] 0.167504%	[3] 0.134351%	[4] 0.147819%	[2] 0.104303%	[1] **0.084415%**
Brazil	[6] 0.206977%	[1] **0.047485%**	[4] 0.066193%	[5] 0.165993%	[2] 0.049643%	[3] 0.053557%
France	[6] 0.761604%	[1] **0.291291%**	[4] 0.454209%	[5] 0.622381%	[3] 0.408169%	[2] 0.315537%
Germany	[5] 0.628394%	[2] 0.296073%	[3] 0.486435%	[6] 0.888655%	[4] 0.512760%	[1] **0.268329%**
UK	[5] 0.158320%	[3] 0.044199%	[4] 0.073488%	[6] 0.211555%	[1] **0.036055%**	[2] 0.037579%
Russia	[3] 0.084860%	[4] 0.092779%	[6] 0.283885%	[1] **0.063388%**	[5] 0.100405%	[2] 0.074391%
Italy	[6] 0.525774%	[1] **0.032344%**	[5] 0.471992%	[4] 0.366857%	[2] 0.168702%	[3] 0.210783%
Turkey	[5] 0.232556%	[3] 0.141059%	[6] 0.587429%	[4] 0.214168%	[2] 0.118658%	[1] **0.085518%**
Spain	[6] 0.559273%	[1] **0.249926%**	[4] 0.460365%	[5] 0.535909%	[3] 0.387958%	[2] 0.308229%
Vietnam	[3] 1.157108%	[1] **0.691349%**	[6] 2.957375%	[2] 0.946599%	[4] 1.699260%	[5] 2.379124%
Argentina	[6] 0.245255%	[4] 0.089956%	[3] 0.068889%	[5] 0.207647%	[2] 0.053661%	[1] **0.048188%**
Avg. Ranking	[4.92]	[2.58]	[4.50]	[4.00]	[2.83]	[2.17]
Median	0.239%	0.139%	0.451%	0.213%	0.118%	0.101%

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models.

TABLE X. ANALYSIS RESULTS OF RMSE PERFORMANCE METRICS FOR COVID-19 DEATH CASES ACROSS TWELVE COUNTRIES

Country	ACF-LSTM	PACF-LSTM	ACF-PACF-LSTM	ACF-GRU	PACF-GRU	ACF-PACF-GRU
USA	[2] 0.001234	[5] 0.001526	[6] 0.004537	[1] **0.001177**	[4] 0.001384	[3] 0.001328
India	[6] 0.002388	[5] 0.002115	[4] 0.001708	[3] 0.001626	[2] 0.001515	[1] **0.001312**
Brazil	[6] 0.002159	[1] **0.000653**	[4] 0.000831	[5] 0.001750	[3] 0.000743	[2] 0.000734
France	[6] 0.007181	[1] **0.002826**	[4] 0.005207	[5] 0.006047	[3] 0.004367	[2] 0.003495
Germany	[5] 0.006297	[2] 0.002971	[4] 0.005446	[6] 0.009024	[3] 0.005061	[1] **0.002624**
UK	[5] 0.001620	[3] 0.000460	[4] 0.000947	[6] 0.002270	[1] **0.000385**	[2] 0.000441
Russia	[4] 0.000958	[3] 0.000890	[6] 0.002853	[1] **0.000733**	[5] 0.001071	[2] 0.000764
Italy	[6] 0.006214	[1] **0.000435**	[5] 0.005660	[4] 0.003911	[2] 0.001755	[3] 0.002111
Turkey	[5] 0.002708	[3] 0.002088	[6] 0.006989	[4] 0.002211	[2] 0.001767	[1] **0.001222**
Spain	[5] 0.005452	[1] **0.003074**	[4] 0.005129	[6] 0.005542	[3] 0.004095	[2] 0.003252
Vietnam	[3] 0.010946	[1] **0.006718**	[6] 0.029411	[2] 0.008528	[4] 0.015452	[5] 0.022089
Argentina	[6] 0.003026	[4] 0.001157	[3] 0.001002	[5] 0.002087	[1] **0.000726**	[2] 0.000909
Avg. Ranking	[4.92]	[2.50]	[4.67]	[4.00]	[2.75]	[2.17]
Median	0.003	0.002	0.005	0.002	0.002	0.001

Note: Bold numbers indicate the best results. Superscript numbers in brackets indicate ranking among six proposed models.

Brazil and India.

When comparing the LSTM-based and GRU-based models, the proposed GRU models generally have a lower average ranking than the LSTM models. Among the eight models, the top three in terms of average MAE ranking are ACF-PACF-GRU (2.25), PACF-LSTM (2.58), and PACF-GRU (2.75). The MAE performance results show that incorporating the PACF series into LSTM models enhances forecasting performance more effectively than using the standard LSTM alone. However, integrating both ACF and PACF series into GRU models improves forecasting accuracy compared to using the standard GRU.

As shown in Table VIII, the R-squared ranking for the standard LSTM and GRU models across twelve countries generally falls between 7^{th} and 8^{th}. For the proposed LSTM-based models, the ACF-LSTM model typically ranks between 5^{th} and 6^{th}, with exceptions in the USA (2^{nd}), Germany (3^{rd}), and Turkey (3^{rd}). The PACF-LSTM model ranks between 1^{st} and 3^{rd}, except for Argentina (4^{th}), the USA (5^{th}), India (2^{nd}), and Turkey (5^{th}). The ACF-PACF-LSTM model usually ranks between 4^{th} and 6^{th}, with the exception of Argentina (3^{rd}).

For the proposed GRU-based models at the country level, the ACF-GRU model generally ranks between 3^{rd} and 6^{th}, except for the USA and Vietnam, where it ranks 1^{st}. The PACF-GRU model typically ranks between 1^{st} and 3^{rd}, except for the USA (4^{th}) and Germany (5^{th}). The ACF-PACF-GRU model, which achieves superior results, ranks between 1^{st} and 3^{rd} in most countries, except for Vietnam (4^{th}). This model demonstrates the highest forecasting accuracy in six countries: India, Germany, Turkey, and Spain.

Among the various models, the proposed GRU-based models tend to have a lower average R-squared ranking compared

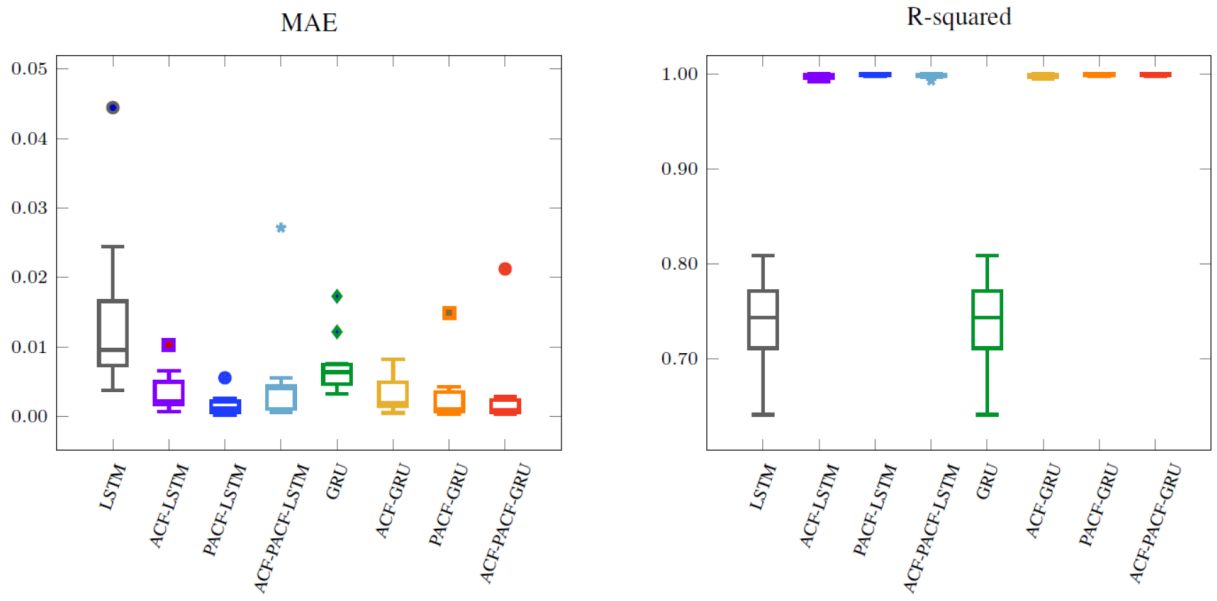


Fig. 10. Analysis results of MAE and R-squared for death cases, compared eight proposed models with the standard deep learning models.

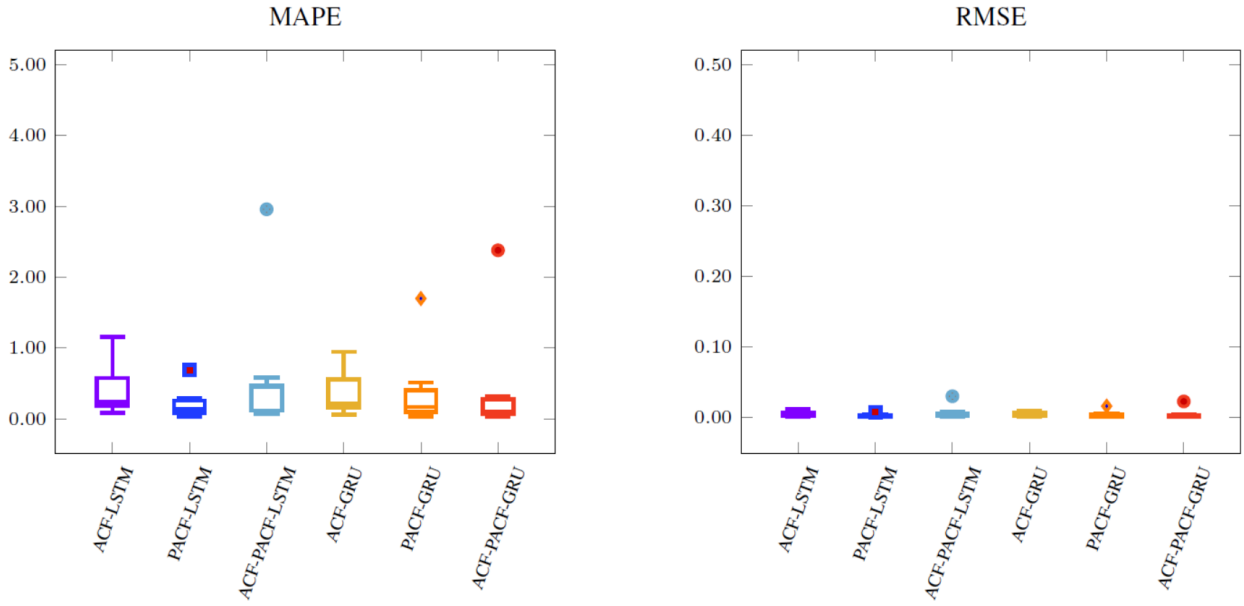


Fig. 11. Analysis results of MAPE and RMSE for death cases compared among six proposed models.

to the LSTM-based models. The top three models based on average R-squared ranking are ACF-PACF-GRU (2.17), PACF-GRU (2.50), and PACF-LSTM (2.67). The R-squared values indicate that incorporating both ACF and PACF series into GRU models can significantly enhance forecasting accuracy compared to models based solely on LSTM or GRU, similar to the improvement observed when adding the PACF series to either LSTM or GRU.

To provide a comprehensive overview of the results for all twelve countries, Fig. 10 presents the median values and boxplots for eight models, organized by performance metrics. The median values in Fig. 10 indicate that the ACF-PACF-GRU model for death cases has the highest predictive ability, with the lowest MAE value of 0.0009 and the highest R-squared value of 0.9996. However, the PACF-GRU model has a slightly higher median R-squared value of 0.9997.

Next, we examine which proposed model demonstrates the greatest forecasting accuracy for COVID-19 death cases. The MAPE and RMSE values for the six proposed models were ranked for each country, as shown in Table IX and Table X. Overall, the top 3 models based on average rankings for MAPE and RMSE are ACF-PACF-GRU (2.17, and 2.17, respectively), PACF-LSTM (2.58, and 2.50), and ACF-GRU (2.83, and 2.75). In Table IX, the ACF-PACF-GRU model, with an average MAPE ranking of 2.17, consistently outperforms the other models in four countries, including India, Germany, Turkey,

and Argentina. Similarly, in Table X, the ACF-PACF-GRU model, with an average RMSE ranking of 2.17, is the best model in six countries, including India, Germany, and Turkey.

To provide a comprehensive performance comparison across all twelve countries, Fig. 11 illustrates the median values and corresponding boxplots for the six proposed models, evaluated using key performance metrics. The ACF-PACF-GRU model consistently outperformed other models, achieving the lowest median MAPE of 0.101% and RMSE of 0.001. These results highlight the superior forecasting capability of the ACF-PACF-GRU model, demonstrating the effectiveness of integrating autocorrelation and partial autocorrelation features in enhancing prediction accuracy for COVID-19 death cases.

Therefore, the forecasting performance of the LSTM model, when only the PACF series is added, is better than when either the ACF series or both series are incorporated into the same algorithm. On the other hand, adding only the ACF series or both series into the GRU model leads to better forecasting accuracy compared to adding only the PACF series. However, the MAPE and RMSE results for death cases confirm that the ACF-PACF-GRU model has the highest potential for accurately forecasting COVID-19 cases.

V. Discussions

This study introduces a novel approach to improve the forecasting accuracy of COVID-19 confirmed and death cases by integrating ACF and PACF series into LSTM and GRU models. The key motivation for choosing this approach is to address the limitations of existing models, which primarily rely on standard LSTM and GRU networks. While LSTM and GRU models are adept at capturing long-term dependencies in time-series data, they fall short in modeling non-linear relationships and temporal dependencies [20], particularly in the context of irregular and complex epidemic data like COVID-19. The incorporation of ACF and PACF enhances the model's ability to capture both short-term and long-term dependencies, providing a more accurate representation of the underlying data dynamics, which is essential for infectious disease forecasting.

This study demonstrates that our proposed models, especially the ACF-PACF-GRU model, outperform standard LSTM and GRU models in forecasting accuracy, as evidenced by the improved performance across key metrics such as MAE, R-squared, MAPE, and RMSE. The incorporation of ACF and PACF features provides additional insights into the temporal dependencies of the data, which were absent in previous studies, including the work by Zhou et al. (2023). This approach not only improves the predictive accuracy for COVID-19 but is also applicable to other epidemic forecasting tasks where similar temporal dependencies exist.

A. Implications of Results

The findings of this study have significant implications for the forecasting of infectious disease trends. The unique structure of LSTM and GRU models, when combined with ACF and PACF series, enhances the ability of the models to effectively capture both linear and non-linear temporal dependencies present in epidemic data. The median MAPE values for the ACF-PACF-GRU model, which are 0.592% for confirmed cases and 0.101% for death cases, underscore the efficacy of the proposed method. These values highlight how the integration of ACF and PACF into deep learning models can significantly improve the forecasting performance over existing models. Furthermore, the improved accuracy achieved by our models supports previous research emphasizing the importance of time-series dependencies in epidemic forecasting. Our proposed methodology, which extends the work of Zhou et al. (2023) by incorporating ACF and PACF features, offers a more robust framework for predicting COVID-19 trends and can be adapted to other infectious diseases in future research.

B. Practical Implications

The proposed deep learning models provide valuable insights for forecasting the cumulative number of confirmed and death cases of COVID-19 by incorporating ACF and PACF series into LSTM and GRU models. By treating the forecasting task as a time-series problem, leveraging historical data to predict future trends, the models can provide more accurate predictions, thereby aiding public health decision-making. The integration of ACF and PACF into GRU models, in particular, improves model performance, making them more suitable for real-time epidemic prediction. The enhanced accuracy of the ACF-PACF-GRU model, when compared to the methods used by Zhou et al. (2023), demonstrates the potential for further improving deep learning-based forecasting models with additional time-series features. These forecasts are crucial for governments and public health officials, enabling them to make informed decisions about epidemic control measures, resource allocation, and intervention strategies.

C. Limitations

While LSTM and GRU models excel at capturing long-term dependencies, they do not inherently account for the temporal dependencies critical to epidemic forecasting. The addition of ACF and PACF significantly enhances the models' ability to capture these temporal patterns, which are often key to understanding disease transmission. However, this study focused solely on time-series data of the disease, excluding other variables such as vaccination rates and country-specific interventions, which could also influence epidemic outcomes. Future research could incorporate these additional features to provide a more comprehensive forecasting model. Additionally, the complex architecture of deep learning models requires careful hyperparameter selection, which may vary depending on the dataset and problem. Thus, the proposed models in this study, while more accurate than existing models, still require optimization for different contexts and datasets.

VI. Conclusions

In this study, we introduced a novel deep learning approach that incorporates ACF and PACF to improve time-series forecasting models for predicting COVID-19 confirmed and death cases. By leveraging the ACF and PACF series in conjunction with original time-series data, our models were able to capture both short-term and long-term temporal dependencies, leading to improved forecasting accuracy. The ACF-PACF-GRU model, in particular, demonstrated superior performance across a range of metrics compared to both standard LSTM and GRU models. These findings suggest that incorporating time-series dependencies, such as ACF and PACF, significantly

enhances the performance of deep learning models in epidemic forecasting. Future research should explore the application of this methodology to other infectious diseases, such as influenza and dengue fever, to further validate the robustness and generalizability of the approach.

ACKNOWLEDGMENT

This research was part of the graduation requirement for the philosopher degree at College of Public Health Sciences, Chulalongkorn University, which was supported by multidisciplinary researchers. The authors extend their heartfelt thanks to the anonymous reviewers for their insightful and constructive feedback. The authors acknowledge the assistance of Gemini in refining the grammar and clarity of this manuscript. However, the scientific content and conclusions are entirely the authors' responsibility.

1) Credit authorship contribution statement:

a) Warapree tangseefa: Conceptualization, Data curation, Methodology, Software, Formal analysis, Validation, Writing – original draft, Writing – review & editing.

b) Tepanata pumpaibool: Conceptualization, Review & editing.

c) Paisit khanarsa: Conceptualization, Software, Formal analysis, Visualization, Writing – review & editing.

d) Krung sinapiromsara: Conceptualization, Methodology, Writing – review & editing, Funding acquisition, Investigation, Supervision.

2) Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

REFERENCES

[1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, "The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2," *Nature microbiology*, vol. 5, no. 4, pp. 536–544, 2020.

[2] B. Vellingiri, K. Jayaramayya, M. Iyer, A. Narayanasamy, V. Govindasamy, B. Giridharan, S. Ganesan, A. Venugopal, D. Venkatesan, H. Ganesan *et al.*, "Covid-19: A promising cure for the global panic," *Science of the total environment*, vol. 725, p. 138277, 2020.

[3] T. M. Abd El-Aziz and J. D. Stockand, "Recent progress and challenges in drug development against covid-19 coronavirus (sars-cov-2)-an update on the status," *Infection, Genetics and evolution*, vol. 83, p. 104327, 2020.

[4] D. Cucinotta and M. Vanelli, "Who declares covid-19 a pandemic," *Acta bio medica: Atenei parmensis*, vol. 91, no. 1, p. 157, 2020.

[5] World Health Organization, "Covid-19 epidemiological update, edition 174, 24 december 2024," 2024, Accessed: March. 12, 2025. [Online]. Available: https://iris.who.int/handle/10665/379898

[6] H. Abbasimehr and R. Paki, "Prediction of covid-19 confirmed cases combining deep learning methods and bayesian optimization," *Chaos, Solitons & Fractals*, vol. 142, p. 110511, 2021.

[7] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang, "Presumed asymptomatic carrier transmission of covid-19," *Jama*, vol. 323, no. 14, pp. 1406–1407, 2020.

[8] Z. Zuo, M. Wang, H. Cui, Y. Wang, J. Wu, J. Qi, K. Pan, D. Sui, P. Liu, and A. Xu, "Spatiotemporal characteristics and the epidemiology of tuberculosis in china from 2004 to 2017 by the nationwide surveillance system," *BMC Public Health*, vol. 20, pp. 1–12, 2020.

[9] V. Rahmanian, S. Bokaie, K. Rahmanian, S. Hosseini, and A. T. Firouzeh, "Analysis of temporal trends of human brucellosis between 2013 and 2018 in yazd province, iran to predict future trends in incidence: A time-series study using arima model," *Asian Pacific Journal of Tropical Medicine*, vol. 13, no. 6, pp. 272–277, 2020.

[10] K. Shashvat, R. Basu, and A. P. Bhondekar, "Application of time series methods for dengue cases in north india (chandigarh)," *Journal of Public Health*, vol. 29, pp. 433–441, 2021.

[11] Y. Chen, K. Leng, Y. Lu, L. Wen, Y. Qi, W. Gao, H. Chen, L. Bai, X. An, B. Sun *et al.*, "Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of shenyang, china, 2010–2018," *Epidemiology & Infection*, vol. 148, p. e29, 2020.

[12] W. Hu, S. Tong, K. Mengersen, and D. Connell, "Weather variability and the incidence of cryptosporidiosis: comparison of time series poisson regression and sarima models," *Annals of epidemiology*, vol. 17, no. 9, pp. 679–688, 2007.

[13] L.-P. Yang, S.-Y. Liang, X.-J. Wang, X.-J. Li, Y.-L. Wu, and W. Ma, "Burden of disease measured by disability-adjusted life years and a disease forecasting time series model of scrub typhus in laiwu, china," *PLoS neglected tropical diseases*, vol. 9, no. 1, p. e3420, 2015. [Online]. Available: https://doi.org/10.1371/journal.pntd.0003420

[14] M. S. Park, K. H. Park, and G. J. Bahk, "Combined influence of multiple climatic factors on the incidence of bacterial foodborne diseases," *Science of the total environment*, vol. 610, pp. 10–16, 2018.

[15] X. Zhang, Y. Liu, M. Yang, T. Zhang, A. A. Young, and X. Li, "Comparative study of four time series methods in forecasting typhoid fever incidence in china," *PloS one*, vol. 8, no. 5, p. e63116, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0063116

[16] I. CHOUJA, S. SAOUD, and M. SADIK, "New sarima approach model to forecast covid-19 propagation: case of morocco," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021.

[17] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert systems with applications*, vol. 140, p. 112896, 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2019.112896

[18] J. Sun, "Forecasting covid-19 pandemic in alberta, canada using modified arima models," *Computer methods and programs in biomedicine update*, vol. 1, p. 100029, 2021. [Online]. Available: https://doi.org/10.1016/j.cmpbup.2021.100029

[19] M. Zhai, W. Li, P. Tie, X. Wang, T. Xie, H. Ren, Z. Zhang, W. Song, D. Quan, M. Li *et al.*, "Research on the predictive effect of a combined model of arima and neural networks on human brucellosis in shanxi province, china: a time series predictive analysis," *BMC Infectious Diseases*, vol. 21, pp. 1–12, 2021.

[20] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200209, 2021. [Online]. Available: https://doi.org/10.1098/rsta.2020.0209

[21] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting methods and applications*. USA: John wiley & sons, 2008.

[22] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern recognition letters*, vol. 42, pp. 11–24, 2014.

[23] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[24] N. Dharani and P. Bojja, "Analysis and prediction of covid-19 by using recurrent lstm neural network model in machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022.

[25] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm," *Chaos, Solitons & Fractals*, vol. 140, p. 110212, 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.110212

[26] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms," *Results in Physics*, vol. 27, p. 104462, 2021. [Online]. Available: https://doi.org/10.1016/j.rinp.2021.104462

[27] K. ArunKumar, D. V. Kalaga, C. M. S. Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (arima) and seasonal auto-regressive integrated moving average (sarima)," *Applied soft computing*, vol. 103, p. 107161, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2021.107161

[28] H. Verma, S. Mandal, and A. Gupta, "Temporal deep learning architecture for prediction of covid-19 cases in india," *Expert Systems with Applications*, vol. 195, p. 116611, 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.116611

[29] N. Ayoobi, D. Sharifrazi, R. Alizadehsani, A. Shoeibi, J. M. Gorriz, H. Moosaei, A. Khosravi, S. Nahavandi, A. G. Chofreh, F. A. Goni *et al.*, "Time series forecasting of new cases and new deaths rate for covid-19 using deep learning methods," *Results in physics*, vol. 27, p. 104495, 2021. [Online]. Available: https://doi.org/10.1016/j.rinp.2021.104495

[30] S. Mehrmolaei, M. Savargiv, and M. R. Keyvanpour, "Hybrid learning-oriented approaches for predicting covid-19 time series data: A comparative analytical study," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106754, 2023. [Online]. Available: https://doi.org/10.1016/j.engappai.2023.106754

[31] F. Kamalov, F. Thabtah, and I. Gurrib, "Autocorrelation for time series with linear trend," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. IEEE, 2021, pp. 181–185.

[32] L. Zhou, C. Zhao, N. Liu, X. Yao, and Z. Cheng, "Improved lstm-based deep learning model for covid-19 prediction using optimized approach," *Engineering applications of artificial intelligence*, vol. 122, p. 106157, 2023. [Online]. Available: https://doi.org/10.1016/j.engappai.2023.106157

[33] A. Ntakaris, G. Mirone, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Feature engineering for mid-price prediction with deep learning," *Ieee Access*, vol. 7, pp. 82 390–82 412, 2019.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013. [Online]. Available: https://arxiv.org/pdf/1308.0850

[36] 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," *arXiv preprint arXiv:1406.1078*, 2014. [Online]. Available: https://arxiv.org/pdf/1406.1078

[37] J. D. Cryer, *Time series analysis*. Boston: Duxbury Press, 1986.

[38] W. Yan, "Toward automatic time-series forecasting using neural networks," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 7, pp. 1028–1039, 2012.

[39] World Health Organization, "Daily frequency reporting of new covid-19 cases and deaths by date reported to who," Accessed: May. 22, 2024. [Online]. Available: https://srhdpeuwpubsa.blob.core.windows.net/whdh/COVID/WHO-COVID-19-global-daily-data.csv

[40] J. Brownlee, "How to use standardscaler and minmaxscaler transforms in python," 2020, Accessed: March. 12, 2025. [Online]. Available: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/

[41] F. Chollet, "Keras," 2015, Accessed: May. 22, 2024. [Online]. Available: https://keras.io // https://github.com/fchollet/keras

[42] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *International journal of forecasting*, vol. 8, no. 1, pp. 69–80, 1992. [Online]. Available: https://doi.org/10.1016/0169-2070(92)90008-W

[43] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[44] H. Zang, L. Cheng, T. Ding, K. W. Cheung, M. Wang, Z. Wei, and G. Sun, "Application of functional deep belief network for estimating daily global solar radiation: A case study in china," *Energy*, vol. 191, p. 116502, 2020.

[45] S. G. Gouda, Z. Hussein, S. Luo, and Q. Yuan, "Model selection for accurate daily global solar radiation prediction in china," *Journal of cleaner production*, vol. 221, pp. 132–144, 2019.

[46] J. Fan, L. Wu, F. Zhang, H. Cai, X. Ma, and H. Bai, "Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of china," *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 168–186, 2019.

[47] K. Bakirci, "Correlations for estimation of daily global solar radiation with hours of bright sunshine in turkey," *Energy*, vol. 34, no. 4, pp. 485–501, 2009.

Cyber Deception Across Domains: A Comprehensive Survey of Techniques, Challenges, and Perspectives

Amal Sayari, Slim Rekhis
University of Carthage, Higher School of Communication of Tunis (SUP'COM),
LR11TIC04, Communication Networks and Security Research Lab. & LR11TIC02,
Green and Smart Communication Systems Research Lab, Tunisia

***Abstract*—Cloud environments (CE), wireless networks (WN), cyber-physical systems (CPS), industrial control systems (ICS), smart grids (SG), internet of things (IoT), internet of vehicles (IOV), and unmanned aerial vehicles (UAV), are currently popular targets for cyberattacks due to their inherent limitations and vulnerabilities. Each domain has its own attack surfaces, weaknesses, and areas for implementing defense strategies appropriate to its specific conditions. Among the various defense mechanisms discussed in previous years, cyber deception has appeared as a very promising method. This approach allows the defenders to steer the attackers in the wrong direction, get threat intelligence, and at the same time, increase security by engaging with adversaries in deception environments in a proactive manner. Cyber deception has been a topic of investigation in several studies, where specific frameworks and techniques were proposed to identify, delay, or disrupt adversarial behavior. Nevertheless, the contributions of earlier works are frequently limited or missing a unified framework that makes a thorough and comparative study necessary. This survey investigates the cyber deception techniques used in various domains. The first part is about the cores of deception and its background. Next, it presents a summary of the available deception techniques with their modeling by different frameworks like MITRE ATT&CK, D3FEND, and Engage, and intelligent orchestration using reinforcement learning (RL) and game theory (GT). Then, it serves as a thorough systematic review of each selected paper, going over the system design, used deception techniques, evaluation metrics, and limitations on each scheme. The achieved results are compiled into a unified summary table to enable a quick and effective comparison across the domains. It concludes, therefore, by discussing the main challenges, open issues, and areas of research that have not yet been explored, thus making it a valuable source for future research on cyber deception.**

***Keywords*—Cyber defense; cyber deception; cloud environments; wireless networks; cyber-physical systems; industrial control systems; smart grids; internet of things; internet of vehicles; unmanned aerial vehicles**

I. Introduction

The growth of cyber-attacks in terms of frequency and sophistication has outstripped the ability of conventional security measures to be the sole protector of the complex and dynamic digital infrastructures. The adversaries are continually adapting their tactics, techniques, and procedures (TTPs) to evade traditional defense mechanisms like firewalls, intrusion detection/ prevention systems, and antivirus software. The reactive schemes act typically after the initiation of an attack, makes the need for different, proactive, and intelligent defense strategies even more urgent.

One promising strategy is cyber deception, which is currently widely known in both research and industry. Cyber deception is putting uncertainty and misdirection in the environment by intentionally misguiding adversaries; thus, their progression becomes slower or they get fictitious information. The method is based on failure or turning away the attacker, providing first warning through the engagement of the attacker, and collecting information about their capabilities. Deception methods, from honeypots and honeynets to honeytokens [1], are successful complementary tools that shift the advantage back to defenders.

At the same time, the wide digital landscape has extended to encompass a wide variety of interconnected and heterogeneous systems, including CE, WN, CPS, ICS, SG, IOT, IOV, and UAV. Apart from the fact that the aforementioned domains have achieved a previously unattainable level of scalability and interconnection, they are also the main sources of new vulnerabilities. Each domain, besides the attacks, also comes with its unique features such as different architectural structures, operational constraints, communication protocols, and security requirements that determine the nature and intensity of the cyber threats that it has to deal with.

Although the individual fields such as WN, CPS, or IOT are seeing a significant increase in the use of cyber deception techniques, the current literature is still fragmented. At present, there is no unified study that examines how deception has been adapted and applied across such diverse technological environments. In particular, research on honeypots and honeynets tends to remain siloed, without much cross-domain analysis and integration.

Previously conducted surveys on cyber deception [2] [3] [4] [5] [6] [7] have introduced foundational taxonomies, described deception techniques, and outlined conceptual models. However, those are mainly dedicated to technique-level analysis (e.g., GT, RL, or Honey-X strategies) and they are not systematic comparing across different operational domains. Although some works [4] [6] did provide specific domains (e.g., in IoT, ICS, CPS, and CE), the exploration is still fragmented, and it has not been focused enough on how architectural constraints, attacker models, and environmental characteristics influence the design and deployment of deception strategies.

Based on those limitations, we have prepared a thorough and comparative review of the cyber deception methods in several critical domains. In contrast to previous studies that either dwell upon particular techniques or theoretical models, our survey stresses the domain-aware analysis of deception

strategies, elaborating on customization to different system architectures and operational contexts. Here are our main contributions:

- We present the fundamental concepts, models, and strategies of cyber deception to provide a clear background for the survey.
- We meticulously extract and synthesize the deception technique, architecture, evaluation metrics, and limitations from each selected paper.
- We categorize the findings by domain and summarize them in standardized tables for cross-domain comparison.
- We identify common challenges, open issues, and trends, which could serve as a guide for further research in deception-based cyber defense in various contexts.

The remainder of this survey is structured as follows. Section II presents the adopted methodology. Section III addresses a comparison with existing surveys. Section IV provides the background and key concepts of cyber deception. Section V describes the intelligent orchestration of cyber deception. Sections VI to XIII present the exploration of cyber deception in CE, WN, CPS, ICS, SG, IoT, IoV, and UAV, respectively. Section XIV discusses open research issues and key insights. Finally, Section XV concludes the study.

II. Adopted Methodology

A systematic literature review was performed to investigate the use of cyber deception techniques in various domains. The review process was structured in a multi-step manner to achieve comprehensiveness, relevance, and quality.

1) Database search and initial collection: We queried high-impact databases like IEEE Xplore, DBLP, and Google Scholar using various combinations of both general and domain-specific keywords such as “cyber deception”, “honeypot”, “honeytoken”, “fake data”, and other domain-specific terms like “IoT”, “ICS”, “UAV”, “Smart Grid”, and “Cloud”.
2) Inclusion and exclusion criteria: Only the articles that described defensive deception techniques were kept. Articles that were about offensive deception techniques were removed from the list. Besides that, we also filtered out duplicates and inaccessible entries.
3) Reference expansion and survey integration: To make sure all important works were included, we examined and integrated the previous surveys along with their key referenced papers. Each past survey was examined to find the gaps, list the most important works, and to make sure there were no overlaps. The articles that were already included in the previous surveys were also taken into account in order to ensure continuity and progression.
4) Screening and filtering: We did the screening of titles and abstracts first and made the full-text reviews to confirm the relevance of the articles to cyber deception. During this process, the studies were indicated according to the techniques used for deception, the target domain, the architectural framework adopted, evaluation metrics, and the limitations identified. Thus, ∼140 papers were selected.
5) Domain relevance-based filtering: We pursued a filtering strategy based on the articles’ dates of publication that would be on a period extending approximately six years (2019 to 2024), alongside a selective insertion of emerging contributions from 2025. Then, from the set of ∼140 papers, 46 papers were selected as relevant to cyber deception exploration by domain, as illustrated in Fig. 1. The others were used to properly define the cyber deception concepts, techniques, and models.

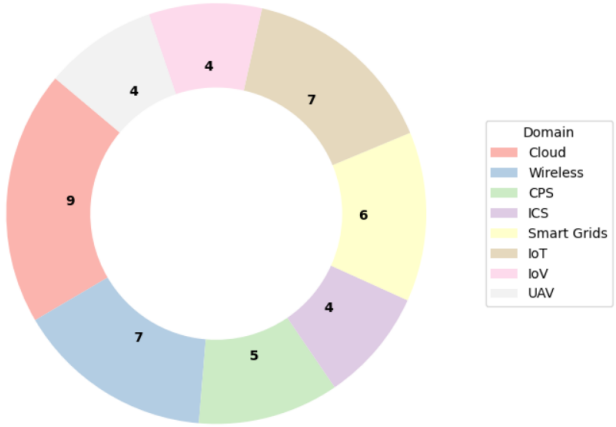


Fig. 1. Distribution of selected papers by domain.

III. Comparison with Existing Surveys

In [2], the authors explored the classification of cyber deception, which cuts across four dimensions: the deceit unit, application layer, defense goal, and deployment mode. They presented techniques like honeypots, honeyfiles, and fake topologies across the network, system, application, and data layers. They further outlined the deployment models as process-based, probabilistic, graph-based, and game-theoretical approaches. The major challenges that were mentioned are the evaluation of deception effectiveness and the automation of dynamic deployment in complicated environments.

In [3], the authors explored 24 studies regarding the application of game theory in the representation of cyber deception and created a taxonomy that contains six types of deception: perturbation, moving target defense (MTD), obfuscation, mixing, honey-x, and attacker engagement. The taxonomy is based on game-theoretic components, including agents, actions, and information structure. Specific deception types were related to the appropriate models, such as stackelberg, Nash, and signaling games. Furthermore, the research gaps were identified by the authors as mimetic deception and practical implementation.

In [4], the authors put forth a completely different classification framework to use for cyber deception based on the following features: conceptual categories, physical and virtual artifacts, intended effects, goals (protection or detection), and

deception activeness. They used game theory (GT) and machine learning (ML) for the discussion of deception techniques, as they pointed out GT's strategic modeling and ML's feature of generating realistic decoys. Their review was complemented by applications in different settings such as CPS, cloud, IoT, and software-defined networking (SDN), and they performed an analysis of evaluation metrics and testbeds. They ended their work by pointing out the existing limitations and urging further exploration of the combination of GT and ML for more adaptive deception strategies.

In [5], the authors made a review of different frameworks of cyber deception in terms of network security, assessing their strengths and weaknesses. They declared that the most effective deception requires network topologies that can adapt, which include scalability, dynamic reconfiguration, and fast responses. They pointed out SDN is the principal facilitator to the deployment of various techniques of deception such as honeypots, honeynets, and Moving Target Defense. Moreover, they assigned the challenges of real-time agility, DDoS resilience, and user impact minimization, and offered directions for the future, such as ML-driven adaptive deception and enhanced network virtualization for APT defense.

In [6], the authors were engaged in scrutinizing Honey-X-techniques such as honeypots, honeynets, honeytokens, and honeywebs. Besides, they also got acquainted with the mechanisms of defense like MTD, ML, and GT, and their works in synergy with them. They studied Honey-X- based strategies in different sectors such as ICS, IoT, critical infrastructure, and web applications, stressing the fact that they can be used in different environments. The effectiveness of these strategies was calculated through metrics such as Mean Time to Compromise (MTTC) and Mean Time to System Failure (MTTSF). The final point they made was the presentation of key research challenges, such as improving honeypot realism and scalability, deception placement optimization through GT, and combining MTD with ML for enhanced threat detection and response.

In [7], the authors investigated deception techniques conceived to boost honeypots and honeynets. These techniques include advanced mimicking, manipulative cooperation, fake databases, honeytokens, and traffic redirection. For this purpose, authors have proposed a categorization system for honeypots based on the following criteria: purpose, interaction level, implementation type, and activity level. They also shared a mathematical model that is supposed to aid in the optimization of both the configuration and the deployment. The evaluation of their approach was realized by comparison with other approaches in terms of deception metrics like discrepancy, wasted time, and adversaries that were returned. The shortcomings they uncovered were in areas such as dynamic, 5G, and SDN-based honeypots, and they suggested future research on ML-driven honeypot optimization and also on more advanced deception strategies.

IV. Background on Cyber Deception

A. Key Concepts of Cyber Deception

Deception originated in military contexts, where it is used to mislead adversaries regarding one's strengths, weaknesses, intentions, and tactics [8]. This tactic includes forms of battlefield deception like camouflage, feints (pretend attacks), ruses (tricks), demonstrations (fake force deployment), and displays (like showing fake military equipment, for example, inflatable tanks). This concept has extended into the cyber domains as cyber deception, refined by Almeshekah and Spafford [9] as "planned actions taken to mislead and/or confuse attackers and to thereby cause them to take (or not take) specific actions that aid computer security defense", and categorized by Whaley [10] into two taxonomies: dissimulation and simulation.

Dissimulation or hiding is a trick of information or system states to prevent hackers from understanding or discovering the true nature of the system. And the purpose of this idea is to mislead hackers or unauthorized users by showing incorrect information, thus making it difficult for them to guess what they should decide. This method is opposite to the traditional protection mechanisms like firewalls, access control, and encryption that mainly resist by prohibiting access or concealing data rather than by misleading the attacker. Simulation or showing is concerned with the making of environments, systems, services, and data that look authentic and operational and thus deceive the attacker into thinking he is dealing with real assets. Such defensive environments are meant to deceive threats, involve enemies, and watch their actions and designs, all without paving the way for real systems or data to be exposed. Simulated components are off the grid yet unbelievably convincing. For instance, after being locked out of the system several times, the system could keep showing the login prompts without ever giving the user permission to log in. The hacker would then think that they still have a chance to enter the system.

Dissimulation and simulation can be used either separately or cooperatively, in accordance with the objectives of the strategy. Dissimulation is the primary means of reducing one system's visibility and appeal, particularly in environments focused on confidentiality and stealth. On the other hand, simulation is the best solution for threat detection by inducing adversaries into a false environment. Simultaneously acting, these two methods create total deception of the mechanism that both distracts and engages adversaries to be able to protect from and gather credible information on the threat.

One of the instances of using deception techniques for cybersecurity can be found in Cliff Stoll's book "The Cuckoo's Egg" [11], where he created a fake system environment with a user account containing fake documents to deceive and stall attackers in order to track their activities and reveal their identities. In the 2000s, honeypots became an adopted method for observing, analyzing, understanding, and modeling attackers' actions [12] [13]. Subsequently, various honey-related approaches, called honeytokens [1] have emerged. These deception techniques fall under six principal tactics [14]. Three of them fall under dissimulation: masking (conceals real systems, data, or network attributes to prevent detection or mislead attackers about their true nature), repackaging (alters the appearance or structure of data or code, such as changing formats or using wrappers, to evade detection and analysis), and dazzling (generates excessive noise such as fake traffic, logs, or alerts to obscure valuable information and confuse attackers). The other three tactics fall under simulation: mimicking (imitates the behavior of real systems,

users, or data to make deceptive assets appear authentic and lure attackers), inventing (creates entirely fictitious systems, data, or user identities that have no real counterpart, designed to mislead and trap attackers), and decoying (deploys simulated assets such as honeypots, honeynets to attract, engage, and analyze attackers while protecting real systems).

Moreover, deception techniques can be leveraged to achieve four principal cybersecurity goals [9]: detection (help identify malicious activity where any interaction is inherently suspicious, by spamming fake assets), prevention (attackers can be deterred or slowed down using creating uncertainty and increasing the perceived risk of being detected), response (deceptive systems can be securely detached and examined post-attack, thus, making the process of incident response and forensic analysis efficient), and research (provide regulated settings to examine the behavior of attackers, create threat intelligence, and study newly developed malware).

B. Layer-Based Cyber Deception Techniques

For the sake of generality and consistency across domains, we adopt the classification proposed by the authors in [2] and further detailed by the authors in [15], which categorizes deception techniques into a four-layer deception stack:

- Network-based deception techniques: are made to deceive attackers in the reconnaissance and exploitation phases that involve manipulation of the network-level information. Examples include altering network topology [16] [17] [18], randomizing IP addresses [19] [20] [21], tarpits [22] [23], deploying network honeypots [24] [25] [26] [27], and using honeyports [28] [29]. Such strategies delay or misguide attackers by presenting a false view of the network infrastructure, thereby increasing their time, cost, and reducing the effectiveness of their probing activities.
- System-based deception techniques: focus on deceiving adversaries within the host or operating system environment. Techniques include honeypatches [30] [31], Ghost Patches [32] [33], deceptive OS fingerprinting [34] [35], deceptive system calls [36][37], and honey RAM / memory injection [38], which allow defenders to monitor attacker behavior without exposing real systems.
- Application-based deception techniques: operate at the software or application interface layer, where user or attacker interaction takes place. They include honeypermissions [39], honeyaccounts [28], honeyprofile (OSN) [40] [41], honeyemail [42], decoy web forms/ fields [43], delayed response (fake feedback) [44] [45], fake API endpoints [46], and decoy hyperlinks [47] [48]. Such methods are effective at detecting and delaying attackers during application-level attacks like SQL injection, credential theft, or web exploitation.
- Data-based deception techniques: target the attacker's goal, access to sensitive data, by injecting deceptive or false data into the system. Examples include decoy document [49], honey URL [50] [51], honeyentries [52] [53] [54], honeyfiles [55] [56] [57] [58], honeywords [59] [60] [61] [62], honey encryption [63], honey database/ metadata [64] [65], and decoy source code [66], all designed to mislead, trace, or waste the resources of attackers attempting to exfiltrate valuable information.

The placement of these deception techniques within a target domain can either be a part of the existing security solution or can be isolated and used as a standalone mechanism. In our classification, we adopt the logic explored in [2], which defines four main deployment modes based on security objectives, system architecture, and operational constraints:

- Built-in deception/ part of: embedded directly into the system during the design phase, such as deceptive logic in source code or modified system responses.
- Add-on deception/ part of: incorporated into an operational system at runtime, through the insertion of honeytokens, fake files, or configuration traps.
- In-front deception/ part of/ intermediary: positioned as a proxy or gateway in front of the real system to intercept and manipulate incoming interactions, allowing early engagement with potential threats.
- Isolated deception/ standalone: deployed in separate, decoupled environments such as honeynets, decoy servers, that mimic real systems to lure attackers and analyze their behavior without impacting production assets.

In Table I, we provide a consolidated overview of representative cyber deception techniques, highlighting their core characteristics in terms of taxonomy, tactics, goals, application layers, and deployment modes as extracted from the surveyed literature. The majority of the cyber deception techniques outlined are designed for dual purposes, simulation and dissimulation, because building false artifacts while concealing true ones at the same time is the precondition of effective deception in most cases. The dual effectiveness increases both the credibility and overall impact of that deception. Also, a lot of the techniques can be utilized on different layers, such as network, system, and application, as they are designed to be adaptable and provide flexibility in their implementation. Therefore, the same technique may cover different security objectives, such as prevention, detection, response, and research, based on how and where it is implemented.

A multi-layer deception system is depicted in Fig. 2. It is a combination of some of the techniques that have been previously discussed. The multi-layer architecture, comprising the network, system, application, and data layers, encompasses both real and honey elements. Each of the honey components (e.g., honeyserver, honeyfile, honeydatabase) imitates a corresponding real counterpart (e.g., server, file, database) to mislead and monitor the attacker's behavior. The analyst server functions as a pivotal element by collecting alerts and logs from all honey elements, which provide a basis for detection, analysis, and response to malicious activities triggered by interactions with the deceptive assets.

C. Key Evaluation Metrics for Cyber Deception

To evaluate the efficiency of cyber deception strategies, researchers have suggested a variety of quantifiable metrics

TABLE I. OVERVIEW OF CYBER DECEPTION TECHNIQUES

Deception technique	Ref	Taxonomy	Tactic	Goal	Layer	Deployment mode
Altering Network Topology	[16]	Simulation + Dissimulation	Inventing, Dazzling	Prevention, Detection	Network	In-front
Reconnaissance Deception System (RDS) using Virtual Network Views	[17]	Simulation + Dissimulation	Inventing, Dazzling, Decoying	Detection, Prevention	Network	In-front, Add-on, Isolated
Virtual Network Topology Deception Defense (VNTDD)	[18]	Simulation + Dissimulation	Inventing, Dazzling, Decoying	Detection, Prevention	Network	In-front, Add-on, Isolated
Randomizing IP Addresses	[19]	Simulation + Dissimulation	Dazzling, Mimicking, Decoying	Detection, Prevention	Network	In-front, Add-on, Isolated
	[20]	Simulation + Dissimulation	Dazzling, Mimicking	Detection, Prevention	Network	Add-on, In-front
Path Randomization	[20]	Simulation	Dazzling	Prevention	Network	Add-on, Isolated
Flexible Random Virtual IP Multiplexing (FRVM)	[21]	Simulation + Dissimulation	Dazzling, Mimicking	Prevention, Detection	Network	Add-on, In-front
Tarpits	[22]	Simulation + Dissimulation	Dazzling, Decoying	Detection, Prevention	Network	Add-on, Isolated
	[23]	Simulation + Dissimulation	Dazzling, Decoying, Mimicking	Detection, Prevention, Research	Network	Add-on, Isolated, In-front
Honeypot	[24]	Simulation + Dissimulation	Decoying, Inventing	Detection, Prevention, Response	Network	Add-on, In-front, Isolated
	[25]	Simulation + Dissimulation	Decoying, Mimicking	Detection, Prevention, Research	Network, Application	Add-on, In-front, Isolated
Network Honeypots	[26]	Simulation + Dissimulation	Inventing, Decoying, Dazzling	Detection, Prevention	Network, Application	Add-on, Isolated
Modular High-Interactivity Honeypot System	[27]	Simulation + Dissimulation	Inventing, Decoying, Dazzling, Mimicking	Detection, Prevention, Response, Research	Network, System, Application	Add-on, Isolated, In-front
Honeytoken-based deception framework	[28]	Simulation + Dissimulation	Inventing, Decoying, Masking, Repackaging	Detection, Response	System, Application, Data	Add-on, Isolated
HoneyPort (meta-honeypot system)	[29]	Simulation + Dissimulation	Inventing, Decoying, Mimicking, Dazzling	Detection, Prevention, Research	Network, System	Add-on, Isolated, In-front
Honeypatches	[30]	Simulation + Dissimulation	Mimicking, Decoying, Dazzling	Detection, Prevention, Response, Research	Application, System	Add-on, In-front, Isolated
	[31]	Simulation + Dissimulation	Mimicking, Decoying, Dazzling	Detection, Prevention, Response, Research	System, Application	Add-on, Isolated, In-front
Ghost Patches	[32]	Simulation + Dissimulation	Mimicking, Decoying, Dazzling	Detection, Prevention, Research	System, Application	Add-on, Isolated, In-front
Deceptive Patch Model (Faux, Obfuscated, Active Response Patches)	[33]	Simulation + Dissimulation	Mimicking, Inventing, Decoying, Dazzling, Repackaging	Detection, Prevention, Response, Research	System, Application	Add-on, In-front, Isolated
Deceptive OS Fingerprinting	[34]	Simulation + Dissimulation	Mimicking, Dazzling, Repackaging	Detection, Prevention, Research	Network, System, Application	Add-on, In-front, Isolated
	[35]	Simulation + Dissimulation	Dazzling, Mimicking, Repackaging	Detection, Prevention, Research	Network, System	Add-on, In-front, Isolated
Deceptive System Calls	[36]	Simulation + Dissimulation	Mimicking, Masking, Dazzling	Detection, Prevention	System	Add-on, Isolated
Honey RAM / Memory Injection	[37]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	System	Add-on, Isolated
	[38]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	System	Add-on, Isolated
Honeypermissions	[39]	Simulation + Dissimulation	Inventing, Masking	Detection, Prevention	Application	Add-on, Isolated
HoneyID	[28]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	Application	Add-on, Isolated
Honeyprofile (OSN)	[40]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	Application	Add-on, Isolated
	[41]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	Application	Add-on, Isolated
Honeyemail	[42]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	Application	Add-on, Isolated
Decoy Web Forms/Fields	[43]	Simulation + Dissimulation	Inventing, Mimicking, Decoying	Detection, Prevention	Application	Add-on, Isolated
Delayed Response (Fake Feedback)	[44]	Simulation + Dissimulation	Mimicking, Decoying, Inventing	Detection, Prevention	Application	Add-on
	[45]	Simulation + Dissimulation	Dazzling, Masking	Detection, Prevention	Application	Add-on, In-front
Fake API Endpoints	[46]	Simulation	Inventing, Decoying, Masking	Detection	Application	Add-on
Decoy Hyperlinks	[47]	Simulation + Dissimulation	Inventing, Decoying, Masking	Detection	Application	Add-on, In-front
	[48]	Simulation + Dissimulation	Inventing, Decoying, Masking	Detection	Application	Add-on, In-front
Decoy Documents	[49]	Simulation + Dissimulation	Inventing, Mimicking, Masking	Detection, Prevention	Data	Add-on, Isolated
Honey URL	[50]	Simulation + Dissimulation	Inventing, Masking	Detection, Response	Data	Add-on, Isolated
	[51]	Simulation + Dissimulation	Inventing, Masking	Detection, Response	Data	Add-on, Isolated
Honeyentries	[52]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
	[53]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
	[54]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
Honeyfiles	[55]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
	[56]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
	[57]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Response	Data	Add-on, Isolated
	[58]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Research	Data	Add-on, Isolated
Honeywords	[59]	Simulation	Inventing	Detection, Prevention	Data	Add-on
	[60]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Prevention	Data	Add-on, Isolated
	[61]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Prevention	Data	Add-on, Isolated
	[62]	Simulation + Dissimulation	Inventing, Masking, Repackaging	Detection, Prevention	Data	Add-on, Isolated
Honey Encryption	[63]	Simulation + Dissimulation	Inventing, Repackaging, Masking	Detection, Prevention, Research	Data	Add-on, Isolated
Honey Database/Metadata	[64]	Simulation + Dissimulation	Inventing, Repackaging, Masking	Detection, Prevention	Data	Add-on, Isolated
Fake Document Infilling (FDI)	[65]	Simulation + Dissimulation	Masking, Repackaging, Mimicking	Detection, Prevention, Research	Data	Add-on, Isolated
Decoy Source Code	[66]	Simulation + Dissimulation	Inventing, Repackaging, Masking	Detection, Prevention, Research	Data	Add-on, Isolated

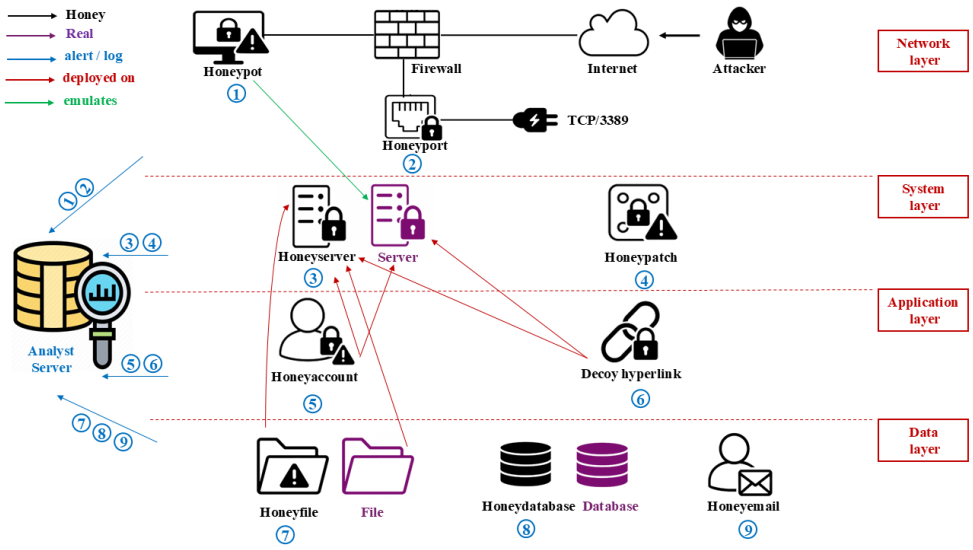


Fig. 2. An example of a multi-layered cyber deception architecture.

that can represent different components of system resilience, attacker behavior, and defense performance. The performance of the system and the progress of the attacker are evaluated with the help of temporal metrics like the Mean Time to Security Failure (MTTSF) [67] which symbolizes the period during which the system remains secure before it reaches a predefined security failure threshold due to a successful compromise, and the Mean Time to Compromise (MTTC) [67] where it represents the time required by the attacker to add a successful compromise to the network. Moreover, the metric of Vulnerability Analysis Time [68] assists in quantifying the amount of time the attacker has to spend identifying and then exploiting weaknesses present in the system, that is, from the complexity and skill level of each attacker. On the contrary, the Decision and Deployment Time (DDT) [69] is meant to show the period of time that passes between making a defensive decision and implementing it, such as deploying a honeypot.

Defense effectiveness is best evaluated through the use of probability-based indicators like the Defense Success Probability (DSP) [70] and the Success Rate of Attacks [71], which act as the basic tools to measure the quantitative success or failure of attaching defense devices under controlled experimental conditions. The measurement of detection metrics also focuses on Detection Accuracy [72], which is typically expressed in terms of the AUC of the ROC curve to judge the correct value of identifying malicious behavior. The other metric is the Mean Time to Detect Attacks [73], which is used for the evaluation of the efficiency of the deception-based detection system. In game-theoretic models of deception, the concept of Utility (or Payoff) is used most often to express the cost-benefit balance for both attacker and defender [74]. Moreover, the Attack Cost [75] evaluates the effort or risk that the attacker gets, including the chance of getting caught or wasted resources.

Resilience and operational continuity are represented by metrics such as Survival Rate [69], which refers to the ratio of the legitimate system components that have remained intact during attacks, and Round-Trip Time [69], which is the time taken by deception mechanisms (for example, code obfuscation or network slowdowns) to delay the attackers' progress.

D. Cyber Kill Chain for Phase-aware Deception Planning

The effectiveness of cyber deception strategies lies in the appropriate selection and positioning of the described deception techniques, which are customized to the particular-domain environment (e.g., UAV, IoT, ICS, Cloud) and the threat actors' behavior. Instead of randomly inserting deceptive assets, a threat-driven modeling approach is a must to prove that both deception strategies are useful and impactful. Two pertinent frameworks uphold this goal: the Cyber Kill Chain (CKC) and the MITRE ATT&CK framework. By using them together, they provide adversary tactics and lifecycle progression, and thus, an integrated view that helps in planning the optimization of deception for when, where, and how to deploy it.

The Cyber Kill Chain (CKC), originally introduced by Lockheed Martin, divides cyber-attacks into sequential phases: reconnaissance, weaponization, delivery, exploitation, installation, command and control (C2), and actions on objectives. This phase-based structure aids deception planning by helping defenders: i) determine the optimal timing for deploying deception (e.g., before delivery versus after exploiting), ii) align techniques to specific stages, for instance, deceptive network banners during delivery or honeyfiles during exfiltration, and iii) layer deception strategically across the kill chain, like using honeypots for exploitation, MTD for delivery, and fake data for exfiltration. Following these lines, the authors in [1] suggested an enhanced CKC model, meant to be used for deception and formed of three sub-kill chains: external, internal, and target manipulation. This refined model not only identifies the multi-stage progression of modern intrusions but also serves as a tool for mapping deception techniques (e.g., honeypots, honeytokens, MTDs) to both the attack phase and the affected system layer (network, system, software, data). By incorporating MITRE ATT&CK tactics into this structure, the model acts as an extensive framework for the orchestration of deception throughout the whole attack lifecycle.

E. MITRE ATT&CK for Tactic-technique Driven Deception

The MITRE ATT&CK [76] framework is based on the tactic-technique-procedure (TTP) representation of the behavior of adversaries, that is firmly based on the real-life observations on a wide range of platforms, namely, Enterprise, ICS, Mobile, and Cloud. ATT&CK plays a significant role in providing defenders with the following benefits: i) modeling attacker behavior by domain: for example, the ATT&CK for ICS stressed techniques such as "Inhibit Response Function" and "Manipulation of Control" which can also be found in the cyber-physical systems (CPS) and smart grid paradigms, ii) identifying deception targets: techniques like "Remote System Discovery", "Command and Scripting Interpreter", and "Credential Dumping" can be countered with honey APIs, honey credentials, or deceptive command environment, iii) building attack graphs: authors of [77] illustrated how techniques can be linked by logical preconditions and thus form attack paths. The graphs assist the defenders to optimize the positioning of decoys by focusing on the nodes that are most likely to be compromised with the least amount of resources.

In [78], the authors suggested a multi-layer graph-based way to model cyber-attacks and strategically allocate deception resources in alignment with threat modeling principles inspired by MITRE ATT&CK. The integration of network, service, and attack layers makes possible the detailed mapping of adversarial paths and their dependencies across system components. The model is capable of risk-based prioritization and preemptive allocation of deception assets such as honeypots and honeytokens at critical nodes. These strategies will contribute to the re-routing of attack chains and misleading of adversaries. This type of planning also drives the threat-informed deception by making the placement of the deception close to the attacker's behavior and technique paths that are illustrated in frameworks like MITRE ATT&CK. Authors in [79] presented a system that automatically chooses cyber deception strategies based on the information it obtains from unstructured CTI reports through NLP. It first maps the extracted Indicators of Compromise (IoCs) to the relevant MITRE ATT&CK tactics and techniques, which then allows for the prediction of the attacker's actions. The deception actions proposed by this framework are in line with the MITRE D3FEND [80] and MITRE Engage [81] frameworks, which are based on the aforementioned profiling.

By aligning deception strategies with specific ATT&CK techniques, defenders can deploy assets that disrupt adversary decision-making, enhance situational awareness, and collect high-fidelity threat intelligence. Fig. 3 presents an example of the Enterprise ATT&CK matrix techniques, annotated with corresponding deception technique, as illustrated by MITRE D3FEND, and also mapped to the corresponding engagement techniques, as described by MITRE ENGAGE.

V. Intelligent Orchestration of Cyber Deception through Reinforcement Learning and Game Theory

The Cyber Kill Chain and MITRE ATT&CK frameworks serve as a guide for the mapping of adversary behaviors to suitable deception techniques, but the actual deception deployment in dynamic environments would need continuous adaptation and decision-making under uncertainty. In this way, reinforcement learning (RL) and game theory (GT) become central for the intelligent orchestration of cyber deception. Their embedding into the threat modeling process not only helps in contextualizing but also promotes the continuous improvement of deception strategies, thus providing smart and proactive defense across a wide range of areas.

Game-theoretic models capture the strategic interactions between attackers and defenders, allowing defenders to find the optimal countermeasures by taking the adversarial payoffs, costs, and bounded rationality into account. Authors in [82] introduced a game-theoretic framework to optimize the deployment of honeypots and software diversity in a network under resource constraints. It presents two-layer deception tactics: the first layer determines the position of honeypots on network edges using a zero-sum game based on attack graph analysis and node importance, while the second layer employs a non-zero-sum game to manage software diversity across honeypots, thereby preventing the attackers from recognizing the uniform deception. The model takes into consideration the attacker-defender interactions and provides the mixed-strategy Nash equilibria to attain the maximum defender payoff. The experimental results indicate that the strategic allocation of honeypots with deception diversity significantly enhances the defense effectiveness against adaptive adversaries.

In [24], the authors proposed a novel game-theoretic framework for honeypot allocation in dynamic tactical networks. It makes use of the fact that node mobility changes network topology and attack paths. The interaction modeled herein is between the defender and the attacker in a two-player zero-sum Markov game on changing attack graphs. The defender is the one who wants to place honeypots in a way that the attacker's deception is maximized and the costs for reconfiguration in the network states are minimized, while the attacker is the one who picks the right attack paths. A predictive model has been incorporated and a solution has been provided for stationary Nash equilibria through the application of a Q-minimax algorithm. The results of the investigators show that projecting the future mobility of the system would be the most effective defense measure, while it also reduces the rewards of the attacker's efforts. Thus, it becomes a method that is both scalable and adaptive for the plan of deception in a mobile and resource-constrained environment.

In [83], the authors provided a foundational game-theoretic framework for applying cyber deception through evidence-based signaling games. They provide mathematical models that capture the dynamics of strategy between a deceiver and a deceived under uncertainty, including the probabilistic evidence of intrusion detection outputs. Also, they analyze: a binary state deception, a continuous state deception with a detection cost, and a multi-level deceptive modeling of APTs. Perfect Bayesian Nash Equilibrium (PBNE) is used to illustrate the optimal deception strategies. This research introduces the term of deceivability in a formal way, and states certain conditions for the deception to be not only sustainable but also effective. It supports the design of adaptive, long-term deception policies by considering how attackers learn and update beliefs over time, making it a theoretical cornerstone for strategic deception planning in adversarial settings.

In [84], the authors proposed a game-theoretic framework

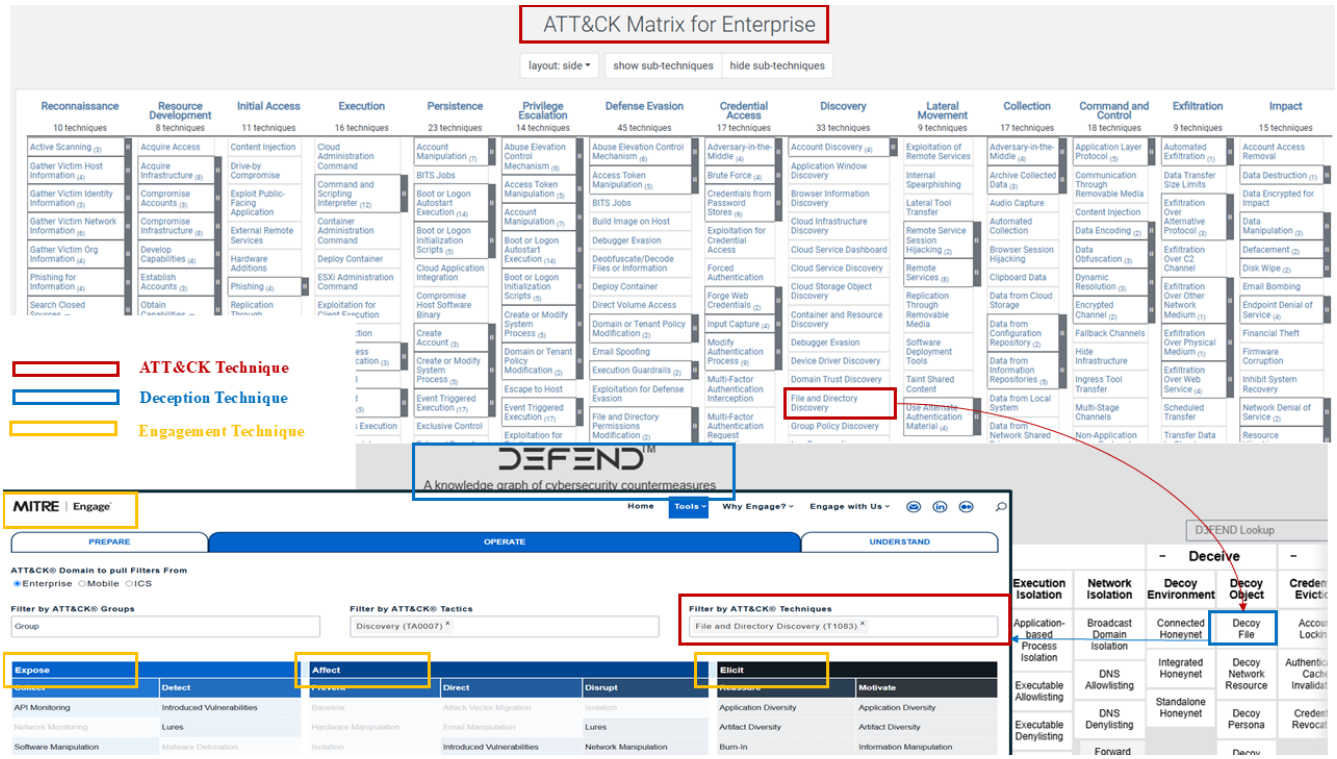


Fig. 3. Mapping of an enterprise MITRE ATT&CK technique to corresponding deception and engagement techniques.

that has multiple levels as well as learning and control theories. The main application is to counter deceptive information attacks in Intelligent Transportation Systems (ITS). The introduction of the PRADA (Proactive Risk Assessment and Mitigation of Misinformed Demand Attack) framework, which regards the attacker-defender dynamics as a stackelberg game, and quantifies both local and network-wide effects of demand manipulation, is the most important contribution. Besides, the paper also proposes trust-constrained recommendations as a tool to ensure system reliability in case of attack. The approach integrates strategic modeling, user equilibrium routing, and predictive resilience mechanisms, where game theory emerges as the core instrument in facilitating adaptive deception in the ITS environments.

On the other hand, reinforcement learning facilitates data-oriented, self-adaptive decision-making by obtaining deception policies over time through trial-and-error interaction with simulated environments. The techniques, such as Deep RL and Q-learning provide defenders with the opportunity to dynamically modify the placement of the deception, the interaction fidelity, and the signaling strategies according to the new threats. Authors in [85] proposed an adaptive Q-learning-based framework for orchestrating cyber denial and deception strategies in enterprise networks, by using a detailed attacker-aware attack graph. The model integrates MITRE ATT&CK tactics and CVSS-derived metrics to evaluate vulnerabilities in terms of importance, exploitability, and damage. Two different attacker profiles: opportunistic and strategic, are used to model different behaviors, and the defender learns optimal responses (none, denial, deception) via Q-learning. The approach enables a dynamic and tailored honeypatch placement that is built to the attacker's strategy and system context. Besides, to enhance decision making, some works combine reinforcement learning with game theory that enables defenders to learn from gained experience and think strategically about the attacker's behavior.

In [86], the authors suggested a reinforcement learning approach to dynamically select and orchestrate deception and Moving Target Defense (MTD) strategies. They developed a confrontation model, which is a prototype to simulate the interaction between an attacker and a defender, taking into account perception asymmetry and phase progression via a simplified Cyber Kill Chain (CKC). They employ Deep Q-Learning to train a defensive agent capable of adapting its strategy based on observed attacker actions and system states. The major contribution is showing how deep reinforcement learning can enable adaptive, perception-aware, and cost-sensitive orchestrations of deception and MTD in complex environments, which are better than static or naive strategies in all tested scenarios.

In [87], the authors developed a deep reinforcement learning framework, a novel solution for the deployment of honeypot strategies in complex networks represented using Bayesian attack graphs. They model a stackelberg game that involves a defender (the one who deploys a honeypot) and a hacker, where the defender takes the lead by deploying high- or low-interaction honeypots to manipulate the hacker. The setting is represented as a Markov Decision Process (MDP), and the search for the strategy space is carried out using Actor-Critic and Proximal Policy Optimization (PPO) algorithms. Their method contributes to the dynamic, state-aware decision making and is reported to outperform random and value-weighted baseline strategies significantly in experiments. The study presents the effectiveness of RL in learning robust and

cost-efficient deception strategies under uncertainty.

In [88], the authors proposed a hybrid game-theoretic and reinforcement learning method to enable quantum superdense coding security against advanced attacks, including scrambling and bijection. The model introduced decoy qubits as deception elements so that eavesdropping attempts can be detected, and the interaction between the sender/receiver (Alice and Bob) and the attacker (Eve) is framed as a non-cooperative game. The authors used custom utility functions: Return on Protection and Return on Attack, analyzed both pure and mixed strategy Nash equilibrium, and used a Q-learning approach to optimize decoy insertion strategies under uncertainty. The simulation results confirm the model's capability to converge to equilibrium and minimize the false negatives, which illustrates the efficiency of game-theoretic reasoning coupled with RL-driven strategy adaptation in quantum security.

VI. Cyber Deception in Cloud Environments

Cloud computing has brought a drastic change in how organizations can deploy and scale IT infrastructure, providing elasticity, multi-tenancy, and on-demand resource provisioning. However, many of these are the same design features that create unique security problems, such as co-residency attacks, lateral movement, insecure APIs, and dynamic reconnaissance. The dynamic, decentralized, and virtualized characteristics of the cloud infrastructure make it an attractive target for attackers, while legacy-based perimeter security approaches frequently do not secure the evolving threats adequately [89]. Deception has been the proactive cyber defense technique to address these threats with the help of misleading artifacts, such as honeypots, decoy virtual machines, fake credentials, and deceptive storage buckets. Novel methodologies are being employed nowadays, such as IP and VM shuffling, AI-driven decoy placement, and adaptive honeynet orchestration for the purpose of confusing respective attackers, postponing their advancement, while collecting useful threat information. Installation of the deception in cloud environments is not only through this promise but also through the involved challenges, such as ensuring the scalability and realism of deceptive elements, avoiding interference with legitimate operations, and integrating virtualized orchestration layers in a seamless manner. This section covers cyber deception in the cloud setting, particularly focusing on deception-aware architectures, the corresponding cloud (IaaS, PaaS, SaaS) threat coverage, and the limitations and deployment trade-offs that current methods come with.

In [90], the authors proposed an integrated cloud security framework based on a mixture of signature-based NIDS (SNORT) with distributed honeypot networks (Glastopf, Cowrie, Dionaea) in an OpenStack environment. The architecture can be seen in Fig. 4, where authors have strategically placed SNORT at network controller nodes (to deal with external traffic) and compute nodes with OpenV Switches (for internal/local traffic), while honeypots emulate services in order to capture interactions of the attacker. The data obtained is then analyzed in a sandbox environment (Cuckoo Sandbox) to dynamically extract behaviors of the malware, which then triggers an automated use case to add new and update the current SNORT rules. The system shows an increase of unknown attacks detected (for example, XSS, malware variants) as well as causing redundant alerts, among which, we should mention the overhead of distributed NIDS instances and that the research is focused mainly on network-layer attacks, but the added value of integrating host-based defenses also like VM introspection is still valid.

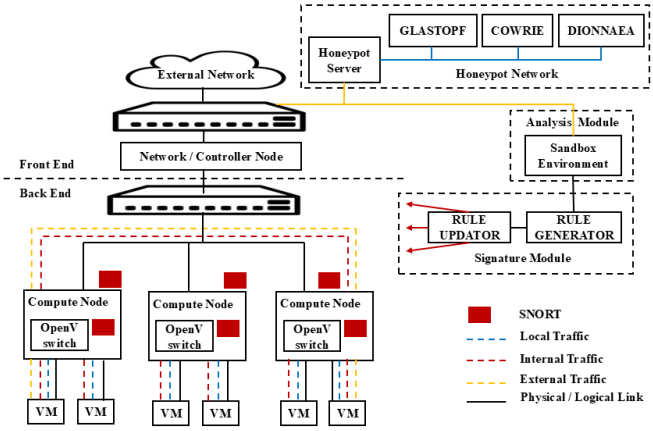


Fig. 4. The overall architecture showing the different modules and types of traffic in [90].

In [91], the authors provided a game-theoretic Moving Target Defense (MTD) framework that tactically combines both real and deceptive virtual machine (VM) migrations to ensure security of the cloud environments against side-channel and co-residency attacks. They proposed a new model that represents the security interaction as a signaling game, in which defenders can either perform live/non-live migrations or simulate them by carefully crafted traffic patterns, creating uncertainty for the attackers while minimizing costly real migrations. The analysis shows that Nash equilibria are the only beyond which deceptive signaling offers the same security level as physical migrations, the network controller operating on detection alerts and cost parameters that dynamically set best strategies. This study of the symbolic trade-off of migration overhead and security gains, thus admitting both a theoretical model for the deceptive MTD in clouds and a practicable decision algorithm for the implementation, clearly underlines its contribution, whereas questions regarding performance in realistic situations and adversarial learning remain unaddressed.

In [71], the authors suggested the Automated Honeynet Deployment Strategy (AHDS), which is a dynamic security framework for container-based clouds that adaptively deploys honeypots to prevent emerging threats. AHDS uses an Attack Graph (AG) to chart hypothetical routes of attacks, and it also comes up with a Deceptive Exploring Surface (DES), a network of strategically placed source honeypots to divert potential enemies from critical assets. The algorithm gives priority to high-value nodes based on the degree of centrality and optimizes the layout with the help of a set cover-based algorithm, which has the effect of not wasting resources and having the maximum possible coverage. The system is built upon three main building blocks, which are: 1) a Monitor Engine for recording the infrastructure changes and costs, 2) a Decision Engine that adapts the strategies of the honeynet in real time based on the AG assessment, and 3) a Deployment Handler for automatic changes of configurations. Evaluations

reveal that AHDS is 83% more efficient at decreasing the possible attack emergence while outgunning the traditional strategies (e.g., "Max" and "Most") both in effectiveness and in no resource waste, and is seamlessly adopted by datacenter sizes. Nonetheless, its effectiveness hinges on precise AG modeling and it relies on assumptions that intruders follow predictable propagation patterns, potentially limiting robustness against novel or adaptive threats.

In [92], the authors suggested a cloud-based deception framework that can disrupt DDoS reconnaissance by using SDN/NFV technology to generate virtual reflection networks in the cloud, which send back to the attackers false topology data but still keep correct details for the legitimate users. The system is using GRE tunnels for the rerouting of the reconnaissance traffic (traceroute/ping) to the virtual SDN switches, which are programs that implement IP hopping and path mutation, with the physical and cloud networks being synchronized by dual SDN controllers and NFV orchestrator. The solution was tested on the GENI platform and only added a 25-35ms delay, decreased packet loss to below 1%, and reflected network alterations in 75ms, while it was >99% cheaper than a physical infrastructure, but it is reliant on the accuracy of the synchronization that is required to connect between the physical and the virtual networks and also it can be difficult in the situation of very frequent or large-scale topology changes.

In [93], the authors proposed a deception-based defense framework for private cloud infrastructures that is based on using honeypots in an enterprise IT environment to identify and contain internal threats by deploying decoy virtual machines (VMs) with emulated services (e.g., routers, databases) that run parallel to the production VMs. The setup, as illustrated in Fig. 5, includes Snort for intrusion detection and Sebek for activity logging, which together log attacker interactions and pass the data to a Multiclass Support Vector Machine (MSVM) classifier that applies feature selection (Gini index) and dimensionality reduction (PCA) methods to traffic analysis. With the help of KDDCup99, NSL-KDD, and Gure-KDD datasets, the classifier achieves a remarkable 95% accuracy in distinguishing between different types of attack (DoS, U2R, R2L, probing) and normal behavior by thus effectively isolating production systems and profiling intruder activity during the process. The application, while functional against most network-layer threats, is limited by the framework's signature-based detection approach that is inadequate for dealing with advanced attacks, which suggests improvement through infrastructure integration with anomaly detection and threat intelligence feeds for a more comprehensive security coverage.

In [94], the authors introduced a novel AI-based defensive deception framework to secure multi-tenant cloud environments from reconnaissance attacks. The strategy overcomes the issues brought about by static, uniform configurations in cloud networks, which add simplicity to reconnaissance for the adversaries. To this end, the framework exploits deep reinforcement learning (DRL) to formulate the optimal deployment of decoys. A utility function models common OS vulnerabilities as a threat, while a DRL agent who has been trained with Proximal Policy Optimization (PPO) is the one to produce strategies for configuring and distributing decoy virtual machines in a fine-grained manner. These strategies aim to hide actual assets

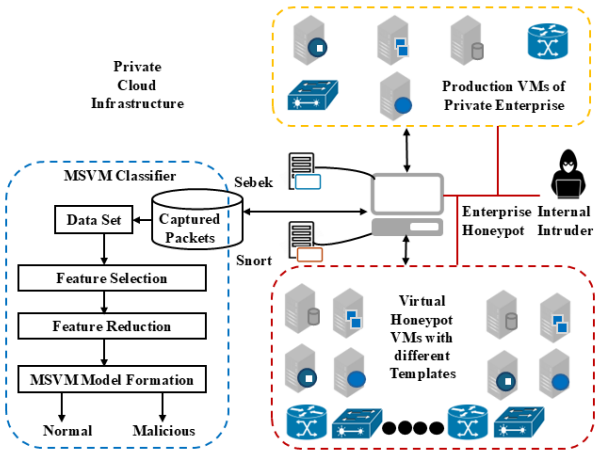


Fig. 5. The private cloud infrastructure with enterprise honeypot in [93].

and thereby increase the attacker's resource consumption. Conducted simulations proved the framework's skill in improving asset concealment by 20.58% and raising attack costs by 40.4% as compared to the traditional approach. This work challenges the idea that DRL constitutes the only probability of deception being dynamic and scalable in the clouds, thus allowing for accurate protection against reconnaissance.

In [95], the authors proposed the use of IRDS4C, a multilayer deception framework for the aim of detecting zero-day ransomware and also intrusions across cloud environments such as IaaS, PaaS, and SaaS. Through deploying of strategically positioned decoys, including high-interaction honey files/tokens (to mislead the attackers) and low-interaction canary resources (for early detection), the system can use the ASCII-based naming scheme to make the ransomware interact with the decoys first. The behavioral monitoring that it uses to track unauthorized access, does not however, disturb the legitimate operations. The testing was performed on Google Cloud with seven ransomware families (e.g., REvil, WannaCry) and IRDS4C was found to achieve 100% detection accuracy, while faster than traditional methods like file hashing. Besides that, the framework also successfully recognized the human intruders in simulations of different breaches carried out, being a lightweight, scalable solution designed for cloud-native deployment, with future extensions planned for Linux and macOS, compatibility.

In [96], the authors introduced a novel containerized honeypot-based deception system which is capable of real-time tracking and profiling of cyber adversaries in the cloud setting. Docker containers are the framework's key feature that are used to establish a variety of honeypots (e.g., Cowrie, Dionaea, DDoSPot, RedisHoneypot) in different Microsoft Azure regions, under the support of ELK stack (Elasticsearch, Logstash, Kibana), which will be used for data aggregation and visualization. The system looks like a typical cloud service and to create open ports for the attackers to connect to it uses a Suricata, a tool for p0f, and FATT, along with some interesting extras whose CVEs, operating systems, and behavioral patterns are being analyzed. The deployment of the work was successful for the reason that it caught a large number of attacks, such as the exploitation of the old CVEs,

the botnet-based DDoS traffic, and the suspected URL usage, which provided intelligence on the adversary techniques and the infrastructures. The experiment setup has shown competitiveness in processing data acquisition, forensics, and attack identification, with plans for the next stage to include the early-warning system upgrade and the threat coverage expansion to the zero-day detection. This work is a representative example showing how scalable and cloud-native honeypot deployments can act as a strategic deception tool for proactive cyber defense.

In [97], the authors suggested a resource-aware cyber deception framework for microservice-based cloud-native applications. This is another option in the case of traditional pre-built decoys that are not effective in dynamic service environments because of their inability to blend. Their design, using the distribution of a non-linear mixed integer programming problem, employs the production microservices cloned as high-fidelity decoys to catch lateral movements of the attacker that are being made along the attack path, at the same time maximizing the intercepted attack paths with respect to CPU/RAM restrictions and also considering the attack graph topology. To solve the problems related to computational complexity, they have offered an approach that enables the prioritization of the most critical microservices based on the betweenness centrality measure. The algorithm provided is highly scalable, and the realistic placement of the decoys is achieved. Their results clearly show that, even with fewer resources, the coverage gained from deceptive means was more than with the basic approach; thus, the framework can be confidently implemented in modern cloud spaces where contextual and resource-efficient deception are a must.

Table II summarizes all the discussed works in cloud environments, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

VII. Cyber Deception in Wireless Networks

Wireless networks are largely under threat due to the inherent challenges posed by their broadcast nature, limited resources, and dynamic topologies, which make them susceptible to a range of cyber-attacks including reconnaissance, eavesdropping, jamming, spoofing, and lateral movement attacks. Security mechanisms that are based on encryption, authentication, and intrusion detection are the most commonly used traditional safety measures. They are not enough, in decentralized networks, as open networks are, with their limitations in computational power and energy efficiency of many wireless devices [98]. In response, this has been countered by the emergence of cyber deception, where this countermeasure aims to create trouble for attackers, ensure the communication of critical information, and collect data on potential threats. For instance, the defenders do not just deploy decoy nodes, fake access points, adaptive honeypots, but also the communication protocols that mislead the attacker. Such practices take the place of the real nodes and thus prevent the attacker from gaining the proper situational awareness. Recent works focused on the recent developments, especially in software-defined and virtualized wireless networks, where the use of AI and machine learning has been instrumental in supporting the dynamic traffic redirection, strategic decoy deployment, and behavior modeling of the attacker. Despite its large potential, the operationalization of deception in wireless networks also encounters issues like keeping the decoys realistic, reducing the interference with the real communication, and ensuring rapid adaptation to network change. This section reviews the most important works in the field that utilize deception as an adaptation mechanism to the constraints and shifting hostile environment in wireless communication systems.

A trilogy of deception-driven architectures is introduced to secure the virtualized wireless networks against both RF-based and infrastructure-layer threats. In the first work [99], the authors introduced a cyber deception framework for virtualized wireless networks that counters unauthorized access and DoS attacks through SDN-controlled dynamic redirection. An SDN controller observes the network traffic and redirects adversarial flows to a Deception mobile virtual network operators (DMVNO), which acts as a decoy through the emulation of legitimate services, while the availability for benign users is preserved. The defense mechanism characterizes the attack dynamics as a birth-death process and assesses them using Monte Carlo simulations with metrics such as attack redirection rate, deployment timeliness, and the probability of legitimate users being blocked. Primary limitations are noted, including the failure to inject real-time adaptability for unidentified attack patterns and the issue related to the delays in the detection-to-deployment pipelines not being mentioned. These limitations are suggested as future work. To address this, the second paper [100] introduced Deceptor-in-the-Middle (DitM), as illustrated in Fig. 6, which is a cyber-deception architecture that can be found in virtualized wireless networks and provides concurrent real-time attacker detection and dynamic adversarial redirection. At the beginning of the process, the system detects adversaries by utilizing RF energy sensing and Wald's Sequential Probability Ratio Test (SPRT), then, it transparently routes them to a Deception VNO (D-VNO) through ARP cache poisoning, by mimicking a Man-in-the-Middle (MitM) attack in the opposite way. The framework is built using an SDN controller and network aggregator that operate in a "sense-observe-manipulate" mode, with the added benefit of ensuring that genuine users do not experience any service disruption. The distribution reflects a comparison between the detection rates (false alarm and miss-detection rates) at different thresholds, with the only limitation being that it lacks formal guidelines for the optimal resource utilization and leaves scalability, as well as, adaptive adversary management to future work. The third paper [101] addressed the virtualized wireless networks cyber deception optimization using a Stackelberg game-theoretic model, in which an SDN controller, under budget constraints, strategically maps actual RF configurations to deceptive observable configurations while interacting with either naive or rational attackers. The work demonstrates the NP-hardness of the optimal strategy selection and presents a polynomial-time algorithm for naive attackers as well as a scalable greedy heuristic (GMS) for rational ones, proving successful attacker utility minimization and computationally efficient deception deployment through simulations. While the steps are made from static to adaptive and after that, to game-theoretically optimized deception, the method is still limited by its assumptions about rational attacker behavior and also faces challenges with optimal Mixed Integer Linear Programming (MILP) solutions on large networks that have scalability issues, thereby leaving dynamic configuration adaptation and real-

TABLE II. SUMMARY OF CYBER DECEPTION TECHNIQUES IN CLOUD ENVIRONMENTS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[90]	Mitigate known and unknown attacks in cloud environments	SQL Injection, XSS, SSH attacks, Malware injection	High/low interaction honeypots	NIDS Module (SNORT), honeypot network module (Glastopf, Dionaea, Cowrie), analysis module (Cuckoo Sandbox), signature module (Rule Generator and Updater)	Reduction in generated alerts, detection of unknown attacks via custom rules, dynamic malware analysis results	Overhead due to multiple NIDS instances not studied, focus limited to network-based attacks, no integration with HIDS, ACLs, or firewalls, performance-security trade-off not quantified
[91]	Mitigate cloud-side attacks (e.g., side-channel) while minimizing migration costs	Side-channel attacks (intra/inter-host), VM co-residency threats	Signaling game-based deceptive VM migration (live/non-live or fake signals)	Network Controller (alerts and migration decisions), VMs (defenders in the game), attacker (monitors VM signals), signaling Module (generates live/non-live migration or deceptive signals)	Cost savings, security gain, equilibrium stability	Focuses on VM-level attacks, ignores host/network-layer threats, no real-world validation (simulation only), unquantified impact of deception on performance, requires attacker belief modeling (p,q), which may not hold in practice
[71]	Mitigate dynamic threats in container-based clouds	Multi-stage lateral movement	Automated honeypot placement using Attack Graph and Deceptive Exploring Surface	Monitor engine (tracks system changes), decision engine (optimizes honeypot placement), deployment handler (configures/ deploys honeypots), honeypot nodes	Attack success rate reduction, resource efficiency, scalability, flexibility	Depends on attack graph accuracy, static assumptions on attacker model
[92]	Disrupt reconnaissance for DDoS attacks (e.g., Crossfire, Coremelt) without affecting users	Reconnaissance phase of DDoS	Cloud-based reflection network: GRE tunnels forward probes to virtual topology, SDN-enabled IP hopping for fake paths, "Reflected" virtual topology mimics physical network	Physical SDN switches, cloud virtual switches, local/Cloud SDN controllers, NFV management host	response time, packet loss, update latency, cost	Requires tight sync between physical and virtual topologies
[93]	Divert and detect internal threats in private clouds	Internal reconnaissance, probe, DoS, User to Root (U2R), Remote to Local (R2L)	Enterprise honeypot with VM-level emulation	Honeypot VMs, Snort, Sebek, MSVM classifier	Classification accuracy, alert generation	Limited to internal threats, dependence on known attack patterns
[94]	Proactively counter reconnaissance attacks in cloud environments	Network scanning, probing, OS fingerprinting, exploitation of common OS vulnerabilities	DRL-based decoy VM placement	DRL agent (PPO), OS vulnerability model, simulated tenants	Deception entropy, attack cost, Joint-Defense Goal	Limited to OS-layer vulnerabilities, scalability
[95]	Detect zero-day ransomware attacks and intruders in cloud environments (IaaS, PaaS, SaaS)	Ransomware (e.g., REvil, WannaCry, NotPetya) and intrusion attempts (e.g., unauthorized access to cloud storage/ servers)	High-interaction decoys (fake files/tokens), low-interaction decoys (canary files/tokens), ANSI/ASCII-named decoy files	Decoy resources (files, tokens, partitions, folders, servers), event handler watcher (monitors interactions with decoys via cloud API hypervisor)	Detection rate, detection time, speed vs. hashing/entropy methods	Lacks compatibility with Linux/MacOS, focus on Windows-based cloud, file system reliance (requires NTFS for positioning technique)
[96]	Monitor and attribute cloud attacks in real time	Botnets, CVE exploits, DDoS	Containerized honeypots, high-interaction (e.g., Cowrie SSH/ Telnet), low-interaction (e.g., Dionaea, ADBHoney) decoys	Azure cloud, ELK stack, Suricata, P0f, FATT, honeypots	Real-time logging, CVE capture, attacker attribution	Manual signature extension, needs zero-day attack adaptation
[97]	Strengthen security in cloud-native microservice architectures by intercepting and mitigating malicious lateral movements of attackers	Lateral movement, container compromise	High-fidelity decoys (clones of legitimate microservices)	Microservice architecture, decoy containers	Computation time, decoy interaction probability, average decoys per attack path	Assumes attack graph availability, computational complexity, node constraints

world adversarial creativity as open challenges for future work. This illustrates a transition from simple redirection to strategic deception in wireless virtualization security.

In [102], the authors suggested counteracting the jamming (which means that an adversary injects noise to decrease of signal-to-noise ratio (SNR) and disrupt communication) attacks on wireless networks using game theory. Their approach concerns a dual channel system, which has a real and a fake transmitter, and a receiver pair, where fake transmissions mislead the jammer into splitting its power of jamming to both channels. The interaction is formulated by a Stackelberg game, in which the system (the leader) performs the optimization of power allocation first, and the jammer (the follower) afterwards senses and reacts to the channel activity. There are two cases discussed: 1) a non-strategic jammer with fixed behavior,

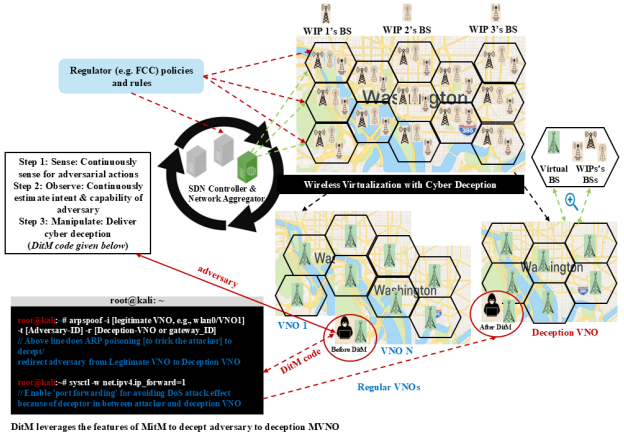


Fig. 6. An example scenario for cyber deception in wireless virtualization in [100]: logical redirection of adversary from VNO N to deception VNO via subleased base station.

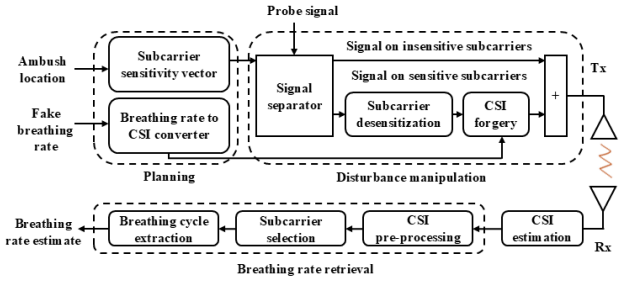


Fig. 7. The flow chart of the proposed HoneyBreath defense in [103].

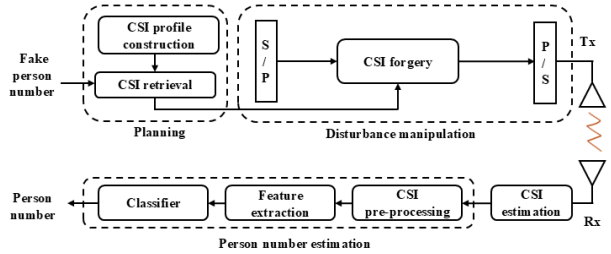


Fig. 8. The flow chart of the proposed Ghost defense in [103].

and 2) a strategic jammer, for which they can establish a Subgame Perfect Nash Equilibrium (SPNE) for optimal power allocation. The simulations validate that the technique reduces the jammed signal and boosts the legitimate channel's transmission rate. However, the work presuming the total efficacy in jammer sensing is limited to just two orthogonal channels and also does not deal with involving scalability in multi-node or dynamic environments. Despite these shortcomings, the paper underscores the fact that power-based deception is a way of significantly increasing resistance to reactive jamming even in simple wireless systems.

In [103], the authors presented two deception-based techniques, depicted in Fig. 7 and Fig. 8 that are used to withstand wireless privacy threats such as the Channel State Information (CSI) that are used by the attacker. HoneyBreath is a method that gives the wrong breathing rate by first hiding the sensitive subcarriers and then sending forged sinusoidal CSI patterns, while Ghost controls all the subcarriers by using previously recorded CSI to fool crowd counting classifiers. These strategies have been applied on USRP X310 platforms and they have given good evaluation results. HoneyBreath has reached the state at which the errors of breathing rate estimation go below 1.2 bpm; on the contrary, Ghost has had complete occupancy detection success. However, they require attackers' location beforehand and pre-collected CSI profiles, which might hinder their use in some dynamic cases. The study illustrates that strategic CSI manipulation could be a viable method for achieving privacy against wireless inference attacks while being largely unnoticed and practical.

A two-part contribution to deception-based defense strategies in wireless sensor networks (WSNs), specifically targeting energy depletion and battery drain denial-of-service (DoS) attacks is presented. The first study [104] presented a deception framework designed to assist cluster heads (CHs) in wireless sensor networks (WSNs) against energy depletion attacks using a game-theoretic approach. The strategy involves both high- and low-interaction honeypots, which are placed in a secure zone, governed by an intrusion detection system (IDS), that counts attackers (sophisticated/ non-sophisticated) and directs them to suitable honeypots. An incomplete information Bayesian game model derives Nash equilibrium strategies for CH energy preservation and attack mitigation. In MATLAB, simulations were carried out, and the approach was validated, which showed a decrease in attack success rates through the reallocating of honeypots dynamically. Major constraints include a static attacker assumption and a sole reliance on predefined honeypot configurations, thus demonstrating the scalability and real-time adaptability issues for the next work. The second study [105] builds on this framework by proposing a lightweight defense mechanism to battery-draining DoS attacks in wireless sensor networks (WSNs) through a signaling game framework where sensor nodes smartly release deceptive energy-level signals in order to misdirect attackers. The authors depicted the relationship among nodes and attackers as a two-player Bayesian game to derive the optimal defense strategies through Perfect Bayesian Nash Equilibrium analysis, showing that the nodes can effectively save energy by trying to conceal the true state when the costs of lying are low. The simulations show that by using the deceptive signaling, the attacker's utility is reduced while the nodes are being protected. However, the approach only considers single-round interactions and does not include coordinated attacks or network failures that could be implemented in future work on dynamic game models and real-world implementation in platforms such as NS3 or Cooja.

Table III summarizes all the discussed works in wireless networks, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

VIII. Cyber Deception in Cyber-physical System

Cyber-Physical Systems (CPS) are systems that consist of computational (cyber) elements and physical processes and are done almost as one, which enables real-time monitoring, control, and interaction with the physical world. These systems are virtually the key to numerous essential fields

TABLE III. SUMMARY OF CYBER DECEPTION TECHNIQUES IN WIRELESS NETWORKS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[99]	Isolate adversarial users in virtualized wireless networks	Unauthorized access, DoS, dynamic attacks	Deception MVNO (DMVNO): a decoy virtual network operator	SDN controller, Wireless Infrastructure Providers (WIPs), legitimate MVNOs, deception MVNO	Blocking probability, attacker arrival rate, deception deployment time	Detection delay (time gap between attack detection and DMVNO deployment)
[100]	Protect virtualized wireless networks from adversarial attacks while maintaining QoS for legitimate users	Impersonation attacks, location falsification attacks	Deceptor-in-the-Middle (DitM): SDN-controlled redirection (legitimate VNOs act as "middle-men" to reroute adversaries to deception VNOs)	SDN controller, wireless infrastructure providers (WIPs), legitimate VNOs, deception VNOs	Detection range, false alarm rate, throughput	Lacks optimal strategy selection mechanism
[101]	Secure virtualized wireless networks by deceiving attackers into interacting with decoy systems	Jamming/DoS attacks, reconnaissance attacks	Stackelberg game-based mapping of true to fake RF bands	SDN controller, wireless infrastructure providers (WIPs), VNO, deceptor VNO	Attacker utility, deception cost, feasibility constraints	Scalability, computational complexity
[102]	Mitigate jamming attacks to protect the transmission of real information between a transmitter-receiver pair	Jamming attacks	Fake signal transmission on orthogonal channel	Twotransmitter-receiver pairs, orthogonal channels, jammer	Transmission rate, optimal power allocation, utility equilibrium	Assumes perfect sensing, static two-channel setup
[103]	Protect user privacy by preventing accurate inference of breathing rates and crowd counts via wireless signals	Channel state information (CSI) wireless inference attacks	HoneyBreath (transmits fake CSI to deceive eavesdroppers into inferring incorrect breathing rates), Ghost (manipulates CSI to fabricate false crowd counts in empty rooms)	Software-defined radio (USRP X310) platforms, CSI profile library, multiple antennas	Breathing rate error, crowd counting	Static or predictable attacker positions, pre-collected CSI profiles
[104]	Protect cluster heads (CHs) in Wireless Sensor Networks (WSNs) from energy depletion attacks	Energy depletion attacks	High- and low-interaction honeypots mimic CHs to lure attackers	Cluster-based WSN with CHs and member nodes (CMs), secure zone with IDS, honeypots	Nash equilibrium analysis, probability of attack strategies, probability of defender strategies	Static attacker beliefs, single-round interaction
[105]	Mitigate battery-draining DoS attacks in WSNs	Battery drain DoS attacks	Bayesian signaling game with deceptive energy state broadcasting	Regular node (sender), attacker (receiver), signaling framework	Attack probability, PBNE types (pooling/separating)	Does not address cooperative attackers or network failures

from self-driving vehicles to energy and health care, factory robots, and industrial automation. A typical CPS is structured in the following form: sensors and actuators interface the system with the physical environment, embedded controllers are responsible for executing control algorithms, networked communication is in charge of data exchange; cloud or edge infrastructure are resources for computation and analytics. More CPS, in general, are more vulnerable to attacks as they are focusing on wireless communications, open protocols, and internet connectivity. This entails the introduction of such vulnerabilities as time-line constraints, limited computational resources, insecure firmware, and insufficient authentication. Threats specific to CPS can be the origin of physical damage, privacy issues, or huge business interruption. Some of the most notorious cases, like remote hijacking of vehicles or manipulation of smart grid components, are a lesson on how traditional security practices can't protect CPS satisfactorily by themselves [106]. Cyber deception finds a place here as a likely solution by putting the attacker off track with signal, data, node misdirection, or wrong behavior models that divert the attacker's perception and planning. This Disinformation can slow down enemies, divulge intelligence operations, and aid in adjustments to threats. However, the usage of untruth in CPS is complicated as the systems' timely response and safety-criticality requirements make it hard, and they need to be taken into account during the design phase. Moreover, the fact that adversaries are sophisticated enough to detect badly implemented disinformation is an additional inconvenience. This section presents cyber deception methods contextualized to CPS environments with an emphasis on deception-aware system design, attacker engagement strategies, and domain-specific constraints.

In [107], the authors suggested the use of a deception-as-defense framework for CPS, in which strategic misinformation is utilized to change the perception of rational adversaries who are trying to estimate the states of the system. The interaction is modeled as a stackelberg game, in which the defender (leader) first of all commits to a signaling policy while the adversary (follower) reacts optimally. The paper shows the results under Gaussian-distributed system states and quadratic cost functions that linear signaling rules (which include optional noise) are the ideal means to indirect adversarial behavior without direct enforcement. The solution, obtained through semi-definite programming (SDP), is also valid for dynamic situations, partial observations, and unknown attacker goals, and it is robust under both Bayesian and non-Bayesian uncertainties. The study provides a foundational game-theoretic approach in protecting CPS roads against adversaries, focusing on the indirect channel perceptual control, which is a key for resilient systems that require both stealth and accuracy.

In [108], the authors introduced a honeypot-based decep-

tion framework for CPS that optimizes the defensive strategies model under limited human analysis resources by modeling attacker-defender interactions, which are represented as a Bayesian game with missing knowledge. The proposed method involves the use of both low- and high-interaction honeypots integrated into the CPS cyber layer, and using the derived optimal configurations for defense and resource allocation to ensure the highest defensive payoff, proving the inclusion of several Bayesian-Nash equilibria under different analysis budgets. Through the game-theoretic approach, they clarify the distinction between the types of attackers (weak vs. strong offensive access) and defense mechanisms (service provision vs. non-provision), thanks to the simulations they carried out, which revealed that turning attention to the honeypots that provide a greater unit analysis utility increases the efficiency of the capturing process. Up to now, the framework has improved the performance of practical deception in CPS, though it is believed that the strategies of attackers should remain static and interactions are to be single-stage, it has put off the adaptation to dynamic and multi-stage attacks for the upcoming tasks.

In [109], the authors presented a honeypot-based intrusion detection system, specifically for CPS, which addresses threats such as spoofing, code injection, malware, and DoS that can disrupt the ordinary physical operations. The architecture they used diverts all the traffic into a virtual or hybrid honeypot that is placed in a strategic location and exploits the high availability of real-time attack detection, logging, and blocking, thus achieving a cost-effective and scalable deployment (e.g., Cisco, Conpot). The work makes a comparison of the interaction level, scalability, and resource requirements of the various honeypot types, which is a good point to present them as the main instruments for the early threat detection of CPS systems, given the fact that conventional methods often fail in real-time conditions. Nevertheless, the present version is applied to single systems only, and there is a need for multi-device testbed with dynamic traffic patterns, which would be the practical deception-based approach for the security of CPS.

In [110], the authors presented a resilient and smart technique to identify the honeypots in softwarized industrial CPS that largely depend on virtualization and network function slicing. Its main architecture combines a secure fuzzy testing approach with a deep learning-based classification model to address the problem of recognizing real devices from honeypots in heavily virtualized network environments. The first module of the architecture, as illustrated in Fig. 9, creates probe packets using multi-object mutation strategies based on security rules, which guarantee that probe queries are safe but rich in features that are not handled correctly between real devices and honeypots. These are the optimized packets that will be used to scan online devices. The additional responses that are returned are then analyzed for features that are found only in the honeypots and are classified by a CNN model that has been trained before. The overall design provides the packet creation and identification phases separately in order to enhance the scalability and minimize risks. Tests carried out demonstrate that, in addition to being fast and safe, the method has outshone the usual identification techniques by achieving both higher precision and lower false positives. The fuzzy testing module conveniently generates discriminative features while the CNN model provides trustworthy classification. This paper proposes a flexible and safe method to the disturbing problem of deception-aware adversaries detecting the honeypots in industrial CPS, but it has the assumption of pre-labeled data availability and the knowledge of error-handling signatures that can limit adaptability in dynamic or heterogeneous environments.

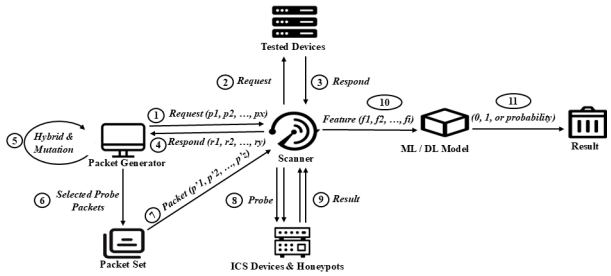


Fig. 9. The architecture of industrial CPS honeypot identification in [110].

In [111], the authors introduced a two-layer deception strategy for CPS. By depicting defender-attacker interactions as a signaling game with incomplete information, the model allows coordinated deception to be implemented both at the application layer (e.g., fake user interfaces) and at the network layer (e.g., falsified messages), which secures detection avoidance through consistency. The defender's strategies, communicating either honest or misleading signals, are improved using Perfect Bayesian Nash Equilibrium (PBNE), while the attacker checks the system type (normal/ honeypot) to determine whether to attack. Experiments carried out in an automotive CPS setting proved that two-layer deception not only increased the uncertainty of the attacker but also brought down the payoffs, and made the cost-deterrence more effective in comparison to single-layer techniques. This work, which is based on the Purdue Enterprise Reference Architecture, is the first to scientifically model multi-layer deception in CPS through game theory, thus presenting a solid framework against such attacks.

Table IV summarizes all the discussed works in cyber-physical systems, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

IX. Cyber Deception in Industrial Control Systems

Industrial Control Systems (ICS), which are specialized cyber-physical systems, are designed to monitor and control the operations of industries such as manufacturing, energy, water, and transportation. The components of ICS include SCADA systems, Distributed Control Systems (DCS), and Programmable Logic Controllers (PLCs). An ICS architecture typically consists of field-level devices, controllers (PLCs/RTUs), supervisory systems (SCADA), HMIs, and communication networks that use industrial protocols such as Modbus and DNP3. As ICS were originally designed for isolated, deterministic environments, they have gradually become interconnected with IT networks, which has led to the problems of legacy components, weak authentication, and insecure communication protocols. They are exposed to sophisticated cyberattacks

TABLE IV. SUMMARY OF CYBER DECEPTION TECHNIQUES IN CYBER-PHYSICAL SYSTEMS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[107]	Control adversarial actions to align with system objectives by crafting information strategically	Advanced adversaries seeking system-related information	Strategic signaling (linear or "linear plus noise" rules) to manipulate adversary perception	Sender and receiver agents, Gaussian information, Quadratic cost functions	Covariance of posterior estimate, semi-definite programming (SDP) equivalence, stackelberg equilibrium outcomes	Limited to Gaussian information, non-classical information schemes in control settings, adversaries are rational
[108]	Optimize defensive strategies against network attacks in CPS by balancing honeypot deployment and human analysis costs	Network attacks (e.g., probing, exploitation of vulnerabilities)	Low- and high-interaction honeypots	Cyber layer (control systems, communication networks), physical layer (sensors, actuators, plants), honeypots (low/high-interaction), firewall, router, servers	Bayesian-Nash equilibria, defense payoff maximization, escape probabilities of attacks, human analysis cost allocation efficiency	Static attackers, static deployment, dynamic adaptation not addressed
[109]	Detect and block CPS-targeted cyberattacks using honeypots	DoS, spoofing, control hijacking, malware, code injection	Low- and high-interaction honeypot-based intrusion detection	Cyber layer (sensors, network), physical layer (actuators, processes), honeypot (virtual/physical), security rules engine, traffic monitoring system	Attack capture rate, scalability, interaction level	Single-system use, Static security rules may not adapt to evolving threats
[110]	Detect honeypots to protect deception infrastructure integrity	Honeypot fingerprinting in softwarized CPS	Fuzzy probe generation + CNN-based honeypot classification	Secure Fuzzy testing module, scanner, response analyzer, deep learning model, dynamic packet set	Accuracy/ precision/ recall, effectiveness, diversity of response packets	Labeled training data, focused on Modbus protocol
[111]	Strengthen CPS resilience by implementing a dual-layer deception strategy	CPS attacks across application/ network layers	Coordinated multi-layer deceptive signaling	Coordinated signal senders (application/ network), belief update model	Defender's payoff, deterrence effectiveness	Increased deployment complexity and cost, fixed attacker beliefs

like unauthorized command injection, manipulation of control logic, stealthy data tampering, and firmware exploitation. Stuxnet, Triton, and BlackEnergy incidents are examples of severe physical disruption and operational downtime through intentionally attacking ICS. Conventional security mechanisms generally do not have the contextual awareness, adaptability, or low-latency operation required in real-time ICS settings, and in addition, frequent patching as a remediation is often impractical due to availability and safety constraints [112]. In response, cyber deception is a good choice not only because it is a proactive and stealthy approach but also it is a strategic alternative such as inserting misleading elements, decoy sensors, fake field devices, deceptive control logic and protocol-emulating honeypots, in the ICS environment to mislead adversaries, trigger early alerts, delays attacker progress, and aid forensic investigations. Design of deception in ICS is a difficult case due to the necessity of preserving protocol fidelity, safety, and the need for realism under the threat of expert attackers. This section explores deception strategies specifically tailored to ICS, emphasizing their architectural integration, effectiveness against known threat models, and deployment limitations.

In [113], the authors suggested the implementation of Honeyd+, an applicable, scalable, and cost-effective implementation of high-interaction ICS honeypots through only one PLC and a proxy-based engine. For example, by changing the network/device identifiers (IP/MAC addresses, hostnames, serial numbers) dynamically and manipulating protocol payloads, Honeyd+ can simulate multiple physical PLC devices while preserving the culture of identity and realism. Tests on Raspberry Pi and high-end laptops certainly indicated its capacity to register more than 75 virtual networks, while, under other moderate traffic, it maintained low error rates, making it effective for reconnaissance deception. Though weaknesses include no support for formatted encrypted protocols, scalability constraints due to PLC features, and poor performance under extremely high traffic. The work shows possible improvements such as protocol-specific adjustments and the integration of diverse ICS components, to extend it to the production environment.

In [114], the authors presented HoneyPLC, a high-interaction honeypot that closely simulates programmable logic controllers (PLCs) in ICS. In contrast to conventional ICS honeypots such as being low interaction fidelity, limited protocol support, and poor malware capture, HoneyPLC, as illustrated in Fig. 10, is equipped with a modular framework that consists of a PLC profile repository, a personality engine (for protocol spoofing and fingerprinting), network services (HTTP, SNMP, S7comm), and an interaction data module to log attacks and capture malicious ladder logic injections. The key innovation is a PLC profiling toolkit that dynamically, for example, configures the Nmap fingerprints and SNMP MIBs to imitate real devices such as Siemens S7-1200 or Allen-Bradley MicroLogix 1100. Beyond that, HoneyPLC's S7comm server and TCP/IP stack operate realistically to trick tools such as Nmap, PLCScan, and Siemens Step7 Manager, whereas its ladder logic capture module records the injected malware for future analysis. The tests demonstrate that there is a high fidelity (>90% Nmap Confidence, Shodan scores matching real PLCs) and scalability (75+ simulated hosts per PLC). It is worth mentioning that the main shortcomings are the lack of physical process simulation and the fact that there is no support for encrypted protocols (e.g., Modbus Secure).

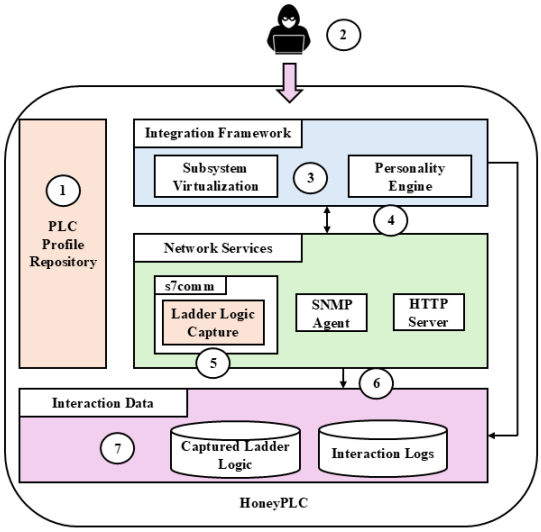


Fig. 10. The architecture of HoneyPLC in [114].

In [115], the authors introduced two deception methods for the identification of stealthy malware in the ICS, which avoid the traditional sensors through the exploitation of internal host discovery mechanisms such as NetBIOS browse lists and ARP tables. The methodology is based on the addition of sensor information to these lists in a way that causes malware to be attracted to self-disclosure. Proposal 1 is aimed at malware such as Conficker through the broadcasting of fake NetBIOS advertisements that are continuously emitted, which makes sure that the sensor's hostname is listed in the network browse lists, a tactic that is most advantageous under Windows-based systems, particularly in old versions of the program. In contrast, Proposal 2 incorporates the scheme which ensures that the sensor's IP-MAC pairing is embedded in the ARP caches of all hosts by sending cyclical ICMP or ARP requests, which, in turn, counteract malware that uses ARP tables for lateral movement. The techniques are designed for simple and non-intrusive application in the vulnerable ICS networks. The evaluations validate that Proposal 1 was successful in identifying Conficker in a laboratory setting, whereas the practicality of Proposal 2 is denoted in a theoretical proposal, which remains to be verified with real ARP-based malware. Despite their efficacy, the techniques, on the other hand, are based on assumptions relating to the behavior of the attacker, and involve a need for trial-scale deployment, decreasing the number of false positives, and integration with further detection frameworks.

In [116], the authors introduced a deception defense framework for ICS leveraging adaptive Hidden Markov Models (HMM) for the detection and mitigation of cyber threats, which include DDoS attacks, malicious program injection, and false data injection. The framework models typical ICS behavior using multidimensional data and detects anomalies by comparing observed behaviors with expected HMM state sequences. It adapts to evolving threats through incremental updates of HMM parameters. While primarily focused on anomaly detection, the framework, as depicted in Fig. 11, integrates a passive form of deception by analyzing attacker interactions within virtual environments, enabling misdirection without exposing the real system. Detected attacks (e.g., TCP SYN floods, ladder logic injections) trigger appropriate responses such as node isolation or administrator alerts. Through experiments conducted on an OpenPLC-based testbed, the solution was found to have very high rates of detection (98% for DDoS, 95% for malware injection, and 90% for false data attacks), very low percentages of false alarms (1-3%), and quick response times (1-2.7 seconds). Whereas the outcome of these tests was successful, it is mentioned that scalability and denial of multi-stage attack coverage are the shortcomings, which place emphasis on broadening the scope of the framework's deployment in larger platforms and its interfacing with other security orchestration platforms.

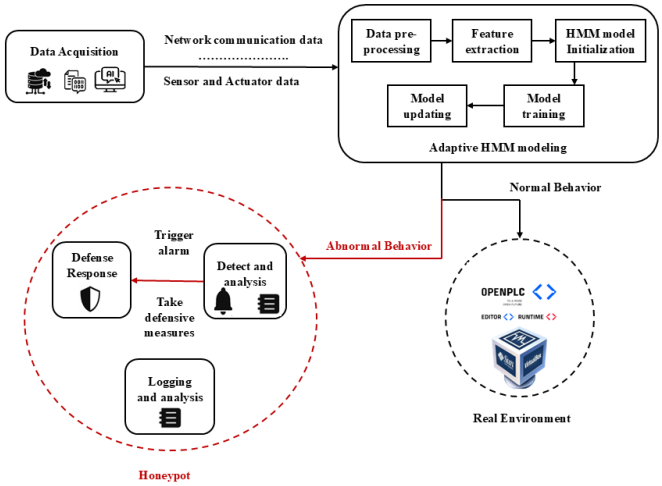


Fig. 11. The adaptive HMM-based ICS deception framework in [116].

Table V summarizes all the discussed works in industrial control systems, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

X. Cyber Deception in Smart Grids

Smart grids are the modernization of traditional electrical networks by integrating advanced information and communication technologies to make power systems more efficient, reliable, and sustainable. The systems comprise components like smart meters, sensors, automated control systems, and communication networks that together deliver the real-time monitoring and control of energy flow innovation. However, the high degree of networking and the dependence on digital technologies have made the trouble spots more and more numerous, and that is the reason the smart grids have become vulnerable to a strong variety of cyber threats including for example, false data injection attacks, denial-of-service attacks, and embedded malware attacks. The cyber-attacks on Ukraine's power grid, which were particularly prominent cases, showed the high impact of such vulnerabilities, bringing about the consequences of significant grid disruptions as a result of that so demonstrating the potential for major disturbances. The traditional measures of cybersecurity are often inefficient in the face of the ever-changing, elusive characteristics of these

TABLE V. SUMMARY OF CYBER DECEPTION TECHNIQUES IN INDUSTRIAL CONTROL SYSTEMS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[113]	Create low-cost, authentic, and targetable honeypots for Industrial Control Systems (ICS) to detect and analyze attacker TTPs	Reconnaissance, enumeration, initial access, and persistent attacks (e.g., via TCP/UDP port scanning, protocol exploitation)	Proxy-based high-interaction honeypots with Honeyd+	Physical PLC, Honeyd+ host, search/ replace Engine, evaluation platforms	Cost-effectiveness, error rate, scalability	Limited support for encrypted protocols, high load handling, PLC bottlenecks
[114]	Understand attack methods targeting Programmable Logic Controllers (PLCs), by deceiving attackers into interacting with a honeypot	Malware injection (e.g., Stuxnet, Crashoverride), reconnaissance, and disruption of physical processes	High-interaction honeypot simulating real PLCs with advanced protocol simulations	PLC profile repository, personality engine, network services module, interaction data module, PLC profiler tool	Stealthiness, compatibility with proprietary tools, ladder logic capture	Limited protocol coverage, lack of physical interaction modeling
[115]	Detect malware infections in Industrial Control Systems (ICS)	Malware infections (e.g., WannaCry, Conficker)	Embeds sensor host names in browse lists, embeds sensor IP addresses in ARP tables	Sensor, host announcement module, ARP/ICMP module, alert system	Detection capability, false positives, coverage vs. traffic	Deployment impact (Proposal 2's continuous ARP requests may affect network performance), evasion by advanced malware, protocol support
[116]	Protect ICS from cyber-attacks by detecting and responding to anomalous behavior in real-time	DDoS attacks, malicious program injection, false data injection	Adaptive Hidden Markov Model (HMM) for modeling normal behavior and detecting anomalies	Data acquisition, adaptive HMM, honeypot-based response, OpenPLC testbed	Detection rate, false positive rate, response time	Dependence on HMM training data, multi-stage attacker handling

threats [117]. For that reason, a new type of defense was introduced, which involves cyber deception. By using deceptive elements like honeypots, decoy systems, and wrong data, cyber deception aims at disorienting attackers, slowing them fast, and earlier recognizing malicious activities. The implementation of cyber deception technology in smart grids, however, gives rise to difficulties, such as the assurance of component deceptiveness, maintenance of system performance, and seamless integration of deception strategies into the already existing infrastructures. In this section, the subject of cyber deception in smart grids is studied by considering methods, advantages, and limitations that the implementation brings.

In [118], the authors discussed the application of honeypot-based defense scheme to deal with the Distributed Denial of Service (DDoS) attacks aimed at the Advanced Metering Infrastructure (AMI) in the smart grids using a game-theoretic approach. The model suggested by the authors envisions the AMI as a tree like network consisting of smart meters, aggregators, and a central headend system. It is shown in Fig. 12 that the honeypots are to be embedded inside the firewalls of the WAN segment to make the traffic divert and to secure the headend and the back office servers. This strategy is framed formally as a Bayesian honeypot game, considering the presence of the users, the attackers, and the service providers, which include the strategies of the attackers, the effectiveness of the honeypots, and the anti-honeypot mechanisms that can be utilized. The model gives the specification of the payoff function and the resolution of the problem, finding multiple Bayesian Nash Equilibria (BNEs), which typically imply the optimization of the placement of honeypots so as to reduce energy consumption and improve detection rates. In the process of verification by experimental setups employing an OPNET-based AMI testbed, it has been established that the model surpasses the others in terms of higher detection rates (up to 85%) and more stable energy efficiency. The model operation deals with what is typically assumed as a rational and static attacker and simply presumes a deployment scenario.

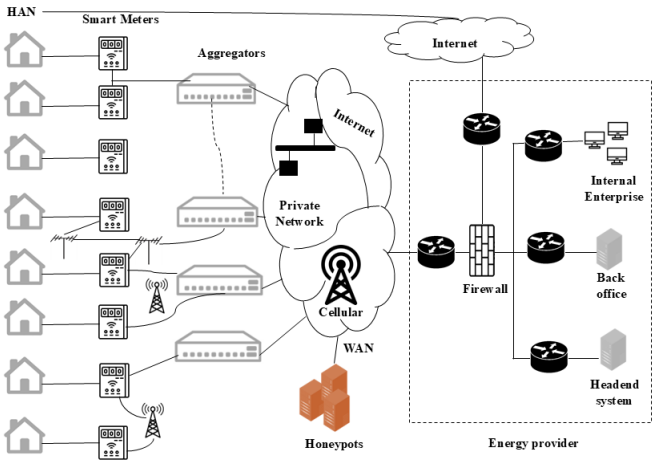


Fig. 12. The AMI network infrastructure with honeypot deployment in [118].

In [119], the authors thoroughly discussed and analyzed the honeypot and honeynet technologies that are utilized in the smart grid for security purposes. They have done an analysis of the factors such as the interaction levels (low versus high), purposes (production versus research), and deployment models (virtual versus physical) while evaluating the protocol supported by the industrial standards such as Modbus, IEC-104, and IEC-61850. The investigation has listed Conpot as the most versatile because of its modular design and multi-protocol capabilities as compared to twelve honey-x solutions. Even though Conpot is lacking in native support for advanced protocols like GOOSE/MMS, DNP3, it is the most versatile

choice. The EU SPEAR project has set up a smart home testbed, which implemented a practical example of Conpot. Conpot has been proven to be effective by applying modified iptables rules and JSON-based logging in the process, while also outlining its limitations in high-interaction fidelity, IPv6 adoption, and dynamic threat adaptation.

A two-phase investigation into enhancing threat intelligence collection for smart grid systems using honeypots was conducted. In the first study [120], low-interaction honeypots were implemented across five AWS regions (Singapore, US, Canada, Germany, Brazil) to simulate ICS protocols IEC 61850, IEC 60870-5-104, DNP3, Modbus TCP, and BACnet. They succeeded in remaining undetected by fingerprinting tools like Shodan's Honeyscore while collecting a total of 6GB of network traffic over six months. Their analysis shows indicators of probing behavior specific to the protocol, geographical correlations on scanning (with 0.97 cross-correlation between Germany and Brazil instances), and the presence of persistent threat actors such as 54 IPs that were detected monthly and executed SYN-flood attacks in a coordinated manner, all of which provided a valuable reference for IDS and firewall tuning. On the other hand, the study has some limitations, such as the fact that the honeypots are of low interaction, which leads to a lack of realistic behavior, and that post-compromise attacker activities are unable to be captured, which is a strong reason to think about the use of more complex, high-fidelity honeypots in other smart grid security studies. To address these limitations, their second study [121] demonstrated a realistic smart grid honeypot system by creating a full-scale infrastructure that includes SCADA HMIs, historians, IEDs, protocol gateways, and firewalls. It employs a decoupled architecture that is integrated with various tools such as Honeyd for anti-fingerprinting, Cowrie for SSH deception, SoftGrid for power-flow simulation, and Mininet for virtual IEDs. The system achieves practical protocol spoofing (e.g., IEC 61850 MMS) and cyber-physical consistency through these tools. Security testing has proven the system's capability to fingerprint devices strongly, log multi-phase attacks by using transparent proxies, and be resistant to probing by adversaries. Yet, difficulties persist in lateral movement modeling, automating dynamic deception, as well as adapting to the strategies of the attacker.

In [122], the authors proposed DecIED, a scalable deception framework for smart grid substations that is based on the idea of k-anonymity to conceal the identity of actual Intelligent Electronic Devices (IEDs) by introducing several indistinguishable decoy IEDs. Tailored for the IEC 61850-compatible environments, DecIED mimics the actions and communication patterns of the genuine devices by implementing the protocols, such as MMS, GOOSE, and SV. The outline, as shown in Fig. 13, involves a virtual IED program which consists of an IED server and logic, GOOSE/SV subscribers and publishers that are all integrated with Honeyd to spoof OS fingerprints and port behaviors. Each decoy IED is provided a unique IP/MAC address, and by following the same timing and content of the messages, it can replicate real devices. The system is designed to be resource-efficient, being able to commit over 200 deception instances on a single industrial PC; thus, scalability is achieved without compromising fidelity. Evaluations prove that deception was achieved without the capacity of real devices in passive and active reconnaissance. Indeed, DecIED out of the box maintained realistic control logic responses and nevertheless succeeded in the deception of scanning tools. Though issues prevail in the process of really exacting OS-level functionality and diverse logic types problems not addressed, yet DecIED is a solid ground for deception in today's substations and, it can be a stepping stone for further tactics like moving target defense.

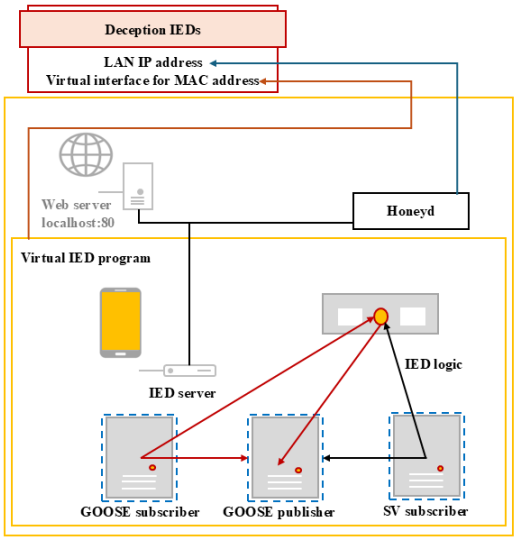


Fig. 13. The module architecture of DecIED in [122].

In [123], the authors proposed a defense strategy framework for smart grids with the help of the honeypot model. This is made with an incomplete information stochastic game that is used for describing the dynamic interactions between the attacker and the defender. The system consists of low-interaction and high inside honeypots in the networks like smart grid servers (web, FTP, OPC, and application servers), where it is isolated through firewall rules. A six-tuple game model that describes the defense mechanism is set up partly by attack/ defense actions, privilege escalation states, transition probabilities, and host-specific metrics (e.g., importance and risk level). The model of Nash equilibrium is obtained by solving it through Gambit software, and as a result, the framework comes up with state-optimal defense strategies that enable adaptation to the attacker through privileges or honeypots. The evaluations show that they achieved good deception rates and resource allocation, yet the model was derived assuming the topology is static and the state transitions are known perfectly, which in turn limits its scalability in a dynamic environment.

Table VI summarizes all the discussed works in smart grids, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

XI. Cyber Deception in Internet of Things

The Internet of Things (IoT) is one of the numerous interconnections that link up physical objects, devices such as sensors, actuators, gateways, and cloud-based analytics platforms, which will work together to monitor, control, and automate different domains like healthcare, smart homes, industrial systems, and smart cities in real time. Rendering a typical IoT

TABLE VI. SUMMARY OF CYBER DECEPTION TECHNIQUES IN SMART GRIDS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[118]	DDoS mitigation in AMI networks	Distributed Denial of Service (DDoS)	Honeypot-based game-theoretic decoy placement	Smart meters, data aggregators, central head-end, honeypots behind firewalls	Detection rate, energy efficiency, BNE solution space	Rational attackers, limited to small-scale testbeds
[119]	Secure smart grids by diverting attackers, collecting threat intelligence, and preserving forensic evidence	Broad coverage, including SCADA exploits, DDoS, and malware targeting ICS protocols (Modbus, IEC-104)	honeypots, honeynets	Emulated ICS devices (PLCs, RTUs), protocol servers (Modbus, SNMP), virtual/ physical honeypot nodes, tools (Conpot, GridLAB-D (power flow simulation), Mininet (SDN))	Detection Rate, realism scalability	Protocol gaps, resource intensity, dynamic threats
[120]	Collect threat intelligence on smart grid attackers	Scanning/probing, SYN-flood DoS, protocol specific exploits (IEC 60870-5-104, DNP3, Modbus)	Low-interaction honeypots emulating ICS protocols	AWS instances, TCP listeners, wireshark, ELK stack	Packet counts (TCP/ICMP), geographic correlation of attacks, source IP persistence	Lacks high-fidelity engagement, low-interaction design limits realism of emulated devices, limited protocol coverage
[121]	Collect threat intelligence, delay attacks, and deceive attackers into engaging with the honeypot	Attacks via VPN/SSH/RDP, malicious IEC 60870-5-104/IEC 61850 commands, lateral movement	High-fidelity honeypots with OS fingerprint spoofing and decoupled logging	Control center (SCADA HMI, Historian DB, VPN Server), substation gateway, Virtual IEDs, external (Jumpbox, transparent proxies)	Realism (Nmap/P0f fingerprint similarity to real devices, detection of virtualization), logging (captured attack phases)	Passive devices: IEDs (passive servers) easier to mimic than active devices (PLCs/ gateways)
[122]	Detect reconnaissance activities, prevent attackers from identifying real IEDs	Passive monitoring, active probing, lateral movement	*k*-anonymous smokescreen (decoy IEDs)	Honeyd (anti-fingerprinting), virtual IED module, Nginx web server, transparent proxies	Indistinguishability ratio, device similarity, scalability	Imperfect OS fingerprinting, requires access to real IED logic for full imitation
[123]	Optimize defense strategies against smart grid attacks	Multi-stage attacks with privilege escalation	Honeypot deployment guided by stochastic game model	Web/ FTP/ OPC/ application servers, honeypots, firewall rules, attack/ defense action sets	Game-theoretic Nash equilibrium, host significance, security risk level, state transition probabilities, fefense success rate	Convergence time, requires predefined vulnerability data

architecture model, one can think of it as a layered framework, for example, comprising the perception layer (sensors and actuators), the network layer (Wi-Fi, Zigbee, 5G), and the application layer (cloud or edge services). Be that as it may, IoT systems are attacked with major security problems such as device heterogeneity, resource limitations, root defaults, and lack of standardization. Such weaknesses permit IoT networks to come under attack from an array of cyber threats, such as device hijacking, eavesdropping, man-in-the-middle attacks, and DDoS amplification through botnet devices like Mirai. In such constrained and distributed environments, traditional defense mechanisms, like encryption and authentication, often seem to be inefficient [124]. Therefore, the cyber-deception approach has been chosen as the effective way. Now defenders deploy honeypots, fake sensors, deceptive APIs, or synthetic data flows to mislead the adversaries, gather threat intelligence, and delay attacks. Among the recent frameworks that have emerged are AIIPot and IoTFlowGenerator that exploit machine learning to synthesize realistic decoy behavior and traffic patterns; thus, they increase IoT deception fidelity. However, IoT is full of particular challenges in the deployment of effective solutions for example the need for major realism at scale, load overhead maintenance on the constrained devices, and the need for seamless integration of deception into the edge/cloud management platform. This section is focused on the deception techniques that are designed specifically for IoT environments, mainly dealing with component-level placement and adaptive threat engagement.

In [125], the authors introduced HoneyCloud, which is a scalable honeypot framework that is meant for analyzing fileless attacks on Linux-based IoT devices. The setup employs four hardware honeypots (like Raspberry Pi, BeagleBone) and 108 software honeypots distributed over eight public clouds. Compared to other hardware honeypots, they are superior in terms of high-fidelity interactions. However, the cost and maintenance overhead issues drove the authors to use QEMU-based software honeypots running OpenWrt, which were enhanced with realism-preserving techniques (like CPU masking, bus emulation). Over 12 months, HoneyCloud amassed 26 million attacks, including 1.5 million fileless attacks, which made it possible to create the first taxonomy of IoT fileless threats such as credential theft, configuration tampering, and SSH tunneling. The major findings illustrate that 65.7% of fileless attacks utilize default shell commands (like rm, kill, passwd) and demonstrate a trade-off: read-only filesystems are resistant to malware persistence, but they also inhibit shell-history auditing for fileless detection. The paper advances IoTCheck, a workflow for defense that entails the hardening of devices, though shortcomings are the inability to decrypt SSH tunnels and absence of support for newer IoT interfaces (like ZigBee). HoneyCloud broadens the perception of stealthy, non-malware threats, as these are the kinds of threats that can evade traditional defenses.

In [126], the authors analyzed empirically the IoT botnet

behavior by using three medium-interaction Cowrie honeypots programmed to emulate SSH/Telnet services and IoT file systems. The honeypots (which were scaled with Apache/Postfix services and ELK stack logging) observed for more than 40 days and recorded malware downloads during login sessions, which were later clustered by the Levenshtein edit distance on command sequences. According to this study, the majority of sessions that were related to malware were from Mirai variants at 97% with the loader patterns being two dominant ones (74.3% and 23.4% prevalence). The authors can confirm that these specific IP addresses operate within a small group of loaders distributing various strains of Mirai, which are often just slightly modified (e.g., by using UPX packing, filler commands, or compiler flags). Furthermore, the system registered some threats that were scarce (e.g., IRC botnets), although the limitations included Cowrie's static emulation (which was lacking high-interaction behaviors), being biased towards x86/64 architecture, and not being able to analyze non-file-downloading sessions.

In [127], the authors presented a scalable and hybrid architecture for IoT honeynet, which consists of three types of honeypots to enhance attack capture and deception fidelity: a medium-high interaction honeypot emulating the CVE-2017-17215 vulnerability, a high-interaction honeypot using real IoT firmware to handle unprocessable requests, and a multiport honeypot simulating SOAP services across exposed ports (e.g., 37215, 52869) with dynamic banner generation based on real device fingerprints. As shown in Fig. 14, the system is distributed over several physical nodes, each one being managed by a local master program and coordinated through a central control node. The architecture takes advantage of Docker both for fast deployment and isolation, being complemented with self-check mechanisms that ensure honeypot stability. Traffic is first routed to medium-interaction honeypots and deployed to high-interaction nodes if a deeper emulation is required. Deployed in the US and Canada, the honeynet remained stable for six months, capturing 332 unique attacker IPs and malware samples that went undetected by VirusTotal (e.g., "Ixa", "gyv"). Geographical analysis disclosed that scanning behaviors stemmed mainly from Japan and injections from the US. Although they are effective, the authors still mark the limitations in the automation issue and post-compromise behavior modeling and propose future work in these areas.

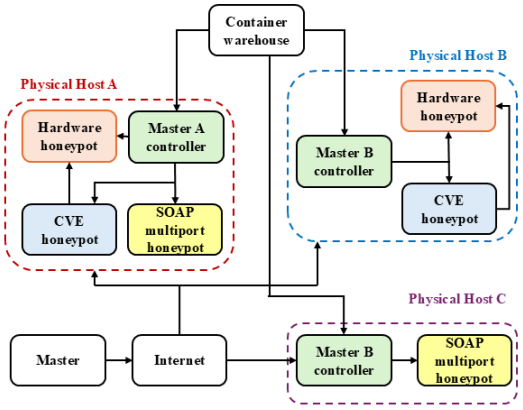


Fig. 14. The composite honeynet system architecture in [127].

In [69], the authors introduced a proactive defense framework, which is a software-defined networking (SDN) enabling technology specifically developed to combat Distributed Denial-of-Service (DDoS) attacks in IoT networks by using the integration of Moving Target Defense (MTD) and cyber deception techniques within a programmable architecture. The framework, as illustrated in Fig. 15, consists of three main parts: the first one is the monitoring agent who is responsible for the detection of attack behaviors, and the second one is the decision module which utilizes a multistage signaling game for modeling the attacker-defender interactions and predicting the adversarial strategies, and the last one is the deployment module that carries out such deception tactics such as random traffic delays, rerouting, and IP randomization. The system can deceive attackers through the use of a switching mechanism that dynamically changes the network properties and thereby alters the perception of the attacker. The manipulation includes operating under true services or honeypots depending on the behavior and confidence of the attacker. The signaling game analysis generates Perfect Bayesian Nash Equilibria (PBNE), with the Optimal Strategy Algorithm (OSA) on top of that, which is used to optimize defense actions that are under the condition of uncertainty. Performance evaluations on both Mininet and a real SDN testbed also show that the improved method, besides survival rates, has controlled CPU usage in the range of 65%, and reduced latency and packet loss when compared to baseline measures. Although it is a very efficient framework, the improvement should be the deployment of security mechanisms and assets that are adaptable and proactive to combat insider threats, which are usually tough to differentiate from legitimate users.

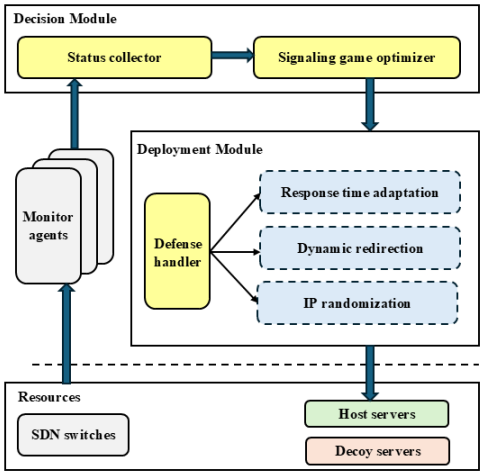


Fig. 15. The proposed framework in [69].

In [128], the authors introduced a cyber deception framework especially designed to fight epidemic botnet propagation in IoT networks using game-theoretic approach. They represented the interaction between attackers and defenders as a zero-sum, one-sided Partially Observable Stochastic Game (OS-POSG). In the proposed model, the defender places some honeypots strategically to monitor and disrupt botnet communication, and the attackers, in turn, know everything about the network structure except the actions of the defender, and their primary goal is to maximize the number of infections. The

defender's partial observability comes from being notified of the reactions of IoT users only (e.g., patching or password changes). Two deception strategies are examined, which are: a randomized deception strategy (RDS) and a k-smart deception strategy (k-SDS), where honeypots are positioned based on the infected devices. The attacker uses straightforward strategies (unicast, half, or broadcast) instead of Q-learning. Simulations on synthetic IoT topologies measure the maximum proportion of infected devices and Time to Extinction (TTE). The results show that k-SDS is significantly better than RDS, especially in high degree connectivity networks. The framework is basically an affirmation of the power of the dynamic deception that is applicable under partial observability. Further actions include the derivation of Nash Equilibrium strategies and the exploration of state-dependent deception techniques.

In [129], the authors presented HoneyComb, a cyber deception framework that has a proactive character and combines a /8 darknet network (16.7 million IPs) with mitigating IoT malware forensic artifacts at scalability. The main distinguishing feature of HoneyComb from the traditional honeypots, which passively wait for attacks, is that it actively detects malware-infected IoT devices by inspecting unsolicited darknet scans and sends then crafted TCP SYN-ACK packets to exploit stateless scanning flaws in IoT malware (e.g., Mirai's dst.IP == TCP.seq signature). The structure, as illustrated in Fig. 16, contains: 1) a darknet traffic analyzer (CAIDA data), 2) a cloud-based honeypot interface, and 3) a pool of virtualized ARM-based IoT devices (OpenWrt) emulating Telnet/SSH services. Once the connected infected devices are tricked, HoneyComb logs malware binaries (1,398 unique URLs), HexString dump commands, C&C communications, busybox commands, and credentials. Throughout a 48-hour evaluation period, HoneyComb reported a total of 1.4 million interactions, which were successfully carried out by 37,323 infected devices, attributing them to 40+ malware variants (e.g., Hajime, Mozi, ZHTRAP). Noteworthy statistics show a 22.07% response rate (37K/169K scanned IPs) and 90% of responses in 750ms or less. Some of the limitations are: 1) ARM-only virtualization (excluding MIPS/x86), 2) low SSH engagement (5.55% vs. Telnet's 25.88%), and 3) proxy/NAT evasion, and improvements will support a broader range of CPUs, simulate other services, and improve proxy penetration, ultimately achieving a real-time IoT malware knowledge graph for collaborative defense.

In [130], the authors introduced IoTFlowGenerator, which is a cyber deception framework based on deep learning that synthesizes realistic IoT traffic flows to increase honeypot believability against adversaries who monitor network activity. Being different from traditional IoT honeypots, which lack the real traffic patterns and can be easily detected, IoTFlowGenerator employs a fine-tuned SeqGAN and VQ-STAE (Vector Quantized-Sequence Transformer AutoEncoder) to produce multivariate, time-consistent packet metadata that imitates real user-device interactions. The system processes PCAP files (using Wireshark) to extract features such as packet length, direction, and timing, clusters packet-level signatures using DBScan, and encodes packets into discrete tokens. These tokens are adversarially trained into synthetic sequences, which are then reconstructed into traffic flows with noise injection for realism. Using 18 IoT devices for evaluation, IoTFlowGenerator not only outperforms DoppelGANger but also reduces the success rate of the adversarial ML models in the differentiation of synthetic traffic from real traffic. Particularly, this reduction is in synthetic-data-aware attack scenarios. Despite its high potential, the main weakness is the dependence on pre-collected traffic data and device-specific modeling, which might pose scalability challenges. The research highlights how sequence-based GANs can be used to increase the deception of encrypted IoT environments.

Table VII summarizes all the discussed works in internet of things, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

XII. Cyber Deception in Internet of Vehicles

The Internet of Vehicles (IoV) is a transforming new phase of intelligent transportation systems, which is now become possible due to the distinct features of the network that allows vehicles to communicate with each other (V2V), with infrastructure (V2I), and with broader networks (V2X) thus improving safety, efficiency and user experience. During this time, increased connectivity has also brought security problems, which are challenging to solve. Protection measures against the spectrum of attacks facing IoV systems, such as spoofing, eavesdropping, denial-of-service attacks, and unauthorized access, have compromised both data integrity and physical safety. Traditional security measures are not capable of tackling the challenges of these compound threats [131]. Furthermore, recent works [132] [133] reported the provision of intelligent sensors in IoV via reinforcement learning based techniques, where virtualized sensors are dynamically assigned to achieve the best possible service and resource optimization. Nevertheless, the main issue of security in IoV persists, and such provisioning schemes must be examined about their potential role in supporting defensive strategies such as cyber deception, which has emerged as a proactive defense strategy. Through the use of such deceptive elements as honeypots, decoy nodes, and faulty communication protocols, the defenders can mislead attackers, make them late, and procure important information on their strategies. We have the latest findings that bring sophisticated frameworks to the IoV setting. To illustrate, the DECEPTWIN framework leverages digital twins and blockchain to create a decoy environment that looks real, enhancing the detection and analysis of malicious activities. Similarly, Honey-Car framework is a great example of the use of game-theoretic models in the optimization of honeypot settings; thus, the attackers are involved well, and we can take actionable threat intelligence. Even if these ideas are great, the process of introducing cyber deception into IoV has its problems. It is vital to create decoys that are real, as well as sustain system performance and incorporate strategies of deceit in a smooth manner into the already established setup. This section considers the cyber deception play in IoV, looking into its methodologies, advantages, and the issues raised during the process of execution.

In [134], the authors introduced NHBADI, a newly introduced honeypot-based scheme for identifying and isolating Black Hole attacks in MANETs. The system utilizes AODV protocol modification so that it sends out RREQ packets to nonexistent nodes with the help of the destination IDs and TTL=1 Flag; thus, the system identifies the malicious nodes

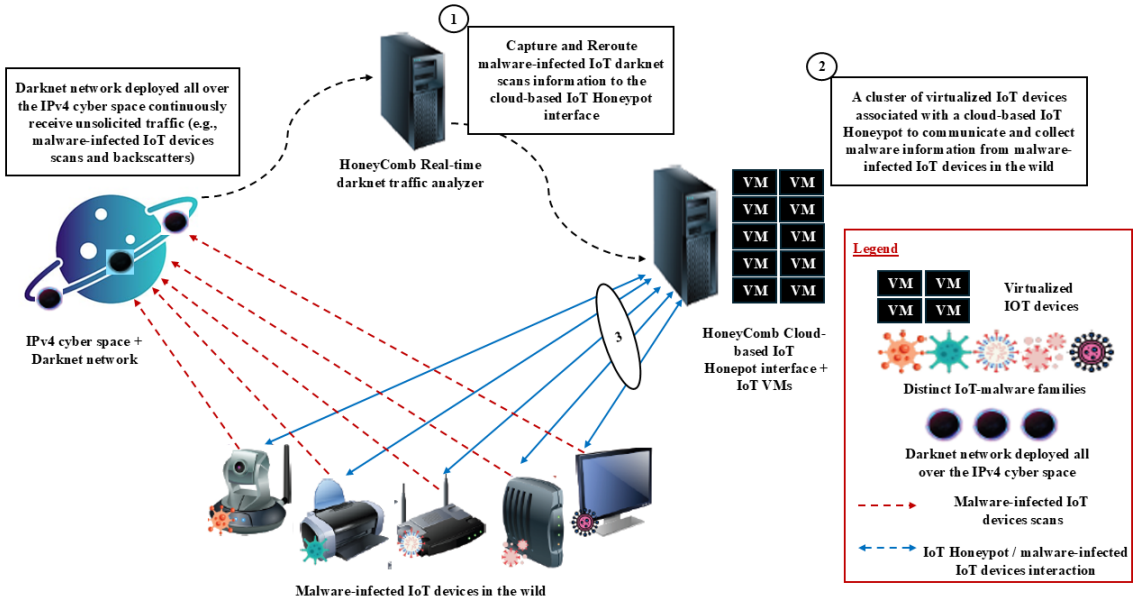


Fig. 16. The overall architecture of the proposed HoneyComb in [129].

TABLE VII. SUMMARY OF CYBER DECEPTION TECHNIQUES IN INTERNET OF THINGS

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[125]	Understand and defend against fileless attacks on Linux-based IoT devices	Fileless attacks (e.g., SSH tunneling, system data damage, de-immunization)	Hardware and software honeypots (HoneyCloud) with high fidelity emulation	Hardware/ software honeypots, shell interceptor & inference terminal, access controller, reset manager	Number of suspicious connections, number of effective attacks, cost comparison	No ZigBee support, visibility limits for encrypted tunnels
[126]	Track and analyze variations in IoT botnets (e.g., Mirai variants)	Mirai-based botnets (SSH/Telnet), SSH brute-force attacks, IRC-based botnets, coin-mining malware	Cowrie-based medium interaction honeypots	Cowrie honeypot, apache webserver, postfix mailserver, ELK stack, filebeat	Edit distance clustering, IP/file hash mappings, ssdeep similarity, session success rate	Limited to x86/64 binaries (Cowrie's emulated architecture), ssdeep inaccuracies
[127]	Capture and analyze IoT-specific attacks (e.g., exploiting UPnP/ SOAP vulnerabilities)	SOAP/UPnP-based attacks, port scanning, CVE-2017-17215 exploitation	Medium-high interaction honeypot, high-interaction honeypot, multiport honeypot	Honeypot core, docker containers, ELK stack, control center, QEMU	Attack volume, geographic distribution, malware samples, functional validation	Limited protocol coverage, static emulation, dependency on Docker
[69]	Proactively mitigate DDoS attacks in IoT networks by dynamically optimizing defense strategies	DDoS, attacks launched via IoT botnets (e.g., Mirai)	MTD techniques (IP/port shuffling, dynamic redirection, response time adaptation), honeypot	Monitor agents, decision module, deployment module, SDN controller	Survival rate, time Overhead, performance metrics (throughput, packet loss rate, CPU utilization, Round-Trip Time, requests per second)	Limited for powerful attackers, limited insider threat handling
[128]	Mitigate botnet propagation in IoT networks	Epidemic botnet (e.g., Mirai) launching DDoS	Honeypot deployment with strategic placement (e.g., k-SDS, RDS)	Defender (honeypot allocator), attacker (botnet propagator), IoT devices (S/I/R states), network topology (low/high connectivity)	Botnet Time to Extinction (TTE), maximum proportion of infected devices, defender's utility (recovery vs. infection rate)	Limited to random/ simple attacker strategies, assumes static game strategies
[129]	Proactive collection of IoT malware forensic artifacts to analyze and mitigate threats	IoT botnet propagation (e.g., Mirai variants)	Darknet-powered SYN-ACK deception with backend honeypots	Darknet traffic analyzer, virtualized IoT devices, cloud-based honeypot interface	Response rate, artifacts collected, response time, geographical/ASN distribution of infections	Limited to ARM architecture, limited protocols, proxy evasion
[130]	Generate realistic synthetic IoT traffic flows to deceive attackers into mistaking honeypots for real devices	Passive network sniffing to identify IoT device types	GAN-based synthetic flow generation (SeqGAN, VQ-STAE)	Traffic analyzer, VQ-STAE autoencoder, SeqGAN, packet fuzzing	Adversarial classification accuracy, traffic fidelity, response variability	Device-specific modeling, dependence on packet-level signatures

when they reply to these decoys. As for the results of the simulation in NS-2, NHBADI provides a good performance with 89.03% of the packets being delivered after the intervention (67.73% of packets with vulnerable AODV) and a routing load which is normalized to just 0.62 (110.79 in AODV with 4 Black Holes), in addition to these, the protocol has cut the end-to-end delay by 23%. The three-layer architecture (detection, verification, and isolation) efficiently quarantines attackers but is built on the assumption that they will react to the specific requests. The strategy might thus be limited in its effectiveness to more advanced adversaries. Through this, the authors show that tactical deception can actually be utilized to improve the security of MANETs more effectively with less effort in spite of the initial resource cost.

In [135], the authors provided a detailed survey on Intrusion Detection Systems (IDS) for Vehicular Ad Hoc Networks (VANETs) and VANET Cloud, dealing with such issues as dynamic en-route node resources, high mobility, protocol differences, and security threats (e.g., Denial of Service (DoS)). The paper condenses into various categories requesting IDS by deployment (distributed, centralized, hybrid), detection techniques (signature-based, anomaly-based, watchdog), and validation strategies while reviewing a total of nearly 25+ different alternatives in a critical manner to achieve trade-offs on accuracy, latency, and overhead. The proactive Honeypot-Optimized IDS (HPIDS) is prominently and solely presented here, which uniquely incorporates bait-based deception along with lightweight detection in order to defend both known and zero-day threats. The article enlists some challenges left unresolved, including a high number of false positives, mobility resilience, and VANET Cloud-specific gaps. Meanwhile, it suggests the implementation of adaptive, low-latency frameworks, along with the privacy of personal data preserved.

In [136], the authors introduced HoneyCar, a game-theoretic framework dedicated to the optimal honeypot configurations in the IoV context, by modeling adversarial interactions as a repeated imperfect-information zero-sum game between a Defender (network administrator) and an Attacker. Using vulnerabilities data from CVE database and CVSS metrics, HoneyCar in its evaluation explores two scenarios: HCG-a, which ignores reconfiguration costs and considers maximal engagement of the attacker through Low-Interaction Honeypots (LIH), and HCG-b, which incorporates reconfiguration costs to optimize cyber threat intelligence using High-Interaction Honeypots (HIH). According to the results of a case study that belongs to LIH, it is displayed that this option can significantly cost-effectively make an attacker waste time, while HIH, along with the selective reconfiguration, can reach the maximum intelligence gain. The framework is designed to be in equilibrium regarding deception costs and security benefits, while improvements are directed towards adaptive strategies and real-time reconfiguration.

In [137], the authors designed DECEPTWIN, a proactive deception framework for the IoV, which is a combination of Digital Twins (DTs) and blockchain, to create realistic, interactive decoys for attackers while securely logging their tactics. The architecture of the framework, as illustrated in Fig. 17, consists of six layers: 1) internet/networking for baiting via decoys, 2) physical system modeling with High-Fidelity DTs, 3) deception environments using Low-Fidelity DTs (LDTs) embedded with deceptive elements to mimic compromised IoV networks, 4) monitoring/tracking of attacker actions across kill-chain stages (reconnaissance, weaponization, exploitation, command and control, exfiltration, and covering tracks), 5) analysis/reporting of threat intelligence, and 6) blockchain-backed storage for immutable logs. Fig. 18 illustrates that DECEPTWIN, after understanding the attacker's interaction, adapts LDTs that show attacker interactions, all through deceptive commands, decoy communications, and falsified logs, enhancing realism while gathering actionable threat data. While it promises proactive IoV security, the framework remains conceptual; its scalability and practical implementation limitations are emphasized as major challenges for future work.

Table VIII summarizes all the discussed works in internet of vehicles, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified limitations.

XIII. Cyber Deception in Unmanned Aerial Vehicles

Unmanned Aerial Vehicles (UAVs) are widely used in various fields such as observation, transportation, and military. They include various devices inside like sensors, flight controllers, communication modules, and ground control stations (GCS). Hence, these can be the targets for hackers and cyber attacks, causing problems like GPS spoofing, jamming, or unauthorized access that result in compromised mission security and safety issues. Conventional protective measures are often insufficient to tackle these dynamic threats [138]. As a result, cyber deception has been developed as an effective solution. Counter-measures have been put in place such that by joining the honeypots with decoy nodes, the defenders can mislead the attackers, postpone their success, and collect useful data on their techniques. Recent developments have brought forth the use of clever deception schemes designed specifically for UAV environments. For example, the HoneyDrone framework is based on game-theoretic models to optimize the honeypot setups and to engage the attackers effectively with the goal of collecting actionable threat intelligence. Even with all the potential, integrating cyber deception in UAV systems is a tough process involving the identification of realist-looking decoys, performance continuity, and the difficulty of intertwining the deception tactics into pre-existing frameworks. This section discusses the utilization of cyber deception within UAVs, outlining its strategies, advantages, and the limitations experienced in its use.

In [139], the authors presented HoneyDrone, the first honeypot created specifically to simulate Unmanned Aerial Vehicles (UAVs) and effectively detect attacks targeting drone protocols. This medium-interaction honeypot also supports UAV-specific protocols such as Telnet, SSH, FTP, and the essential MAVLink standard, in addition to its low-cost deployment option on devices like Raspberry Pis. Its architecture comprises the following key features: 1) a Network Interface Emulator (NIE) which allows the simulation of wireless interfaces (Wi-Fi/Bluetooth/SiK); 2) a UAV Emulation module like a real drone with visualization of protocol behavior, filesystems, and telemetry; 3) MongoDB logging backend; and 4) Ardupilot SITL simulator for the realistic MAVLink flight data emulation. Two attack scenarios, the Telnet-based

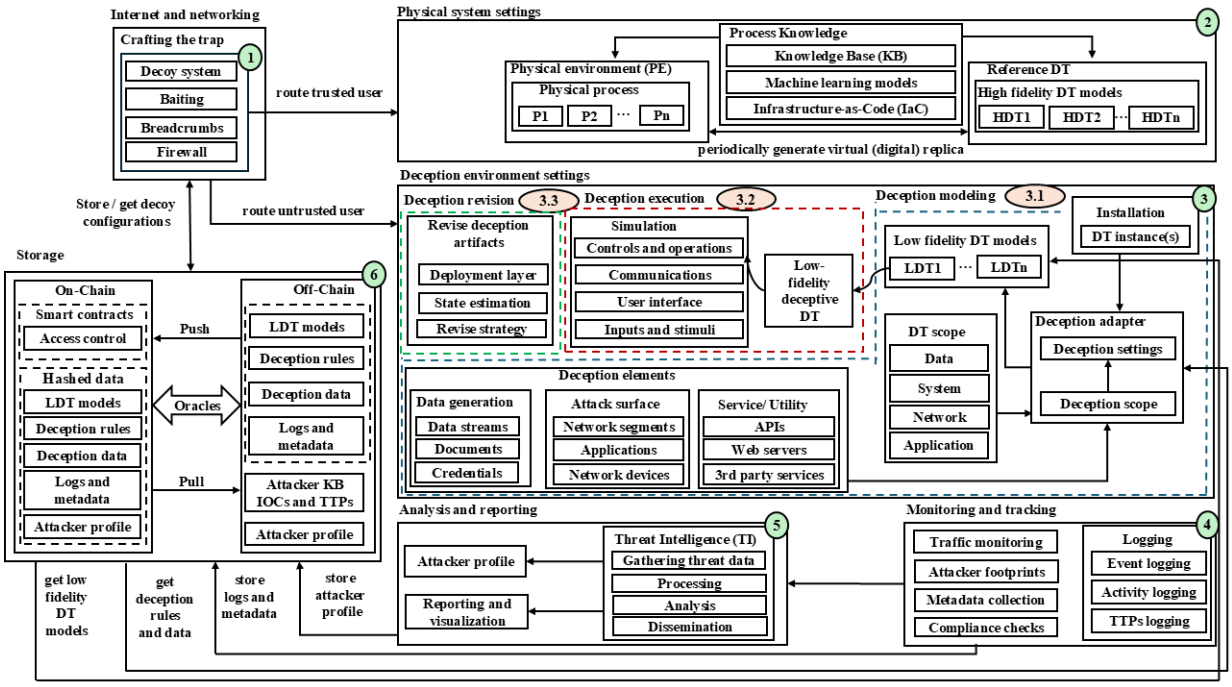


Fig. 17. The proposed architecture of DECEPTWIN in [137].

TABLE VIII. SUMMARY OF CYBER DECEPTION TECHNIQUES IN INTERNET OF VEHICLES

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[134]	Detect and isolate Black Hole attacks in MANET/IoV	Black Hole attacks (forged RREP messages in AODV)	Honeypot-based spoofed RREQ packets	Malicious node detection layer, route lookup layer, isolation layer	Packet Delivery Fraction (PDF), Normalized Routing Load (NRL), End-to-End Delay (EED), Packet Drop Ratio (PDR)	Assumes passive attackers, limited to AODV protocol
[135]	Detect and mitigate malicious nodes and attacks in VANETs and VANET Cloud	DoS, DDoS, Hidden Vehicle, Tunnel, Wormhole, Blackhole, Location Spoofing, Privilege Escalation, alteration, sybil, social Attacks	Honeypot-optimized IDS (HPIDS)	Signature-based IDS, Honeypot nodes, rule generation engine, distributed detection	Detection rate, false positive rate, detection time, resource overhead	Limited VANET Cloud-specific solutions, challenges in optimal placement and density of honeypot nodes
[136]	Develop deception strategies in IoV to engage attackers, gather intelligence, and optimize honeypot use within budget limits	CVE-based exploitation	honeypots, honeypatches	Low-interaction and high-interaction honeypots, vulnerability database, game-theoretic engine	Exploitation time, re-configuration Cost, game utility	Assumes rational attackers, parameter heuristics
[137]	Proactive deception and attacker profiling in IoV	Remote hijacking, data breaches, unauthorized access (e.g., OTA injection, CAN bus attacks)	Digital Twin (DT)-based deception with blockchain integration	Internet/ Networking (decoys, baiting, breadcrumbs, firewalls), physical System, deception environment	Cost-effectiveness, adaptiveness to evolving threats, realism of deception, immutability of attacker logs/ TTPs (via blockchain)	No real-world testing, context-dependent applicability, scalability challenges with blockchain storage

filesystem manipulation and the MAVLink flightpath hijacking via QGroundControl were the basis for its evaluation. HoneyDrone once again came off well against the attackers through mirroring while recording all the interactions. Still, a possible drawback of being identified as a honeypot is the static signal characteristics, so the intention is to dynamically address this issue with the upcoming work on signal emulation and better scalability for concurrent connections.

In [140], the authors presented a scheme based on deep reinforcement learning (DRL), which serves as a deception strategy to successfully fight reactive jamming in IRS-assisted UAV communication systems. The system instead uses a UAV as an aerial base station in the air and trees the control element as the Intelligent Reflecting Surfaces (IRS) to connect the signal better through the controllable non-line-of-sight (NLOS) and direct (LOS) paths. The adversary in this case is a stealthy reactive jammer that selects the high-power transmissions

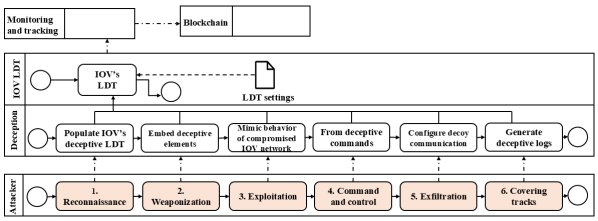


Fig. 18. The realization of DECEPTWIN during the remote access attack on a vehicle within IOV environment in [137].

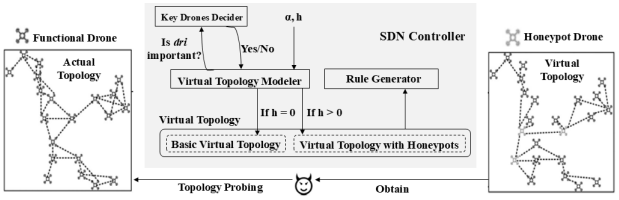


Fig. 19. The designed SDN-based topology deception scheme in [141].

after it has sensed the channel activity. To counter this, the framework includes a Deep Q-Network (DQN) that supports dynamic power allocation and IRS optimization (e.g., phase shifts, positioning), deceiving the jammer through a two-layer strategy: dummy packet transmission to profile the jammer's behavior and power-boosted decoy channels to misdirect attacks toward non-critical frequencies. The DQN agent learns an optimal policy via Q-learning, balancing energy efficiency and transmission rates. Simulations have shown the improvement in Total Received Power (TRP), the reduction of transmit power, and the enhancement of resilience against the baseline methods (e.g., Q-learning only or non-IRS systems). Apart from limitations, authors have made a single UAV/jammer, perfect environmental knowledge, and unexplored real-world deployment challenges (e.g., IRS mobility, multi-jammer scenarios).

In [141], the authors addressed an underexplored challenge of protecting key drones in UAV networks from targeted attacks. These drones, besides being tracers or service hosts, may be a means for some malicious elements to compromise the drone network. The authors begin with the attacker strategy analysis (which may include degree, betweenness, and closeness centrality), and the result is that even targeting a few key drones can cause significant performance degradation. To avert such a situation, they proposed an SDN-based topology deception scheme, which is made up of the following three parts: 1) a Key Drone Privider that decides by identifying critical drones with the help of a custom metric to quantify the connectivity loss; 2) a Virtual Topology Modeler that constructs the deceptive graphs by using the minimum spanning trees and the strategically placed honeypot drones; 3) a Rule Generator that makes recourse to the SDN controller by redirecting the probing attempts (e.g., via TTL-limited traceroute) to faked topologies. It is evident from the illustration in Fig. 19 that non-critical probes get a basic virtual topology, while the critical ones respond in a way that misleads the attackers, by guiding them to honeypots. The plan is executed in a Mininet-WiFi environment, controlled by a Ryu controller, and is notable for its high resilience as exhibited by statistics like Connectivity Loss (CL) and the success rate of attackers, which show even a small difference in honeypot deployment reduces attack impact. Nevertheless, the method presupposes the presence of all drones with OpenFlow support and runs into scalability problems when the network is larger.

In [142], the authors proposed a protection system for UAV networks that avails of honeypot-based deception mechanism and uses learning-based game theory to entice UAVs into sharing their Valid Defense Data (VDD) with the other UAVs. In the frame of this idea, every UAV carries low/ or medium-interaction honeypot modules that catch attack data, with a ground control station (GCS) designing VDD-reward contracts for UAVs to deal with issues like information asymmetry and UAVs' selfish behaviors. The contract is based on partial information asymmetry, and contract theory is utilized to ensure that the incentives are optimal, fair, and budget-feasible, with guaranteed theoretical feasibility and fairness for contracts. For scenarios where information is completely asymmetric, the framework uses a multi-agent Markov game model, and a two-tier Policy Hill-Climbing (PHC) reinforcement learning algorithm that is designed to optimize the strategies dynamically to the UAVs and GCS, enabling adaptive decision-making in the time-varying environments. As illustrated in Fig. 20, the system makes use of UAV private information like VDD volume and communication delays in contract menu, that is GCS is cutting the rewards based on the different UAV types the GCS has and is using Air-to-Air (A2A) and Air-to-Ground (A2G) links for collaborative defense. Simulation experiments have shown that the proposed method noticeably elevates the UAV utility, the rates of participation, and the defensive effectiveness compared to standard mechanisms. Nevertheless, the framework takes a GCS to be trusted yet it does not question the possibility of UAV collusive attacks. Improvements will focus on understanding the activities of hostile UAVs, combining active network conditions and seeing whether stealthy methods such as federated learning might also work and the implementation of a blockchain model on which the free of trust data exchange can take place to make the confidentiality and expansiveness even more secure.

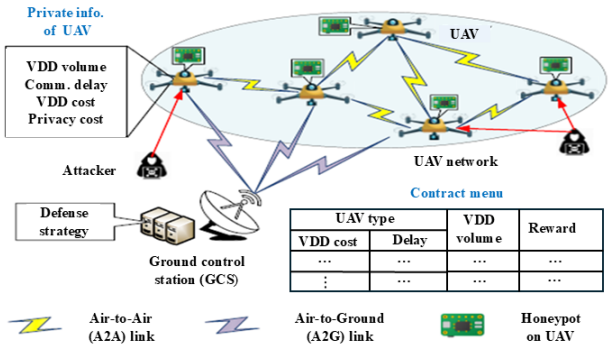


Fig. 20. The illustration of the incentive-driven honeypot game for collaborative defense in [142].

Table IX summarizes all the discussed works in unmanned aerial vehicles, providing a comparative overview of their defense goals, targeted attack types, deception methods, architectural components, evaluation metrics, and identified lim-

itations.

XIV. Open Issues and Key Insights

Despite significant progress in cyber deception research across diverse domains, several issues are still unresolved. The main reasons for this are the rising complexity of infrastructures, the changing patterns of attacker behavior, and the poor implementation of deception frameworks. This section reviews the main open questions accompanied by the most representative findings from the surveyed literature.

A. Incomplete or Simplistic Attacker Modeling

Deception strategies often have overly simplified attacker models, which usually include the assumptions of rational, single-agent, or static behaviors. Game-theoretic approaches in cloud settings, as in [91], are a good example as they usually take into account cost-sensitive attackers but fail to include adaptive adversaries. Advanced modeling is exemplified in [102] and [100], which discuss dynamic and context-aware behaviors in wireless networks. Moreover, adaptive attacker responses are included in [140], which augment the realism. However, there are very few systems that can mirror cognitive biases, collaborative threats, or deception-aware APTs. Cognitive, probabilistic, and multi-agent models should be used to represent the decision of the attacker under advanced uncertainty.

B. Limited Coverage of Post-compromise, Multi-Stage Attacks

Deception strategies mainly find their place in the primary attack phases, namely, scanning, probing, and initial access. Many of the attackers, however, tend to be open to the traps and they execute lateral movement or data exfiltration on their own without even being noticed. The deceiving attack approach is mostly implemented in the beginning phase in ICS and smart grids (for example, [114] and [118]). The same is true for IoT (for example, [129] and [130]), where early traffic analysis is the main focus. A limited number of systems include deception in the runtime workflows or administrative interface. The proposed solutions should cope with final phases of the attack kill chain using decoys designed for the post-compromise phases like privilege escalation and persistence.

C. Limited Adaptivity and Real-time Reconfiguration

The current systems are mostly characterized by static honeypots and inflexible deception logic. There are only a few exceptions, such as [116] and [140], which adapt their deception policies dynamically based on the feedback. The cloud environment has [96] its virtual decoy redeployment solution. Nonetheless, the majority of the systems do not have real-time environmental awareness and self-reconfiguration. It is the application of AI, mainly through reinforcement learning, behavior modeling, and predictive analytics, that is the key to flexible adaptation of deception strategies.

D. Scalability and Fidelity in Resource-constrained Environments

Trade-offs between scalability and deception realism are often mandated by resource constraints. One example is [105] where signaling games are employed for a reduction in energy consumption, whereas [139] and [136] get lightweight deception appropriate for mobile systems. In the cloud environment, [97] and [122] show dynamic orchestration of decoy under resource limitations. Combining low- and high-interaction techniques in a honeypot and dynamically changing to fit the perceived threat level are very good but come with the challenge of creating formal optimization frameworks.

E. Lack of Standardized Metrics and Quantification Frameworks

Current assessment methods vary across different fields. Metrics like engagement time, attacker delay, and resource cost are usually used separately. Some works, such as [122] in smart grids, introduce innovative privacy-oriented metrics (e.g., k-anonymity), while others like [114] and [123] apply reward-based metrics from game theory. Similarly, measuring deception success, attacker confusion, or resilience improvements is not achievable as long as there is no universally adaptable and context-aware framework. The immediate requirement is to normalize testing processes and set up reproducible benchmarks.

F. Limited Testbeds and Validation Platforms

Validation of the majority of deception systems is accomplished through simulation, which leads to the diminished credibility for their real-world deployment. Just like [107] [142] and [137], they put forward interesting ideas but do not demonstrate a physical testbed. Moreover, even in safety-critical areas such as smart grids that have no choice but to depend on simulated environments, there are risks of deployment. Modular scalable testbeds emulating multi-layer architecture and realistic user and attacker behavior are prerequisites for credible evaluation and operational trust.

G. Security and Privacy Trade-Offs in Collaborative Deception

Collaboration through deception, particularly in distributed systems (such as UAV swarms, IoV), brings forth coordination, trust, and privacy challenges. In the case of UAVs, [142] looks at role-switching and shared intelligence among drones. In IoV, [137] uses digital twins for proactive deception. Nevertheless, the majority of the systems don't address claims like adversarial control injection, secure communication, and privacy-preserving coordination. Countermeasures like federated learning, blockchain-based consensus, and zero-trust coordination models are still under-exploited in cyber deception.

H. Integration Gaps with Existing Defense Ecosystems

Deception platforms are frequently set up as isolated systems, which means they do not communicate with SIEM, IDS/IPS, and cyber threat intelligence (CTI) systems. Apart from the tools given, such as [114] or [129] that have shown effectiveness in the engagement of the attacker, they usually miss the mechanisms to put the ideas acquired through this process back into the general defense workflow systematically. The cloud-based models, such as [95], that suggest improved integration potential are left behind due to the unavailability of standardized APIs, logging formats, and data models that are

TABLE IX. SUMMARY OF CYBER DECEPTION TECHNIQUES IN UNMANNED AERIAL VEHICLES

Reference	Defense Goal	Attack Type	Deception Method	Architecture Components	Evaluation Metrics	Uncovered Limitations
[139]	Detect UAV attacks, gather threat intelligence, divert attackers from real drones	Wi-Fi hijacking, Telnet/SSH/FTP exploits, MAVLink protocol abuse	Medium-interaction honeypot	Network Interface Emulator (NIE), UAV emulation, File System Emulator, configuration file	Performance, effectiveness, realism	Static signal behavior, limited to emulated protocols
[140]	Mitigate jamming attacks in IRS-assisted UAV communications	Reactive jamming (intelligent, stealthy)	dummy packets, decoy channels	IRS with passive reflecting elements, UAV as aerial BS, DRL-driven power allocation, Q-learning and DQN for dynamic resource allocation	Rate of communication under jamming, transmit power efficiency, training time of DRL methods	Increased resource usage, assumes jammers attack highest power channels
[141]	Protect key drones in UAV networks from targeted attacks	Topology probing-based targeted attacks	Virtual topology generation (spanning trees) to hide real key drones, Honeypot drones	SDN controller, Key Drone Decider, Virtual Topology Modeler, Rule Generator, OpenFlow-enabled drones, Honeypot drones	Connectivity loss, normalized communication connectivity decrement, importance degree from attacker's view, optimal transmit power and resource usage	Assumes OpenFlow-enabled drones, scalability issues
[142]	Protect UAV networks by incentivizing collaborative defense via honeypot data sharing	Cyber threats (e.g., DoS, hijacking, data theft), free-riding attacks	Contract theory + multi-agent learning for honeypot sharing	Low/ medium-interaction honeypots, Ground Control Station (GCS), UAVs	Connectivity loss, social surplus, defense effectiveness, UAV/GCS utilities, Attack detection rate, CPU utilization of honeypots	Scalability challenges

valid for interoperability. The operational impact of deception systems will be maximized only when they are designed with interoperability as a guiding principle. The introduction of a common and standardized framework, like STIX in the CTI, makes it possible to create actionable deception-derived feedback that could be accessed and used by various other security solutions.

I. Emerging Technologies as Catalysts for Next-Gen Deception

Digital twins, large language models (LLMs), and multi-agent AI have a vast potential for the development of cyber deception. In the IoT sector, dynamic flow redirection has been put forward by [69] as the way to mislead the intruders. Digital twins have been used in the process of anticipatory deception in [137], which provided an opportunity to the users for simulating and predicting the attacker's behavior. Likewise, in UAV networks [140] and smart grids [122], AI-driven and privacy-preserving deception architectures have also been the focus of research. Yet, these things remain primarily theoretical or applicable to particular areas only. To be queried within a full range, these key players should be utilized in a cross-domain, full-stack deception model. As an illustration, digital twins can be used to replicate real environments, thus allowing the defenders to either emulate the vulnerable system in the physical absence of assets or simulate the real-world scenarios with the data of actual threats. LLMs can be incorporated into generating deceptive responses, analyzing adversarial traces, coordinating multi-agent deception strategies, and determining adaptive actions to disrupt, tolerate, or mislead attacks. Joining these two technologies together can greatly modify the flexibility, realism, and intelligence of the deception systems in various functional situations.

J. Challenges in Realism and Fingerprint Resistance

Over time, as the decoys have become more sophisticated, adversaries have also advanced when it comes to the detection of such decoys by means of timing anomalies, low-level operating system artifacts, or service inconsistencies. Researches like [110] and [99] show that the mid-interaction honeypots can be identified by their fingerprints. The presence of behavioral mimicry, dynamic response crafting, and protocol emulation with the help of artificial intelligence makes it possible to stay one step ahead of the attacker reconnaissance. In most systems, these capabilities are still not utilized adequately.

XV. CONCLUSION

This survey sorts the cyber deception techniques in a comparative manner for different domains like CE, WN, CPS, ICS, SG, IOT, IOV, and UAV. The study rigorously studied and communicated the architectural structures, attack types, deception techniques, evaluation metrics, and limitations present in each domain. The findings, which are collected and represented in the unified tables, pave the way for an easy comparison. Besides, the survey deals with the most frequently met challenges and the research directions that remain open, which gives practical insights and support in the design of adaptive and effective deception-based defenses tailored to different operational environments. In our future work, we intend to overcome some of the limitations discussed in the previous section, and our focus will be on utilizing Large Language Models (LLMs) for cyber deception. We aim to examine the efficacy of the LLMs in the generation of deceptive responses, the interpretation of adversarial behavior, and the decision-making of adaptive strategies for disruption, misinformation, or acceptance of cyber threats.

ACKNOWLEDGMENT

This project is carried out under the MOBIDOC scheme, funded by the Ministry of Higher Education and Scientific Research through the PromEssE project and managed by the ANPR.

REFERENCES

[1] L. Zhang and V. L. Thing, "Three decades of deception techniques in active cyber defense-retrospect and outlook," *Computers & Security*, vol. 106, p. 102288, 2021.

[2] X. Han, N. Kheir, and D. Balzarotti, "Deception techniques in computer security: A research perspective," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–36, 2018.

[3] J. Pawlick, E. Colbert, and Q. Zhu, "A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–28, 2019.

[4] M. Zhu, A. H. Anwar, Z. Wan, J.-H. Cho, C. A. Kamhoua, and M. P. Singh, "A survey of defensive deception: Approaches using game theory and machine learning," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2460–2493, 2021.

[5] M. A. Sayed, M. Rahman, M. A. I. Khan, and D. Tosh, "A survey of network requirements for enabling effective cyber deception," *arXiv preprint arXiv:2309.00184*, 2023.

[6] X. Qin, F. Jiang, M. Cen, and R. Doss, "Hybrid cyber defense strategies using honey-x: A survey," *Computer Networks*, vol. 230, p. 109776, 2023.

[7] A. Javadpour, F. Ja'fari, T. Taleb, M. Shojafar, and C. Benzaïd, "A comprehensive survey on cyber deception techniques to improve honeypot performance," *Computers & Security*, p. 103792, 2024.

[8] M. Johnson and J. Meyeraan, "Military deception: Hiding the real-showing the fake," *USAF Joint Forces Staff College, Joint and Combined Warfighting School*, vol. 7, 2003.

[9] M. H. Almeshekah and E. H. Spafford, "Cyber security deception," *Cyber Deception: Building the Scientific Foundation*, pp. 23–50, 2016.

[10] B. Whaley, "Toward a general theory of deception," *The Journal of Strategic Studies*, vol. 5, no. 1, pp. 178–192, 1982.

[11] C. Stoll, "The cuckoo's egg: Tracing a spy through the maze of computer espionage, 1989," *Google Scholar Google Scholar Digital Library Digital Library*.

[12] L. Spitzner, "The honeynet project: Trapping the hackers," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 15–23, 2003.

[13] N. Provos *et al.*, "A virtual honeypot framework." in *USENIX Security Symposium*, vol. 173, no. 2004, 2004, pp. 1–14.

[14] M. H. Almeshekah and E. H. Spafford, "Planning and integrating deception into computer security defenses," in *Proceedings of the 2014 New Security Paradigms Workshop*, 2014, pp. 127–138.

[15] A. Aly, M. Fayez, M. M. Al-Qutt, and A. Hamad, "Navigating the deception stack: In-depth analysis and application of comprehensive cyber defense solutions," *International Journal of Intelligent Computing and Information Sciences*, vol. 23, no. 4, pp. 50–65, 2023.

[16] S. T. Trassare, R. Beverly, and D. Alderson, "A technique for network topology deception," in *MILCOM 2013-2013 IEEE Military Communications Conference*. IEEE, 2013, pp. 1795–1800.

[17] S. Achleitner, T. La Porta, P. McDaniel, S. Sugrim, S. V. Krishnamurthy, and R. Chadha, "Cyber deception: Virtual networks to defend insider reconnaissance," in *Proceedings of the 8th ACM CCS international workshop on managing insider security threats*, 2016, pp. 57–68.

[18] B. Wang and B. Lu, "A network deception defense mechanism based on virtual topology generation," in *International Conference on Intelligent Systems, Communications, and Computer Networks (ISCCN 2023)*, vol. 12702. SPIE, 2023, pp. 641–647.

[19] A. Clark, K. Sun, and R. Poovendran, "Effectiveness of ip address randomization in decoy-based moving target defense," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 678–685.

[20] A. R. Chavez, W. M. Stout, and S. Peisert, "Techniques for the dynamic randomization of network attributes," in *2015 international carnahan conference on security technology (ICCST)*. IEEE, 2015, pp. 1–6.

[21] J. Narantuya, S. Yoon, H. Lim, J.-H. Cho, D. S. Kim, T. Moore, and F. Nelson, "Sdn-based ip shuffling moving target defense with multiple sdn controllers," in *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks–Supplemental Volume (DSN-S)*. IEEE, 2019, pp. 15–16.

[22] L. Alt, R. Beverly, and A. Dainotti, "Uncovering network tarpits with degreaser," in *Proceedings of the 30th Annual Computer Security Applications Conference*, 2014, pp. 156–165.

[23] L. Shing, "An improved tarpit for network deception," Ph.D. dissertation, Monterey, California: Naval Postgraduate School, 2016.

[24] M. A. Sayed, A. H. Anwar, C. Kiekintveld, and C. Kamhoua, "Honeypot allocation for cyber deception in dynamic tactical networks: A game theoretic approach," in *International Conference on Decision and Game Theory for Security*. Springer, 2023, pp. 195–214.

[25] Z. Morić, V. Dakić, and D. Regvart, "Advancing cybersecurity with honeypots and deception strategies," in *Informatics*, vol. 12, no. 1. MDPI AG, 2025, p. 14.

[26] L. Patil, S. Desai, and A. Singh, "Honeypot based secure network system," 2021.

[27] X. Yang, J. Yuan, H. Yang, Y. Kong, H. Zhang, and J. Zhao, "A highly interactive honeypot-based approach to network threat management," *Future Internet*, vol. 15, no. 4, p. 127, 2023.

[28] D. Fraunholz, D. Krohmer, F. Pohl, and H. D. Schotten, "On the detection and handling of security incidents and perimeter breaches-a modular and flexible honeytoken based framework," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2018, pp. 1–4.

[29] E. Filippi, "Honeyport-a scalable meta-honeypot system for security applications," Ph.D. dissertation, Politecnico di Torino, 2019.

[30] F. Araujo, K. W. Hamlen, S. Biedermann, and S. Katzenbeisser, "From patches to honey-patches: Lightweight attacker misdirection, deception, and disinformation," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 942–953.

[31] F. Araujo, M. Shapouri, S. Pandey, and K. Hamlen, "Experiences with {Honey-Patching} in active cyber security education," in *8th Workshop on Cyber Security Experimentation and Test (CSET 15)*, 2015.

[32] J. Avery and E. H. Spafford, "Ghost patches: Fake patches for fake vulnerabilities," in *ICT Systems Security and Privacy Protection: 32nd IFIP TC 11 International Conference, SEC 2017, Rome, Italy, May 29-31, 2017, Proceedings 32*. Springer, 2017, pp. 399–412.

[33] J. K. Avery, "The application of deception to software security patching," Ph.D. dissertation, Purdue University, 2017.

[34] M. Albanese, E. Battista, and S. Jajodia, "A deception based approach for defeating os and service fingerprinting," in *2015 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2015, pp. 317–325.

[35] M. A. Rahman, M. M. Hasan, M. H. Manshaei, and E. Al-Shaer, "A game-theoretic analysis to defend against remote operating system fingerprinting," *Journal of Information Security and Applications*, vol. 52, p. 102456, 2020.

[36] N. C. Rowe, J. Rrushi *et al.*, *Introduction to cyberdeception*. Springer, 2016.

[37] F. Araujo, *Engineering cyber-deceptive software*. The University of Texas at Dallas, 2016.

[38] F. Araujo, W. Kevin *et al.*, "Compiler-instrumented, dynamic {Secret-Redaction} of legacy processes for attacker deception," in *24th USENIX Security Symposium (USENIX Security 15)*, 2015, pp. 145–159.

[39] P. Kaghazgaran and H. Takabi, "Toward an insider threat detection framework using honey permissions." *J. Internet Serv. Inf. Secur.*, vol. 5, no. 3, pp. 19–36, 2015.

[40] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 1–9.

[41] N. Virvilis, B. Vanautgaerden, and O. S. Serrano, "Changing the game: The art of deceiving sophisticated attackers," in *2014 6th International Conference On Cyber Conflict (CyCon 2014)*. IEEE, 2014, pp. 87–97.

[42] P. Peng, C. Xu, L. Quinn, H. Hu, B. Viswanath, and G. Wang, "What happens after you leak your password: Understanding credential sharing on phishing sites," in *Proceedings of the 2019 ACM Asia conference on computer and communications security*, 2019, pp. 181–192.

[43] C. Katsinis and B. Kumar, "A framework for intrusion deception on web servers," in *2013 International Conference on Internet Computing, ICOMP'13*, 2013.

[44] D. P. Julian, "Delaying-type responses for use by software decoys," Ph.D. dissertation, Monterey, California. Naval Postgraduate School, 2002.

[45] S. Sugrim, S. Venkatesan, J. A. Youzwak, C.-Y. J. Chiang, R. Chadha, M. Albanese, and H. Cam, "Measuring the effectiveness of network deception," in *2018 IEEE international conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 142–147.

[46] M. M. Islam and E. Al-Shaer, "Active deception framework: An extensible development environment for adaptive cyber deception," in *2020 IEEE Secure Development (SecDev)*. IEEE, 2020, pp. 41–48.

[47] D. Gavrilis, I. Chatzis, and E. Dermatas, "Flash crowd detection using decoy hyperlinks," in *2007 IEEE International Conference on Networking, Sensing and Control*. IEEE, 2007, pp. 466–470.

[48] D. Brewer, K. Li, L. Ramaswamy, and C. Pu, "A link obfuscation service to detect webbots," in *2010 IEEE International Conference on Services Computing*. IEEE, 2010, pp. 433–440.

[49] B. M. Bowen, S. Hershkop, A. D. Keromytis, and S. J. Stolfo, "Baiting inside attackers using decoy documents," in *Security and Privacy in Communication Networks: 5th International ICST Conference, SecureComm 2009, Athens, Greece, September 14-18, 2009, Revised Selected Papers 5*. Springer, 2009, pp. 51–70.

[50] C. M. McRae and R. B. Vaughn, "Phighting the phisher: Using web bugs and honeytokens to investigate the source of phishing attacks," in *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 2007, pp. 270c–270c.

[51] A. R. Petrunić, "Honeytokens as active defense," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2015, pp. 1313–1317.

[52] M. Bercovitch, M. Renford, L. Hasson, A. Shabtai, L. Rokach, and Y. Elovici, "Honeygen: An automated honeytokens generator," in *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2011, pp. 131–136.

[53] M. Msaad, S. Srinivasa, M. M. Andersen, D. H. Audran, C. U. Orji, and E. Vasilomanolakis, "Honeysweeper: Towards stealthy honeytoken fingerprinting techniques," in *Nordic Conference on Secure IT Systems*. Springer, 2022, pp. 101–119.

[54] M. Kahlhofer, S. Achleitner, S. Rass, and R. Mayrhofer, "Honeyquest: Rapidly measuring the enticingness of cyber deception techniques with code-based questionnaires," in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, 2024, pp. 317–336.

[55] J. Yuill, M. Zappe, D. Denning, and F. Feer, "Honeyfiles: deceptive files for intrusion detection," in *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.* IEEE, 2004, pp. 116–122.

[56] J. Voris, J. Jermyn, N. Boggs, and S. Stolfo, "Fox in the trap: Thwarting masqueraders via automated decoy document deployment," in *Proceedings of the eighth European workshop on system security*, 2015, pp. 1–7.

[57] M. Lazarov, J. Onaolapo, and G. Stringhini, "Honey sheets: What happens to leaked google spreadsheets?" in *9th Workshop on Cyber Security Experimentation and Test (CSET 16)*, 2016.

[58] R. C. Timmer, D. Liebowitz, S. Nepal, and S. Kanhere, "Tsm: Measuring the enticement of honeyfiles with natural language processing," *arXiv preprint arXiv:2203.07580*, 2022.

[59] M. H. Almeshekah, "Using deception to enhance security: A taxonomy, model, and novel uses," Ph.D. dissertation, Purdue University, 2015.

[60] H. Bojinov, E. Bursztein, X. Boyen, and D. Boneh, "Kamouflage: Loss-resistant password management," in *Computer Security–ESORICS 2010: 15th European Symposium on Research in Computer Security, Athens, Greece, September 20-22, 2010. Proceedings 15*. Springer, 2010, pp. 286–302.

[61] A. Juels and R. L. Rivest, "Honeywords: Making password-cracking detectable," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pp. 145–160.

[62] J. Dani, B. McCulloh, and N. Saxena, "When ai defeats password deception! a deep learning framework to distinguish passwords and honeywords," *arXiv preprint arXiv:2407.16964*, 2024.

[63] A. Juels and T. Ristenpart, "Honey encryption: Security beyond the brute-force bound," in *Advances in Cryptology–EUROCRYPT 2014: 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings 33*. Springer, 2014, pp. 293–310.

[64] O. T. Taofeek, M. Alawida, A. Alabdulatif, A. E. Omolara, and O. I. Abiodun, "A cognitive deception model for generating fake documents to curb data exfiltration in networks during cyber-attacks," *IEEE Access*, vol. 10, pp. 41 457–41 476, 2022.

[65] Y. Hu, Y. Lin, E. S. Parolin, L. Khan, and K. Hamlen, "Controllable fake document infilling for cyber deception," *arXiv preprint arXiv:2210.09917*, 2022.

[66] Y. Park and S. J. Stolfo, "Software decoys for insider threat," in *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, 2012, pp. 93–94.

[67] M. Ge, J.-H. Cho, D. Kim, G. Dixit, and I.-R. Chen, "Proactive defense for internet-of-things: moving target defense with cyberdeception," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 1, pp. 1–31, 2021.

[68] S. Wang, Q. Pei, Y. Zhang, X. Liu, and G. Tang, "A hybrid cyber defense mechanism to mitigate the persistent scan and foothold attack," *Security and Communication Networks*, vol. 2020, no. 1, p. 8882200, 2020.

[69] Y. Zhou, G. Cheng, and S. Yu, "An sdn-enabled proactive defense framework for ddos mitigation in iot networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 5366–5380, 2021.

[70] S. Wang, Q. Pei, J. Wang, G. Tang, Y. Zhang, and X. Liu, "An intelligent deployment policy for deception resources based on reinforcement learning," *IEEE Access*, vol. 8, pp. 35 792–35 804, 2020.

[71] T. Kong, L. Wang, D. Ma, Z. Xu, Q. Yang, Z. Lu, and Y. Lu, "Automated honeynet deployment strategy for active defense in container-based cloud," in *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2020, pp. 483–490.

[72] N. C. Abay, C. G. Akcora, Y. Zhou, M. Kantarcioglu, and B. Thuraisingham, "Using deep learning to generate relational honeydata," *Autonomous Cyber Deception: Reasoning, Adaptive Planning, and Evaluation of HoneyThings*, pp. 3–19, 2019.

[73] A. Basak, C. Kamhoua, S. Venkatesan, M. Gutierrez, A. H. Anwar, and C. Kiekintveld, "Identifying stealthy attackers in a game theoretic framework using deception," in *Decision and Game Theory for Security: 10th International Conference, GameSec 2019, Stockholm, Sweden, October 30–November 1, 2019, Proceedings 10*. Springer, 2019, pp. 21–32.

[74] O. Thakoor, M. Tambe, P. Vayanos, H. Xu, C. Kiekintveld, and F. Fang, "Cyber camouflage games for strategic deception," in *Decision and Game Theory for Security: 10th International Conference, GameSec 2019, Stockholm, Sweden, October 30–November 1, 2019, Proceedings 10*. Springer, 2019, pp. 525–541.

[75] A. H. Anwar, C. Kamhoua, and N. Leslie, "Honeypot allocation over attack graphs in cyber deception games," in *2020 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2020, pp. 502–506.

[76] Mitre att&ck. Accessed on 03-01-2025. [Online]. Available: https://attack.mitre.org/

[77] M. Zambianco, C. Facchinetti, and D. Siracusa, "A proactive decoy selection scheme for cyber deception using mitre att&ck," *arXiv preprint arXiv:2404.12783*, 2024.

[78] A. Sayari, Y. Djemaiel, S. Rekhis, A. Mabrouk, and B. Jerbi, "Attack modeling and cyber deception resources deployment using multi-layer graph," in *International Conference on Advanced Information Networking and Applications*. Springer, 2022, pp. 560–572.

[79] A. Sayari, S. Rekhis, Y. Djemaiel, A. Mabrouk, and W. Mahouachi, "Decept-cti: A framework for enhancing cyber deception strategies through nlp-based extraction of cti from unstructured reports," in *2024 22nd International Symposium on Network Computing and Applications (NCA)*. IEEE, 2024, pp. 286–293.

[80] Mitre d3fend. Accessed on 17-01-2025. [Online]. Available: https://d3fend.mitre.org/

[81] Mitre engage. Accessed on 17-03-2025. [Online]. Available: https://engage.mitre.org/matrix/?phase=operate

[82] A. H. Anwar and C. A. Kamhoua, "Cyber deception using honeypot allocation and diversity: A game theoretic approach," in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2022, pp. 543–549.

[83] T. Zhang, L. Huang, J. Pawlick, and Q. Zhu, "Game-theoretic analysis of cyber deception: Evidence-based strategies and dynamic risk mitigation," *Modeling and Design of secure Internet of Things*, pp. 27–58, 2020.

[84] Y.-T. Yang and Q. Zhu, "Game-theoretic foundations for cyber resilience against deceptive information attacks in intelligent transportation systems," *arXiv preprint arXiv:2412.04627*, 2024.

[85] A. Sayari, S. Rekhis, Y. Djemaiel, and A. Mabrouk, "An adaptive reinforcement learning-based approach for effective cyber denial and deception strategies finding," in *International Conference on Advanced Information Networking and Applications*. Springer, 2025, pp. 224–234.

[86] A. Charpentier, N. Boulahia Cuppens, F. Cuppens, and R. Yaich, "Deep reinforcement learning-based defense strategy selection," in *Proceedings of the 17th International Conference on Availability, Reliability and Security*, 2022, pp. 1–11.

[87] Y. Zhang, F. Liu, and H. Chen, "Optimal strategy selection for cyber deception via deep reinforcement learning," in *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*. IEEE, 2022, pp. 1841–1847.

[88] A. Sayari, S. Rekhis, and A. Mabrouk, "A game-theoretic approach with decoy qubits for quantum superdense coding security," in *2024 International Conference on Ubiquitous Networking (UNet)*, vol. 10. IEEE, 2024, pp. 1–8.

[89] U. A. Butt, R. Amin, M. Mehmood, H. Aldabbas, M. T. Alharbi, and N. Albaqami, "Cloud security threats and solutions: A survey," *Wireless Personal Communications*, vol. 128, no. 1, pp. 387–413, 2023.

[90] V. Mahajan and S. K. Peddoju, "Integration of network intrusion detection systems and honeypot networks for cloud security," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 829–834.

[91] M. T. Adili, A. Mohammadi, M. H. Manshaei, and M. A. Rahman, "A cost-effective security management for clouds: A game-theoretic deception mechanism," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, 2017, pp. 98–106.

[92] A. Aydeger, N. Saputro, and K. Akkaya, "Cloud-based deception against network reconnaissance attacks using sdn and nfv," in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*. IEEE, 2020, pp. 279–285.

[93] K. D. Singh, "Securing of cloud infrastructure using enterprise honeypot," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 2021, pp. 1388–1393.

[94] H. Li, Y. Guo, S. Huo, H. Hu, and P. Sun, "Defensive deception framework against reconnaissance attacks in the cloud with deep reinforcement learning," *Science China Information Sciences*, vol. 65, no. 7, p. 170305, 2022.

[95] A. El-Kosairy and N. Abdelbaki, "Deception as a service: intrusion and ransomware detection system for cloud computing (irds4c)," *Advances in Computational Intelligence*, vol. 3, no. 3, p. 9, 2023.

[96] V. D. Priya and S. S. Chakkaravarthy, "Containerized cloud-based honeypot deception for tracking attackers," *Scientific Reports*, vol. 13, no. 1, p. 1437, 2023.

[97] M. Zambianco, C. Facchinetti, R. Doriguzzi-Corin, and D. Siracusa, "Resource-aware cyber deception for microservice-based applications," *IEEE Transactions on Services Computing*, 2024.

[98] R. Nazir, A. A. Laghari, K. Kumar, S. David, and M. Ali, "Survey on wireless network security," *Archives of Computational Methods in Engineering*, pp. 1–20, 2021.

[99] D. B. Rawat, N. Sapavath, and M. Song, "Performance evaluation of deception system for deceiving cyber adversaries in adaptive virtualized wireless networks," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 401–406.

[100] A. Adebayo and D. B. Rawat, "Deceptor-in-the-middle (ditm): cyber deception for security in wireless network virtualization," in *2020 IEEE 17th annual consumer communications & networking conference (CCNC)*. IEEE, 2020, pp. 1–6.

[101] A. A. Adebayo and D. B. Rawat, "Cyber deception for wireless network virtualization using stackelberg game theory," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2021, pp. 1–6.

[102] S. Nan, S. Brahma, C. A. Kamhoua, and N. O. Leslie, "Mitigation of jamming attacks via deception," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2020, pp. 1–6.

[103] Q. He, E. Yang, S. Fang, and S. Zhao, "Revisiting wireless breath and crowd inference attacks with defensive deception," *IEEE/ACM Transactions on Networking*, 2024.

[104] I. C. K. Sihomnou, A. Benslimane, A. H. A. Hemida, G. Deugoue, C. Kamhoua, and C. So-In, "Mitigating energy attacks in wireless sensor networks using deception: A game theoretic approach," in *GLOBECOM 2024-2024 IEEE Global Communications Conference*. IEEE, 2024, pp. 4792–4797.

[105] I. C. K. Sihomnou, A. Benslimane, A. H. Anwar, G. Deugoue, and C. Kamhoua, "Cyber deception against battery drain dos attacks in wireless sensor networks using signaling game," *IEEE Access*, 2025.

[106] W. Duo, M. Zhou, and A. Abusorrah, "A survey of cyber attacks on cyber physical systems: Recent advances and challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 5, pp. 784–800, 2022.

[107] M. O. Sayin and T. Başar, "Deception-as-defense framework for cyber-physical systems," in *Safety, Security and Privacy for Cyber-Physical Systems*. Springer, 2021, pp. 287–317.

[108] W. Tian, X. Ji, W. Liu, G. Liu, R. Lin, J. Zhai, and Y. Dai, "Defense strategies against network attacks in cyber-physical systems with analysis cost constraint based on honeypot game model." *Computers, Materials & Continua*, vol. 60, no. 1, 2019.

[109] G. K. Edwin, S. V. Ewards, G. J. W. Kathrine, G. M. Palmer, A. Bertia, and S. Vijay, "Honeypot based intrusion detection system for cyber physical system," in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. IEEE, 2022, pp. 958–962.

[110] Y. Sun, Z. Tian, M. Li, S. Su, X. Du, and M. Guizani, "Honeypot identification in softwarized industrial cyber–physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5542–5551, 2020.

[111] P. C. Kamdem, A. Zemkoho, L. Njilla, M. Nkenlifack, and C. Kamhoua, "Two-layer deception model based on signaling games against cyber attacks on cyber-physical systems." *IEEE Access*, 2024.

[112] M. Nankya, R. Chataut, and R. Akl, "Securing industrial control systems: Components, cyber threats, and machine learning-driven defense strategies," *Sensors*, vol. 23, no. 21, p. 8840, 2023.

[113] M. Winn, M. Rice, S. Dunlap, J. Lopez, and B. Mullins, "Constructing cost-effective and targetable industrial control system honeypots for production networks," *International Journal of Critical Infrastructure Protection*, vol. 10, pp. 47–58, 2015.

[114] E. López-Morales, C. Rubio-Medrano, A. Doupé, Y. Shoshitaishvili, R. Wang, T. Bao, and G.-J. Ahn, "Honeyplc: A next-generation honeypot for industrial control systems," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 279–291.

[115] T. Machida, D. Yamamoto, Y. Unno, and H. Kojima, "Novel deception techniques for malware detection on industrial control systems," *Journal of Information Processing*, vol. 29, pp. 559–571, 2021.

[116] J. Luo, T. Liu, M. Liang, and N. Hu, "A hmm-based ics adaptive deception defense framework," in *2023 8th International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2023, pp. 359–366.

[117] B. Paul, A. Sarker, S. H. Abhi, S. K. Das, M. F. Ali, M. M. Islam, M. R. Islam, S. I. Moyeen, M. F. R. Badal, M. H. Ahamed *et al.*, "Potential smart grid vulnerabilities to cyber attacks: Current threats and existing mitigation strategies," *Heliyon*, vol. 10, no. 19, 2024.

[118] K. Wang, M. Du, S. Maharjan, and Y. Sun, "Strategic honeypot game model for distributed denial of service attacks in the smart grid," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2474–2482, 2017.

[119] C. Dalamagkas, P. Sarigiannidis, D. Ioannidis, E. Iturbe, O. Nikolis, F. Ramos, E. Rios, A. Sarigiannidis, and D. Tzovaras, "A survey on honeypots, honeynets and their applications on smart grid," in *2019 IEEE conference on network softwarization (NetSoft)*. IEEE, 2019, pp. 93–100.

[120] D. Mashima, Y. Li, and B. Chen, "Who's scanning our smart grid? empirical study on honeypot data," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[121] D. Mashima, D. Kok, W. Lin, M. Hazwan, and A. Cheng, "On design and enhancement of smart grid honeypot system for practical collection of threat intelligence," in *13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20)*, 2020.

[122] D. Yang, D. Mashima, W. Lin, and J. Zhou, "Decied: Scalable k-anonymous deception for iec61850-compliant smart grid systems," in *Proceedings of the 6th ACM on Cyber-Physical System Security Workshop*, 2020, pp. 54–65.

[123] B. Li, Y. Shi, Q. Kong, C. Zhai, and Y. Ouyang, "Honeypot-enabled optimal defense strategy selection for smart grids," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[124] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, and B. Sikdar, "A survey on iot security: application areas, security threats, and solution architectures," *IEEe Access*, vol. 7, pp. 82 721–82 743, 2019.

[125] F. Dang, Z. Li, Y. Liu, E. Zhai, Q. A. Chen, T. Xu, Y. Chen, and J. Yang, "Understanding fileless attacks on linux-based iot devices with honeycloud," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 482–493.

[126] B. Lingenfelter, I. Vakilinia, and S. Sengupta, "Analyzing variation among iot botnets using medium interaction honeypots," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2020, pp. 0761–0767.

[127] W. Zhang, B. Zhang, Y. Zhou, H. He, and Z. Ding, "An iot honeynet based on multiport honeypots for capturing iot attacks," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3991–3999, 2019.

[128] O. Tsemogne, Y. Hayel, C. Kamhoua, and G. Deugoué, "Game-theoretic modeling of cyber deception against epidemic botnets in internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2678–2687, 2021.

[129] M. S. Pour, J. Khoury, and E. Bou-Harb, "Honeycomb: A darknet-centric proactive deception technique for curating iot malware forensic artifacts," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–9.

[130] J. Bao, M. Kantarcioglu, Y. Vorobeychik, and C. Kamhoua, "Iot-flowgenerator: Crafting synthetic iot device traffic flows for cyber deception," *arXiv preprint arXiv:2305.00925*, 2023.

[131] S. M. Karim, A. Habbal, S. A. Chaudhry, and A. Irshad, "Architecture, protocols, and security in iov: Taxonomy, analysis, challenges, and solutions," *Security and Communication Networks*, vol. 2022, no. 1, p. 1131479, 2022.

[132] S. Abbes and S. Rekhis, "Reinforcement learning-based virtual sensors provision in internet of vehicles (iov)," in *2022 IEEE 21st International Symposium on Network Computing and Applications (NCA)*, vol. 21. IEEE, 2022, pp. 217–224.

[133] A. Slim and S. Rekhis, "Reinforcement learning for intelligent sensor virtualization and provisioning in internet of vehicles (iov)," *IEEE Access*, 2024.

[134] M. Rajesh Babu and G. Usha, "A novel honeypot based detection and isolation approach (nhbadi) to detect and isolate black hole attacks in manet," *Wireless Personal Communications*, vol. 90, pp. 831–845, 2016.

[135] S. Sharma and A. Kaul, "A survey on intrusion detection systems and honeypot based proactive security mechanisms in vanets and vanet cloud," *Vehicular communications*, vol. 12, pp. 138–164, 2018.

[136] S. Panda, S. Rass, S. Moschoyiannis, K. Liang, G. Loukas, and E. Panaousis, "Honeycar: a framework to configure honeypot vulnerabilities on the internet of vehicles," *IEEE Access*, vol. 10, pp. 104 671–104 685, 2022.

[137] M. Iqbal, S. Suhail, and R. Matulevicius, "Deceptwin: Proactive security approach for iov by leveraging deception-based digital twins and blockchain," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, 2024, pp. 1–11.

[138] Z. Wang, Y. Li, S. Wu, Y. Zhou, L. Yang, Y. Xu, T. Zhang, and Q. Pan, "A survey on cybersecurity attacks and defenses for unmanned aerial systems," *Journal of Systems Architecture*, vol. 138, p. 102870, 2023.

[139] J. Daubert, D. Boopalan, M. Mühlhäuser, and E. Vasilomanolakis, "Honeydrone: A medium-interaction unmanned aerial vehicle honeypot," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2018, pp. 1–6.

[140] F. O. Olowononi, D. B. Rawat, C. A. Kamhoua, and B. M. Sadler, "Deep reinforcement learning for deception in irs-assisted uav communications," in *MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM)*. IEEE, 2022, pp. 763–768.

[141] Y. Tan, J. Liu, and J. Wang, "How to protect key drones in unmanned aerial vehicle networks? an sdn-based topology deception scheme," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 12, pp. 13 320–13 331, 2022.

[142] Y. Wang, Z. Su, A. Benslimane, Q. Xu, M. Dai, and R. Li, "Collaborative honeypot defense in uav networks: A learning-based game approach," *IEEE Transactions on Information Forensics and Security*, 2023.

Gender and Age Estimation from Facial Images Based on Multi-Task and Curriculum Learning

Toma Brezovan, Claudiu Ionuț Popîrlan
University of Craiova, Romania

***Abstract*—This study presents a multi-task deep learning approach for predicting age and gender attributes from facial images, with the aim of obtaining a robust dual classifier. The proposed system uses the pre-trained EfficientNet-B4 model as the feature extractor of the main model and incorporates a two-branch architecture, where the output of the gender classification branch informs the age prediction branch. This means a conditional feature learning with an explicit injection mechanism, by injecting gender information into the age field of the dual-task model, which is one of the novelties of our proposal. A curriculum learning strategy is applied during training to progressively improve the model's performance using various datasets, such as UTKFace, MORPH-II, and Adience. The proposed multi-phase curriculum learning strategy, which uses both multi-task learning and multi-dataset training, is another novelty of our proposal. Experimental results show that the model achieves high accuracy in both age and gender classification tasks while maintaining low inference latency. Furthermore, the experiments highlighted that the classification accuracy values of the proposed method, both for gender and age, as well as in all datasets used, are close to the best state-of-the-art results, which validates the robustness of the proposed classifier.**

***Keywords*—Age estimation; gender classification; multi-task learning; curriculum learning**

I. Introduction

Human face analysis is an important area of research in computer vision, for performing and optimizing various facial perception tasks, such as face recognition, age estimation, gender determination, etc. In addition, biometric identification has been increasingly used recently in various fields, such as automatic detection of physical presence and confirmation of the identity of individuals through image analysis, and gender and age are two of the main biometric attributes [1].

Predicting age and gender from facial images is still an active task in the field of computer vision, having a wide range of applications in areas such as human-computer interaction, image retrieval, security, surveillance and web content filtering [2].

Recent advances in deep learning have significantly improved the performance of age and gender recognition systems, particularly through convolutional neural networks (CNNs) and multi-task learning (MTL) frameworks [3]. MTL approaches are especially attractive in this domain, as age and gender share underlying visual features that can be exploited to improve generalization.

However, existing models usually use either a specialized dataset of facial images using age and gender attributes for training, or other models trained on very large facial datasets (e.g., FaceNet [4], VGGFace [5], or ImageNet [6]), and then, through transfer learning, the training is refined also on a specialized dataset of facial images using age and gender attributes. In both cases, the training and then estimation process is limited to a not very large number of facial images, which depends on the size and diversity of the specific dataset used for training (or fine tuning).

One goal of our approach is to create a dual classifier for the age and gender of people in facial images, which is robust and can estimate the two attributes in various images. For this, we will use several public datasets for training, which contain images in various conditions, which will allow for more accurate estimation of both age and gender.

The tasks related to the prediction of age and gender of individuals are not completely independent, because there are studies that have demonstrated that incorporating gender information can improve the accuracy of age estimation from facial images [7]. Another goal of our approach is to exploit the dependency between the two tasks, which allows for increased accuracy for both age and gender estimation.

Although there are proposals for the correlated treatment of gender and age in facial images, these refer either to different ways of training the models or to modifying the structure of the models, but not to the direct dependence between the gender and age tasks. We use a simpler variant, regarding the modification of the model structure, resulting in a conditional or hierarchical model with dual tasks, as presented in Fig. 1.

In this study, we propose an efficient multitask deep learning model to predict age and gender in facial images. Although there are various proposals for this problem, the presented system incorporates efficient methods from the field of multi-task learning, which allow increasing both the robustness of the system and the accuracy of the estimates.

The model employs *EfficientNet-B4* [8] as a pre-trained backbone, which is known for its balance between accuracy and computational cost. EfficientNet introduces a new strategy, known as Composite Scaling. It allows balancing and adjusting the depth, width, and resolution of the network, thus allowing the model to scale to arbitrary sizes. In addition, EfficientNet maintains a balance between performance and efficiency. EfficientNet (B0-B7) networks have been widely used recently in various fields of computer vision, especially in image classification [9]. They have also been used for age estimation in facial images [10].

A two-branch architecture is constructed, in which the age and gender branches are not totally independent. According to [7], gender classification output is fed into the age estimation branch to improve age prediction performance through feature sharing.

In the case of multi-task classification systems, there can be problems with the loss function, because the tasks usually differ in scale, complexity, or quality of the labels. Instead of using the weighted linear sum of losses for each individual task, we use a learnable task uncertainty by modeling homoscedastic uncertainty, as in [11].

Even if an efficient pre-trained model is used as feature extractor, the robustness of a dual-task model for simultaneous gender and age recognition depends largely on the dataset on which the dual model is trained. To increase the robustness of such a dual classifier, we used several established datasets for training and adapted the curriculum learning strategy [12].

Our work builds on these foundations, which are also the main contributions:

- Using the pre-trained *EfficientNet-B4* model as a backbone, we propose a dual-task classifier for gender and age in facial images, where the output of the gender branch is used in the input of the age branch, so that the gender prediction output is used to improve age estimation.
    - Unlike previous studies, such as [7], [13]–[15], where multiple subnetworks for the two tasks are created (and possibly trained separately) and interconnected, we propose a conditional feature learning with explicit injection mechanism, by injecting gender information into the age head of the dual-task model.
- Due to inconsistent age distributions and labeling schemes in the training datasets, we use a unified age classification strategy for all datasets, as well as a unified data loader for the datasets, which can be later used in the curriculum learning strategy. In addition to the unified age classification strategy, we propose an adaptive algorithm that learns age categories based on the age distribution in the datasets.
    - As far as we know, there are no proposals that automatically generate age groups based on information in datasets.
- We use a learnable task uncertainty so that the contribution of the task losses in the total loss is adaptive during training and the accuracy of the estimates is increased. This method is based on the method proposed in [11].
- We implement a multi-phase curriculum learning strategy for the training operation, based on several elements, including: a) age range coverage, b) age label granularity, and c) image quality. This strategy allows for increased prediction accuracy, as well as eliminating the catastrophic forgetting effect due to the sequential use of samples from the datasets.
    - Although it is not an adaptive method like in [16], unlike the previous methods [17], [18], the proposed method refers to both multi-task learning and multi-dataset training. To our knowledge, there have been no other proposals regarding multi-task learning and multi-dataset training

The remainder of the study is organized as follows: Section II provides a review of the related work. The proposed method is described in Section III which also includes the dual-task classifier and the curriculum learning strategy. The experimental results are presented in Section IV. The study is summarized, with some future directions, in Section V.

II. Related Work

Deep learning methods for age and gender recognition mainly use convolutional neural networks (CNN) trained from scratch or with minimal learning transfer. These models focused on learning hierarchical features directly from facial images. Levi and Hassner [19] introduced a CNN architecture for age and gender classification, demonstrating the effectiveness of CNNs in capturing age- and gender-relevant facial features. The paper [20] proposes an age and gender prediction method from face images using convolutional neural network (CNN). In this paper [21] authors utilized Haar Cascades for face detection and a CNN for classification, achieving real-time age and gender prediction.

A. Architectural Models

Recent advancements have incorporated sophisticated architectures, transfer learning, and attention mechanisms to enhance accuracy and robustness in age and gender recognition. The research [22] uses pre-trained models such as VGG19 and VGGFace, fine-tuned for age and gender prediction, using the MORPH2 dataset. The authors [23] compared custom CNN architectures with pre-trained models such as VGG16, ResNet50, and SE-ResNet50. The study found that transfer learning significantly improved performance over training from scratch. The paper [24] introduces an ensemble of attentional and residual CNNs, leveraging attention mechanisms to focus on informative facial regions, thereby enhancing prediction accuracy.

B. Multi-Task Learning

Multi-Task Learning (MTL), or *multi-attribute learning* (MAL) aims to improve the generalization of models by leveraging domain-specific information contained in the training signals of related tasks. In computer vision, this often involves simultaneously learning tasks such as object detection, semantic segmentation, and depth estimation.

There are two main categories for the MTL, related to the structure of these neural networks [25]:

1) Hard parameter sharing: where the initial layers of the network are shared among all tasks, while the task-specific layers are kept separate.

2) Soft parameter sharing: where each task has its own model with its own parameters, but regularization techniques are used to keep the parameters of different models similar. This approach can lead to high inference cost [26].

MTL networks were used for face attribute estimation [7], as well for other tasks, such as human pose estimation [27], or face alignment [28].

In MTLs, in addition to network structure, recent advances also relate to loss weighting. Several loss weighting methods are proposed to automatically combine the losses, such as: *task prioritization* [29], which weights task losses according to more difficult tasks during training, *uncertainty weighting* [11], which models loss weights as task-dependent but data-agnostic uncertainty, or *gradient normalization* [30], which learns loss weights to enforce that the gradient norms for each task are close

C. Curriculum Learning

Curriculum learning (CL) was introduced by Bengio et al. [31] and involves training models by gradually increasing the difficulty of training data. This method contrasts with traditional training, in which data is presented in random order. Easy-to-difficult ordering can also be used in multi-task learning, by determining a learning order of tasks to maximize the final result [17], [18]. There are several categories of curriculum learning [12], such as: *Teacher-student CL* [32], or *Implicit CL* [33]. In [34], a curriculum learning approach for classification tasks on small to medium-sized datasets is proposed, based on the order of samples.

III. Proposed Method

A. Dual-Task Architecture

Since hard parameter sharing allows multiple tasks to share low-level parameters [26], this structure has advantages such as lower storage costs. For this reason, the proposed deep multi-task network is based on *hard parameter sharing*, as illustrated in Fig. 1, where X represents the input tensor (images), while X_1 and X_2 represent the outputs of the model (predicted age and predicted gender, respectively). In addition, for a faster training and improved accuracy, the proposed dual-task classifier uses *EfficientNet-B4* as a shared backbone for feature extraction. The shared layer before the branches forces the model to compress shared task-based information into a lower-dimensional vector, encouraging learning of shared features and acting as an inductive bias.

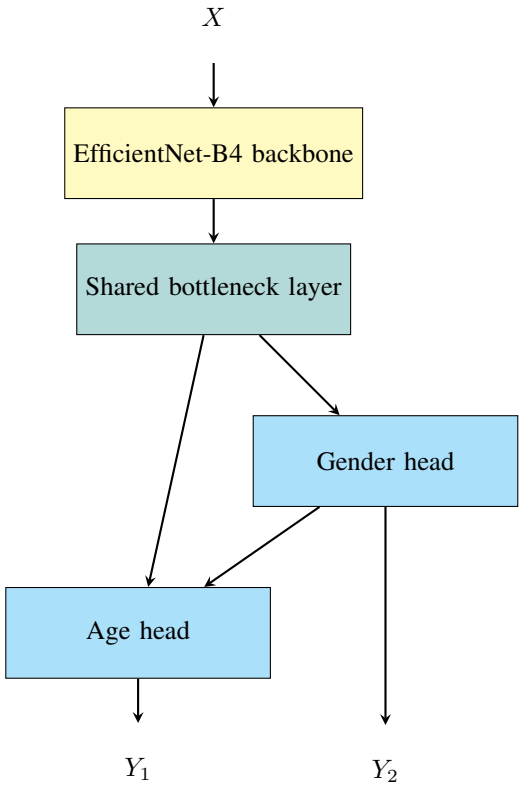


Fig. 1. Architecture of the dual-task classifier

Several studies across biomedical research and computer vision confirm that aging patterns differ significantly between men and women, which supports the integration of gender cues in age estimation models. Biological factors such as hormonal differences lead to gender-specific aging trajectories in skin elasticity, bone structure, and facial fat distribution [35]. In computer vision, deep learning-based age estimation models consistently achieve higher accuracy when gender information is included as input, indicating that aging cues vary in distribution across genders [36].

There are not many proposals for the correlated treatment of gender and age in facial images. In the study [13], a fusion layer between gender-based output and age-based output is proposed, which generates age prediction. In the study [14] a cascaded framework containing a parent network and multiple subnetworks is presented. For example, when using a gender classifier trained on gender information, the other two subnetworks are trained on male and female samples, respectively. In the study [15], age grouping for male and female faces is processed separately: one model is trained to group male faces into different age groups, and another model is learned to group female faces. In the study [36], a hybrid deep learning architecture is proposed for facial gender and age classification: a deep random forest is used to estimate facial gender, and then age is recognized under the conditional probability of gender alignment. In the study [7], it is proposed four two-level CNN models: the first level for gender classification, which contains one CNN, while the second level contains three CNNs, one for each age category (children, middle-aged people, adults, and the elderly).

In this proposal, a different approach has been chosen. We designed a *conditional* or *hierarchical model* with dual task, in which the gender prediction comes first, and the age classifier benefits from this prediction, as presented in Fig. 1.

More exactly, to inject gender information into the age head, we use a *hard conditioning* (post-prediction) method:

- Run gender head first.
- Concatenate predicted gender with shared features.
- Feed them into the age head.

The structure of the proposed architecture, as shown in Fig. 1, has several advantages:

- The model can learn gender-specific age patterns.
- It also encourages the shared feature extractor to learn disentangled but informative representations.

- Conditioning can reduce the variance in age prediction.

B. Training Phase

In the training process, we use two main improvements to increase the accuracy of the estimation operations:

- An adaptive method to determine task-specific weights when determining the total loss.
- A multi-dataset curriculum strategy by designing a training schedule, which is similar to domain adaptation and multi-source learning.

In multi-task learning models, despite the advantages offered by the hard parameter sharing structure, it presents a major drawback: for a global optimization, all task-specific loss objectives must be combined, requiring task-specific weights. Selecting these weights can be difficult and expensive [25]. Even though the simplest way is to sum partial task losses, this often leads to unbalanced training.

For this reason, in this proposal we will use an adaptive method to determine task-specific weights, based on the *Multi-Task Uncertainty Weighting* method [11]. This *first improvement* of the training process allows automatic learning of task weighting by modeling homoscedastic uncertainty.

The *second improvement* in the training process is related to how a deep learning model learns from its input data. Bengio [31] introduced the term *curriculum learning*, which is an optimization strategy for handling a minimization problem. When training a deep neural network, instead of using random samples, it is better to organize these samples so that the less complex examples are presented first.

Since a curriculum is implemented by ordering training data according to difficulty, we will need to generalize this to be able to handle training with multiple datasets, taking into account the following information:

- The quality of the images in the datasets,
- Age range coverage,
- Label granularity.

1) Dynamic learning of task weighting: Multi-task learning involves optimizing a model based on multiple objectives (determining a total loss based on the partial task losses) [see Eq. (1)]:

$$\mathcal{L} = \sum_i w_i \mathcal{L}_i, \tag{1}$$

where, task weights w_i are difficult to estimate.

In the proposal [11], the estimation problem of these weights is viewed as an aleatoric uncertainty, which is task-dependent (called *homoscedastic uncertainty*). In this case, the total loss is derived from maximum likelihood estimation, and for classification tasks, the minimisation objective (loss) has the following form (only two taks are considered here):

$$\mathcal{L} = \frac{1}{\sigma_1^2}\mathcal{L}_1 + \log(\sigma_1) + \frac{1}{\sigma_2^2}\mathcal{L}_2 + \log(\sigma_2), \tag{2}$$

where, σ_1 and σ_2 are observation noise scalars, representing the noise parameters for the two tasks. Using classical notation, σ_i represents the standard deviation, while σ_i^2 represents the variance.

Remark. There are some small differences from the relation in the work [11], because we are performing classification tasks, not regression tasks:

- $\mathcal{L}_1$ and $\mathcal{L}_2$ are cross-entropy losses, not Euclidean losses.
- Coefficients $\frac{1}{2\sigma_i^2}$ were replaced by $\frac{1}{\sigma_i^2}$.

As in [11], in practice the dual-task model should be trained to predict the log variance,

$$s = \log(\sigma^2), \tag{3}$$

Because it is more numerically stable.

To achieve this, two trainable scalar parameters, p_1 and p_2, will be defined as Eq. (4),

$$p_i = \log(\sigma_i^2),\ i \in \{1, 2\}, \tag{4}$$

which represents the log of the variance (uncertainty) of each task. These trainable parameters:

- Will be initialized to 0 and updated in the training phase through gradient descent, just like any other weight.
- They will be used as uncertainties associated with the tasks when calculating the total loss:

$$\mathcal{L} = \mathrm{e}^{-p_1}\mathcal{L}_1 + p_1 + \mathrm{e}^{-p_2}\mathcal{L}_2 + p_2 \tag{5}$$

Remark. Using Eq. (3), the Eq. (2) and Eq. (5) are the same.

2) Multi-dataset curriculum strategy: Typically, deep classification networks are trained using a specific dataset, which is relevant to the respective domain, and the method used for training is to randomly select samples at each iteration step.

This proposal attempts to extend classical training methods in the following way:

- Instead of using a single dataset to train the dual-task classifier, we use multiple datasets that are common in the field of estimating the age and gender of people in images (UTKFace, MORPH-II, and Adience)
- We use a learning strategy based on the *curriculum learning* method [31], which involves training models by gradually increasing the difficulty of the training data.

When a neural network is trained sequentially on multiple datasets, a phenomenon called *catastrophic forgetting* can

occur, in which the neural network, after being trained on one dataset and reaching a certain level of performance, loses that knowledge when trained on a new dataset. There are several proposals to reduce forgetting in curriculum learning, such as:

- Regularization-based methods [37], where methods are applied that penalize changes in important weights for previously learned data
- Progressive training with overlap [38], which involves gradually increasing the proportion of more difficult samples, while the model continues to train on the previous ones

We will define a multi-phase curriculum training strategy that allows training the dual-task model using multiple datasets, according to the *progressive training with overlap* method. In other words, an ordered sequence of examples, from simple to complex, will need to be determined, which will be exposed in the training process.

To be consistent with [31], the following questions must be answered:

- What does a sample mean, given that multiple datasets are used?
- How can the difficulty of a sample be estimated?
- How can the training schedule be specified so that the model is exposed from easier to more complex samples?

Of the two classification tasks, the one associated with age is the most difficult, because the facial aging process is random, both for each individual person and for different categories of people. For this reason, age-related characteristics contribute to determining the difficulty of a sample. To be able to efficiently and robustly learn the age classifier:

- A uniform age category scheme will need to be created, depending on the structure of the datasets used for training.
- A generic, instantiable data loader will need to be created that allows the return of samples from a specified dataset, which refers to a specified list of age categories (age bins).

The answer to the first question is the following: a sample from a multi-dataset curriculum strategy is a sequence of samples from a list of specified datasets, associated with a specified list of age bins. This can be formalized as follows.

Let N_b be the number of age categories, and

$$\mathcal{B}_a = \{0, 1, \ldots, N_b - 1\},$$

the set of age bins. Let $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{N_d}$ be the set of training datatsets, where each dataset $\mathcal{D}_i$, $1 \leq i \leq N_b$, is specified as:

$$\mathcal{D}_i = (Name_i, \mathcal{S}_i),$$

where, $Name_i$ is the name of the dataset, and $\mathcal{S}_i \subseteq \mathcal{B}_a$ is the set of age bins the dataset $Name_i$ contains.

Denoting by $\mathcal{DL}$ the generic data loader, a k^{th}-*sample* from a multi-dataset curriculum strategy is associated to a sequence of datasets, $DSeq_k \subseteq \mathcal{D}$, and a set of age bins, $BSet_k \subseteq \mathcal{B}_a$, and it can be defined as a sequence (a list) of *elementary samples*:

$$\begin{aligned} \mathcal{X}_k &= \mathcal{X}(DSeq_k, BSet_k) = [\mathcal{X}_k^1, \ldots, \mathcal{X}_k^p], \\ \mathcal{X}_k^j &= \mathcal{DL}(Name_{k_j}, BSet_k),\ 1 \leq j \leq p, \end{aligned} \tag{6}$$

where, $Name_{k_j} \in DSeq_k(\mathbb{N})$ for $1 \leq j \leq p$.

Remark. The order of the datasets in Eq. (6) matters because they form a sequence (a list) and not a set.

The difficulty of the samples depends on the datasets used for training, and the calculation of this score considers elements such as :

- Image quality in datasets,
- Age range coverage,
- Number of samples per age range.

We will denote such a scoring function with $score()$.

A *schedule training* is a list of ordered samples [see Eq. (7)],

$$\begin{aligned} \mathcal{ST} = [\mathcal{X}(DSeq_1, BSet_1), \ldots, &\mathcal{X}(DSeq_t, BSet_t)] = \\ &= [\mathcal{X}_1, \ldots, \mathcal{X}_t], \end{aligned} \tag{7}$$

which meet the following conditions:

- $score(\mathcal{X}_i) < score(\mathcal{X}_{i+1}),\ \forall i \in \{1, \ldots, t-1\}$,
- $DSeq_i$ is a subsequence of $DSeq_{i+1}$,
- $BSet_i \subseteq BSet_{i+1}$,
- $DSeq_t = \mathcal{D}$,
- $BSet_t = \mathcal{B}_a$.

The first condition has its source in the curriculum learning method, while the others have their source in the progressive training with overlap method.

Remark. Some examples of curricula (list of multi-dataset samples) are presented in Section IV-D.

3) The training algorithm: The function that describes the training method is described in Algorithm 1, where $model$ is the dual-task classifier, while $scheduleList$ is the list of training samples.

In Algorithm 1:

1) The parameter $model$ represents an instance of the class $DualClassifier$, which implements the dual task classifier, as presented in Fig. 1. The class constructor uses only two parameters: the number of classes for the age attribute, and for the gender attribute:

   $model \leftarrow DualClassifier(ageClasses, genderClasses)$

2) The parameter $scheduleList$ is a dictionary implementing the multi-dataset curriculum strategy. The

Algorithm 1 Training with Multi-Dataset Curriculum Strategy

Require: $model$, $scheduleList$
function TRAINCURRICULUM($model$, $scheduleList$)
  for *schedule* $\in$ *scheduleList* **do**
    for $(datasets, bins, epochs) \in schedule$ **do**
      for *dataset* $\in$ $datasets$ **do**
        for *epoch* $\in$ $epochs$ **do**
          TRAINONEEPOCH($model, trainLoader$)
        end for
      end for
    end for
  end for
end function

algorithm that generates the dictionary is adaptive and does not require parameters (the datasets *MORPH-II*, *UTKFace* and *Adience* are predefined).

3) Each element in a training list has an additional element, $epochs$, which specifies the number of training epochs for that sample.
4) The parameter $trainLoader$ has the following meaning:

$$filteredDataset \leftarrow FilteredDataset(dataset, bins)$$
$$trainLoader \leftarrow DataLoader(filteredDataset)$$

The function TRAINONEEPOCH is briefly described in Algorithm 2.

Algorithm 2 Training One Epoch

Require: $model$, $trainLoader$
function TRAINONEPOC($model$, $trainLoader$)
  for $(images, labels) \in trainLoader$ **do**
    $out \leftarrow model(images)$
    $\mathcal{L}_{age} \leftarrow$ ENTROPYLOSS($out_{age}, labels_{age}$)
    $\mathcal{L}_{gender} \leftarrow$ ENTROPYLOSS($out_{gender}, labels_{gender}$)
    $\mathcal{L} \leftarrow e^{-p_{age}}\mathcal{L}_{age}+p_{age}+e^{-p_{gender}}\mathcal{L}_{gender}+p_{gender}$
    BACKWARD($\mathcal{L}$)
  end for
end function

In Algorithm 2:

1) Loss values for age and gender are calculated using a *cross entropy loss* function to be consistent with classification tasks.
2) The total loss value is calculated using Eq. (4) and Eq. (5).
3) p_{age} and p_{gender} are two trainable scalar parameters, which are initialized to 0 and updated in the training phase at each epoch.

IV. EXPERIMENTS AND RESULTS

A. Datasets

We evaluated several publicly available facial datasets, both controlled and uncontrolled, that allow for the estimation of both gender and age of individuals based on face images, in the field of multi-task learning. A controlled dataset is generated in a controlled environment with some limited variability during image capture, while an uncontrolled dataset involves high variability in real-life image capture [3].

As mentioned in Section III, we chose to use three widely used public domain facial datasets, *MORPH-II* [39], *UTKFace* [40] and *Adience* [41]:

- *MORPH-II* is the *most commonly used* dataset in literature. It is a controlled dataset with some environmental variability (the images are captured in a constrained environment).
- *UTKFace* is a controlled dataset with the *widest age range* (0 to 116 years). Similar to MORPH, the images in UTKFace are captured in a constrained environment.
- *Adience* is the *most challenging* (and uncontrolled) dataset [3].

A part of these datasets are used for training and validation, while the other part is used for testing.

1) UTKFace [40]: UTKFace is a large-scale facial dataset containing over 20,000 images, spanning a long age range from 0 to 116 years. In addition to age labels (which represent the exact age), the images in the dataset also contain binary gender labels (0 - Male, 1 - Female), as well as ethnicity labels. The images are 200×200 pixels in size and are aligned and cropped.

Some qualitative elements related to UTKFace:

- Images are generally front-facing.
- Lighting and background are generally consistent, with low variation.
- Label quality: because they are derived from filename, not manually verified, can contain noise, especially in age.
- UTKFace is more balanced than other datasets in terms of age range but it underrepresents elderly and infants

2) MORPH-II [39]: MORPH-II is a very large public dataset of non-celebrity individuals, containing approximately 55,000 facial images from over 13,000 subjects, with an age range of 16 to 77 years. The images have a variable size, often over 200×240, and they have three labels: exact age, gender (binary label), and race (black, white, others).

Qualitative elements related to MORPH-II:

- Images are frontal well-aligned.
- Lighting and background is controlled (studio-like) and uniform.
- Label quality is high, with metadata.
- MORPH-II has a strong gender imbalance ($\approx 85\%$ Male) and racial imbalance ($\approx 77\%$ African-American).
- Temporal data: It contains multiple images per person over time.

3) Adience [41]: Adience is a challenging dataset containing approximately 26,000 facial images of 2,284 subjects, collected from difficult real-world scenarios. Each image from the dataset is labelled with a gender class label (binary), and an age group class label. There are eight age groups (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60+), and the distribution of face images across age group and gender class labels in Adience is shown in Table I.

TABLE I. THE AGE AND GENDER DISTRIBUTION IN ADIENCE

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+	Total
Male	4.69%	5.60%	5.72%	4.52%	14.04%	7.65%	2.49%	2.75%	47.18%
Female	4.20%	7.58%	7.32%	5.59%	16.47%	6.48%	2.59%	2.59%	52.82%
Total	8.89%	13.18%	13.05%	10.11	30.52%	14.14%	5.08%	5.34%	100.00%

Qualitative elements related to Adience:

- Images unconstrained, non-frontal, real-world poses.
- High variation of lighting and background (wild dataset).
- Label quality images are human-labeled, but age groups, not exact ages. In addition, there are gaps between age groups.
- The distribution of labels is more balanced across genders and ages than MORPH-II, but still uneven between some age groups (as in Table I).

B. Preprocessing Operations

To achieve the research goals, several preprocessing steps were performed on the datasets.

1) Preprocessing file names in UTKFace dataset: : A step specific only to the UTKFace dataset, consisted of uniformly processing the image file names, to eliminate labeling noise (since labels are extracted from file names):

- Incorrect age labels: Some file names contain unrealistic ages due to estimation or annotation errors, and there is also a mismatch between the apparent age and the actual age written in the file name.
- Ambiguous gender labels: The labels are 0 for male and 1 for female, but some are mislabeled (especially for children or when gender is difficult to determine visually).
- Non-standard or corrupted filenames: Some images are mislabeled or don't follow the standard file naming pattern, and a few are even missing tags completely.

2) Face detection and cropping: To obtain a robust classifier, but also because there are large differences between datasets in terms of framing faces in images, before the training, evaluation and testing operations, the two operations must be performed.

3) Face alignment and resizing: The face alignment operation is useful for improving classification accuracy in face-based tasks, such as age and gender classification, because:

- It reduces variability, as faces in datasets may be skewed, rotated, or improperly centered.
- Improves feature learning, as aligned faces lead to better feature consistency across different images, facilitating convergence and generalization.

Image resizing is important because CNN-based backbones (such as EfficientNet-B4) accept images of predefined sizes (380×380 in the case of EfficientNet-B4).

Remark. Face detection, cropping, face alignment are operations included in the generic class used for loading custom datasets in the training, validation and testing operations.

C. Age Labeling Strategy and Data Augmentation

- Label binning for age
- Handling class imbalance

Since the training operation is performed using three datasets, a unified age labeling scheme will need to be designed. To achieve this, the following elements were taken into account:

- Adience is the only dataset that has age groups, the others use exact age.
- The age range for the three datasets differs: 0 to 116 for UTKFace, 16 to 77 for MORPH-II and 0 to 100 for Adience.
- The age ranges at Adience are not consecutive (for example, the range for the second group is 4 to 6, and for the third group is 8 to 13).
- The distribution of images for age is generally imbalanced for all datasets.

This scheme will have eight consecutive age groups, which will cover the entire age range between 0 and 100 years (in the UTKFace dataset there are very few images of people over 100 years old, so they can be included in the last group). We will denote the age groups as follows:

$$\begin{aligned} &g_1 = [0, c_1], \\ &g_i = [c_{i-1} + 1, c_i],\ i \in \{2, 3, \ldots, 7\}, \\ &g_8 = [c_7 + 1, 100]. \end{aligned} \tag{8}$$

Denoting the Adience age groups as,

$$[a_1, b_1], [a_2, b_2], \ldots, [a_8, b_8], \tag{9}$$

we will need to determine the values for the variables $c_1, \ldots, c_7$, so that:

1) Age ranges from Adience groups to be included in the new ranges.
2) The number of images in all datasets, for the eight age groups, should be as balanced as possible.

Because the values for the boundaries $c_1, \ldots, c_7$ are integers, this is an integer optimization problem. We use an age histogram function:

$$f(x,y) = \sum_{i=x}^{y} n(i),$$

where, $n(i)$ is a function that counts in all three datasets images with age i. In addition, we will denote by $f_i = f(x_i, y_i)$ the number of images in group g_i [Eq. (8)].

The optimization problem can be formulated as following:

Problem 1: Given the Adience age groups as in Eq. (9), and the new age groups as in Eq. (8), find all integer values for the variables $c_1, \ldots, c_8$ that minimize the mean squared deviation of group sizes:

$$\frac{1}{8}\sum_{i=1}^{8}(f_i - \overline{f})^2,$$

subject to the following constraints:

$$c_i \in [b_i, a_{i+1} - 1], \ \forall i \in \{1, \ldots, 7\}.$$

This discrete optimization problem does not always have a unique solution, due in particular to the age distribution of the datasets. This is not a problem, because there is no need for just one canonical solution. In conclusion, we will choose the first solution, if there are multiple possible solutions.

In this case we use the following solution:

$$(c_1, \ldots, c_7) = (2, 6, 14, 23, 34, 45, 59),$$

noindent which means that the age groups are as follows:

$$0-2, 3-6, 7-14, 15-23, 24-34, 35-45, 46-59, 60-100.$$

Even after finding the solution, the age groups still remain unbalanced, as seen in Fig. 2, due to age group constraints in the Adience dataset.

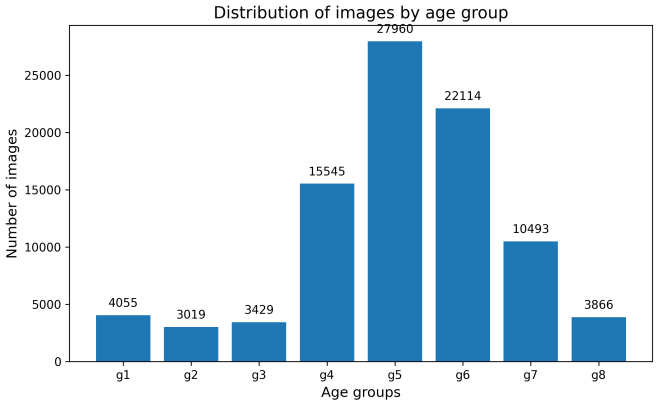


Fig. 2. Distribution of images in datasets by age groups.

To further reduce the imbalance, two operations will be performed:

- Randomly removing images from those age groups that have too many images.
- Generating new images in age groups that have too few images.

The generation of new images is based on augmentation operations, such as: rotation, skewing, and random distortion. In addition, these methods can enhance the dataset without introducing distortions that degrade classification accuracy.

Above the median value, drop-out operations will be performed, and below this value, augmentation operations. After performing these operations, the histogram of images by age group is shown in Fig. 3.

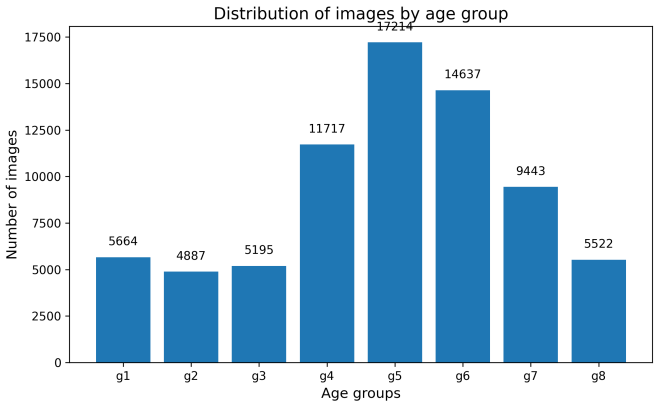


Fig. 3. Distribution of images in datasets by age groups after augmentation2.

The $\frac{1}{9}$ ratio has been reduced to $\frac{1}{3.5}$. To avoid overfitting, we do not believe this ratio should be reduced further. When necessary, a weighted sampling method will be used.

D. Implementing a Multi-Dataset Curriculum Strategy

To implement a *multi-dataset curriculum strategy*, we use the following elements:

- A multi-dataset sample is represented by a list of samples from a list of datasets, related to a list of age bins, as described in Eq. (6).
- A score related to a a multi-dataset sample specifies the training difficulty of that sample, which comes from:
  - The quality of images from each dataset
  - The degree of imbalance of age bins
- A training schedule is a list of multi-dataset samples ordered by their score, as in Eq. (7).

Since we are not yet ready to present an adaptive algorithm that automatically learns a multi-dataset curriculum, regardless of the list of datasets used for training, we will describe the static creation of such a strategy for the three datasets used in this research. Such an algorithm will be the subject of a future study.

According to the quantitative elements of the datasets presented in Section IV-A, the following three difficulty criteria

TABLE II. DATASETS DIFFICULTY CRITERIA

Image quality	Age bins imbalance	Gender imbalance
UTKFace	Adience	UTKFace, Adience
MORPH-II	UTKFace	MORPH-II
Adience	MORPH-II	-

are presented in Table II, in increasing order of difficulty for each dataset:

Remark. The most important difficulty criterion is *Gender imbalance*, because the dual-task model learns gender first, and then age. The difficulty order is as follows:

- *Gender imbalance*: important since the dual-task model depends on gender output for age prediction.
- *Image quality*: impacts feature extraction and prediction robustness.
- *Age imbalance*: makes learning certain age bins harder.

Considering the above remarks and the difficulty criteria, as shown in Table II, below is an ordered list of multi-dataset samples:

1) Multi-dataset sample 1 (Gender-balanced, medium-age, decent quality):
```
{'datasets': ['utkface', 'adience'], '
bins': [3, 4, 5]}
```
2) Multi-dataset sample 2 (Gradual age complexity increase):
```
{'datasets': ['utkface', 'adience'], '
bins': [3, 4, 5, 6, 7]}
```
3) Multi-dataset sample 3 (Full age distribution):
```
{'datasets': ['utkface', 'adience'], '
bins': [0, 1, 2, 3, 4, 5, 6, 7]}
```
4) Multi-dataset sample 4 (Incorporate harder dataset gradually):
```
{'datasets': ['utkface', 'adience', '
morph2'], 'bins': [3, 4, 5]}
```
5) Multi-dataset sample 5 (Adds aging extremity):
```
{'datasets': ['utkface', 'adience', '
morph2'], 'bins': [3, 4, 5, 6, 7]}
```
6) Multi-dataset sample 6 (Full complexity and diversity):
```
{'datasets': ['utkface', 'adience', '
morph2'], 'bins': [0, 1, 2, 3, 4, 5, 6,
 7]}
```

Remarks:

- The previous samples were written as dictionaries in the *Python* language.
- Instead of the age groups $g_1, \ldots, g_8$, were used age bins: $[0, 1, \ldots, 7]$.

E. Results

In this section, the results of experiments performed on the test parts of the UTKFace, MORPH-II, and Adience datasets are presented.

The proposed method is implemented in PyTorch. The training was performed on the Google Colab framework, with an NVIDIA Tesla T4 graphics card with 16 GB of GPU memory. The maximum number of epochs was set to 45, but an early stopping mechanism was used to avoid overfitting. A dynamic learning rate was used via the ReduceLROnPlateau learning rate scheduler in PyTorch. The total training time was approximately 8 hours.

The tests were performed on a Windows PC with an Intel Core Ultra 7, 16-core, 5 GHz processor and 32 GB of CPU memory. The total number of test images in the 3 datasets is approximately 9000. Using a batch size of 64, the total testing time was approximately 5 minutes.

The *accuracy* is used to measure the performance of the model, both for age and gender estimation. The accuracy metric is calculated based on the true positive (TP), false positive (FP), true negative (VN), and false negative (FN) values, as follows:

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

Table III presents the estimation performance results for gender and age for all three datasets.

TABLE III. GENDER AND AGE ACCURACY FOR ALL DATASETS

	Gender	Age
Adience	92.25	73.12
MORPH-II	97.92	77.17
UTKFace	97.74	69.83

It is observed that, although the MORPH-II dataset was introduced last in the training process and has a strong gender imbalance, the classification performances are the best in this case (both in gender and age). Although it is the most gender-unbalanced, the image quality is better than other datasets, which makes the learning process for images in MORPH-II better.

Regarding the other two datasets, Adience and UTKFace, the results are different for gender and age: the classification accuracy for age is better in Adience, while the accuracy for gender is better in UTKFace.

This can be explained for the following reasons:

- Although the images in Adience are collected from difficult real-world scenarios, the image quality is better than in the case of UTKFace, and the image sizes are quite large, over 200×200 (most are around 600×600).
- In UTKFace, all images are 200×200 in size and are aligned and cropped. But because they need to be scaled to the size of 380×380 (which is the input size for EfficientNet-B4) the scaling operation introduces noise.

In conclusion, age estimation for the UTKFace dataset is more difficult than for the other datasets.

In the following, we will compare the results of the proposed method with state-of-the-art works on facial gender and age group estimation.

Table IV presents the results of estimating the performance of the proposed method on the Adience dataset and compares this with other proposed approaches based on deep learning.

TABLE IV. GENDER AND AGE ACCURACY FOR ADIENCE DATASET

Method	Gender accuracy	Age accuracy
Zhang et al., 2017 [42]	93.24	66.74
Duan et al., 2018 [43]	88.20	52.30
Gurnani et al., 2019 [44]	91.80	62.11
Khan et al., 2020 [45]	**93.60**	69.40
Garain et al., 2021 [46]	81.80	66.10
Saha et al. 2023 [47]	84.94	**75.10**
Proposed method	92.25	73.12

The methods specified in Table IV use the 8 age groups as defined in the Adience dataset (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60+).

Table V presents comparisons of performance estimation on the MORPH-II dataset, while Table VI presents comparisons of performance estimation on the UTKFace dataset.

TABLE V. GENDER AND AGE ACCURACY FOR MORPH-II DATASET

Method	Gender accuracy	Age accuracy
Guo et al., 2014, [48]	**98.50**	70.00
Han et al., 2015 [49]	97.60	77.40
Wang et al., 2017 [50]	98.00	**85.30**
Khan et al., 2020 [45]	96.70	75.00
Han et al., 2018 [51]	81.80	66.10
Proposed method	97.92	77.17

TABLE VI. GENDER AND AGE ACCURACY FOR UTKFACE DATASET

Method	Gender accuracy	Age accuracy
Krizhevsky et al., 2017, [52]	85.10	55.96
Huang et al., 2017 [53]	87.28	59.22
Das et al., 2018 [54]	**98.23**	**70.10**
Yuan et al., 2024 [55]	90.91	64.74
Proposed method	97.74	69.83

The vast majority of the methods specified in Table VI use 7 age groups: a) baby: 0 to 3 years, b) child: 4 to 12 years, c) teenagers: 13 to 19 years, d) young: 20 to 30 years, e) adult: 31 to 45 years, f) middle aged: 46 to 60 years, and g) senior: 61 and above years.

It is observed that, in all three tables, the classification results of our method do not exceed the best results of the other proposals, but the following elements should be noted:

- The classification accuracy values for ours method are close to the best results of the other methods, both for gender and age.
- Related the classification accuracy for age groups, the values are not always conclusive, because each proposal had its own age grouping scheme. With the exception of the Adience dataset, many other methods use between 3 and 6 age groups, while our method used 8 age groups, which led to a decrease in accuracy.
- Each of the other proposals used a single dataset for training and testing, while our proposal used all 3 datasets.

One of the goals of our method is to create a robust classifier for the gender and age of individuals in facial images, which means that testing the classifier on unknown images should have high accuracy, both for estimating age and gender.

F. Ablation Study

There are two important factors in our proposal. One is dynamic learning of task weighting when calculating the total loss. The other is the use of multi-dataset curriculum strategy in the training phase. We investigate the effects of the two factors on the accuracy of classification by gender and age.

1) Dynamic learning of task weighting: In multi-task learning, the total loss must be determined based on the partial losses in the task, as presented in Eq. (1). Usually the total loss is simply the sum of partial task losses:

$$\mathcal{L} = \sum_i \mathcal{L}_i.$$

Another approach is to manually determine normalized weights w_i, depending on the difficulty of the tasks:

$$\mathcal{L} = \sum_{i=1}^{n} w_i \mathcal{L}_i,$$

such that $\sum_{i=1}^{n} w_i = 1$.

We used a modified verison of the proposal from [11],

$$\mathcal{L} = \sum_{i=1}^{n} \mathrm{e}^{-\log(\sigma_i^2)} \mathcal{L}_i + \log(\sigma_i^2) \tag{10}$$

As described in Eq. (3), Eq. (4) and Eq. (5), which learns the log of the variance, $\log(\sigma_i^2)$, for each classification task i.

The results shown in Tables IV, V and VI use the Eq. (10), for $n = 2$ [or similarly, Eq. (5)].

To see the importance of using dynamic learning of task weighting, we then retrained of the dual classifier. We used the multi-dataset curriculum strategy in the training part, but in Algorithm 2, we used the following two variants for calculating the total loss value:

$$\mathcal{L} = \mathcal{L}_{age} + \mathcal{L}_{gender}, \tag{11}$$

and:

$$\mathcal{L} = 0.6 \cdot \mathcal{L}_{age} + 0.4 \cdot \mathcal{L}_{gender}. \quad (12)$$

Remark. In Eq. (12), the weight for age classification is higher than for gender classification, because this task is more difficult.

In testing, the values obtained for accuracy were quite close when Eq. (11) and Eq. (12) were used. The differences were, however, greater between using Eq. (11) [or Eq. (12)] and Eq. (5).

Table VII presents the classification accuracy results for gender and age when using both the original Eq. (5), and Eq. (11) to calculate the total loss. An increase of approximately 2 per cent in accuracy is observed comparing the use of Eq. (11) with the use of Eq. (5).

TABLE VII. COMPARISON OF MODEL PERFORMANCE DEPENDING ON THE CALCULATION METHOD OF TOTAL LOSS

Dataset	Total loss determination	Gender	Age
Adience	Equation 5	**92.25**	**73.12**
	Equation 11	91.27	71.65
MORPH-II	Equation 5	**97.92**	**77.17**
	Equation 11	97.23	75.31
UTKFace	Equation 5	**97.74**	**69.83**
	Equation 11	94.68	66.25

2) Multi-dataset curriculum strategy: To determine the contribution of the training method, we will perform several experiments in which the dynamic learning of task weighting is used as in Eq. (5). Since there are many possible combinations, we will only test two combinations, which seem to be more important for the study.

- Since the accuracy values for the MORPH-II dataset are the highest, and it uses only age groups $g4$, $g5$ and $g6$ (bins $[3-5]$), the samples will be ordered according to the age group balance and the results obtained in the main ordering scheme. We will use the following ordered list of multi-dataset samples (will be noted as **list 1**):
  - (1.1) Core mid-age learning: `{'datasets': ['morph2'], 'bins': [3, 4, 5]}`
  - (1.2) Extended core mid-age: `{'datasets': ['morph2', 'utkface', 'adience'], 'bins': [3, 4, 5]}`
  - (1.3) Young age focus: `{'datasets': ['utkface', 'adience'], 'bins': [0, 1, 2]}`
  - (1.4) Elderly focus: `{'datasets': ['utkface', 'adience'], 'bins': [6, 7]}`
  - (1.5) Mid-age re-focus: `{'datasets': ['morph2', 'utkface', 'adience'], 'bins': [3, 4, 5]}`
  - (1.6) All datasets: `{'datasets': ['morph2', 'utkface', 'adience'], 'bins': [0, 1, 2, 3, 4, 5, 6, 7]}`
- Another variation of the curriculum is based on the uneven age distribution and imbalances of the Adience, UTKFace, and MORPH2 datasets. This time MORPH-II is no longer considered an advantageous dataset. We will denote this curriculum as **list 2**:
  - (2.1) Core mid-age learning: `{'datasets': ['adience', 'utkface', 'morph2'], 'bins': [3, 4, 5]}`
  - (2.2) Young age specialization: `{'datasets': ['adience', 'utkface'], 'bins': [0, 1, 2]}`
  - (2.3) Elderly age focus: `{'datasets': ['adience', 'utkface'], 'bins': [6, 7]}`
  - (2.4) Balanced reintroduction: `{'datasets': ['adience', 'utkface'], 'bins': [0, 1, 2, 3, 4, 5, 6, 7]}`
  - (2.5) Fine tuning (all datasets): `{'datasets': ['adience', 'utkface', 'morph2'], 'bins': [3, 4, 5]}`

Table VIII presents the classification accuracy results for gender and age when using three lists of multi-dataset samples in training phase: the original list and the two lists described above (**list 1** and **list 2**).

TABLE VIII. COMPARING MODEL PERFORMANCE BASED ON THE LIST OF MULTI-DATASET SAMPLES

List of multi-dataset samples	Dataset	Gender	Age
Original list	Adience	92.25	73.12
	MORPH-II	97.92	77.14
	UTKFace	97.74	69.32
List 1	Adience	81.25	60.37
	MORPH-II	99.08	87.25
	UTKFace	85.32	58.45
List 2	Adience	92.75	73.82
	MORPH-II	97.12	76.87
	UTKFace	96.21	67.24

Based on the values in this table, the following remarks can be drawn:

- In the case of **list 1**, the accuracy values are extremely high for the MORPH-II dataset, but extremely low for the other two datasets. The model learns classifications for MORPH-II very well because this dataset is the first in the list of datasets. And because of this, the learnings for the other two datasets are at a disadvantage.
- The original list and **list 2** generate predictions with similar accuracy: the accuracy is slightly better in the case of the original list for the MORPH-II and

UTKFace datasets, and in the case of **list 2** the accuracy is slightly better for Adience.

In conclusion, evaluating the difficulty of multi-dataset samples is a complex and difficult operation. It is expected that in the future the manual design of a curriculum will be replaced with an adaptive algorithm that automatically learns a multi-dataset curriculum.

V. Conclusion

This study presents a deep multi-task learning approach for facial gender and age recognition. Facial features are extracted using a pre-trained model, EfficientNet-B4, which has a balanced ratio between accuracy and computational cost. A two-branch architecture is proposed, where the output of the gender branch is fed into the age estimation branch to improve age prediction.

Besides the mechanism for injecting gender information into the age branch of the dual-task model, which is in the architectural part of the proposal, the other contributions of the proposal reside mainly in the training part of the dual classifier.

For robust classification, instead of training the model on a single dataset, three of the most well-known datasets were used simultaneously and a unified strategy for determining age groups was proposed for all datasets. To improve training efficiency and final model performance, a multi-dataset curriculum strategy was proposed, which is based on the curriculum learning method for single datasets, which it extends to multiple datasets.

Another improvement to increase the accuracy of the estimation operations consists of implementing an adaptive method for determining the task-specific weights when determining the total loss. This uses two trainable scalar parameters, representing the logarithm of the variance of each task, which are used to calculate the total loss.

The experimental study conducted on three datasets used (Adience, MORPH-II and UTKFace) validated our contributions. Comparing the results of the proposed method with the state-of-the-art works on gender and age estimation, the classification accuracy values of our method are close to the best results of the other methods, both for gender and age. These results are valid for all datasets used, which demonstrates the robustness of the proposed dual classifier.

In its current form, the proposed method has some limitations and constraints, which could be remedied in the future:

- The multi-dataset curriculum method, although theoretically described, relies on manual curriculum creation.
- Optimal generation of age groups is done automatically, but it is based on predefined age groups in the Adience dataset.

Future research are in several directions:

- Developing an adaptive algorithm that automatically learns a a multi-dataset curriculum, based on loss values, instead of a static and manually developed one.
- Development of an adaptive algorithm that automatically determines age groups for the datasets used, based only on the annotation .csv files.
- Developing a similar lightweight model (e.g. based on a pre-trained model such as EfficientNet-B0), which can be transformed into a model for iOS applications.

References

[1] P. Terhorst, N. Damer, and A. K. F. Kirchbuchner, "Suppressing Gender and Age in Face Templates Using Incremental Variable Elimination," in *2019 International Conference on Biometrics (ICB)*, Crete, Greece, 2019, pp. 1–8.

[2] M. Benkaddour, "CNN Based Features Extraction for Age Estimation and Gender Classification," *Informatica*, vol. 45, no. 5, pp. 697–703, 2021.

[3] S. Gupta and N. Nain, "Review: Single attribute and multi attribute facial gender and age estimation," *Multimedia Tools and Applications*, vol. 82, pp. 1289–1311, 2023.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, X. Xie, M. Jones, and G. Tam, Eds., Boston, MA, USA, 2015, pp. 815–823.

[5] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, X. Xie, M. Jones, and G. Tam, Eds., Swansea, UK, 2015, pp. 41.1–41.12.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challeng," *International Journal of Computer Vission*, vol. 115, pp. 211—252, 2015.

[7] H. Liao, L. Yuan, M. Wu, L. Zhong, G. Jin, and N. Xiong, "Face Gender and Age Classification Based on Multi-Task, Multi-Instance and Multi-Scale Learning," *Applied Sciences*, vol. 12, no. 23, p. 12432, 2023.

[8] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)*, Long Beach, California, USA, 2019, pp. 6105–6114.

[9] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, p. 99, 2024.

[10] I. Aruleba and S. Viriri, "Deep Learning for Age Estimation Using EfficientNet," in *Advances in Computational Intelligence. IWANN 2021*, ser. Lecture Notes in Computer Science, I. Rojas, G. Joya, and A. Catal, Eds. Springer, 2021, vol. 12861, pp. 407–419.

[11] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7482–7491.

[12] P. Soviany, R. Ionescu, P. Rota, and N. Sebe, "Curriculum Learning: A Survey," *International Journal of Computer Vision*, vol. 130, pp. 1526–1565, 2022.

[13] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognition*, vol. 66, pp. 106–116, 2017.

[14] J. Wan, Z. Tan, G. Guo, S. Li, and Z. Lei, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2531–2541, 2018.

[15] K.-H. Liu and T.-J. Liu, "A Structure-Based Human Facial Age Estimation Framework Under a Constrained Condition," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5187–5200, 2019.

[16] Y. Kong, L. Liu, J. Wang, and D. Tao, "Adaptive Curriculum Learning," in *EIEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 5047–5056.

[17] A. Pentina, V. Sharmanska, and C. Lampert, "Curriculum learning of multiple tasks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 5492–5500.

[18] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris, "Curriculum Learning for Multi-task Classification of Visual Attributes," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Austin, Texas, SUA, 2017, pp. 2608–2615.

[19] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, MA, USA, 2015, pp. 34–42.

[20] K. Ito, H. Kawai, T. Okano, and T. Aoki, "Age and Gender Prediction from Face Images Using Convolutional Neural Network," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 2018, pp. 7–11.

[21] B. Abirami, T. Subashini, and V. Mahavaishnavi, "Gender and age prediction from real time facial images using CNN," *Materials Today: Proceedings*, vol. 33, no. 7, pp. 4708–4712, 2020.

[22] P. Smith and C. Che, "Transfer Learning with Deep CNNs for Gender Recognition and Age Estimation," in *2018 IEEE International Conference on Big Data*, Seattle, WA, USA, 2018, pp. 2564–2571.

[23] V. Sheoran, S. Joshi, and T. Bhayani, "Age and Gender Prediction Using Deep CNNs and Transfer Learning," in *Computer Vision and Image Processing. CVIP 2020*, ser. Communications in Computer and Information Science, S. Singh, P. Roy, B. Raman, and P. Nagabhushan, Eds. Springer, 2021, vol. 1377, pp. 293–304.

[24] A. Abdolrashidi, M. Minaei, E. Azimi, and S. Minaee, "Age and Gender Prediction From Face Images Using Attentional Convolutional Network," 2020. [Online]. Available: https://arxiv.org/abs/2010.03791

[25] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1706.05098

[26] L. Liu, Y. Li, Z. Kuang, J.-H. Xue, Y. Chen, W. Yang, Q. Liao, and W. Zhang, "Towards Impartial Multi-task Learning," in *9th International Conference on Learning Representations, ICLR 2021*, Austria, Virtual Event, 2021.

[27] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," *International Journal of Computer Vision*, vol. 113, no. 1, pp. 19–36, 2015.

[28] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Learning Deep Representation for Face Alignment with Auxiliary Attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 113, no. 1, pp. 918–930, 2016.

[29] M. Guo, A. Haque, D. Huang, S. Yeung, and L. Fei-Fei, "Dynamic Task Prioritization for Multitask Learning," in *CComputer Vision. ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, 2018, vol. 1122, pp. 282–299.

[30] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International Conference on Machine Learning, ICML 2018*, Stockholm, Swede, 2018, pp. 794–803.

[31] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, Montreal, Quebec, Canada, 2009, pp. 41–48.

[32] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 2535–2544.

[33] S. Sinha, A. Garg, and H. Larochelle, "Curriculum by smoothing," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)*, vol. 97, Vancouver, BC. Canada, 2020, pp. 21 653–21 664.

[34] S. Chaudhry and A. Sharma, "Data Distribution-Based Curriculum Learning," *IEEE Access*, vol. 12, pp. 138 429–138 440, 2024.

[35] S. Fitzmaurice and H. Maibach, "Gender Differences in Skin," in *Textbook of Aging Skin*, M. Farage, K. Miller, and H. Maibach, Eds. Springer, Berlin, Heidelberg, 2010, pp. 282–299.

[36] S. Dammak, H. Mliki, and E. Fendri, "Gender effect on age classification in an unconstrained environment," *Multimedia Tools and Applications*, vol. 80, pp. 28 001–28 014, 2021.

[37] F. Benzing, "Unifying Importance Based Regularisation Methods for Continual Learning," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (PMLR)*, 2022, pp. 2372–2396.

[38] H. Fayek, L. Cavedon, and H. Wu, "Progressive learning: A deep learning framework for continual learning," *Neural Networks*, vol. 128, pp. 345–357, 2020.

[39] K. Ricanek and T. Tesafaye, "MORPH: a longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, Southampton, UK, 2006, pp. 341–345.

[40] Z. Zhang, Y. Song, and H. Qi, "Age Progression/Regression by Conditional Adversarial Autoencoder," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 4352–4360.

[41] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[42] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. Han, Z. Zhao, and B. Li, "Age Group and Gender Estimation in the Wild With Deep RoR Architecture," *IEEE Accesss*, vol. 5, pp. 22 492–22 503, 2017.

[43] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.

[44] A. Gurnani, K. Shah, V. Gajjar, V. Mavani, and Y. Khandhediyae, "SAF-BAGE: Salient Approach for Facial Soft-Biometric Classification - Age, Gender, and Facial Expression," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2019, pp. 839–847.

[45] K. Khan, M. Attique, R. Khan, I. Syed, and T.-S. Chung, "A Multi-Task Framework for Facial Attributes Classification through End-to-End Face Parsing and Deep Convolutional Neural Networks," *Sensors*, vol. 20, no. 2, p. 328, 2020.

[46] A. Garain, B. Ray, P. Singh, A. Ahmadian, N. Senu, and R. Sarkar, "GRA_Net: A Deep Learning Model for Classification of Age and Gender From Facial Images," *IEEE Access*, vol. 9, pp. 85 672–85 689, 2021.

[47] A. Saha, S. Kumar, and P. Nithyakani, "Age and Gender Prediction using Adaptive Gamma Correction and Convolutional Neural Network," in *International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2023, pp. 1–5.

[48] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image and Vision Computing*, vol. 32, no. 10, pp. 761–770, 2014.

[49] H. Han, C. Otto, X. Liu, and A. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, 2015.

[50] F. Wang, H. Han, S. Shan, and X. Chen, "Deep Multi-Task Learning for Joint Prediction of Heterogeneous Face Attributes," in *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, Washington, DC, USA, 2017, pp. 173–179.

[51] H. Han, A. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2597–2609, 2018.

[52] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 284–90, 2017.

[53] G. Huang, Z. Liu, L. V. D. Maaten, and K. Weinberger, "Densely Connected Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261–2269.

[54] A. Das, A. Dantcheva, and F. Bremond, "Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-task Convolution Neural Network Approach," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 573–585.

[55] H. Yuan, Y. He, P. Du, and L. Song, "Multi-task learning using uncertainty to weigh losses for heterogeneous face attribute estimation," 2024. [Online]. Available: https://arxiv.org/abs/2403.00561

Improving Cross-Patient Epilepsy Detection via EEG Decomposition into Canonical Brain Rhythms with Deep Learning

Jose Yauri, Elinar Carrillo-Riveros, Edith Guevara-Morote, Juan Carlos Carreño-Gamarra,
Karel Peralta-Sotomayor, Pelayo Quispe-Bautista
Dept. of Mathematics and Physics, Universidad Nacional de San Cristóbal de Huamanga, Ayacucho, Peru

***Abstract*—Epilepsy affects more than 50 million people worldwide, and almost 80% of them live in low-income countries with limited access to medical and public services. Beyond these challenges, epileptic patients also face other problems, such as stigma and social exclusion due the misunderstanding of epilepsy. Thus, epilepsy has become a major public health problem with a high social impact. Electroencephalography (EEG) remains the primary tool for diagnosing epilepsy; however, the traditional procedure of reviewing long EEG recordings is time-consuming, error-prone, and highly dependent on the neurologist's experience. Recent advances in deep learning (DL) have driven the development of new methods for automatic epilepsy detection. Despite these advances, most methods are not generalizable to all patients, limiting their clinical applicability in real-life cases. In this work, we present a cross-patient method capable of improving epilepsy detection by spectral decomposition of EEG signals into canonical brain rhythms. These spectral bands improve the signal significance and the model performance. The proposal was evaluated in a cross-patient validation scheme on the CHB-MIT dataset and proved superior performance using EEG signals from the interictal and ictal epilepsy stages. The model achieved of 100% of sensibility and specificity using the theta band, outperforming the state-of-the-art methods and offering a promising step towards real-world clinical implementation.**

***Keywords*—EEG signals; EEG signal decomposition; canonical brain rhythms; deep learning; convolutional neural network; transformer neural network**

I. Introduction

Epilepsy is a neurological disorder that is one of the most severe and affects the normal functioning of the brain [1]. Epilepsy is characterized by abnormal electrical activity in neurons that initiates seizures, which is the observable manifestation of this disease [2]. Seizures or convulsions consist of movements of the body, arms, and hands, and are often accompanied by loss of consciousness, fainting, and salivation [3]. The degree of seizures can vary, from subtle to strong manifestations. When seizures occur more than once in a day, the quality of life of patients deteriorates dramatically [4]. Furthermore, numerous epileptic patients experience social exclusion and stigmatization [5].

Epilepsy can affect anyone, regardless of age, gender, race, social status, or geographic location; it affects both children and adults [6]. Current research is centered on identifying the causes, medical treatments, and diagnosis of epilepsy since there is no definite cure [7]. The World Health Organization states that epilepsy affects over 50 million people worldwide, and most of them are living in poor or developing countries without access to medical services or treatments [8]. Because epilepsy strongly changes the lives of people, it has turned into a global public health problem with high social impact that deserves attention [9].

But not all seizures are epilepsy itself. Other diseases also cause seizures (e.g., Alzheimer's disease [10], stroke [11], diabetes [12]). Therefore, the goal of physicians and neurologists is to obtain a precise diagnosis of epilepsy to provide the best treatment [13]. Poor treatment of epilepsy could even worsen the disease [14].

Due to its low cost and non-invasive nature, the electroencephalogram (EEG) has become the standard method for diagnosing epilepsy [15]. EEG captures the electrical activity of neurons using electrodes placed on the scalp of the head [1]. In order to diagnose epilepsy, the physician visually observes the EEG signals to identify patterns of spikes, sharp, and slow waves that are characteristic of an epileptic seizure [16]. Although it may seem very simple, this work is visually-intense, time-consuming, and error-prone, depending on neurologist expertise [17]. So, it is crucial to have a proper diagnosis of epilepsy in order to provide an adequate treatment.

In the past two decades, there have been numerous research efforts conducted to create automatic tools for detecting epileptic seizures [18], [19], [20] and for predicting seizure episodes [21], [22], [23]. The majority of research studies rely either on machine learning (ML) or deep learning (DL) algorithms to detect seizure patterns in EEG signals. Among them, DL is one of the most popular due to its ability to model a more flexible feature space [24].

In spite of the recent progress, the classification of epileptic seizures still faced some challenging problems. Two main issues are addressed in this work:

- The majority of methods for developing a DL classifier to detect epileptic seizures commonly use raw EEG signals as input data. However, decomposing the EEG signal into major brain rhythms is more suitable for clinical diagnosis, as it enhances the interpretation of the signal [25]. Can classification performance be improved by the brain rhythms of EEGs?
- The majority of methods to detect epileptic seizures usually assess their model's performance at a cross-validation and intra-subject level. However, cross-patient assessment remains limited due to the difficulty

of achieving a good generalization model. Could it be feasible to create a population-level classifier by analyzing specific EEG brain rhythms/bands?

Therefore, the main contribution of this work is twofold:

- An approach that decomposes the raw EEG signal into the four main brain rhythms: theta (θ), alpha(α), beta(β), and gamma (γ) waves. A basic transformer neural network proposed in [26] is employed to objectively assess the performance caused by the decomposition of EEG signals .
- Two ways to evaluate the model's performance. First, a k-fold cross-validation level to determine how well the model works with the maximum number of raw data . Second, to determine how effectively the model works with the decomposed brain bands on a cross-patient assessment level.

The remainder of this study is organized as follows: Section II presents the fundamentals about epilepsy, as well as the related work. Section III exposes our proposal to detect epileptic seizures. Section IV shows the achieved results and provides a comparison against previous work. Lastly, Section V summarizes the findings of this study and the upcoming investigations.

II. Background

Since its invention, EEG has become the standard device for diagnosing epilepsy and investigating other brain disorders [1]. The neuronal electrical activity is recorded by the EEG through electrodes placed over the head. Thereby, EEG allows a real time investigation of what is happening in the brain due to the fact that each electrode registers data from a specific region of the head. As a result, the EEG produces a recording file that can be further analyzed by the neurologist [27]. Fig. 1 illustrates an EEG headset and its electrode placements over the head. The arrows indicate the EEG montage used during analysis.

Traditional epilepsy diagnosis involves human visual analysis of EEG registers, which is both time-consuming and susceptible to misdiagnosis. To overcome these issues, in the last years, many methods have been proposed for automatic detection of epilepsy [18], [19], [20]. Therefore, the goal of epileptic seizure detection is to classify EEG time windows as normal (non-seizure) or abnormal (seizure) through supervised binary classification.

To build a classifier, it is desirable to have a large amount of EEG data [24]. However, researchers select EEG data to reduce imbalance between seizures and non-seizures classes due to their strong unbalance. So, given a long-time EEG recording, investigators extract specific parts of the signal. These parts are taken from the stages of epilepsy that the patient experiences over time: the interictal, preictal, ictal, and postictal stages [1], [2].

The ictal stage refers to the patient's seizure episode, and the other stages are placed in relation to this stage over time. Fig. 2 shows the four phases of epilepsy in an long-time EEG recording. Note that the ictal stage lasts very short compared to other stages. Fig. 3 illustrates the ictal or seizure episode (red shaded), showing variability in the transition between episodes (green shaded).

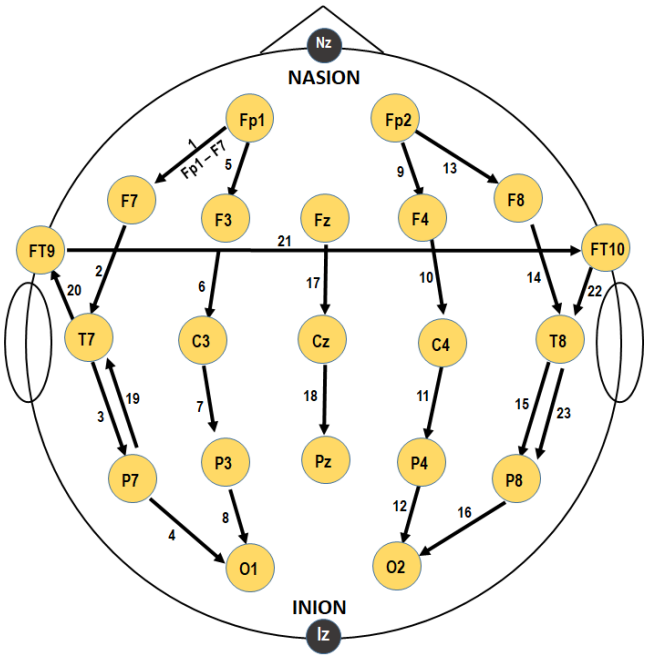


Fig. 1. An EEG device and its electrode placements according to the 10–20 system.

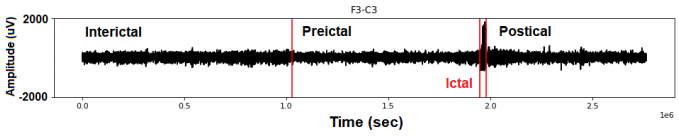


Fig. 2. Epilepsy phases in a long time EEG recording. For convenience, only a single channel is plotted.

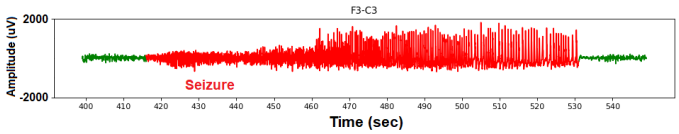


Fig. 3. An epilepsy seizure stage (red shaded). For convenience, only a single channel is plotted.

The interictal stage occurs a few hours before a seizure episode and the patient experiences a normal condition; the preictal stage occurs a few minutes preceding a seizure (e.g., 30 minutes); and the postictal stage occurs some minutes after a seizure (e.g., 5 minutes). However, it is worth mentioning that, yet there is no consensus about the duration of the interictal, preictal, and postictal stages due to the high variability of epilepsy symptoms between patients [28], [29], [30].

In order to construct their classification model for seizure detection, most researchers use signals from interictal and ictal stages as input data [31], [32], [33], [34], [35], and a few investigators use signals from preictal and ictal stages

[36], [37]. Regardless of the use of interictal versus ictal or preictal versus ictal sources to build the classification model, the authors have reported high accuracy results; however the lack of consensus on the signals used to train and test the model often prevents these results from being reproducible.

In order to examine how well models detect epilepsy using EEG signals, researchers have used various validation schemes to assess their generalizability and clinical relevance [38], [39]. One common approach is the k-fold cross-validation, that involves randomly partitioning an entire dataset composed of EEG recordings from multiple patients into training and testing subsets. The use of all available data for both training and testing in this method, typically results in highly optimistic performance results due to the possibility of EEG patterns of specific patients can be leaked between training and test folds. While this setup is useful for benchmarking algorithms in early development stages, it lacks the rigor needed to evaluate real-world deployment in clinical settings [26], [39].

Aiming to better approximate practical use cases, researchers also use two more realistic validation schemes: the intra-patient and cross-patient (or patient-independent) evaluation. In one hand, in the intra-patient mode [31], the model is both trained and tested exclusively on EEG data from the same individual. This approach produces a model that is tailored to the patient and can effectively capture their distinct seizure characteristics, but they are limited in scalability and do not generalize to new patients. Furthermore, as epilepsy symptoms can change over time even in the same patient, the model should be updated as well [40].

On the other hand, the cross-patient mode [36], [30] is considered as an ideal evaluation setting. This approach uses data from one patient as the test set, while the model is trained on data from other patients. This scheme examines the generalizability of the model across individuals, which highlights its potential for clinical application in real-world. However, achieving high performance in this experimental setup remains challenging due to inter-patient variability in EEG patterns, making it a key focus of ongoing research [38], [39].

Ultimately, although DL approaches have achieved significant success in epilepsy detection using raw EEG signals, the exclusive reliance on unprocessed data may overlook clinically informative structures inherent in the signal. A notably underexplored yet promising alternative is the decomposition of EEG signals into canonical brain rhythms, such as thetaθ, alphaα, betaβ, and gammaγ bands, which are foundational to neurophysiological interpretation [1]. Moreover, from a clinical perspective, such decomposition aligns with well established diagnostic practices and facilitates the identification of abnormal frequency components associated with epileptic activity [2]. In addition, rhythm-based decomposition can enhance data quality by reducing noise and concentrating features that are more stable in the context of inter-patient variability [41].

In the context of DL, these frequency-specific components could serve as semantically enriched inputs, potentially improving both model generalization and interpretability. So, we hypothesize that selecting and integrating brain rhythm into the DL pipeline not only offers a biologically grounded data representation but also contributes to bridging the gap between computational learning strategies and clinical reasoning [42]. Therefore, it is expected to achieve an improvement in the effectiveness and the trustworthiness of diagnostic models to detect epilepsy.

This work studies the influence of brain rhythms on the performance of the model in the cross-patient evaluation scheme. In this way, although there have been many advances in epilepsy detection, the following literature revision is restricted to studies that have been experimented in cross-patient schemes that use preferably the largest public CHB-MIT EEG dataset [43]. Moreover, the basic assumption is that the extraction of spatial and temporal features is excellent and that the unbalanced nature of datasets is manageable. The studies of [38], [39] provide a deep insight into these problems.

The following are the most relevant works:

The first study to address universal generalization in cross-patient assessment scheme was presented by Hossain et al. [36]. The authors propose the EEG channel fusion in the first layer before proceeding to extract temporal features in the next layers. The model was trained by extracting data from the preictal and ictal EEG recordings, with a 2-second time window and an 80% overlap. The complete model consists of four CNN blocks (each block containing a convolution, an activation function, and a max-pooling). Evaluation in the CHB-MIT dataset, the models showed sensitivity of 90%, specificity of 91.65%, and accuracy of 98.05%. Despite the achieved high performance, the data selection process is unclear and the number of assessed seizures remains unknown.

Next, Liu et al. [44] also focused on detecting epilepsy in a cross-patient scheme. They propose a CNN-BiLSTM (Bidirectional Long Short Term Memory Network) model enhanced with a novel channel perturbation layer (CPL) during training, which randomizes EEG channel order to improve spatial generalization. First, EEG data is preprocessed using wavelet decomposition, focusing on clinically relevant brain rhythms (4–32 Hz). Evaluation in the CHB-MIT dataset, the models showed sensitivity of 86.5%, and AUC-ROC improvements from 77% to over 90% due to CPL. However, this study lacks interpretability analysis and offers limited insight into physiological relevance of learned features.

Then, Alqirshi et al. [45] proposed a cross-patient seizure detection method that utilizes graph convolutional networks (GCN) with the goal of overcoming the major limitations of conventional EEG analysis: the insufficient attention to inter-channel brain dynamics and reliance on patient-specific training. Hence, the GCN represents the EEG data as graph-structured input to capture spatial dependencies. Evaluated in a custom dataset from four unseen patients, the GCN-based approach achieved a sensitivity of 88.71%, precision of 91.32%, F1-score of 91.57%, and an accuracy of 91.70%. Although the method has reported high detection metrics, the dataset is too limited to prove that the model is highly generalizable. Moreover, it would benefit from a more transparent analysis of how specific graph features contribute to classification, and testing on larger data sets, such as the CHB-MIT database.

Next, Jana et al. [46] presented a channel selection strategy for EEG-based seizure prediction aimed at using only a single optimal channel in a cross-patient assessment. The authors propose a method that ranks each EEG channel based on

its average performance over five evaluation rounds using a CNN-LSTM1D model. The top-performing channel is then selected for prediction, and an ensemble learning framework (integrating CNN1D, DenseNet1D, and CNN-LSTM1D) is employed using majority voting. Evaluation in the CHB-MIT dataset, the model achieves a sensitivity of 94.25%, and a specificity of 92.94%. The study fails to explain why specific channels perform better than others, and assumes that a single channel cannot be generalized to all seizure types or patients with atypical EEG presentations. Furthermore, using ensembles increases complexity and could prevent real-time deployment even with a single-channel design.

Recently, Mohammadpoory et al. [47] proposed a method for epileptic seizure detection based on the weighted visibility graph (WVG). Features are extracted from both raw EEG signals and their decomposed sub-bands (delta to gamma) from the preictal and ictal EEG recordings. A combination of multiple feature selection techniques is used to select key features and evaluate them using five different classifiers. Next, a post-processing step is performed to improve detection accuracy and accurately identify seizure onset and offset. Evaluation in the CHB-MIT dataset, the model detected 163 out of 184 seizures, achieving a sensitivity of 92.31%, a specificity of 94.12%, and an accuracy of 94.02%. Despite the high performance, the main drawback is the complexity of feature engineering, which could hinder real-time applications and scaling over low-power clinical systems.

In summary, there are only a few studies focused on cross-patient or patient-independent settings. The high variability of epileptic seizures, even between the same patients, prevents achieving a generalizable model. It is worth mentioning that sub-bands as input data have only been explored by a small number of researchers, but those who have employed them have shown better results, such as the study of [47]. The main focus of this work is on EEG signal decomposition; we hypothesize it is the primary source of improvement for building a model that is highly generalizable at a cross-patient level.

III. Our Approach

As stated in Section I, this work is an extension of the study of [26], which aimed to detect epilepsy in a k-fold cross-validation scheme. Different from the previous one, this work primarily concentrates on decomposing the EEG signal into canonical brain rhythms and aims to detect epilepsy in a cross-patient validation scheme, with minor changes during application.

Fig. 4 outlines a general pipeline for detecting epileptic seizures in EEG data using DL. First, a short description of the data acquisition is provided to collect a dataset of EEG signals. Then, the signal preprocessing is performed. Next, a neural classifier is trained in order to discriminate between non-seizure and seizure signals. Finally, the trained classifier is evaluated in cross-patient mode to evaluate its generalization capability.

A detailed description of each step of the pipeline is provided as follows:

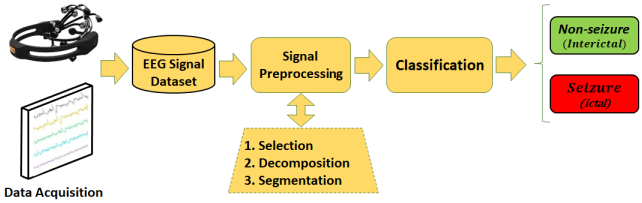


Fig. 4. The general pipeline for epileptic seizure detection.

A. EEG Dataset

The EEG dataset used in this study is derived from the CHB-MIT public dataset that contains more than 980 hours of EEG recordings and 198 seizures [43]. In extent, the dataset contains EEG registers from 23 pediatric patients, 3 to 22 age, with incurable epilepsy. Because another register of patient chb01 was obtained 1.5 years later, the chb21 register is considered a new patient's register.

EEG data was collected using EEGs of different numbers of electrodes, but registers of EEG of 23 channels at 256 Hz are the most common, which are organized in a longitudinal bipolar montage (see Fig. 1). Because the records are of long-term duration, as a result, the dataset is released in registers of one, two, or four hours long recordings. If the register contains at least one seizure, then it is named seizure records; otherwise, it is named non-seizure records. Each seizure register provides the start and end for each seizure episode as ground truth (GT).

The CHB-MIT dataset has been utilized extensively in numerous studies because it has a large amount of EEG recordings per patient and is openly accessible [48], [19]. On the other hand, there are other databases, but they are either small or privately accessible [49]. For instance, the Bonn EEG dataset contains sequences of 23.6 seconds long from a single EEG electrode, which are not suitable for training modern neural models. On the other hand, the EPILEPSIAE dataset is a large EEG database that contains continuous and long-term EEG recordings from 250 epileptic patients, but it is private.

B. Preprocessing

In this stage, four main activities are proposed: data selection, data decomposition, data segmentation, and data augmentation.

- Data selection aims to select and extract EEG registers to be used to train and test the model. In this work, we selected and extracted data from the interictal and ictal stages that best discriminate between non-seizure and seizure signals (see Fig. 2 to illustrate the selected data coming from). The use of interictal and ictal data to train classifiers is widely supported by previous studies [18], [38], [39].
  Therefore, the EEG signals come from EEG registers of 23 channels and contains a total of 181 seizures. The interictal data comes from signals two hours away of a seizure, whereas the ictal data comes from all seizures. Also, the data was downsampled to 128 Hz like in [26].
- Data decomposition aims to disassemble the raw EEG signal into four main brain rhythms: theta (θ: 4–8

Hz), alpha(α:8–12 Hz), beta(β:12–30 Hz), and gamma (γ:30–50 Hz) waves. Signal decomposition was performed via the Fast Fourier Transform (FFT). The delta (δ) The delta signal was excluded from this analysis due to its association with sleep [25].
The process of EEG signal decomposition is performed using the Fast Fourier Transform (FFT), with the goal of isolating and retaining the frequency components of a desired brain band [50]. Once the signal is isolated, it is possible to analyze the neural activity within the selected frequency band.

- Data segmentation aims to split the selected and decomposed EEG signals into small processable time windows. In accordance to the study [26], we use a time window of one second for non-seizure data, while seizure data has an 80% overlap to increase the amount of data.

C. Neural Network

To ensure an impartial evaluation of the proposed hypothesis and compare potential gains in model performance in the cross-patient evaluation scheme, we suggest reusing the model presented in [26]. Fig. 5 illustrated the reused neural architecture in this work.

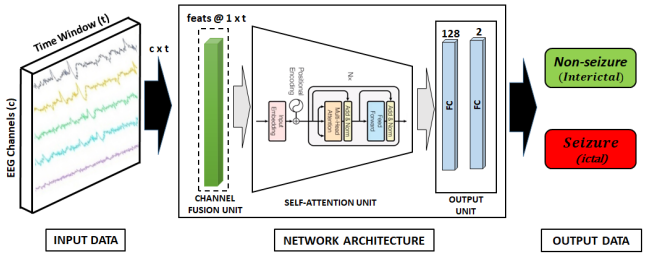


Fig. 5. The neural network architecture.

The proposed neural model has the main advantage of being a state-of-the-art (SoA) model for epilepsy detection in a k-fold cross-validation scheme and leverages the most recent advances in EEG channel fusion and attention mechanism.

D. Classification

After training, the model has the ability to discriminate between non-seizure and seizure signals. To determine the label class of input data, the final layer of the model calculates the probability distribution with the Softmax activation function [24].

E. Experimental Design

To determine the feasibility of the proposed brain rhythms decomposition into sub-bands to improve the model's ability to detect epileptic seizures in multiple patients, we use the leave-one-patient-out (LOPO) validation scheme [24].

In LOPO, the model is trained on EEG data from all patients except one, and then tested on the held-out patient. This process is repeated for every patient to guarantee that their data acts as an independent test sample only once. Finally, LOPO offers a reliable estimation of the model's performance in real-world situations where the system needs to detect epileptic seizures in patients who have never had seizures before [47].

IV. Results and Discussion

In this work, we used the neural architecture presented in [26], which was implemented in the DL Framework PyTorch v.2.4 and Python v.3.9. The experimental platform consisted of a computer desktop Intel CPU I7-8550U, 16GB RAM, equipped with a GPU NVIDIA RTX 2070 with 8GB RAM. The model was trained from scratch with the following hyperparameters: the Adam optimizer, a weighted cross-entropy loss function to deal with unbalanced data samples, a batch size of 128, a learning rate of 0.0001, and 150 epochs for training the model.

The model performance was assessed using the LOPO validation scheme. Metrics such as sensitivity, specificity, precision, F1-Score, and accuracy are employed to display the results in the format mean average plus or minus the standard deviation.

Table I summarizes the achieved results in the CHB-MIT database. To highlight the difference between decomposition and non-decomposition, the non-decomposition (i.e., raw data) is also displayed in the first row of the table. The next rows of the table show the spectral signals that have been disaggregated (theta, alpha, beta, and gamma). The decomposition approach is proven to produce better results than the non-decomposition approach, particularly by utilizing only the theta spectral band (gray shaded row).

Table II shows the achieved results of the model only using the non-decomposed signal (or raw data) as input data. To gain an understanding of individual performance, each row of the table shows the performance of each patient. It is noticeable that there is variability in patients' performance.

Using the decomposed approach, the achieved by the model shown in Table III. In this situation, the input data is represented by the theta band. The table displays the performance of each patient in each row, and it is clear that the variability between patients' performance has decreased.

To compare the achieved results with the proposed approach and related work fairly, we selected the most recent methods of SoA that focus on cross-patient or patient-independent validation schemes (i.e., LOPO). Table IV summarizes the results reported by several SoA works. After analyzing the results, it is evident that our proposed method outperforms other studies by a significant margin. Only the studies of Jana et al. [46] and Mohammadpoory et al. [47] are among the most close to our reported results.

The following points are worth discussing:

First, although many methods have been proposed for detecting epileptic seizures, only a few are capable of generalizing to patient-independent scenarios, due to the challenges in capturing inter-patient variability and the high diversity of seizure patterns (see Table IV). However, the deployment of reliable and trustworthy seizure detection systems in clinical settings will not become a reality unless we continue to address cross-patient variability.

TABLE I. SUMMARY OF THE CLASSIFICATION EFFECTIVENESS OF THE PROPOSED METHOD IN THE LOPO EVALUATION SCHEME ON THE CHB-MIT DATABASE

Brain rhythm	Sensitivity	Specificity	Precision	F1-score	Accuracy
raw data	93.27±11	96.96±7.59	71.87±32.5	75.69±26.6	96.63±7.39
theta	100±0	100±0	100±0	100±0	100±0
alpha	99.98±0.09	100±0	100±0	99.99±0.04	100±0
Beta	99.91±0.26	100±0.01	99.94±0.31	99.92±0.19	100±0.01
Gamma	99.97±0.08	99.95±0.12	98.58±3.81	99.23±2.03	99.96±0.12

TABLE II. CLASSIFICATION RESULTS USING THE NON-DECOMPOSED EEG SIGNAL IN THE LOPO EVALUATION SCHEME ON THE CHB-MIT DATABASE

Patient	Sensitivity	Specificity	Precision	F1-score	Accuracy
chb01	94.12	99.96	98.81	96.41	99.76
chb02	98.26	95.52	23.02	37.31	95.55
chb03	96.52	99.96	98.73	97.61	99.85
chb04	98.94	98.44	62.96	76.95	98.45
chb05	98.03	98.81	78.48	87.17	98.78
chb06	92.16	98.97	37.11	52.91	98.92
chb07	95.38	63.33	5.55	10.48	64.04
chb08	100.00	99.56	95.83	97.87	99.60
chb09	100.00	92.50	20.35	33.82	92.64
chb10	99.33	99.42	84.25	91.17	99.42
chb11	97.15	99.98	99.75	98.43	99.81
chb12	51.97	99.94	99.23	68.21	94.15
chb13	97.95	98.22	81.78	89.14	98.20
chb14	70.41	99.06	50.00	58.48	98.68
chb15	88.40	99.79	98.55	93.20	98.24
chb16	97.10	99.63	58.77	73.22	99.61
chb17	92.49	92.58	22.47	36.16	92.58
chb18	85.80	99.99	99.63	92.20	99.64
chb19	88.56	99.94	96.31	92.27	99.73
chb20	99.66	99.99	99.66	99.66	99.98
chb21	99.50	99.96	97.54	98.51	99.95
chb22	98.04	99.96	97.56	97.80	99.93
chb23	100.00	92.10	27.16	42.72	92.33
chb24	98.63	99.35	91.47	94.92	99.30
	93.27±11	96.96±7.59	71.87±32.5	75.69±26.6	96.63±7.39

TABLE III. CLASSIFICATION RESULTS USING THE THETA BAND SIGNAL IN THE LOPO EVALUATION SCHEME ON THE CHB-MIT DATABASE

Patient	Sensitivity	Specificity	Precision	F1-score	Accuracy
chb01	100.00	100.00	100.00	100.00	100.00
chb02	100.00	100.00	100.00	100.00	100.00
chb03	100.00	100.00	100.00	100.00	100.00
chb04	100.00	100.00	100.00	100.00	100.00
chb05	100.00	100.00	100.00	100.00	100.00
chb06	100.00	100.00	100.00	100.00	100.00
chb07	100.00	100.00	100.00	100.00	100.00
chb08	100.00	100.00	100.00	100.00	100.00
chb09	100.00	100.00	100.00	100.00	100.00
chb10	100.00	100.00	100.00	100.00	100.00
chb11	100.00	100.00	100.00	100.00	100.00
chb12	100.00	100.00	100.00	100.00	100.00
chb13	100.00	100.00	100.00	100.00	100.00
chb14	100.00	100.00	100.00	100.00	100.00
chb15	100.00	100.00	100.00	100.00	100.00
chb16	100.00	100.00	100.00	100.00	100.00
chb17	100.00	100.00	100.00	100.00	100.00
chb18	100.00	100.00	100.00	100.00	100.00
chb19	100.00	100.00	100.00	100.00	100.00
chb20	100.00	100.00	100.00	100.00	100.00
chb21	100.00	100.00	100.00	100.00	100.00
chb22	100.00	100.00	100.00	100.00	100.00
chb23	100.00	100.00	100.00	100.00	100.00
chb24	100.00	100.00	100.00	100.00	100.00
	100±0	100±0	100±0	100±0	100±0

Second, with respect to the type of EEG input data, earlier studies primarily utilized raw signals [36], [45], whereas more recent approaches have begun to incorporate canonical brain rhythms. Hence, the studies of Liu et al. [44] and Mohammadpoory et al. [47] specifically perform EEG signal decomposition into main brain rhythms. But, while previous approaches attempted to combine features from multiple frequency bands using complex algorithms, this work evaluates each band individually, revealing that using isolated frequency bands as input data provides superior performance.

TABLE IV. COMPARISON OF THE PROPOSED METHOD AND RELATED WORK ON THE CHB-MIT DATABASE

Author	Method	Total seizures	Sensitivity	Specificity	Precision	F1-score	Accuraccy
Hossain et al. [36]	CNN	-	90.00	91.65	-	-	98.05
Liu et al. [44]	CNN-BiLSTM	-	86.50	-	-	-	-
Alqirshi et al. [45]	GCN	-	88.71	-	91.32	91.57	91.70
Jana et al. [46]	CNN-LSTM1D	-	94.25	92.94	-	-	-
Mohammadpoory et al. [47]	WVG-RF	163	92.31	94.12	-	-	94.02
This work (θ band)	CNN-Transformer	181	100.00	100.00	100.00	100.00	100.00

Third, the analysis of decomposed EEG signals is useful and intuitive from a clinical diagnostic perspective. For instance, neurologists and physicians are trained to recognize patterns within specific brain frequency bands and associate them with specific neurological conditions (e.g., delta band analysis to diagnose sleep disorders). Therefore, the analysis of canonical brain rhythms offers new opportunities to deepen our understanding of brain function, cognitive processes, and associated diseases [1], [42].

V. CONCLUSION

In this work, a new method to detect epileptic seizures in a cross-patient scenario has been presented. The method is based on the use of isolated frequency bands as input data to achieve superior model performance. This approach can be utilized to fuel any deep neural network, boosting its performance.

The feasibility of the proposed method was validated in the public CHB-MIT EEG dataset using the leave-one-patient-out (LOPO) evaluation scheme with 24 patients. We trained the well-established CNN-Transformer deep neural network proposed in [26], using individual EEG frequency bands (i.e., theta, alpha, beta, and gamma) as input data. Compared to using the raw EEG signal, our approach provides significantly superior performance. Furthermore, our approach outperforms several state-of-the-art methods and offers a promising step toward real-world clinical systems implementation.

For future work, it is necessary to conduct more research to enhance epileptic seizure recognition in clinical setting. First, the proposed method could be validated on other epilepsy datasets, such as the EPILEPSIAE database, to evaluate its generalizability and robustness across a population of patients. Second, instead of fixed band boundaries, some kind of adaptive decomposition may be feasible depending on the type of patient and his illness. Third, it is desirable to employ low-cost devices to implement and deploy the neural model, aiming to emulate clinical environments for epilepsy diagnosis in developing countries.

REFERENCES

[1] G. Cascino, J. I. Sirven, and W. O. Tatum, *Epilepsy*, 2nd ed. Wiley, 2021.

[2] V. S. Wasade and M. V. Spanaki, *Understanding Epilepsy: A Study Guide for the Boards*. Cambridge University Press, 2019.

[3] Mayo Clinic, "Epilepsy - Symptoms and Causes," 2025. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/epilepsy/symptoms-causes/syc-20350093

[4] M. Tombini, G. Assenza, L. Quintiliani, L. Ricci, J. Lanzone, and V. Di Lazzaro, "Epilepsy and Quality of Life: What does Really Matter?" *Neurological Sciences*, vol. 42, no. 9, pp. 3757–3765, sep 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10072-020-04990-6

[5] L. Hohmann, J. Berger, S. U. Kastell, and M. Holtkamp, "Perceived Epilepsy-related Stigma is Linked to the Socioeconomic Status of the Residence," *Frontiers in Public Health*, vol. 10, p. 952585, aug 2022.

[6] S. D. Howard, P.-A. Campbell, C. T. Montgomery, S. B. Tomlinson, D. I. Ojukwu, H. I. Chen, and M. H. Chin, "Effect of Race and Insurance Type on Access to, and Outcomes of, Epilepsy Surgery: A Literature Review," *World Neurosurgery*, vol. 178, pp. 202–212.e2, 2023.

[7] H. Kumar, S. Debnath, and A. Sharma, "Can Epilepsy Be Cured? A Review," *Health Sciences Review*, vol. 5, p. 100062, dec 2022.

[8] World Health Organization, "Epilepsy," 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[9] Center for Disease Control and Prevention, "Epilepsy Is a Public Health Problem," 9 2025. [Online]. Available: https://www.cdc.gov/epilepsy/communications/infographics/cdc-epilepsy-text.htm

[10] J. Vöglein, I. Ricard, S. Noachtar, W. A. Kukull, M. Dieterich, J. Levin, and A. Danek, "Seizures in Alzheimer's Disease Are Highly Recurrent and Associated with a Poor Disease Course," *Journal of Neurology*, vol. 267, no. 10, p. 2941, 10 2020.

[11] J. Phan, M. Ramos, T. Soares, and M. S. Parmar, "Poststroke Seizure and Epilepsy: A Review of Incidence, Risk Factors, Diagnosis, Pathophysiology, and Pharmacological Therapies," *Oxidative Medicine and Cellular Longevity*, vol. 2022, no. 1, p. 7692215, jan 2022.

[12] A. Nangia, J. S. Saravanan, S. Hazra, V. Priya, R. Sudesh, S. S. Rana, and F. Ahmad, "Exploring the Clinical Connections Between Epilepsy and Diabetes Mellitus: Promising Therapeutic Strategies Utilizing Agmatine and Metformin," *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol. 397, no. 12, pp. 9617–9632, 2024. [Online]. Available: https://doi.org/10.1007/s00210-024-03295-1

[13] U. Amin and S. R. Benbadis, "The Role of EEG in the Erroneous Diagnosis of Epilepsy," *Journal of Clinical Neurophysiology*, vol. 36, no. 4, pp. 294–297, 7 2019.

[14] W. Löscher and P. Klein, "The Pharmacology and Clinical Efficacy of Antiseizure Medications: From Bromide Salts to Cenobamate and Beyond," *CNS Drugs*, vol. 35, no. 9, pp. 935–963, 2021.

[15] M. Z. Koubeissi and N. J. Azar, *Epilepsy Board Review : A Comprehensive Guide*, 2017.

[16] N. M. Jadeja, *How to Read an EEG*. Cambridge university press, 2021.

[17] W. O. Tatum, "Mistaking EEG Changes for Epilepsy," in *Common Pitfalls in Epilepsy: Case-Based Learning*, D. Schmidt, W. O. Tatum, and S. Schachter, Eds. Cambridge: Cambridge University Press, 2018, ch. 2, p. 25–42.

[18] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad, A. Khosravi, A. F. Atiya, D. Aminshahidi, S. Hussain, M. Rouhani, S. Nahavandi, and U. R. Acharya, "Epileptic Seizures Detection Using Deep Learning Techniques: A Review," *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 5780*, vol. 18, no. 11, p. 5780, 5 2021.

[19] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, "Epilepsy Detection From EEG Using Complex Network Techniques: A Review," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 292–306, 2023.

[20] A. Lucas, A. Revell, and K. A. Davis, "Artificial Intelligence in Epilepsy - Applications and Pathways to the Clinic," *Nature Reviews Neurology 2024 20:6*, vol. 20, no. 6, pp. 319–336, may 2024.

[21] R. Hussein, S. Lee, and R. Ward, "Multi-Channel Vision Transformer for Epileptic Seizure Prediction," *Biomedicines*, vol. 10, no. 7, p. 1551, 2022.

[22] Z. Altaf, M. A. Unar, S. Narejo, M. A. Zaki, and Naseer-u-Din, "Generalized Epileptic Seizure Prediction using Machine Learning Method," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023.

[23] Z. Wang, X. Song, L. Chen, J. Nan, Y. Sun, M. Pang, K. Zhang, X. Liu, and D. Ming, "Research Progress of Epileptic Seizure Prediction Methods Based on EEG," *Cognitive Neurodynamics*, vol. 18, no. 5, pp. 2731–2750, 2024.

[24] C. M. Bishop and H. Bishop, *Deep Learning : Foundations and Concepts*. Springer, 2024.

[25] R. Sharma and H. K. Meena, "Emerging Trends in EEG Signal Processing: A Systematic Review," *SN Computer Science*, vol. 5, no. 4, pp. 1–14, apr 2024.

[26] J. Yauri, M. Lagos, H. Vega-Huerta, P. De-La-Cruz-VdV, G. L. E. Maquen-Niño, and E. Condor-Tinoco, "Detection of Epileptic Seizures Based-on Channel Fusion and Transformer Network in EEG Recordings," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 1067–1074, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01405110

[27] L. Hu and Z. Zhang, *EEG Signal Processing and Feature Extraction*. Springer Singapore, 1 2019.

[28] A. Shoeb and J. Guttag, "Application of Machine Learning to Epileptic Seizure Detection," in *ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010, pp. 975–982.

[29] S. M. Usman, S. Khalid, R. Akhtar, Z. Bortolotto, Z. Bashir, and H. Qiu, "Using Scalp EEG and Intracranial EEG Signals for Predicting Epileptic Seizures: Review of Available Methodologies," *Seizure*, vol. 71, pp. 258–269, 10 2019.

[30] J. Zhou, L. Liu, Y. Leng, Y. Yang, B. Gao, Z. Jiang, W. Nie, and Q. Yuan, "Both Cross-Patient and Patient-Specific Seizure Detection Based on Self-Organizing Fuzzy Logic," *International Journal of Neural Systems*, vol. 32, no. 6, 2022.

[31] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, "Epileptic Seizure Detection Based on EEG Signals and CNN," *Frontiers in Neuroinformatics*, vol. 12, p. 95, 12 2018.

[32] Y. Li, Z. Yu, Y. Chen, C. Yang, Y. Li, X. Allen Li, and B. Li, "Automatic Seizure Detection using Fully Convolutional Nested LSTM," *International Journal of Neural Systems*, vol. 30, no. 4, 2020.

[33] X. Wang, X. Wang, W. Liu, Z. Chang, T. Kärkkäinen, and F. Cong, "One Dimensional Convolutional Neural Networks for Seizure Onset Detection using Long-term Scalp and Intracranial EEG," *Neurocomputing*, vol. 459, pp. 212–222, 2021.

[34] A. Abdelhameed and M. Bayoumi, "A Deep Learning Approach for Automatic Seizure Detection in Children With Epilepsy," *Frontiers in Computational Neuroscience*, vol. 15, p. 29, 2021.

[35] C. Li, W. Zhou, G. Liu, Y. Zhang, M. Geng, Z. Liu, S. Wang, and W. Shang, "Seizure Onset Detection Using Empirical Mode Decomposition and Common Spatial Pattern," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 458–467, 2021.

[36] M. Shamim Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying Deep Learning for Epilepsy Seizure Detection and Brain Mapping Visualization," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 15, no. 1s, 2 2019.

[37] S. Chakrabarti, A. Swetapadma, and P. K. Pattnaik, "A Channel Independent Generalized Seizure Detection Method for Pediatric Epileptic Seizures," *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106335, 9 2021.

[38] N. Fatma, P. Singh, and M. K. Siddiqui, "Survey on Epileptic Seizure Detection on Varied Machine Learning Algorithms," *https://doi.org/10.1142/S0219467825500135*, aug 2023.

[39] J. Xu, K. Yan, Z. Deng, Y. Yang, J. X. Liu, J. Wang, and S. Yuan, "EEG-based Epileptic Seizure Detection using Deep Learning Techniques: A Survey," *Neurocomputing*, vol. 610, p. 128644, dec 2024.

[40] G. M. Schroeder, P. J. Karoly, M. Maturana, M. Panagiotopoulou, P. N. Taylor, M. J. Cook, and Y. Wang, "Chronic Intracranial EEG Recordings and Interictal Spike Rate Reveal Multiscale Temporal Modulations in Seizure States," *Brain Communications*, vol. 5, no. 5, aug 2023.

[41] K. Lehnertz, T. Rings, and T. Bröhl, "Time in Brain: How Biological Rhythms Impact on EEG Signals and on EEG-Derived Brain Networks," *Frontiers in Network Physiology*, vol. 1, p. 755016, sep 2021.

[42] G. Buzsáki and M. Vöröslakos, "Brain Rhythms Have Come of Age," *Neuron*, vol. 111, no. 7, pp. 922–926, apr 2023.

[43] A. Shoeb and J. Guttag, "CHB-MIT Scalp EEG Database," 2010. [Online]. Available: https://physionet.org/content/chbmit/1.0.0/

[44] G. Liu, L. Tian, and W. Zhou, "Patient-Independent Seizure Detection Based on Channel-Perturbation Convolutional Neural Network and Bidirectional Long Short-Term Memory," *https://doi.org/10.1142/S0129065721500519*, vol. 32, no. 6, nov 2021.

[45] R. Alqirshi and S. B. Belhaouari, "EEG-Based Patient Independent Epileptic Seizure Detection Using GCN-BRF," in *Deep Learning Theory and Applications*, A. Fred, A. Hadjali, O. Gusikhin, and C. Sansone, Eds. Springer Nature Switzerland, 2024, pp. 351–366.

[46] R. Jana and I. Mukherjee, "Patient-Specific and Patient-Independent Seizure Prediction Using Ensemble Learning Technique," in *Lecture Notes in Networks and Systems*, vol. 1165. Springer, Singapore, 2025, pp. 41–52.

[47] Z. Mohammadpoory, M. Nasrolahzadeh, and S. A. Amiri, "Patient-independent Epileptic Seizure Detection using Weighted Visibility Graph Features and Wavelet Decomposition," *Multimedia Tools and Applications*, vol. 84, no. 6, pp. 3197–3221, feb 2025.

[48] A. A. Ein Shoka, M. M. Dessouky, A. El-Sayed, and E. E. D. Hemdan, "EEG Seizure Detection: Concepts, Techniques, Challenges, and Future Trends," *Multimedia Tools and Applications 2023 82:27*, vol. 82, no. 27, pp. 42 021–42 051, apr 2023.

[49] P. Handa, M. Mathur, and N. Goel, "EEG Datasets in Machine Learning Applications of Epilepsy Diagnosis and Seizure Detection," *SN Computer Science*, vol. 4, no. 5, pp. 1–11, sep 2023.

[50] S. Sanei and J. A. Chambers, *EEG Signal Processing and Machine Learning*, 2nd ed. Wiley, 2021.

Evaluating and Interpreting Pooling Techniques in Spectrogram-Based Audio Analysis Using Diverse Metrics

Supun Bandara, Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa, Katubedda, 10400, Sri Lanka

***Abstract*—Audio analysis is a rapidly advancing field that spans various domains, including speech, music, and environmental sound data. Using spectrograms with Convolutional Neural Networks (CNNs) enables the visualization and extraction of critical audio features by combining time-frequency representations with deep learning. Pooling plays a crucial role in this process, as it reduces dimensionality while retaining essential information. However, existing evaluations of pooling methods primarily emphasize downstream task performance, such as classification accuracy, often overlooking their effectiveness in preserving critical signal features. To address this gap, we use 17 distinct metrics, categorized into four domains, to comprehensively assess various pooling operations. Furthermore, we explore the underexamined relationship between specific pooling techniques and their impact on feature retention across diverse audio applications. Our analysis encompasses spectrograms from three audio domains (speech, music, and environmental sound), identifying their key characteristics, and grouping them accordingly. Using this setup, we evaluate the performance of 12 pooling methods across these applications. By investigating the features critical to each task and evaluating how well different pooling techniques preserve them, we give insights into their suitability for specific applications. This work aims to guide researchers in selecting the most appropriate pooling strategies for their applications, enabling more granular evaluations, improving explainability, and thereby advancing the precision and efficiency of audio analysis pipelines.**

***Keywords*—Audio data analysis; pooling; deep learning; dimensionality reduction; spectrograms**

I. Introduction

Audio analysis has become an essential field of study with wide-ranging applications, including audio classification [1], [2], [3], speaker recognition [4], [5], [6], and sound event detection [7], [8], [9], [10], [11]. The ability to accurately process and analyze audio data is crucial for developing systems that can understand and interact with human environments. Over the past few decades, significant advances in machine learning (ML) and deep learning have revolutionized audio analysis, enabling more sophisticated and accurate methodologies.

Feature extraction is a critical step in the analysis of audio signals, where audio signals are transformed into representations that can effectively capture meaningful patterns and features. Over the years, a wide range of techniques have been developed to extract meaningful information from audio data [12]. Time-domain approaches analyze raw waveforms to derive features such as zero-crossing rate and signal energy [12], while frequency-domain methods, including the Fourier Transform [13] and Mel-Frequency Cepstral Coefficients (MFCCs) [14], focus on spectral characteristics. Time-frequency representations, such as spectrograms [15] and chromagrams [16], combine the strengths of both domains, capturing temporal dynamics and spectral content simultaneously. Recent advances in ML have further introduced data-driven methods, such as Convolutional Neural Networks (CNN) [17], Recurrent Neural Networks [18], and transformers [19], which learn task-specific features directly from raw audio or time-frequency representations.

By combining traditional signal processing approaches with advanced deep learning methods, these techniques can enable robust and scalable solutions, driving progress in diverse audio applications and opening the door to innovative research and development. One of those approaches is using spectrograms with CNNs. A significant advantage of using this lies in their ability to harness the strengths of both time-frequency representations and deep learning models [20], [21]. Spectrograms provide a rich and visually interpretable representation that captures both temporal dynamics and spectral patterns of audio signals. CNNs excel at learning hierarchical and spatial features, enabling the identification of complex harmonic structures, frequency modulations, and localized spectral features in the spectrogram. Despite their effectiveness, the integration of spectrograms with CNNs presents challenges due to its high dimensionality, which increases computational demands and memory usage [22]. To address this issue, pooling methods are employed to reduce the dimensionality of spectrograms while preserving essential information. Pooling techniques such as max pooling [1] and average pooling [2] summarize spectral content, thereby improving computational efficiency and enabling the training of deep learning models on large audio datasets.

Traditionally, the effectiveness of pooling methods in audio analysis has been evaluated indirectly using the accuracy of the downstream tasks such as classification [23]. However, this approach may not fully capture how well pooling methods preserve critical features within spectrograms. In this study, we apply 17 evaluation metrics categorized into four distinct groups for evaluating pooling methods used in audio analysis. These metrics can be used to directly assess the ability of pooling methods to preserve critical features and patterns in spectrograms, independent of downstream task results. By focusing on the intrinsic performance of pooling techniques, this approach enables a more granular analysis of their impact on the representation of audio features. Our investigation spans

12 pooling methods, facilitating a standardized comparison of these techniques. This analysis gives insights to identify the pooling methods best suited to specific applications by investigating which audio features should be preserved for specific applications and how well different pooling methods preserve these features. This work provides insights into optimizing the pooling process, helping researchers better understand the features relevant to their applications, and improve accuracy by identifying better pooling methods for their applications. Therefore, it facilitates better model design, improves the explainability of deep audio systems, and supports more informed decisions in the deployment of audio analysis pipelines for real-world applications.

The rest of the study is organized as follows: Section II provides the theoretical background, including a detailed review of pooling methods and evaluation metrics. Section III describes the system architecture. Section IV presents the experimental setup and includes an in-depth spectrogram analysis. Section V outlines the experimental findings across multiple audio domains, highlighting the quantitative performance of various pooling methods, and Section VI interprets the experimental findings, linking observed patterns to the functional demands of each audio task. Section VII concludes the study, highlighting key findings and directions for future work.

II. Theoretical Background

This section outlines the theoretical foundations of spectrogram-based audio analysis, focusing on the characteristics of speech, music, and environmental sounds. It introduces key pooling methods used to reduce spectrogram dimensionality while preserving important features. To assess their effectiveness, we also present evaluation metrics that measure information retention, structural preservation, local detail, and compression, forming a basis for analyzing pooling performance across audio tasks.

A. Local Details and Global Structure of a Spectrogram

Audio signals can generally be categorized into three primary types: speech, music, and environmental sounds [12].

1) Speech: Speech is generated by humans through the combined activity of organs such as the lungs, vocal cords, mouth, nose, and brain. The vocal cords and vocal tract play a crucial role in shaping speech sounds [24]. The frequency range of human speech typically begins at 100 Hz and can extend up to 17 kHz.

2) Music: Musical sounds are created by instruments or the human voice to produce harmony and express emotions. Music can be analyzed based on characteristics such as genre, mood, and tonal quality [25]. Traditional genres include rock, jazz, classical, and pop. The frequency range of music usually falls between 40 Hz and 19.5 kHz.

3) Environmental sounds: The sounds we encounter daily, such as those from vehicles, running water, doorbells, phones, machinery, and animals, are classified as environmental sounds [26]. These sounds often cover the full range of audible frequencies.

Fig. 1 displays the time-domain waveforms of three example sound types: human speech, a guitar note, and a car horn. Speech and music signals exhibit periodicity, whereas environmental sounds generally lack any clear periodic pattern. As shown in Fig. 1, speech signals are typically smooth and continuous, while guitar notes are brief and intermittent. In comparison, sounds such as a fire truck siren appear as noisy signals with high amplitude. Differences between these sounds are also apparent in the time-frequency domain, as illustrated in Fig. 2, which highlights the distinct frequency spectra of each sound type.

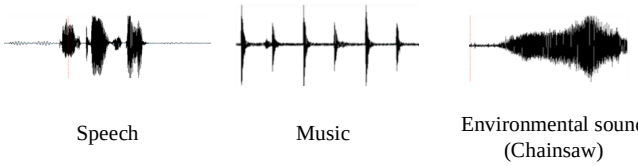


Fig. 1. Time-domain waveforms for speech, music, and environmental sound data.

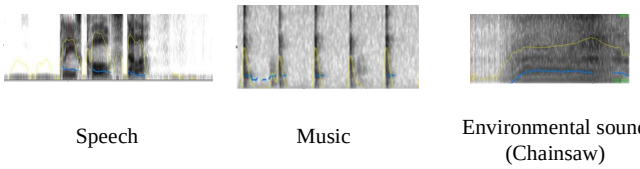


Fig. 2. Time-frequency domain waveforms for speech, music and environmental sound data.

By creating spectrograms for these three audio categories, we can analyze the features that need to be preserved during processing. This analysis is crucial when applying pooling or other transformations to spectrograms, as it helps ensure that the relevant features are retained in the pooled outputs. By analyzing spectrograms, we can determine what features to preserve for different applications. Based on the analysis, these applications can be categorized into two distinct groups: global structure preservation and localized detail retention. This categorization is driven by the need to either retain overall patterns in the spectrogram or emphasize specific localized features.

B. Review of Pooling Methods

Pooling techniques are important to audio analysis, offering a way to condense critical information over time or frequency dimensions. By summarizing features, pooling enhances the robustness of models against variations in speech signals, such as differences in speaking rates, accents, and background noise. This section explores a range of pooling methods, detailing their roles in reducing dimensionality while retaining essential features.

1) Average pooling: Average pooling [2], [27] computes the mean value of elements in a pooling region, and is expressed as in Eq. (1).

$$f_{\text{avg}}(x) = \frac{1}{N} \sum_{i=1}^{N} |x_i| \tag{1}$$

This method smooths feature maps by averaging values, effectively reducing noise. However, since it assigns equal weight to all elements, background regions can dominate, potentially diminishing the model's ability to distinguish between critical and irrelevant features.

2) Max pooling: Max pooling [1], [28], [29] selects the maximum value from a pooling region. Eq. (2) shows how max pooling can be used in a feature map.

$$f_{\max}(x) = \max\{x_i\}_{i=1}^{N} \tag{2}$$

This approach highlights the strongest activations, reducing the influence of background noise and emphasizing prominent features. Despite these advantages, max pooling can overlook subtle but important information and may amplify noisy elements with high values.

3) Min pooling: Min pooling [30] extracts the smallest value in a pooling region, as shown in Eq. (3).

$$f_{\min}(x) = \min\{x_i\}_{i=1}^{N} \tag{3}$$

By focusing on the weakest feature in each region, min pooling suppresses noise and outliers, making it useful in tasks such as anomaly detection. However, it shares limitations with max pooling, such as losing mid-range information and subtle details.

4) Mixed pooling: Mixed pooling [31] combines the strengths of max and average pooling by using a weighted combination as defined in Eq. (4).

$$f_{\text{mixed}}(x) = \alpha \cdot f_{\max}(x) + (1-\alpha) \cdot f_{\text{avg}}(x) \tag{4}$$

The parameter α determines the balance between max and average pooling. For $\alpha = 1$, it behaves as max pooling, and for $\alpha = 0$, it becomes average pooling. Mixed pooling can adapt to different tasks and datasets by capturing both salient features and broader contextual information.

5) Linear softmax pooling: Linear softmax pooling (LSP) [32] assigns weights to features based on their squared values, ensuring prominent features receive higher emphasis, as defined in Eq. (5).

$$f_{LSP} = \frac{\sum_i x_i^2}{\sum_i x_i} \tag{5}$$

where, f_{LSP} represents the output of the softmax function and x_i represents the individual elements of the input vector.

This method is particularly effective in scenarios, where identifying dominant peaks in data distributions is critical, as it balances emphasis on key features while considering all inputs.

6) Exponential softmax pooling: Exponential softmax pooling (ESP) [33] uses an exponential transformation to accentuate differences between input features in accordance with Eq. (6).

$$f_{ESP} = \frac{\sum_i x_i \exp(x_i)}{\sum_i \exp(x_i)} \tag{6}$$

where, $\exp(x_i)$ denotes the exponential function applied to the individual elements x_i in the input vector. By emphasizing higher values, this approach excels in tasks requiring sharp distinctions between features, such as identifying critical events in a data sequence.

7) Learned-norm pooling: As described in Eq. (7), learned-norm pooling (LNP) [34] unifies max and average pooling principles using the p-norm.

$$f_{LNP}^{p}(x) = \left(\frac{1}{n}\sum_{i=1}^{n}|x_i|^p\right)^{1/p} \tag{7}$$

The parameter p, ranging from 1 to infinity, determines the behavior of the pooling. When $p = 1$, it corresponds to mean pooling, while $p \to \infty$ approximates max pooling. LNP provides flexibility in capturing nuanced feature distributions.

8) Log-Sum-Exp pooling: Log-Sum-Exp Pooling(LSEP) [35] offers a smooth interpolation between mean and max pooling. This behavior is captured by Eq. (8).

$$f_{LSEP}^{r}(x) = \frac{1}{r}\log\left(\frac{1}{n}\sum_{i=1}^{n} exp(r \cdot x_i)\right) \tag{8}$$

The parameter r controls the pooling behavior, with $r \to 0$ resembling mean pooling and $r \to \infty$ approximating max pooling.

9) Auto-pooling: Auto-pooling [33] adapts dynamically to data, combining max and average pooling principles as formulated in Eq. (9).

$$f_{auto}^{\alpha} = \frac{\sum_j \mathbf{x}_j \exp(\alpha \mathbf{x}_j)}{\sum_j \exp(\alpha \mathbf{x}_j)} \tag{9}$$

This adaptability is particularly useful in sound event detection, where event duration and intensity vary significantly. Auto-pooling excels in environments with diverse and unpredictable audio patterns.

10) Power pooling: Power pooling [36], [37] introduces a trainable parameter n to adjust frame-level predictions, as defined in Eq. (10).

$$f_{power}^{c} = \frac{\sum_i x_i^f \times (x_i^f)^n}{\sum_i (x_i^f)^n} \tag{10}$$

This method transitions between max and average pooling, making it suitable for applications involving weakly labeled datasets or varying temporal scales.

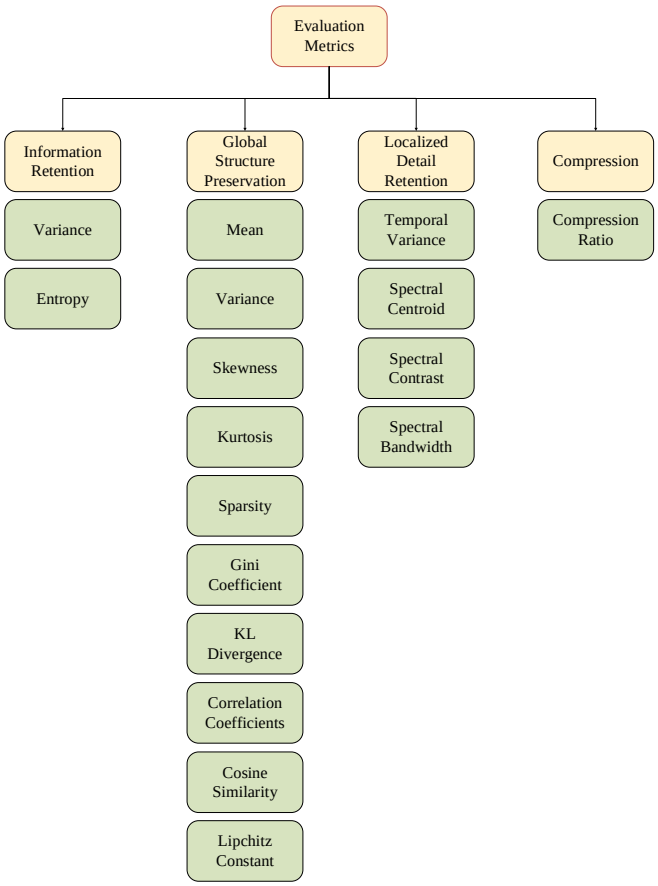


Fig. 3. Types of evaluation metrics categorized based on different usages.

11) Entropy pooling: Entropy pooling [38] selects high-entropy features, ensuring the most informative elements are retained. Eq. (11) demonstrates how entropy pooling works.

$$f_{\text{entr}}(X_r) = X_r[g(P_r)] \tag{11}$$

where, f_{entr} is the entropy-pooled output. By ensuring that the pooled output retains the most informative and least redundant features, entropy pooling reduces redundancy and maintains a more uniform feature distribution, thereby making the model more resilient to noise.

12) Attention pooling: Attention pooling [39], [40], [4], [41], [5], [42] dynamically assigns weights to input features, focusing on the most relevant parts of the audio signal. It uses a learnable weight vector $\mathbf{w}_i$ to emphasize frames critical for the task, creating a weighted average output.

C. Metrics for Evaluating Pooled Spectrograms

We have identified a set of metrics that can be used to evaluate the performance of pooling methods applied to spectrograms. Based on their specific focus, Fig. 3 shows how these metrics can be grouped into four categories: information retention, preservation of global structure, retention of localized details, and dimensionality reduction. These metrics ensure that the pooled spectrograms remain representative of the original data while achieving dimensionality reduction.

1) Information retention: Information retention metrics evaluate whether the pooled spectrogram preserves sufficient variability, complexity, or detail from the original spectrogram. These metrics ensure that pooling does not discard too much essential information.

a) Variance: Variance [43] quantifies the spread of intensity values in the spectrogram. As shown in Eq. (12), it measures how much the intensity values deviate from their mean, representing the variability in the data. In the context of pooled spectrograms, preserving variance ensures that dynamic range and signal variability are maintained. A significant drop in variance after pooling could indicate a loss of essential details, such as transitions or contrasts in intensity.

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{12}$$

where, x_i is intensity value at pixel i in the spectrogram, μ is mean of all intensity values ($\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$), and N is total number of pixels or intensity values.

b) Entropy: Entropy [44] assesses the complexity or randomness of the spectrogram by evaluating the distribution of intensity values, as given in Eq. (13). Higher entropy indicates a richer or more diverse representation of information. If the entropy of the pooled spectrogram is close to the original, it implies that essential patterns and nuances, such as background noise or intricate harmonic content, are preserved. A decrease in entropy suggests that the pooling operation may have oversimplified the spectrogram, discarding finer details.

$$\text{Entropy} = -\sum_{i=1}^{N} p(x_i)\log p(x_i) \tag{13}$$

where, $p(x_i)$ is the probability of intensity value x_i in the spectrogram, and N is the total number of intensity levels.

2) Global structure preservation: Global structure preservation metrics evaluate whether the structural and temporal patterns of a spectrogram are retained after pooling, which is crucial for applications that rely on macro-level spectrogram features. Pooling reduces the size of a spectrogram, but maintaining these patterns ensures that the essential characteristics remain intact. Several types of pattern preservation metrics can be identified, focusing on aspects such as statistical properties, near-zero values, energy distribution, and similarity measures.

a) Statistical properties: Preserving statistical properties ensures that the overall patterns and distributions in the spectrogram remain consistent after pooling [45]. Key statistical metrics include,

- Mean [45], which represents the average intensity of the spectrogram. Preserving the mean ensures that the global energy level is consistent, reflecting the overall sound intensity. Mean is calculated as in Eq. (14)

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{14}$$

where, x_i is the intensity value at pixel i, and N is the total number of pixels.

- Variance [45], which captures the spread of intensity values, preserving the variability and dynamic range of the spectrogram.
- Skewness [45], which measures asymmetry in the intensity distribution, as presented in Eq. (15). Maintaining skewness ensures that pooling does not disproportionately favor high-energy or low-energy regions.

$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \tag{15}$$

where, x_i is the intensity value at pixel i, μ is the mean of intensity values, and σ is the standard deviation ($\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$).

- Kurtosis [45], which reflects the sharpness or flatness of the intensity distributions, in accordance with Eq. (16). Preserving kurtosis retains critical features such as peaks (high energy areas) or flat regions (low activity areas).

$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{16}$$

where, x_i is the intensity value at pixel i, μ is the mean of intensity values, and σ is the standard deviation. Together, these metrics ensure that the original statistical characteristics of the spectrogram are reflected in the pooled version.

b) Low activity regions: Sparsity [46] refers to the proportion of near-zero or inactive regions in a spectrogram, and retaining it is essential in applications, where silent or low-activity regions convey meaningful information. For instance, in speech recognition, silent gaps or pauses are crucial for segmenting phonemes or words, while in speaker recognition, low-energy regions may contain speaker-specific traits. Pooling methods must preserve these sparse areas to ensure critical contextual information is not lost. Sparsity is calculated as shown in Eq. (17), and sparsity ratio, which is given in Eq. (18), is calculated in our experiments.

$$\text{Sparsity} = \frac{\text{Number of near-zero values}}{\text{Total number of values}} \tag{17}$$

$$\text{Sparsity ratio} = \frac{\text{Sparsity of pooled spectrogram}}{\text{Sparsity of original spectrogram}} \tag{18}$$

c) Energy distribution: The Gini coefficient [47] measures the inequality in intensity distribution across a spectrogram, where a lower value indicates a more uniform distribution and a higher value reflects concentrated energy, such as dominant frequency bands. By comparing the Gini coefficients before and after pooling, it is possible to evaluate whether the pooling process has preserved the energy balance across the spectrogram.

d) Similarity measures: Similarity measures directly compare the pooled spectrogram with the original to ensure that key structural and temporal features are preserved. Some of the similarly metrics which might be useful for evaluating the global structure preservation are as follows:

- Kullback-Leibler (KL) Divergence [48] quantifies the difference between intensity distributions, according to Eq. (19). A lower divergence indicates that the pooled spectrogram closely resembles the original in terms of energy patterns.

$$D_{\text{KL Divergence}}(P|Q) = \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{19}$$

where, $p(x_i)$ is the probability distribution before pooling, and $q(x_i)$ is the probability distribution after pooling.

- Correlation Coefficients [49], [50] evaluates the linear relationship between intensity values in the original and pooled spectrograms, ensuring that proportional changes are consistent.
- Cosine Similarity [51] assesses alignment between feature vectors of the original and pooled spectrograms, indicating the degree to which proportional feature retention is achieved, as given in Eq. (20).

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{20}$$

where, x_i is the intensity value in the original spectrogram, and y_i is the intensity value in the pooled spectrogram.

- Lipschitz Continuity [52] ensures that small changes in the original spectrogram correspond to small changes in the pooled spectrogram, preventing distortions that could compromise structural integrity [see Eq. (21)].

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \tag{21}$$

where, $f(x)$ is the mapping after pooling, and L is the Lipschitz constant.

3) Localized detail retention: Feature-specific metrics assess the retention of localized or detailed aspects of the spectrogram. These are crucial for fine-grained tasks such as instrument identification or event detection.

a) Temporal features: Temporal features [53] focus on intensity fluctuations, transitions, and coherence over time. As shown in Eq. (22), temporal variance is a metric that measures how well time-based variations, such as rhythm or speech patterns, are preserved. Temporal feature retention is vital for applications, where timing and transitions play a critical role, such as in rhythm analysis or speech recognition.

$$\text{Temporal Variance} = \frac{1}{T}\sum_{t=1}^{T}(I_t - \mu_t)^2 \tag{22}$$

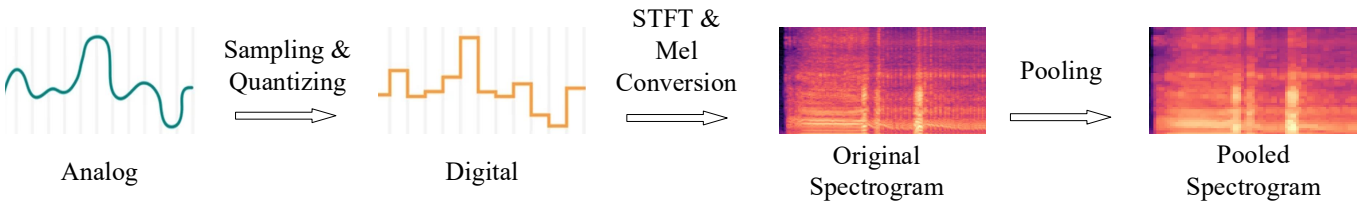


Fig. 4. Illustration of the feature extraction and pooling process of audio analysis applications. Evaluation metrics are used to compare the original and pooled spectrograms.

where, I_t is the intensity at time t, μ_t is the mean intensity over time, and T is the total number of time frames.

b) Spectral features: Spectral metrics evaluate the preservation of frequency-based characteristics. They are calculated as in respective reference papers.

- Spectral Centroid [54]: Represents the "center of mass" of the frequency spectrum, indicating the dominant frequency range. Preserving the spectral centroid ensures that the tonal characteristics remain consistent.
- Spectral Contrast [55]: Captures the differences between peaks and valleys in the frequency spectrum. Retaining spectral contrast ensures the preservation of harmonic content and overall tonal richness.
- Spectral Bandwidth [54]: Measures the spread of frequencies in the spectrum. A high spectral bandwidth indicates a broad frequency range, while a low bandwidth suggests focused frequency content. Preserving bandwidth ensures that the spectral characteristics of the pooled spectrogram match the original.

4) Compression: Compression-focused metrics [56] measure the effectiveness of pooling methods in reducing spectrogram dimensions while preserving essential information. A key metric is the compression ratio, which quantifies the extent of dimensionality reduction and reflects the balance between minimizing size and retaining critical features. The compression ratio is calculated as given in Eq. (23). An effective pooling method achieves this balance by eliminating redundancy without compromising the spectrogram's informational content.

$$\text{Compression Ratio} = \frac{\text{Original size of spectrogram}}{\text{Pooled size of spectrogram}} \tag{23}$$

III. System Architecture

In Fig. 4, the steps used for generating a spectrogram from an audio file are outlined. The process starts with sampling the analog audio signal at a fixed rate of 16 kHz and quantizes the samples into discrete values. The digital signal is divided into overlapping frames, and the Hann windowing function is applied to reduce spectral leakage. Each frame undergoes a Short-Time Fourier Transform (STFT) to convert the time-domain signal into a frequency-domain representation, resulting in a magnitude spectrum. This spectrum is then mapped to the Mel scale to create a mel spectrogram, aligning the frequency bins perceptually. The resulting power spectrogram is converted to a logarithmic scale by normalizing the intensity relative to the maximum value, producing a log-scaled Mel spectrogram suitable for visualization and further analysis. Once the spectrogram is generated from the audio signal, pooling can be applied to the spectrogram itself.

IV. Experiments

A. Experimental Setup

We selected a range of audio classification tasks for our experiments, encompassing all three major audio types: speech, music, and environmental sounds.

1) Speech: Speech Emotion Recognition task is used for the speech data domain. IEMOCAP dataset [57], which contains approximately 10,000 speech utterances labeled with emotions, and focuses on four emotion classes (angry, happy, sad, neutral), is used.

2) Music: For Music Genre Classification, GTZAN dataset [58] is used, which consists of 1000 music excerpts (30 seconds each) spanning 10 musical genres. IMRAS dataset [59] is used for Music Instrument Recognition task.

3) Environmental sounds: Environmental Sound Classification task is used for this domain with ESC-50 dataset [60], which consists of 2000 five-second audio clips in 50 classes. It contains five major categories of animals, natural soundscapes and water sounds, human non-speech sounds, domestic sounds, and exterior sounds, each containing 10 equally balanced classes of sound events.

These datasets were chosen for their standard usage in the respective domains, diversity of audio samples, and relevance to the identification of global versus localized features in spectrogram-based audio analysis. Spectrograms were generated as described in Section III, and analyzed to identify both local and global features. Subsequently, pooling methods described in Section II-B were applied, and the evaluation metrics from Section II-C were used to assess how well each pooling method preserves local and global features. These metrics also help determine which are most effective for measuring feature preservation in spectrogram-based models. Mixed pooling is non-trainable and includes hyperparameters, with its mixing proportion set to a fixed value of 0.5.

B. Analyzing Spectrograms

Spectrograms were generated for each audio type to facilitate visual identification of their characteristic features.

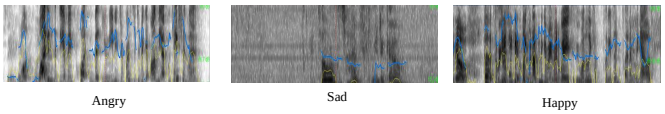


Fig. 5. Voice patterns of the same person with different emotions.

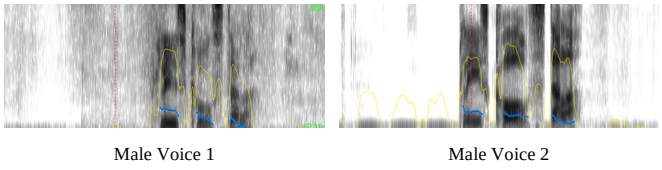


Fig. 6. Voice patterns of different people of same gender talking the same sentence shown in a spectrogram.

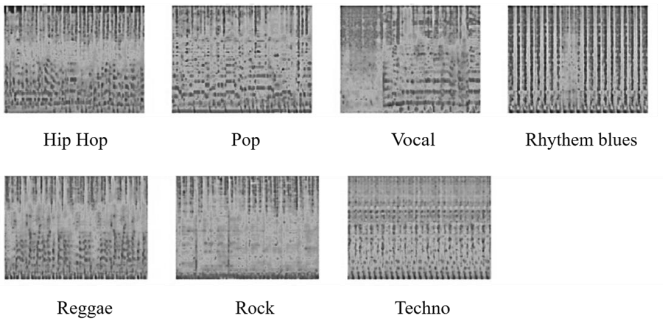


Fig. 8. Spectrogram patterns of different music genres that are used for music genre classification.

1) Speech applications: In speech applications, preserving localized features in the spectrogram is crucial for tasks such as speech emotion recognition, where subtle variations in pitch, intensity, and temporal dynamics carry significant meaning. As shown in Fig. 5, we have identified how these features appear as localized changes in the spectrogram, such as brief intensity surges or shifts in frequency, which are essential for identifying emotional states. For instance, emotions such as anger or happiness are often expressed through louder speech with elevated pitch, while sadness is characterized by softer, flatter intonation patterns. The retention of these intricate features enables a deeper understanding of vocal expressions, making them vital for emotion recognition systems that rely on the nuanced analysis of audio signals.

However, in most speech applications, it is crucial to preserve the overall patterns of the spectrogram. These patterns are influenced by various factors, such as the speaker's unique voice timbre, pitch variations, and speaking style, including how pauses are placed between words. Such variations are particularly significant in speaker recognition tasks, where identifying an individual speaker depends on recognizing unique timbre and consistent frequency patterns over time. For instance, as shown in Fig. 6, two male speakers saying the same sentence exhibit noticeable differences in pauses, frequency patterns, pitch, and intensity variations. Similarly, Fig. 7 highlights the distinctions between male and female speakers for the same sentence. Therefore, it is obvious that when applying pooling, overall patterns of the spectrograms should be preserved in most of the speech applications.

2) Music applications: In music-related tasks, preserving spectrogram patterns is crucial for capturing the intricate details of musical compositions and maintaining audio fea-

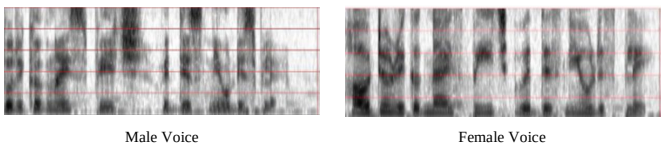


Fig. 7. Voice patterns of male and female voice with the same sentence shown in a spectrogram.

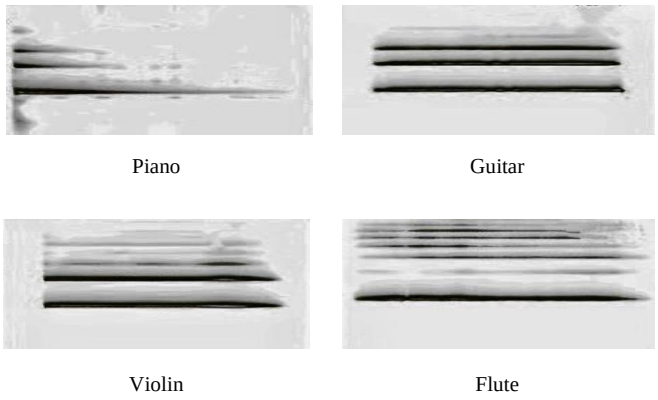


Fig. 9. Spectrogram patterns with different music instruments that is used for music instruments identification.

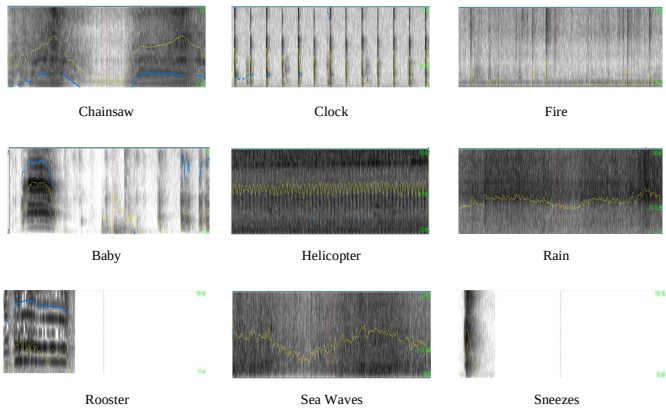


Fig. 10. Spectrogram patterns with different environmental sounds.

tures critical for accurate analysis and classification. We have illustrated in Fig. 8 how different music genres exhibit distinct continuous patterns, which must be preserved to accurately identify those genres. Music genre classification relies on distinguishing broad musical features such as rhythm patterns, harmonic structures, and timbre, each of which reflects unique spectrogram characteristics. Classical music often presents smooth, sustained harmonic patterns, while electronic dance music (EDM) features repetitive beats and sharp intensity changes. These preserved spectrogram features reveal the temporal and spectral patterns necessary for precise classification.

However, as shown in Fig. 9, some music applications such as musical instrument identification depends on preserving timbral qualities, harmonic structures, and frequency ranges unique to each instrument. A piano, for instance, displays evenly spaced horizontal harmonic lines, while a guitar shows richer textures due to its plucked strings. These features, which are often small but distinct, provide critical cues for accurate identification. Additionally, capturing the unique spectral patterns of each instrument, such as the smooth overtones of a violin or the sharp attacks of a drum, is essential for ensuring that their individual characteristics are retained during audio analysis, enabling precise classification in diverse musical contexts.

3) Environmental sound applications: In environmental sound applications, detecting and classifying sounds relies heavily on preserving their distinct temporal and spectral characteristics, which are often brief but highly specific. Environmental sound classification involves recognizing natural sounds such as rain, wind, or traffic noise by retaining their defining spectral features and temporal dynamics such as rain's high-frequency noise bursts or wind's diffuse, broad frequency patterns ensuring these localized elements serve as a foundation for accurate classification. We have highlighted these distinct patterns in Fig. 10 to demonstrate the importance of spectrogram feature preservation in capturing the nuances of various environmental sounds, enabling precise classification across diverse audio tasks.

V. Results

A. General Trends Across Pooling Methods

Across all three domains, several consistent patterns are observed in the behavior of pooling methods. Max pooling consistently preserves the highest variance, mean, skewness, and temporal variance, indicating its strong capacity to retain localized, high-energy features such as transient peaks and short-duration events. While this makes max pooling particularly effective for tasks where such localized features are essential, it often results in the loss of broader structural coherence, as seen in its comparatively lower cosine similarity and correlation coefficients. In contrast, average pooling demonstrates reliable global structure preservation across all datasets. It maintains constant mean retention, cosine similarity, and correlation coefficients, suggesting that it effectively captures the general shape of spectrograms while smoothing out noise and minor variations. This makes average pooling especially suitable for tasks where the continuity of harmonic or rhythmic patterns is more critical than localized precision.

Entropy pooling and ESP consistently exhibit the highest entropy and kurtosis values, reflecting their ability to capture a wide diversity of information. However, these methods simultaneously suffer from poor performance in structure-preserving metrics such as correlation, Lipschitz continuity, and temporal variance, indicating that they degrade interpretability and coherence. This trade-off suggests that while entropy-based methods may retain information richness, they do so at the expense of structural and temporal fidelity.

Auto pooling, mixed pooling, and LSEP emerge as the most balanced approaches. These methods perform well across a broad range of metrics, combining moderate-to-high variance preservation with strong similarity and temporal coherence. Their capacity to retain both global structures and localized features renders them especially well-suited to tasks that require a blend of both information types. Notably, similarity metrics such as cosine similarity and KL divergence show relatively little variation across pooling methods, suggesting that while useful for general alignment assessment, these metrics are less effective for differentiating nuanced pooling behavior compared to statistical or spectro-temporal measures.

B. Task-Specific Observations

1) Speech emotion classification: Speech emotion classification is primarily a localized feature-dependent task. Emotional states are conveyed through subtle and transient modulations in pitch, intensity, and timing, which necessitate the preservation of fine-grained temporal and spectral features. As shown in Table I, max pooling is notably effective in this domain, producing the highest variance (1.1031) and skewness, and excelling in capturing rapid vocal changes associated with expressive emotions such as anger or surprise. However, its emphasis on energetic extremes may occasionally overamplify variations, leading to potential imbalances in emotion recognition.

2) Music genre classification: In music genre classification, the primary analytical focus lies in preserving global spectro-temporal structures such as rhythm, harmony, and tonal balance across longer time scales. As shown in Table II, average pooling consistently performs well in this domain, achieving high mean preservation (0.9988), cosine similarity (0.9922), and correlation (0.9854), making it effective for capturing the overarching structure of musical compositions, particularly in genres characterized by smooth and continuous transitions (e.g., classical, ambient).

3) Environmental sound classification: While environmental sound classification relies on both global and local feature preservation, it primarily depends on local features. Audio scenes in this domain frequently combine ambient background noise with short, high-intensity acoustic events. Table III shows that the effectiveness of max pooling in ESC is evident from its superior variance (1.0477) and temporal variance (1.00), confirming its utility in capturing brief, salient events such as alarms or door slams. However, average pooling also plays a vital role, achieving the highest cosine similarity (0.9876) and correlation coefficient (0.9812), indicating that it effectively preserves background structure and overall scene continuity.

TABLE I. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE SPEECH EMOTION CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.7120	0.9780	0.0911	2.9371	1.3185	0.6410	-0.0704	28.1342	0.9811	0.9732	0.61	0.88	0.86	0.82	0.83	4.1
max	1.1031	1.1224	0.6550	4.6204	1.4869	0.0048	-0.1003	28.0427	0.9620	0.9540	0.93	1.00	1.12	1.26	1.15	4.3
min	0.6028	0.8393	-0.4835	2.1044	1.2910	0.8724	-0.0582	28.2438	0.9367	0.9273	0.36	0.63	0.74	0.61	0.70	4.0
mixed	0.8611	0.9143	0.2489	3.2649	1.3341	0.0190	-0.0922	27.9881	0.9505	0.9405	0.72	0.90	0.96	0.89	0.93	4.2
LSP	0.7803	1.0122	0.1348	3.0715	1.3742	0.0029	-0.0816	28.1605	0.9699	0.9604	0.59	0.82	0.91	0.84	0.87	4.1
ESP	0.0094	0.0015	-1.9341	10.2145	4.5987	0.4042	0.8840	28.1710	0.0084	0.0103	0.03	0.18	0.31	0.24	0.28	3.2
LNP	0.6812	-0.8841	-1.6221	5.3648	1.3372	0.0340	0.1012	28.0347	0.9350	0.9231	0.28	0.55	0.69	0.62	0.67	4.0
LSEP	1.0225	0.9011	0.4082	3.5653	1.4027	0.0233	-0.1081	28.0910	0.9477	0.9389	0.80	0.93	1.01	0.95	0.96	4.3
auto	1.0512	1.0581	1.2431	6.1111	3.7224	0.5143	-0.7523	28.2299	0.8604	0.8511	0.46	0.69	0.76	0.65	0.74	3.8
power	0.6270	1.0721	0.2903	3.3211	1.4198	0.1534	-0.0610	27.9740	0.9801	0.9702	0.68	0.85	0.93	0.88	0.90	4.1
entropy	1.0284	3.9810	2.7629	11.0342	7.0204	0.0275	0.0895	28.0001	0.8023	0.7650	0.11	0.27	0.42	0.33	0.39	3.4
attention	0.1699	0.0664	1.2908	6.2123	3.9011	0.5193	-0.7620	28.1530	0.8790	0.8685	0.52	0.75	0.85	0.78	0.80	3.8

TABLE II. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE MUSIC GENRE CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.8285	0.9988	0.1093	3.0050	1.2711	0.7190	-0.0812	30.0122	0.9922	0.9854	0.67	0.89	0.94	0.90	0.91	4.2
max	1.0775	1.1451	0.4782	4.2239	1.3920	0.0000	-0.1233	30.0988	0.9763	0.9651	0.92	1.00	1.13	1.20	1.12	4.4
min	0.7212	0.8711	-0.4147	2.4568	1.3123	0.9087	-0.0593	30.0775	0.9491	0.9400	0.39	0.60	0.78	0.66	0.71	4.1
mixed	0.9068	0.9394	0.1921	3.1791	1.2922	0.0000	-0.0952	30.0221	0.9655	0.9550	0.74	0.91	0.99	0.92	0.95	4.3
LSP	0.7595	1.0032	0.1582	3.1195	1.3321	0.0000	-0.0744	30.1012	0.9832	0.9728	0.62	0.78	0.92	0.85	0.88	4.2
ESP	0.0118	0.0032	-1.8011	9.8484	4.1311	0.4702	0.8934	30.0440	0.0061	0.0073	0.04	0.14	0.29	0.23	0.27	3.1
LNP	0.7230	-0.9051	-1.5111	5.2457	1.3195	0.0311	0.0955	30.0999	0.9421	0.9304	0.33	0.51	0.68	0.59	0.63	4.1
LSEP	1.0022	0.9199	0.3568	3.4511	1.3855	0.0205	-0.1094	30.0331	0.9703	0.9605	0.78	0.93	1.03	0.94	0.97	4.3
auto	0.5021	0.4777	1.1710	5.7219	3.8110	0.5831	-0.7721	30.1211	0.8891	0.8724	0.48	0.71	0.82	0.75	0.79	3.7
power	0.6872	1.0711	0.2292	3.2744	1.3677	0.1599	-0.0682	30.0114	0.9871	0.9793	0.70	0.84	0.91	0.88	0.90	4.2
entropy	1.0232	3.9822	2.8965	11.7652	7.1333	0.0344	0.0783	30.0552	0.7985	0.7593	0.10	0.26	0.40	0.32	0.37	3.3
attention	0.1799	0.0822	1.1532	5.8799	3.8750	0.5675	-0.7744	30.0901	0.8944	0.8822	0.54	0.73	0.84	0.77	0.81	3.7

TABLE III. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE ENVIRONMENTAL SOUND CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.7919	0.9962	0.1274	3.0256	1.2915	0.7012	-0.0885	32.6579	0.9876	0.9812	0.62	0.85	0.93	0.81	0.87	4.0
max	1.0477	1.1240	0.4873	4.1285	1.3757	0.0000	-0.1294	32.6133	0.9723	0.9621	0.94	1.00	1.11	1.22	1.09	4.5
min	0.6937	0.8690	-0.3982	2.5127	1.3512	0.9041	-0.0679	32.6632	0.9451	0.9370	0.38	0.59	0.78	0.64	0.73	4.1
mixed	0.8937	0.9326	0.2147	3.1846	1.3033	0.0000	-0.1051	32.6127	0.9584	0.9503	0.71	0.88	0.95	0.87	0.91	4.3
LSP	0.7412	1.0087	0.1623	3.1178	1.3457	0.0000	-0.0837	32.6407	0.9792	0.9734	0.58	0.75	0.88	0.79	0.83	4.2
ESP	0.0001	-0.0009	-1.8345	9.7265	4.1431	0.4610	0.9087	32.6390	0.0021	0.0032	0.02	0.09	0.22	0.18	0.25	3.1
LNP	0.7657	-1.0023	-1.5124	5.1473	1.3103	0.0365	0.0861	32.5663	0.9432	0.9315	0.31	0.48	0.66	0.59	0.64	4.0
LSEP	1.0093	0.9129	0.3719	3.4621	1.3852	0.0186	-0.1176	32.5853	0.9615	0.9574	0.81	0.91	1.02	0.95	0.97	4.4
auto	0.9841	0.9759	1.1872	5.8234	3.8288	0.5749	-0.7837	32.6675	0.8754	0.8622	0.48	0.67	0.74	0.63	0.72	3.6
power	0.6691	1.0643	0.2365	3.2783	1.3886	0.1657	-0.0722	32.5257	0.9845	0.9751	0.66	0.83	0.91	0.86	0.89	4.1
entropy	1.0191	3.9962	2.9174	11.8932	7.1555	0.0389	0.0731	32.6040	0.7893	0.7532	0.09	0.22	0.35	0.27	0.33	3.2
attention	0.1841	0.0759	1.1623	5.9128	3.8288	0.5749	-0.7837	32.6395	0.8907	0.8731	0.53	0.71	0.82	0.75	0.78	3.6

VI. DISCUSSION

By looking at the spectrogram analysis in Section IV-B and results in Section V, we can identify some important points. The relationship between the effectiveness of pooling methods and the type of audio task becomes particularly evident when comparing the results across the above datasets. As we identified in Section IV-B, speech emotion classification [61] is predominantly a local-detail-driven task. Emotional expression in speech is often encoded in short-term fluctuations and nuanced changes in tone, pitch, and duration. Pooling methods that preserve such transient dynamics, particularly max, auto, and LSEP pooling, are better suited to this domain. Music genre classification [62], by contrast, places the highest emphasis on global structure preservation. The ability to maintain continuity in tonal progression, harmonic layering, and rhythmic patterns is vital for distinguishing among musical genres. Here, pooling strategies such as average, auto, and LSEP pooling that prioritize smoothness and structural fidelity outperform those that emphasize localized variance. Environmental sound classification [63] stands out as a domain that requires a balanced approach to both local and global feature retention. The simultaneous need to detect transient events and maintain scene-level continuity necessitates pooling methods that can handle both aspects effectively, such as auto, mixed,

and LSEP pooling.

Therefore, when applying pooling in all types of audio applications, it is important to look at the global structure preservation and local detail retention of the pooling methods. In global structure preservation, the primary focus is to retain the overall structure and temporal patterns of the spectrogram. This is particularly relevant in applications, where the recognition of broad patterns or sequences is critical for identifying or classifying audio signals. Localized Detail Retention focuses on retaining localized and distinctive patterns within the spectrogram that are critical for identifying or classifying audio signals. Unlike global patterns, these features are often small and transient but hold significant importance for understanding the finer details of audio characteristics. Beyond merely retaining these features, it is important to evaluate how transformations affect their clarity and interpretability. Overly aggressive pooling can blur critical details, diminishing the spectrogram's discriminative power for downstream tasks. Conversely, pooling methods that balance dimensionality reduction with effective feature retention can improve computational efficiency without compromising performance. The effectiveness of pooling should therefore be assessed in the context of the specific application. Comparative analyses of pooling methods can shed light on their strengths and limitations, ensuring that the preserved spectrogram features align with the requirements of the intended task. This alignment is essential for optimizing accuracy, performance, and robustness in audio analysis applications.

VII. Conclusion

This study presents a comprehensive evaluation of pooling techniques in audio analysis, focusing on their roles in global structure preservation and localized detail retention within spectrograms. By introducing diverse evaluation metrics across four domains this research uncovered the nuanced strengths and limitations of twelve pooling methods. The findings highlight the critical importance of selecting pooling techniques based on task-specific requirements. Max pooling demonstrates effectiveness in capturing localized features essential for tasks such as emotion recognition and transient event detection. In contrast, average pooling excels at preserving global patterns vital for applications such as music genre classification and acoustic scene analysis. Entropy pooling, with its ability to retain diverse and intricate information, emerges as particularly suitable for complex audio tasks. In order to map the use of pooling methods into audio applications, this study categorizes audio applications into two primary focuses, global structure preservation and localized detail retention based on its characteristics. This categorization provides a practical framework for aligning pooling strategies with specific application needs. Beyond task-specific insights, the study emphasizes the importance of innovative evaluation metrics, such as variance, entropy, sparsity, and similarity measures, in assessing pooling methods. These metrics shift the focus from traditional downstream task accuracy to the ability of pooling techniques to preserve critical spectrogram features. In this study, the most important evaluation metrics are evaluated to identify which metrics are more critical in feature extraction. In the future, the development of more adaptive and hybrid pooling techniques holds promise for achieving an ideal balance between global structure preservation and localized feature retention. Furthermore, integrating pooling methods with advanced neural architectures and self-supervised learning frameworks offers significant potential to advance audio analysis systems.

Acknowledgment

The authors acknowledge the support received from the LK Domain Registry in publishing this study.

References

[1] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," pp. 131–135, 2017.

[2] P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient end-to-end audio embeddings generation for audio classification on target applications," pp. 601–605, 2021.

[3] V. Passricha and R. K. Aggarwal, "A comparative analysis of pooling strategies for convolutional neural network based hindi asr," *Journal of ambient intelligence and humanized computing*, vol. 11, no. 2, pp. 675–691, 2020.

[4] W. Lin, M.-W. Mak, and L. Yi, "Learning mixture representation for deep speaker embedding using attention." pp. 210–214, 2020.

[5] P. Safari and J. Hernando, "Self multi-head attention for speaker recognition," 2019.

[6] M. Rouvier, P.-M. Bousquet, and J. Duret, "Study on the temporal pooling used in deep neural networks for speaker verification," pp. 501–505, 2021.

[7] Q. Shi, H. Luo, and J. Han, "Subspace pooling based temporal features extraction for audio event recognition." pp. 3850–3854, 2019.

[8] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," pp. 1–4, 2013.

[9] X. Cai, D. Yu, D. Liu, and M. Wu, "Weakly and semi-supervised learning for sound event detection using image pretrained convolutional recurrent neural network, weighted pooling and mean teacher method," vol. 2010, no. 1, p. 012108, 2021.

[10] D. Yu, X. Cai, D. Liu, and Z. Liu, "Semi-supervised sound event detection using multi-scale convolu-tional recurrent neural network and weighted pooling," Tech. Rep., DCASE2021 Challenge, Tech. Rep., 2021.

[11] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Temporal attentive pooling for acoustic event detection." pp. 1354–1357, 2018.

[12] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[13] E. Salah, K. Amine, K. Redouane, and K. Fares, "A fourier transform based audio watermarking algorithm," *Applied Acoustics*, vol. 172, p. 107652, 2021.

[14] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[15] S. Atito, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "Asit: Local-global audio spectrogram vision transformer for event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[16] A. Lependin, P. Ladygin, V. Karev, and A. Mansurov, "Fourier chromagrams for fingerprinting, verification and authentication of digital audio recordings," pp. 263–275, 2023.

[17] A. Maccagno, A. Mastropietro, U. Mazziotta, M. Scarpiniti, Y.-C. Lee, and A. Uncini, "A cnn approach for audio classification in construction sites," *Progresses in Artificial Intelligence and Neural Systems*, pp. 371–381, 2021.

[18] J.-W. Hwang, R.-H. Park, and H.-M. Park, "Efficient audio-visual speech enhancement using deep u-net with early fusion of audio and video information and rnn attention blocks," *IEEE Access*, vol. 9, pp. 137 584–137 598, 2021.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," pp. 646–650, 2022.

[20] A. K. Das and R. Naskar, "A deep learning model for depression detection based on mfcc and cnn generated spectrogram features," *Biomedical Signal Processing and Control*, vol. 90, p. 105898, 2024.

[21] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Analysis and Applications*, vol. 24, pp. 423–431, 2021.

[22] O. A. Onasoga, N. Yusof, and N. H. Harun, "Audio classification-feature dimensional analysis," pp. 775–788, 2021.

[23] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[24] X. Wang and L. Xu, "Speech perception in noise: Masking and unmasking," *Journal of Otology*, vol. 16, no. 2, pp. 109–119, 2021.

[25] J. Pons, O. Slizovskaia, R. Gong, E. Gomez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," pp. 2744–2748, 2017.

[26] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.

[27] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[28] J. Pons and X. Serra, "Randomly weighted cnns for (music) audio classification," pp. 336–340, 2019.

[29] W. Li, P. Wang, R. Xiong, and X. Fan, "Spiking tucker fusion transformer for audio-visual zero-shot learning," *IEEE Transactions on Image Processing*, 2024.

[30] P. Satti, N. Sharma, and B. Garg, "Min-max average pooling based filter for impulse noise removal," *IEEE Signal Processing Letters*, vol. 27, pp. 1475–1479, 2020.

[31] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," pp. 464–472, 2016.

[32] Y. Han, S. Lee, J. Nam, and K. Lee, "Sparse feature learning for instrument identification: Effects of sampling and pooling methods," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2290–2298, 2016.

[33] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[34] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," pp. 530–546, 2014.

[35] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," pp. 1713–1721, 2015.

[36] Y. Liu, H. Chen, Y. Wang, and P. Zhang, "Power pooling: An adaptive pooling function for weakly labelled sound event detection," pp. 1–7, 2021.

[37] Y. Liu, C. Chen, J. Kuang, and P. Zhang, "Semi-supervised sound event detection based on mean teacher with power pooling and data augmentation," pp. 4–6, 2020.

[38] C. Nalmpantis, L. Vrysis, D. Vlachava, L. Papageorgiou, and D. Vrakas, "Noise invariant feature pooling for the internet of audio things," *Multimedia Tools and Applications*, vol. 81, no. 22, pp. 32 057–32 072, 2022.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[40] K. Zhang, Z. Wu, J. Jia, H. Meng, and B. Song, "Query-by-example spoken term detection using attentive pooling networks," pp. 1267–1272, 2019.

[41] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, and A. Mertins, "Spatio-temporal attention pooling for audio scene classification," 2019.

[42] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," pp. 436–454, 2020.

[43] C. M. Eckhardt, S. J. Madjarova, R. J. Williams, M. Ollivier, J. Karlsson, A. Pareek, and B. U. Nwachukwu, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, no. 2, pp. 376–381, 2023.

[44] S. A. Sepúlveda-Fontaine and J. M. Amigó, "Applications of entropy in data analysis and machine learning: A review," *Entropy*, vol. 26, no. 12, p. 1126, 2024.

[45] L. Tang, H. Tian, H. Huang, S. Shi, and Q. Ji, "A survey of mechanical fault diagnosis based on audio signal analysis," *Measurement*, p. 113294, 2023.

[46] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[47] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[48] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," pp. 1–8, 2021.

[49] Z. Liu, T. Shao, and X. Zhang, "Bcg signal analysis based on improved vmd algorithm," *Measurement*, vol. 231, p. 114631, 2024.

[50] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[51] S. Tanberk, V. Dağlı, and M. K. Gürkan, "Deep learning for videoconferencing: A brief examination of speech to text and speech synthesis," pp. 506–511, 2021.

[52] S.-W. Park, J.-S. Ko, J.-H. Huh, and J.-C. Kim, "Review on generative adversarial networks: focusing on computer vision and its applications," *Electronics*, vol. 10, no. 10, p. 1216, 2021.

[53] T. Rathi and M. Tripathy, "Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review," *Speech Communication*, p. 103102, 2024.

[54] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2007.

[55] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," vol. 1, pp. 113–116, 2002.

[56] I. Bozhilov, R. Petkova, K. Tonchev, and A. Manolova, "A systematic survey into compression algorithms for three-dimensional content," *IEEE Access*, 2024.

[57] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[58] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[59] J. J. Bosch, F. Fuhrmann, and P. Herrera, "Irmas: a dataset for instrument recognition in musical audio signals," 2014.

[60] K. J. Piczak, "Esc: Dataset for environmental sound classification," pp. 1015–1018, 2015.

[61] G. Mohmad and R. Delhibabu, "Speech databases, speech features and classifiers in speech emotion recognition: A review," *IEEE Access*, 2024.

[62] V. Lyberatos, S. Kantarelis, E. Dervakos, and G. Stamou, "Challenges and perspectives in interpretable music auto-tagging using perceptual features," *IEEE Access*, 2025.

[63] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.

Recommendation Engine for Amazon Magazine Subscriptions

Sushil Khairnar[1], Deep Bodra[2]
Virginia Tech, Blacksburg, USA[1]
Harrisburg University of Science and Technology, USA[2]

***Abstract*—Recommender systems play a crucial role in enhancing user experience and engagement on e-commerce platforms by suggesting relevant products based on user behavior. In the context of Amazon's extensive catalog of over 8,000 magazines spanning more than twenty-five categories, providing personalized magazine subscription recommendations poses a significant challenge. This study addresses the problem of identifying potential future associations between magazine reviewers and products using a graph-based approach. Specifically, we aim to predict unseen but likely links between users and magazines to improve recommendation quality. To achieve this, we construct an undirected bipartite network connecting reviewers and magazine products based on review data. We perform network analysis using measures such as centrality, modularity, and clustering, and apply sentiment analysis and topic modeling to extract behavioral and thematic insights from user reviews. These insights inform a series of link prediction techniques including Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment evaluated using cross-validation and ROC curves. Our results show that the Preferential Attachment model outperforms other approaches, attributed to the skewed degree distribution inherent in the dataset's structure.**

***Keywords*—Sentiment analysis; topic modeling; recommender system; link prediction**

I. Introduction

A recommendation system is a data filtering system that generates data predictions based on user data linked with the item. Both merchants and consumers benefit from recommendation systems in internet marketing. On the one hand, recommendation algorithms are employed by sellers in order to maintain high levels of user engagement. Consumers, on the other hand, benefit from recommendation systems because they can quickly find what they're looking for. For example, Amazon suggests products to consumers based on their past purchases to encourage future purchases, while news websites suggest related content to viewers to optimize ad revenue. These businesses rely significantly on recommendation systems to improve user experience.

Both service providers and users benefit from recommender systems. Users are able to lower the resource consumption and expenses of searching for and selecting the right things to buy in an online environment. Also, service providers reduce the cost of marketing large amount of products to large user base. If a person enjoys reading scientific periodicals, he or she might be interested in reading magazines about scientific research. Recommendation systems can help them find related products and improve the quality of their decisions. Recommender systems increase revenue in an e-commerce scenario since they are efficient at selling more things. Recommender systems in scientific libraries assist users by allowing them to go beyond catalog searches. In addition to this, indirect connections between the users can be identified to recommend the products purchased by one user to another user. As a result, the importance of employing efficient and accurate recommendation algorithms inside a system that provides consumers with relevant and dependable recommendations cannot be overstated.

Many e-commerce companies offer a diverse selection of products to their customers. Providing users with the most suited items makes the purchasing process more efficient and boosting user happiness. Improved customer happiness keeps users coming back to the website, increasing sales and profitability for products. As a result, more companies are beginning to recommend products to customers, necessitating the efficient analysis of user product interests. Standard recommendation algorithms, such as content-based filtering and collaborative filtering, use a matrix to model user ratings and anticipate customer ratings for unrated goods based on user/item similarity.

In the context of Amazon's magazine subscription platform, which features more than 8,000 magazines across over twenty-five categories, recommending personalized content is particularly challenging due to the diversity and volume of data. This study focuses on addressing the following research question: How can a graph-based recommendation system leveraging user reviews, sentiment analysis, and topic modeling improve the prediction of potential future links between reviewers and magazine subscriptions?

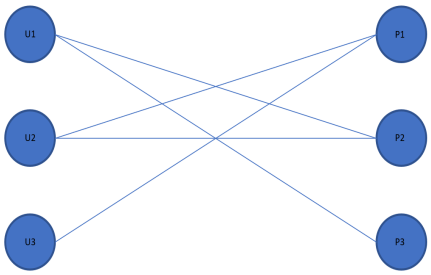


Fig. 1. Bipartite network.

To answer this question, we build a bipartite network (Fig. 1) between users and products using Amazon review data and explore the utility of various link prediction algorithms in identifying new recommendations. We incorporate techniques such as sentiment analysis to assess review polarity, and topic

modeling to identify key themes in user preferences. These components help enrich the graph and guide the recommendation process.

The main contributions of this study are as follows:

- We construct and analyze a large-scale bipartite graph between Amazon reviewers and magazines based on review data.
- Integrate sentiment analysis and topic modeling to extract contextual and thematic features from textual reviews.
- Apply and compare multiple link prediction methods Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment—to identify potential reviewer-magazine associations.
- Evaluate the performance of these models using ROC-AUC and demonstrate that the Preferential Attachment model yields superior results due to the skewed degree distribution in the network.

The implications of this work lie in enhancing personalized recommendation engines for online subscriptions and providing a scalable framework for integrating behavioral and semantic data in graph-based recommender systems.
The rest of the study is organized as follows: Section II reviews the related work. Section III details the methodology, including network construction, sentiment analysis, and topic modeling. Section IV presents experimental results and analysis. Section V presents the discussion of the study. Section VI concludes the study and discusses possible directions for future work.

II. Background and Related Work

Recommender systems have significantly improved user engagement and satisfaction on e-commerce platforms. Particularly, graph-based methods have become prominent for modeling user-product interactions as networks. Li et al. explored the efficacy of bipartite graphs for recommendations, demonstrating their superiority over conventional collaborative filtering. Similarly, Zhang and Chen highlighted the potential of network centrality and clustering to deliver explainable recommendations, enhancing user trust and transparency.

Link prediction is a core component of graph-based recommendations, providing insights into potential future interactions within networks. The foundational works by Adamic and Adar [1], Liben-Nowell and Kleinberg [8], Newman [11], and Barabási et al. [2] introduced seminal link prediction approaches including Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment. Preferential Attachment, in particular, emphasizes that nodes with higher degrees are likelier to form new connections, a concept strongly relevant to real-world scale-free networks.

Incorporating textual data through sentiment analysis and topic modeling has further enriched recommender systems. Pak and Paroubek [20] leveraged sentiment analysis to refine recommendations based on user-generated textual feedback. Additionally, Blei et al. introduced Latent Dirichlet Allocation (LDA) for topic modeling, enabling a more nuanced understanding of user interests through thematic analysis of reviews. McAuley et al. extended these techniques, combining textual and visual features to enhance the contextual relevance of recommendations.

Parallel to recommender system advancements, significant research has been conducted in predictive analytics within other application domains, such as air traffic delay prediction. Flight delays have been a persistent issue, with only 79.63% of flights landing on time over the past decade, according to the Bureau of Transportation Statistics (BTS). Ye's study employed multiple machine learning models (Linear Regression, SVM, ExtraRT, and LightGBM) to forecast flight departure delays at Nanjing Lukou International Airport, identifying critical relationships between meteorological factors and delays [6]. Similarly, Atlioglu [1] analyzed flight operational data using diverse machine learning methods, emphasizing the importance of data selection for accurate delay prediction.

Esmaeilzadeh and Mokhtarimousavi utilized Support Vector Machine (SVM) analysis to examine air traffic delays at major airports in New York City, focusing on operational and flow management factors [3]. Xu et al. [5] applied Bayesian networks for modeling delay propagation across airports, providing insights into the systemic nature of delay spread through an empirically informed Bayesian approach. Bratu and Barnhart [2] introduced the Passenger Delay Calculator, shifting focus from aircraft-centric to passenger-centric delay analysis, thus providing a more nuanced understanding of the impact of delays. Kim and Choi [4] proposed advanced deep-learning models, including deep RNN architectures, for predicting delays with greater accuracy, emphasizing the efficacy of neural networks in capturing complex temporal patterns in delay data [9].

While the application domains of recommender systems and air traffic delay prediction differ substantially, methodologies such as machine learning, sentiment analysis, and data-driven modeling are common threads linking these fields. Our research leverages these interdisciplinary insights, employing advanced analytical techniques such as sentiment analysis, topic modeling, and graph-based link prediction methods to provide personalized magazine recommendations on Amazon. This integrated approach not only enhances recommendation quality but also demonstrates the adaptability and robustness of predictive methodologies across diverse contexts.

III. Approach

We analyze the dataset for amazon magazine subscriptions which contain the data about magazines subscribed by a certain user and reviews written for the particular magazine. We generate a bipartite graph using the reviewerID and ProductID as the two independent sets of vertices [13]. This is a bipartite graph because the edges are between the reviewerID and productID and no edges from and to the same set. Once the graph is generated, we perform network analysis which involves centrality analysis, modularity analysis, finding connected components, small world network analysis and heavy tail analysis. This analysis supplies crucial insights about the graph. After the network analysis, we perform sentiment analysis to capture the sentiment of the reviewers in relation to that productID. This step gives us insights about the relevance of the edges while conducting link prediction. We then perform topic modeling to

capture the group of topics the reviewer is discussing about. We select the highest-ranked reviewer Id and apply topic modeling to the reviews written by that user. Following that, we apply link prediction to find a link between the reviewer and the products with which aim we recommend magazines to users.

A. Network Analysis

The use of networks and graph theory to investigate group structures is known as network analysis [14]. It describes networked structures in terms of nodes, which can be individual actors, people, or items inside the network, and the ties, edges, or connections that connect them, as well as the relationships and interactions that exist between them. It aids us in comprehending the structure of a relationship in social networks, as well as a structure or change process in natural phenomena. Identifying the critical node in a network is an important application of network analysis. This task is called Network Centrality measurement. It can relate to the task of determining the most influential person or the group's representative in social network analysis. We make use of the networkx [10] library in the python to capture the insights about the network. Fig. 2 presents the Gephi visualization.

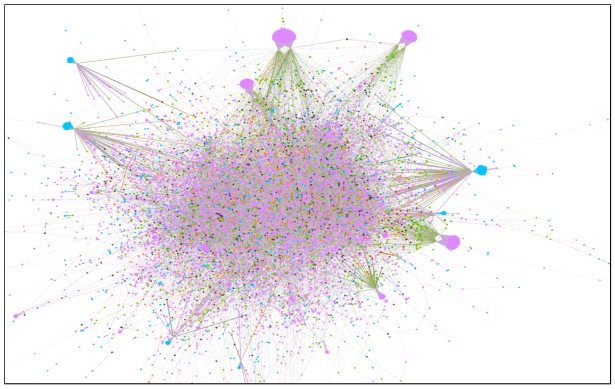


Fig. 2. Gephi visualization.

B. Sentiment Analysis

Sentiment Analysis, also known as opinion mining, is the computational study of opinions, sentiments, subjectivity, or evaluations. It is widely used across domains such as market research, public relations, reputation management, financial trading, and customer experience analysis. By examining the polarity (positive, negative, or neutral) of user-generated text, sentiment analysis helps organizations understand public perception and detect shifts in opinion.

In our study, sentiment analysis was used to determine the tone of user reviews and to assess whether sentiment polarity could improve link prediction performance. Specifically, each review was analyzed to extract a sentiment score, which we considered as a potential weight on the edge between a reviewer and a product in the bipartite graph.

To prepare the text for sentiment analysis, we applied a multi-step preprocessing pipeline: we used regular expressions to remove links and special characters, eliminated stop words using the NLTK corpus, and performed stemming using the Porter Stemmer [15]. These steps ensured that irrelevant tokens and variations of the same word did not distort sentiment scores. We used the VADER sentiment analyzer, which outputs a compound score between -1 (most negative) and +1 (most positive). These scores were mapped onto a 3-class label system (positive, neutral, negative) using VADER's default thresholds. Although we initially considered using these scores as edge weights in the graph for weighted link prediction, the review distribution was heavily skewed: out of 88,318 edges, only 2.8% were negative. This imbalance would have led to biased edge weighting and diminished performance in graph-based models. Therefore, we opted not to use sentiment scores as weights, instead using an unweighted graph for better generalizability and efficiency.

C. Topic Modeling

Topic modeling is a powerful text mining technique used to uncover hidden thematic structures in a corpus. It is especially useful in organizing and summarizing large collections of unstructured text data. We employed Latent Dirichlet Allocation (LDA), a widely adopted probabilistic model that represents each document as a mixture of topics and each topic as a distribution over words. For our study, we applied topic modeling to the set of reviews written by the most active reviewer in the dataset. This helped us gain insights into the types of content that frequent reviewers engage with, which in turn aids in better personalization for recommendations [17]. We selected four topics for the LDA model based on empirical evaluation using coherence scores, which measure the semantic consistency of words within a topic. We evaluated topic numbers ranging from 3 to 7 and found that k=4 yielded the highest coherence score (0.51), indicating that four topics offered a good balance between granularity and interpretability. The topics identified included themes such as food and travel, business and income, technology and education, and writing and products, reflecting the diverse interests of the reviewer. This thematic categorization allowed us to associate reviewers with topic clusters, which could be used for finer-grained recommendation strategies beyond graph-based link prediction.

IV. Experiment

A. Data Analysis

The analysis was conducted using the Amazon magazine subscription dataset. The provided dataset contains data about amazon magazine subscriptions, publications subscribed by a certain user (reviewerID), and reviews published for that magazine (reviewText). The overall score ranges from 0 (worst) to 5 (excellent) (best). Amazon assigns a unique identifying number (Asin) to each magazine. The data set has a total of 89,689 rows, with a total of 72098 distinct reviewers and 2428 unique Asin's.

Features:

- Overall - States the user rating of the product
- Vote - Total votes/ratings given
- Verified - Is the user verified
- reviewTime - Time at which the product was reviewed
- reviewerID - Unique ID of the reviewer

- Asin - Amazon Standard Identification Number (Product ID)
- reviewerName - Name of the review
- reviewText - product review written by the user
- Summary - summary of the user review
- unixReviewTime - timestamp of the review

The bipartite graph was subjected to network analysis, and key parameters such as density, connected components, and clustering coefficient were computed. Below describes a more detailed investigation of network.

B. Preprocessing

Prior to analysis, the dataset underwent several preprocessing steps to ensure data quality and suitability for graph construction and text-based modeling. The original dataset contained 89,689 records. After preprocessing, a total of 88,318 valid review entries remained, which were used for graph construction. The following preprocessing steps were performed:

- Text Cleaning: URLs, HTML tags, special characters, and excessive whitespace were removed using regular expressions.
- Stop Word Removal: Common English stop words were removed using the NLTK stopword corpus.
- Lowercasing: All review text was converted to lowercase to maintain consistency.
- Stemming: Words were reduced to their base/root form using the Porter Stemmer to normalize vocabulary.
- Filtering Empty Reviews: Reviews with fewer than 10 characters after cleaning were discarded.
- Verified Purchases: Only reviews from verified purchasers were retained to improve reliability.
- Reviewer Filtering: For topic modeling, we selected the reviewer with the highest number of reviews, filtering the dataset to just their entries, resulting in 146 reviews for that user.

The remaining dataset was then used to build a bipartite graph between reviewerID and Asin, where each edge represents a review. This cleaned and filtered dataset formed the basis for subsequent network, sentiment, and topic modeling analyses.

C. Graph Analysis

The strength and direction of links between nodes in a graph are determined via graph analysis. The graph analysis in this study began with determining the density and average clustering coefficient of the graph [18]. The centrality analysis was then performed, which determined how essential a node or edge is for the network's connectedness or information flow. The graph's maximum centrality was found to be 0.023. Fig. 3 shows the basic graph analysis.

Following that, small world network identification and heavy tail analysis were performed. Fig. 4 shows the result of the small world network analysis. The modularity analysis was then used to determine how thick the network's connections are. The optimal partition's modularity was found to be 0.8767 and Fig. 5 reflects the outcome.

```
Number of nodes in graph:  74526
Number of edges: 88318
Density:  3.180309778081496e-05
Number of connected components in graph: 798
Longest Component length : 71272
Maximum Degree Centrality of the graph:  0.02302583025830258
Node with maximum degree centrality:  B00005NIOH
```

Fig. 3. Basic graph analysis.

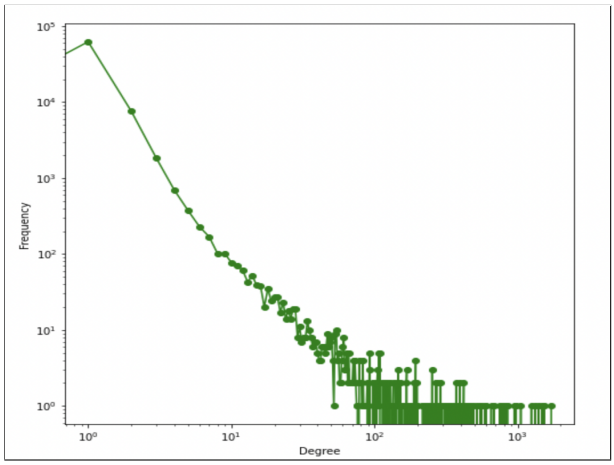


Fig. 4. Small world network analysis.

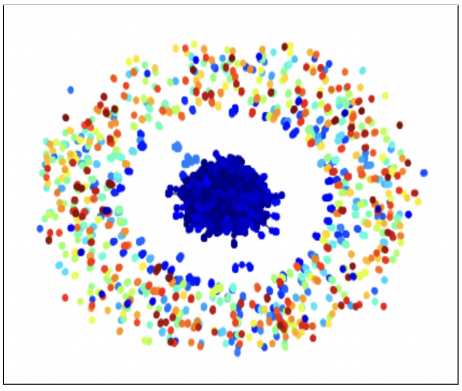


Fig. 5. Modularity plot for the best partition.

D. Topic Modeling

A user with the maximum number of Amazon magazine reviews was chosen for the purpose of topic modeling. The data of reviewers with the most reviews was sorted, and the reviewerID with the most reviews was chosen. The data list that was utilized to select the user is shown in Fig. 6.

```
reviewerID,0
A3JPFWKS83R49V,55
A2OTUWUSH49XIN,26
AEMZRE6QYVQBS,25
A3GA09FYFKL4EY,24
A3R7MXVQRGGIQ9,22
A30H2335OM7RD6,22
A1RPTVW5VEOSI,21
AKMEY1BSHSDG7,21
AVF9FV7AMRP5C,20
A2H3JURQZOHVMB,20
A2O6SU5YDVSDN4,19
A3FVAWZNKW9GX,19
AA14AMM03HMXW,19
```

Fig. 6. List of reviewers with maximum reviews.

Topic modeling was performed on the user (A3JPFWKS83R49V) with the most reviews. We filtered the dataset to capture the reviews written by the most active reviewer. We used Latent Dirichlet allocation (LDA) [6] method for fitting a topic model. This technique treats each document as a mixture of topics, and each topic as a mixture of words. The function parameters to the LDA model - the number of topics and number of words were set to four. Fig. 7 depicts the outcome of the topic modeling. Words like home, food, travel, and life appear in the first topic, implying that the user reviewed more about food and travel. The words business and income appeared in the second topic, writing product appeared in the third, and technology and education appeared in the fourth. This implies that the reviewer is a voracious reader and actively evaluates magazines about these subjects.

```
[(0, '0.031*"home" + 0.018*"travel" + 0.014*"food" + 0.012*"life"'),
 (1, '0.050*"business" + 0.038*"ads" + 0.033*"monthly" + 0.023*"small"'),
 (2, '0.035*"writers" + 0.032*"writing" + 0.024*"tips" + 0.018*"products"'),
 (3,
  '0.031*"technology" + 0.027*"popular" + 0.019*"educational" + 0.019*"reviews"')]
```

Fig. 7. Result of topic modeling on user with maximum reviews.

E. Sentiment Analysis

Out of 88318 edges or reviews, we only got 2567 (2.8 %) (see Fig. 8) negative reviews, the rest were either positive or neutral . In our case, an edge with weight as the sentiment score represents the relation between the reviewer and the product. An edge with positive sentiment score should indicate that the user praised the product and other users having common properties to that user should get recommendation of the product.

However, as observed from the Fig. 8, the number of positive, negative, and neutral reviews are very skewed. This indicates that, when it comes to magazine subscriptions, people do not tend to write negatively about them. Neutral edges can also be considered harmless to link prediction as these are most likely stating facts that summarize what is published in the magazine [18]. Considering these findings, sentiment analysis was effective in deciding not to use the sentiment scores for weighing the graph due to the skewed distribution and high complexity of bipartite weighted graph.

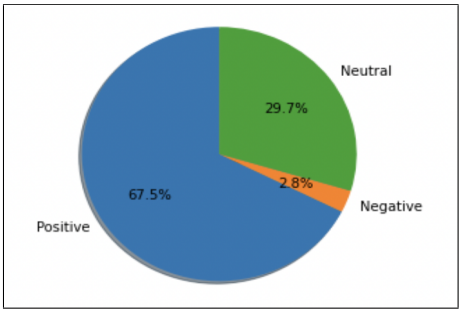


Fig. 8. Sentiment analysis results.

```
df[["cleaned_text","Positive","Negative","Neutral","Compound","Label"]].head(10)
```

	cleaned_text	Positive	Negative	Neutral	Compound	Label
0	computer enthusiast MaxPC welcome sight mailbo...	0.273	0.000	0.727	0.9790	Positive
1	Thank god Ziff Davis publication MaxPC actuall...	0.234	0.109	0.656	0.4588	Neutral
2	Antiques Magazine publication made antique lov...	0.268	0.039	0.693	0.9953	Positive
3	This beautiful magazine work art The quality e...	0.286	0.000	0.714	0.8860	Positive
4	great read every issue	0.577	0.000	0.423	0.6249	Positive
5	This magazine great times technology magazines...	0.348	0.104	0.547	0.7096	Positive
6	read Maximum MPC many years The articles techn...	0.252	0.085	0.662	0.9867	Positive
7	ordered magazine grandson going deploy technol...	0.285	0.000	0.715	0.5563	Positive
8	subscribed nook version years thoroughly enjoy...	0.418	0.000	0.582	0.8016	Positive
9	old computer Any advice help maximize computer...	0.381	0.000	0.619	0.8225	Positive

Fig. 9. Sentiment analysis scores and labels.

F. Link Prediction

In the realm of graphs and networks, link prediction is one of the most important research areas. The goal of link prediction is to find node pairs that will establish a link in the future. The simplest link prediction metrics are similarity-based metrics, which produce a similarity score for each pair x and y. The structural or node properties of the considered pair are used to calculate the score S(x,y). The non-observed linkages are given scores based on how similar they are. The anticipated link between two nodes is represented by the pair with the highest score. In the study [19], the authors details that the structural attribute of the network is one of the properties that may be used to determine the similarity measures between each pair. Local and global scores, node-dependent and path-dependent scores, parameter-dependent and parameter-free scores, and so on can all be categorized using this feature.

1) Common neighbors: The size of common neighbors [11] for a given pair of nodes is determined as the intersection of the two node neighborhoods in a particular network or graph.

$$S(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

where, $\Gamma(x), \Gamma(y)$ are neighbors of the node x and y respectively. With the number of shared neighbors between

them, the likelihood of a link between x and y increases. Newman estimated this quantity in a collaboration network and showed that the likelihood of collaboration between two nodes is determined by their shared neighbors. In a large social network, Kossinets and Watts [7] found that two students who have a lot of similar friends are more likely to be friends. The common neighbor method has been found to outperform other complex methods on most real-world networks. On applying this technique to the amazon dataset, we found the Area under the curve to be 0.65. Fig. 10 shows the ROC curve for the same.

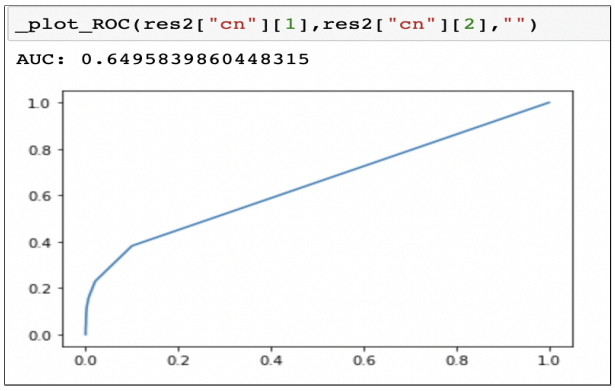


Fig. 10. ROC curve for common neighbors algorithm.

2) Adamic-Adar: Adamic and Adar[1] proposed a measure for calculating a similarity score between two web pages based on shared traits, which Liben-Nowell et al. modified and employed in link prediction.

$$S(x,y) = \sum 1/log(k_z) z \in \Gamma(x) \cap \Gamma(y)$$

where, k_z denotes the node z's degree. The equation clearly shows that the common neighbors with lower degrees are given more weight. This is also intuitive in the real world; for example, a person with a larger number of friends spends less time or resource on a single friend than someone with fewer friends[16]. On applying this technique to the dataset, we found the area under the curve to be 0.65, which is comparable to the results of Common Neighbors. Fig. 11 shows the ROC curve for the same.

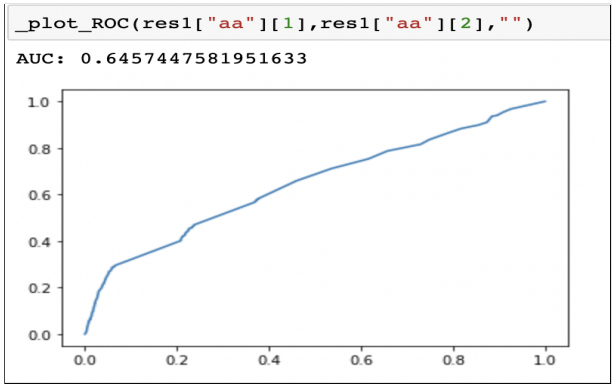


Fig. 11. ROC curve for Adamic-Adar algorithm.

3) Preferential attachment: In this study [2], to create a growing scale-free network, the concept of preferential attachment is used. The phrase "growing" refers to the progressive nature of nodes in a network over time. The likelihood of adding a new link to a node x is related to k_x, the node's degree. The preferential attachment score between two nodes, x and y can be calculated using the formula:

$$S(x,y) = k_x.k_y$$

The key advantages of this metric are its simplicity (since it takes the least amount of information for score calculation) and computing time. It can also be employed in a non-local setting because it simply requires degree as information rather than common neighbors. The PA's effectiveness improves in assortative networks, while it deteriorates in disassortative networks. In other words, if greater degree nodes are densely connected and smaller degree nodes are rarely connected, PA produces superior outcomes. Hasan et al. [5] demonstrated that aggregation functions (e.g., sum, multiplication, etc.) over feature values of vertices might be used to compute link feature value in a supervised learning framework. Fig. 12 presents the ROC curve for Preferential Attachment algorithm.

Summation can be used instead of multiplication as an aggregate function in the preceding equation, and it has been proven to be highly effective. In [5], the authors show that preferential attachment combined with the aggregate function "sum" works effectively for link prediction in co-authorship networks.

Compared to other methods, we got better results for Preferential Attachment model because of the nature of the graph. Degree distribution (Fig. 14) shows that more than half of the nodes are having degree equal to two. Additionally, as observed, very few number of nodes have high degree. Fig. 13 shows that there are some nodes in the network to which many nodes having lesser degree are attached. This situation favors the Preferential Attachment model yielding better results than other models.

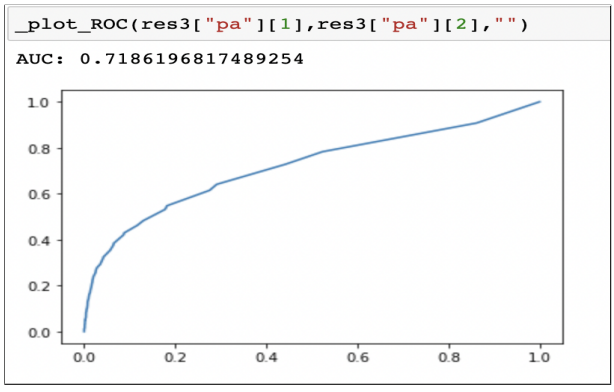


Fig. 12. ROC curve for preferential attachment algorithm.

4) Jaccard Coefficient: This [12] metric is similar to the common neighbor. Additionally, it normalizes the above score, as given below.

$$S(x,y) = |\Gamma(x) \cap \Gamma(y)| \div \Gamma(x) \cup \Gamma(y)|$$

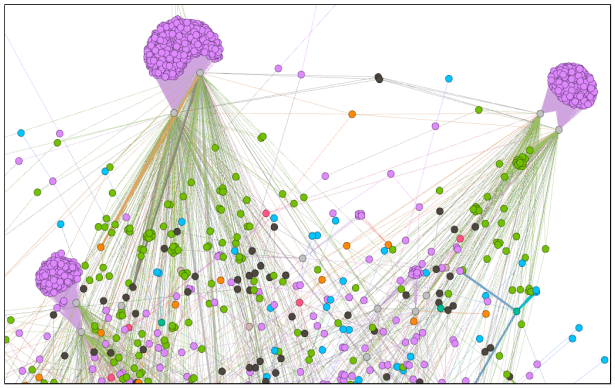


Fig. 13. Gephi visualisation - graph component where the reviewer has written 2 or more reviews.

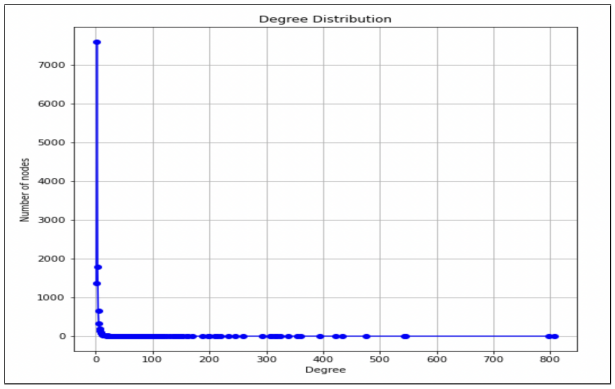


Fig. 14. Degree distribution of the graph.

The probability of choosing common neighbors of paired vertices from all the neighbors of either vertex is described by the Jaccard coefficient. The number of shared neighbors between the two vertices being analyzed improves the pairwise Jaccard score. This similarity metric performs worse than Common Neighbors, as proved by Liben-Nowell et al.[8]. Jaccard Coefficient algorithm performs the worst amongst all algorithms. As observed in Fig. 15, the area under the curve for this technique is 0.57.

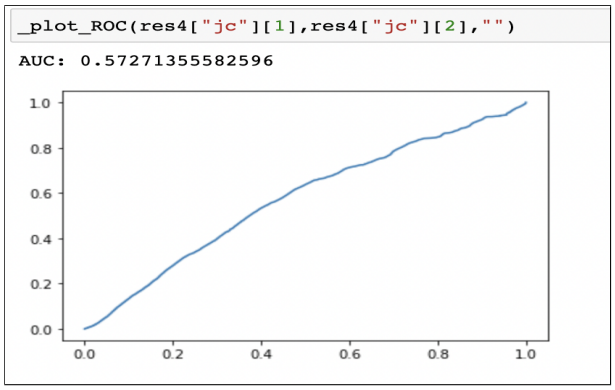


Fig. 15. ROC curve for jaccard coefficient algorithm.

The results in Fig. 16 shows the links obtained using the Preferential Attachment algorithm. As observed, the results indicate the possible links between the reviewer and the productID along with the Preferential attachment score for reviewer and the product. The score indicates how likely it is to establish a link between the two.

```
Preferential attachment Executed in 170.41651725769043 seconds


{('A3JPFWKS83R49V', 'B00005N7Q1'): 17820,
 ('A3JPFWKS83R49V', 'B00005N7OD'): 2970,
 ('A3JPFWKS83R49V', 'B00005N7OU'): 495,
 ('A3JPFWKS83R49V', 'B00005N7QC'): 3795,
 ('A3JPFWKS83R49V', 'B00005N7OV'): 16115,
 ('A3JPFWKS83R49V', 'B00005N7RD'): 5225,
 ('A3JPFWKS83R49V', 'B00005N7OP'): 440,
 ('A3JPFWKS83R49V', 'B00005N7QO'): 1540,
 ('A3JPFWKS83R49V', 'B00005N7RJ'): 1485,
 ('A3JPFWKS83R49V', 'B00005N7NQ'): 1430,
 ('A3JPFWKS83R49V', 'B00005N7O6'): 1485,
 ('A3JPFWKS83R49V', 'B00005N7PG'): 1540,
 ('A3JPFWKS83R49V', 'B00005N7S5'): 715,
 ('A3JPFWKS83R49V', 'B00005N7T8'): 1155,
 ('A3JPFWKS83R49V', 'B00005N7S8'): 495,
```

Fig. 16. Predictions for preferential attachment algorithm.

5) Other Methods: We tried link prediction with the graph distance method, which uses the shortest path length between nodes without an edge. The shortest path lengths are then sorted in ascending order, with the top "K" edges designated as future links. With the help of sentiment scores, we changed our bipartite network into a weighted graph. The edges are weighted using the compound score indicated in the Fig. 9, are used as weights of the edges. To prevent negative weights, these scores were transformed to a range of 0 to 2. Then, to find out the edges which do not exist, it was required to iterate the network of nearly 28000 edges in a quadratic time complexity. Every iteration also included time-consuming processes like identifying the shortest path length. The anticipated time it would take to run this algorithm on the network was roughly 100 days [20]. As a result, the weighted graph and shortest path length method was avoided.

Approaches like user-based collaborative filtering and item-based collaborative filtering were also investigated. Users or items are both considered vectors in these procedures. The user is then recommended an item based on the similarity between these vectors. Due to the significant sparsity of data and edges, the majority of the entries would be 0. Both algorithms will become ineffective as a consequence of this. Hence, these methods were avoided.

V. Discussion

The results from our study reveal that the Preferential Attachment model consistently outperformed other link prediction algorithms such as Common Neighbors, Adamic-Adar, and Jaccard Coefficient. This outcome can be attributed to the inherent skewed degree distribution within the reviewer-product bipartite graph, where a small number of users have disproportionately high activity levels. These highly connected reviewers are more likely to form new links, aligning with the assumptions of the Preferential Attachment model. An important insight gained from the sentiment analysis was the overwhelming positivity or neutrality in user reviews—only 2.8% of the reviews were classified as negative. While sentiment analysis provided valuable context for understanding

edge importance, the sentiment imbalance ultimately made it impractical to use these scores as edge weights for graph modeling. This highlights a broader challenge in applying sentiment-weighted graphs in domains where user feedback is rarely negative. The use of topic modeling also proved insightful in identifying the thematic diversity of active users. For instance, the top reviewer in our dataset wrote about a wide range of topics such as travel, food, business, and education. This reinforces the importance of capturing thematic preferences in designing personalized recommender systems. However, choosing the optimal number of topics remains a subjective decision, and while coherence scores guided our choice of four topics, more sophisticated tuning strategies could further improve interpretability. From a broader perspective, this study demonstrates the value of integrating semantic and behavioral insights into graph-based recommender systems. By combining network structure with textual features, we move closer to personalized and context-aware recommendations. However, computational complexity remains a concern, especially for scaling methods like shortest path-based prediction on weighted graphs. Future work could explore more efficient graph embedding techniques that incorporate both structure and sentiment while maintaining interpretability.

VI. Conclusion

In this study, we constructed a bipartite network connecting Amazon reviewers to magazine subscriptions using a real-world dataset. To extract deeper behavioral and thematic insights, we applied sentiment analysis, topic modeling, and network visualization techniques. These insights were then leveraged to perform link prediction, aiming to recommend relevant magazine subscriptions to users.

We evaluated several link prediction algorithms such as Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment—using K-fold cross-validation and ROC-AUC scores. Among these, the Preferential Attachment model yielded the most promising results, largely due to the skewed degree distribution observed in the graph, where a small number of highly active users dominated the network structure.

Although sentiment analysis offered valuable context, the overwhelmingly positive and neutral nature of the reviews limited its effectiveness as a weighting factor. As a result, link prediction was performed on an unweighted graph to avoid introducing bias from sentiment imbalance.

Looking forward, future research could explore methods to balance or normalize sentiment distributions to support weighted graph modeling. Additionally, incorporating graph embedding techniques or temporal dynamics could further enhance recommendation accuracy and scalability. Overall, this work highlights the potential of combining network structure with natural language insights to improve recommendation systems in content-rich domains like magazine subscriptions.

References

[1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. Social Networks 25, 3 (2003), 211–230. hNps://doi.org/10.1016/S0378-8733(03)00009-1

[2] A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. 2002. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 311, 3 (2002), 590– 614. hNps://doi.org/10.1016/S0378-4371(02)00736-7

[3] J. Kleinberg D. Liben-Nowell. 2003. The Link Prediction Problem for Social Networks. CIKM (2003).

[4] Evan Darke, Zhou Zhuang, and Ziyue Wang. 2017. Applying Link Prediction to Recommendation Systems for Amazon Products.

[5] M.A. Hasan, V. Chaoji, S. Salem, and M. Zaki. 2006. Link prediction using supervised learning. Proc. of SDM 06 Workshop on Link Analysis, Counterterrorism and Security (04 2006).

[6] Hamed Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. 2017. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. (11 2017).

[7] Gueorgi Kossinets and Duncan J. Watts. 2009. Origins of Homophily in an Evolving Social Network. Amer. J. Sociology 115, 2 (2009), 405–450. hNps://doi.org/10.1086/599247 arXiv:https://doi.org/10.1086/599247

[8] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In Proceedings of the Twelfth International Conference on Information and Knowledge Management (New Orleans, LA, USA) (CIKM '03). Association for Computing Machinery, New York, NY, USA, 556–559. hNps://doi.org/10.1145/956863.956972

[9] Yilun Liu, Chunhao Wu, and Xiaohui Tong. 2021. Prediction of Co-purchasing Products. (02 2021).

[10] NetworkX. 2022. https://networkx.org/. (04 2022). hNps://networkx.org

[11] M. E. J. Newman. 2001. Clustering and preferential attachment in growing networks. Phys. Rev. E 64 (Jul 2001), 025102. Issue 2. hNps://doi.org/10.1103/PhysRevE.64.025102

[12] Jaccard P. 2013. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. (06 2013).

[13] Sushil Khairnar, "Application of Blockchain Frameworks for Decentralized Identity and Access Management of IoT Devices" International Journal of Advanced Computer Science and Applications(IJACSA), 16(6), 2025. http://dx.doi.org/10.14569/IJACSA.2025.0160604

[14] Zhang, Y., & Chen, X. (2020) Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1–101. [https://doi.org/10.1561/1500000056]

[15] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[16] McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference, 43–52.

[17] D. Bodra and S. Khairnar, "Comparative Performance Analysis of Modern NoSQL Data Technologies: Redis, Aerospike, and Dragonfly," J. Res. Innov. Technol., vol. 4, no. 2, pp. 193-200, 2025. https://doi.org/10.57017/jorit.v4.2(8).05 [https://doi.org/10.1145/2766462.2767755]

[18] Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer Series in Statistics.

[19] Rajaraman, A., & Ullman, J. D. (2011). Mining of Massive Datasets. Cambridge University Press.

[20] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. LREC 2010.

Empowering Accessibility: IoT-Driven Smart Buildings for Elderly and Disabled Individuals

Zeyad Alshboul[1], Burhan Mahmoud Hamadneh[2]*,
Turki Mahdi Alqarni[3], Bajes Zeyad Aljunaeidia[4], Methaq Khadum[5]
Department of Civil Engineering-Faculty of Engineering, Ajloun National University P.O. Box 43, Ajloun 26810, Jordan[1]
Department of Special Education-Faculty of Educational Sciences,
Ajloun National University, Ajloun, Jordan, P.O. Box 43, Ajloun 26810[2]
Department of Special Education-Faculty of Education, Najran University, Najran, Kingdom of Saudi Arabia[3]
Department of Computer Science-Faculty of Information Technology,
Ajloun National University, Jordan, P.O. Box 43, Ajloun 26810[4]
Department of Electrical Engineering-Faculty of Engineering, Al-Mustansiriyah University, Iraq[5]

***Abstract*—This study aims to examine the attitudes of elderly and disabled individuals in Saudi Arabia toward Internet of Things (IoT)-enabled smart home technologies, with specific attention to the influence of demographic variables. The research employed a descriptive survey design, utilizing an online questionnaire distributed to a stratified random sample of 249 participants. Stratification ensured balanced representation across gender, age, educational attainment, employment status, and economic background. Statistical analyses, including Scheffé's post hoc test, revealed generally positive attitudes toward IoT adoption, primarily driven by perceived benefits related to enhanced quality of life, personal safety, and autonomy. Significant differences were identified across several demographic variables: married individuals, employed participants, those with higher education, higher-income groups, and individuals aged 30 to 45 all reported more favorable attitudes. Similarly, individuals with disabilities expressed stronger acceptance compared to their elderly counterparts. In contrast, gender differences were not statistically significant. These findings highlight the need for targeted, inclusive strategies that promote the adoption of IoT technologies across diverse social groups. The study contributes to a deeper understanding of how demographic characteristics shape technology acceptance and underscores the urgency of designing accessible, user-centered smart home systems. Recommendations emphasize public awareness initiatives, affordability measures, and inclusive design practices contributing to digital equity and aligning with the broader objectives of Saudi Vision 2030 for sustainable urban development.**

Keywords*—*IoT; smart home; sustainability; urban development; quality of life; assistive technologies

I. Introduction

Smart home technology plays a pivotal role in the realm of IoT applications. It transforms a traditional household into a connected environment, where various devices communicate with one another via the Internet. This interconnected setup enhances the home's intelligence, connectivity, and effectively automates various aspects of daily life [1]. The concept of smart home technology began to gain prominence in the early twenty-first century, aiming to network and interconnect household devices and equipment. However, researchers have no consensus on the definition of smart homes [2]. Information and Communication Technologies (ICTs) have been seamlessly integrated into consumer products like phones and televisions, as well as urban infrastructure and networks, significantly improving the quality of life [3]. A smart home represents a fusion of technology and services tailored to enhance the living environment's efficiency, comfort, and security for its occupants [4].

The integration of these home systems allows them to communicate with one another via a central control unit. This enables the execution of pre-programmed scenarios or operational modes, facilitating single-button or voice-controlled management of various home systems, simultaneously [5]. One of the most significant aspects of a smart home is its capacity to provide environmental control, allowing remote or automated management of essential home components like thermostats, ventilation systems, lighting fixtures, kitchen appliances, and various household devices. This automation not only simplifies household management but also reduces the workload of residents, contributing to their overall convenience [6].

Despite the numerous advantages and benefits of smart homes, they are often viewed as somewhat inconvenient and novel technology due to certain drawbacks. These drawbacks encompass a high initial cost, limited availability, and a shortage of experts in the field for installation and maintenance [7]. Moreover, concerns related to weak security and privacy are a significant issue for users or potential users. Information collected by these devices is typically transmitted over open networks like the Internet or via systems provided by service providers. This can potentially result in data breaches and access by hackers, exposing individuals to real risks such as theft or extortion [8].

This study explores the growing role of smart homes in IoT applications while addressing concerns such as privacy violations and weak security due to potential remote breaches [9]. Research in South Korea and Malaysia has shown that many users have negative attitudes toward adopting IoT in smart homes [10]. Additionally, many homeowners remain unaware of the benefits of smart home technology, contributing to reluctance to adopt [11]. Previous studies have primarily focused on younger users while overlooking elderly individuals

*Corresponding Author.

and people with disabilities [12]. Furthermore, research on smart home adoption in Arab countries remains limited. This study seeks to examine attitudes toward IoT in smart homes among elderly and disabled individuals in Saudi Arabia. It also investigates whether demographic factors such as gender, age, education, and economic status influence these attitudes [13].

This study makes several significant contributions to the literature on IoT and smart home systems. First, it investigates the influence of demographic factors, such as gender, age, education, employment, and economic status, on attitudes toward adopting IoT in smart homes. This focus provides insights into how diverse populations perceive and interact with smart home technologies. Second, the study highlights the potential of IoT to enhance the quality of life for elderly individuals and people with disabilities, emphasizing its role in promoting independence, safety, and convenience. This study delivers deep knowledge to user-centered design in smart home systems, aligning with Saudi Vision 2030's goals of fostering sustainable urban development and inclusivity. By addressing the attitudes and motivations of underrepresented groups, this research offers practical recommendations for governments, policymakers, and technology companies to develop inclusive and accessible smart home solutions.

This work is structured to include (see Section II) reviews of related works on smart home technology and IoT adoption, focusing on elderly and disabled populations. A description of the methodology used, including the sample, data collection tools, and analysis techniques, has been presented in Section III. Section IV presents the results, highlighting the influence of demographic factors on attitudes towards IoT adoption. Section V presents the discussion of the study. A conclusion is given to provide recommendations for future research and policy implications in Section VI. Limitations of the study are presented in Section VII.

II. Related Works

A. Societal Influence and the Need for User-Centered Design

The evolution of smart homes is fundamentally driven by societal perceptions, emphasizing the necessity of a user-centered design rather than a purely technological orientation. Smart home systems deliver clear advantages in security, energy efficiency, and accessibility via automated lighting, temperature regulation, and appliance management. For example, occupancy sensors dynamically optimize lighting and HVAC performance to reduce energy consumption while enhancing comfort [8, 10, 11].

A recent systematic review highlights that demographic factors, including gender, age, and education, are consistently among the strongest moderators affecting consumer adoption and acceptance of smart home IoT devices [14]. Another empirical study reports that income and education levels significantly influence intention to use smart home services, whereas gender and age may not always have a consistent direct effect. Specifically, those with higher education and income exhibit greater receptivity to IoT technologies, while adoption intention tends to vary by residential type and prior experience [13]. Furthermore, cutting-edge research from 2024 proposes human-centered explain ability frameworks for AI-driven smart home systems, advocating for interfaces that users can intuitively understand, thereby strengthening trust, transparency, and acceptance [15]. Similarly, another study leveraging large language models (LLMs) to personalize smart home interactions demonstrated a measurable improvement of around a 52% increase in user satisfaction ratings, underscoring the value of preference-based, user-adaptive design [16].

B. Smart Homes in Assisted Living Contexts

Smart home solutions are increasingly being designed within assisted living environments to support active aging, independence, and adaptive care, with a focus on features such as fall detection, emergency response, medication adherence, and sensory integration [12, 13].

Recent developments in smart home platforms have incorporated technologies such as radio frequency identification (RFID), wearable sensors, wireless networks, and artificial intelligence to achieve seamless interoperability and personalized functionality. These integrations enhance user autonomy and minimize dependence on continuous direct caregiving, particularly in assisted living environments [17]. Recent studies expand this field by demonstrating how smart flooring powered by the Internet of Things and deep learning analytics can achieve up to 98% fall detection sensitivity and 99% accuracy using non-intrusive sensors embedded under carpets, an approach that balances effectiveness, privacy, and everyday functionality [18].

Furthermore, innovations in assistive technology now include automated medication dispensers, wearable tracking devices with geofencing capabilities, and AI-powered remote monitoring, which collectively enable immediate support for healthcare adherence and rapid alerts to caregivers in emergencies [19].

For example, specialized IoT-based smart home systems for individuals with dementia have demonstrated their ability to monitor medication use, track wandering behaviors, and support remote healthcare functions through NFC-enabled pillboxes and environmental sensors, facilitating safe and independent living [20]. Advanced commercial solutions highlight the progress made in smart assisted living[21]. This device achieves up to 99.5% accuracy in detecting falls even in the case of partial obstruction, features two-way audio communication, and automatically alerts caregivers or emergency services upon detection, demonstrating the practicality and cost-effectiveness of smart technologies in private residences [21, 22]. Together, these developments reflect a clear evolution in smart home technologies, ranging from general automation to highly personalized, user-centric systems. These technologies are designed not only to meet the functional needs of the elderly or those with cognitive disabilities, but also to preserve their dignity, minimize intrusion, and enhance their independence in their familiar home environments.

C. Barriers to Adoption: Privacy, Security, and Awareness

Research in South Korea and Malaysia has shown that many users have negative attitudes toward adopting IoT in smart homes, often due to concerns about privacy and security [23]. Furthermore, many homeowners remain unaware of the benefits

of smart home technology, which contributes to their reluctance to adopt it [24].

Recent empirical evidence from across Asia reveals a range of critical barriers to the adoption of IoT-based smart home technologies, with privacy concerns and cybersecurity vulnerabilities emerging as the most prominent challenges. In addition to these technical and ethical concerns, a pervasive lack of user awareness regarding the potential benefits of IoT devices significantly contributes to users' reluctance, thereby impeding the broader integration of such technologies into residential settings. Furthermore, limited security literacy among users and inadequate adherence to data confidentiality practices further exacerbate these issues, highlighting the urgent need for targeted awareness initiatives and robust security frameworks to support secure and informed adoption [14, 25, 26].

D. Gaps in Literature: Marginalized Populations and Regional Contexts

Previous empirical studies have predominantly centered on younger, technologically adept users, often overlooking key marginalized groups such as elderly individuals and people with disabilities [3, 27]. This oversight has resulted in a limited understanding of how these populations perceive and interact with smart home IoT technologies. In addition, there is a significant paucity of research focusing on the adoption of smart home innovations within Arab contexts, particularly in Gulf countries such as Saudi Arabia [27, 28]. To address these research gaps, the present study investigates the attitudes of elderly and disabled individuals in Saudi Arabia toward smart home technologies supported by the Internet of Things (IoT). Specifically, the study explores the extent to which various demographic variables influence these attitudes. These variables include gender, marital status, educational level, employment status, age, economic status, previous experience with smart technologies, and user category (e.g., elderly vs. disabled). By examining these dimensions, the research aims to provide a more inclusive and context-sensitive understanding of the factors shaping technology adoption among underrepresented populations in the Arab world.

III. Methodology and Approaches

A descriptive survey method has been conducted to quantitatively depict the current state of affairs. This method is well-suited to the study's nature, enabling it to address its inquiries and fulfill its objectives effectively. The questionnaire served as the primary data collection tool to gather descriptive quantitative data regarding the attitudes of elderly individuals and those with disabilities regarding the utilization of IoT in smart home systems.

A. Sample of the Study

A stratified random sample of 249 elderly people and people with disabilities was selected from the Najran region in the Kingdom of Saudi Arabia. The participants contributed to answering the questionnaire tool in collaboration with Candle of Hope Association for People with Disabilities and Noor Najran Association for Women and People with Disabilities. The questionnaire has been electronically distributed to a study sample via an electronic link by use of Google Forms services. The questionnaire has been published in collaboration with authorities from the previously stated entities through personal email and mobile apps for study participants. The demographic distribution showed significant participation of both women and men in the survey. In Table I, the demographic description of the targeted groups is equally distributed among demographic variables. It depicts the distribution of research sample members based on demographic factors (gender, marital status, educational level, employment status, age, experience, category, and economic status).

TABLE I. Sample Distribution based on Demographic Variables, Namely: Gender, Marital Status, Educational Level, Employment Status, Age, Experience, Category, and Economic Status

Variable	Value Label	No.	Percentage (%)
Gender	Male	106	43
	Female	143	57
Marital status	Married	106	43
	Single	143	57
Educational level	Illiterate	94	38
	School	89	36
	University	66	27
Employment status	Work	128	51
	No work	121	49
Age	15-30	56	22
	30-45	69	28
	45-60	76	31
	+60	48	19
Experience	Likely to use	131	53
	Use	118	47
Category	The Elderly	137	55
	People with disabilities	112	45
Economic status	Average or below	124	50
	Above average	125	50
Total		249	100

B. Tools of the Study

The development of the questionnaire was guided by a clear objective: to assess attitudes toward the utilization of Internet of Things (IoT) technologies in smart home systems, particularly in relation to improving the quality of life for elderly and disabled individuals. This instrument was designed based on a thorough review of relevant theoretical frameworks and empirical studies. Key references included national resources such as the Saudi Association for Individuals with Disabilities and international studies by Marn-Daz [29], Wilson et al. [3] and others [30, 31]. These sources provided validated constructs and item formats that were adapted to the context of this study to ensure both content relevance and conceptual alignment within the target population. The rationale for using a structured questionnaire was to systematically capture subjective attitudes and perceptions in a standardized format that facilitates quantitative analysis. The survey was constructed in two sections: the first collected demographic and contextual

information; the second included 11 Likert-scale items measuring the perceived impact of IoT use on quality of life. The five-point Likert scale (ranging from "strongly agree" to "strongly disagree") was employed to allow for nuanced responses while maintaining analytic simplicity. Score ranges were determined based on prior psychometric studies, enabling consistent interpretation of response trends. However, the design and implementation of the questionnaire are subject to several limitations. First, the reliance on self-reported data inherently introduces potential biases, such as social desirability bias, recall bias, or respondents' limited awareness of IoT functionalities. Although the anonymity of responses was preserved to mitigate some of these effects, such biases may still influence the results. Second, the sample was selected based on accessibility and willingness to participate, which may restrict the generalizability of findings to the broader population of elderly and disabled individuals. This self-selection bias, coupled with possible underrepresentation of individuals with severe disabilities or low technological literacy, limits the extent to which the results can be extrapolated. Third, while the questionnaire underwent expert validation and pilot testing for clarity and relevance, its psychometric properties (e.g., construct validity, test-retest reliability) warrant further investigation in future studies with larger and more diverse samples.

C. Validity of the Tool

The initial version of the study tool has been presented to ten faculty experts, specializing in computer science, networks, information, and communication systems, as well as psychological measurement and evaluation, hailing from various Saudi universities. Experts have examined the relevance of each item to the overall tool, scrutinized the linguistic formulation for accuracy, and assessed the tool's appropriateness for reaching the study's objectives. The experts confirmed that the tool was well-suited for fulfilling the study's aims. The questionnaire was administered to a preliminary sample of thirty elderly individuals and people with disabilities from the study area, a subset outside the main study population. Afterwards, the correlation coefficients of each item with the total tool score were computed. The results are detailed in Table II.

TABLE II. CORRELATION COEFFICIENTS AMONG VARIABLES AND THE TOTAL SCORE OF THE TOOL AT THE SIGNIFICANT LEVEL OF 0.05

Variable no.	Correlation significance /variable	Variable no.	Correlation significance /variable
1	0.31	7	0.54
2	0.49	8	0.58
3	0.41	9	0.61
4	0.43	10	0.45
5	0.66	11	0.55
6	0.35		

The correlation coefficients of the individual items with the entire tool fell within the range of 0.31 to 0.66, demonstrated acceptable and statistically significant levels (see Table II). As a result, there has no need to remove any of the items from the tool. These findings affirm the tool's validity in accurately measuring the intended aspects.

D. Reliability of the Tool

To ensure the tool's reliability, it was administered to a group of 25 individuals with disabilities, who were not part of the study sample. The tool has initially applied and then reapplied after a two-week interval. The Pearson correlation coefficient has been calculated to determine the consistency of responses across the two administrations of the tool. The reliability coefficient, as assessed through the test-retest method, is 0.88 for the entire tool. The tool's reliability has also been assessed using the internal consistency method, calculated through the "Cronbach's Alpha" equation for the tool. According to this method, the reliability coefficient for the entire tool was 0.79. These results collectively affirm the tool's reliability in consistently measuring the intended aspects. The statistical software SPSS version 23 has been employed for data analysis. To address the first research question, means, standard deviations, and rankings were calculated. The second research question has been tackled through multiple analyses of variance, with subsequent multiple comparisons conducted using Scheffé's method.

IV. RESULTS AND DISCUSSION

According to the variable associated with the level of attitudes towards using IoT in implementing the smart home system, according to the estimates of the study sample members of the elderly and people with disabilities as users and potential users?

Mean value, variance, and rank have been calculated for the responses for the first question of this study about the level of attitudes toward using IoT in implementing the smart home system to support the quality of home life as perceived by the elderly and people with disabilities as users and potential users. Mean values ranged between 3.34 and 3.56, with medium and high ratings (see Table III).

TABLE III. MEANS AND STANDARD DEVIATIONS OF THE LEVEL OF ATTITUDES TOWARD USING IOT IN THE SMART HOME SYSTEM TO SUPPORT THE QUALITY OF HOME LIFE AS PERCEIVED BY THE ELDERLY AND PEOPLE WITH DISABILITIES AS USERS AND POTENTIAL USERS

No.	Items	Mean	SD	Rank	Level
5	I see that this system will provide me with belonging to the home in which I live	3.56	1.31	1	High
3	I believe this system will help me be free from anxiety and stress in my home	3.53	1.18	2	High
9	I believe that this system will enable me to carry out work and tasks within my home easily	3.53	1.08	2	High
1	I believe that this system helps me practice motor activities inside the house with complete freedom	3.49	1.26	3	High
11	I believe that this system will help me practice recreational activities at home with ease	3.49	1.19	3	High
2	I believe this system will help me have a healthy atmosphere at home	3.46	1.19	4	High
4	I think this system will help me improve my overall mood at home	3.44	1.32	5	High

7	I believe that this system will help me enjoy a strong social network around my home	3.39	1.16	6	Medium
10	I believe that this system will enable me to go out and return to my home easily and conveniently	3.39	1.18	6	Medium
6	I believe that this system has a high ability to protect my privacy	3.35	1.26	7	Medium
8	I see that this system will keep me close to my family members	3.34	1.20	8	Medium
	Total degree	3.45	0.910		High

The mean value of the total score for the level of attitudes towards using IoT in the smart home system, according to the estimates of the study sample members of the elderly and people with disabilities as users and potential users, was 3.45±0.91. The responses that see that the smart system will provide them with belonging to the home in which they live came in first place with the highest mean value of 3.56±131, with a high degree. Item (3) "I believe that this system will help me to be free from anxiety and psychological pressure in my home", and item (9) "I believe that this system will enable me to carry out work and tasks inside my home easily", scored second with the same mean values of 3.53 and the high level of degree. Item (8) "I see that this system will provide me with closeness to my family members" ranked last with a mean of 3.34 and a medium degree.

Mean values and standard deviations have been calculated for the responses of the study sample members regarding the level of attitudes towards the use of IoT in applying the smart home system to support the quality of home life for the elderly and people with disabilities as users and potential users according to demographic factors (gender, marital status, level Educational, employment status, age, experience, category, economic status). Table IV depicts the results. Results associated with the effect of demographic factors (gender, marital status, educational level, employment status, age, experience, class, economic status) attitudes towards using IoT in the smart home system, according to the estimates of the study's sample of elderly people and disabled people as users and potential users presents evidence of variations among the mean responses of the study sample participants concerning their attitudes toward using IoT in smart home systems (see Table IV).

TABLE IV. MEAN AND VARIANCE OF THE RESPONSES OF THE MEMBERS REGARDING THE LEVEL OF KNOWLEDGE TOWARDS USING IOT IN THE SMART HOME, ACCORDING TO DEMOGRAPHIC FACTORS

Variable	Category	Mean	No.	SD
Gender	Male	3.36	106	0.93
	Female	3.53	143	0.88
Marital status	Married	3.82	106	0.96
	Single	3.18	143	0.76
Educational level	Illiterate	3.37	94	0.80
	School	3.28	89	0.91
	University	3.79	66	0.97
	Work	3.67	128	1.00
Employment status	No work	3.23	121	0.75
Age	15-30	3.58	56	0.82
	30-45	3.70	69	0.96
	45-60	3.25	76	0.88
	+60	3.27	48	0.89
Experience	Likely to use	3.17	131	0.74
	Use	3.77	118	0.97
Category	The Elderly	3.28	137	0.90
	People with disabilities	3.67	112	0.87
Economic status	Average or below	3.10	124	0.75
	Above average	3.80	125	0.92

These variations are associated with various demographic factors, including gender, marital status, educational level, employment status, age, experience, category, and economic status. To determine the statistical significance of these differences at the 0.05 level, multiple analyses of variance were employed. The outcomes are depicted in Table V.

TABLE V. MULTIPLE VARIANCE ANALYSIS OF THE LEVEL OF ATTITUDES TOWARDS THE USE OF IOT IN THE SMART HOME SYSTEMS ACCORDING TO DEMOGRAPHIC FACTORS

Variable	Type I Sum of Squares	df	Square Mean	F	Significance
Gender	1.779	1	1.779	2.884	.091
Marital status	23.700	1	23.700	38.413	.000
Educational level	8.137	2	4.069	6.595	.002
Employment status	4.134	1	4.134	6.700	.010
Age	5.768	3	1.923	3.117	.027
Experience	8.746	1	8.746	14.176	.000
Category	2.534	1	2.534	4.107	.044
Economic status	5.337	1	5.337	8.650	.004
Error	146.221	237	.617		
Total	3175.877	249			

There have been no significant differences in the level of attitudes towards using IoT in the smart home system due to the gender variable (F value = 1.779), which was statistically significant at the level of 0.091. There were statistically significant differences at the level of attitudes towards using IoT in the smart home system due to the marital status variable in favor of the married group (F value = 38.413) was statistically significant. There were statistical differences due to the employment status variable for the group of workers (employees) (F value = 6.700) scored a statistical significance of 0.010; the experience variable for the use category (I use it) (F value = 14.176); the category variable in favor of the category of people with disabilities (F value = 4.107); and the economic status variable in favor of a group above the average (F value = 8.650) was statistically significant at 0.004. There were significant differences (sig. level of 0.05) in the level of attitudes towards using IoT in the smart home system due to the

educational level variable (P value = 6.595) and statistically significant at 0.002; the age variable (F value = 3.117) statistically significant at 0.027. To demonstrate the significance of the differences according to the variables of educational level and age, multiple comparisons were used by Scheffé's method (see Table VI).

TABLE VI. MULTIPLE COMPARISONS USING SCHEFFÉ'S METHOD FOR THE SIGNIFICANCE OF THE DIFFERENCES IN THE LEVEL OF ATTITUDES TOWARD THE USE OF IOT IN THE SMART HOME SYSTEM ACCORDING TO EDUCATIONAL LEVEL AND AGE

Level of Education (I)	level of Education (J)	I-J diff. mean	Significant level
University	School	0.51 (significant level of 0.05)	0.000
	Illiterate	0.42 (significant level of 0.05)	0.004
Age (I)	Age (J)	Mean Difference (I-J)	Sig.
15 -30	(30 -45)	0.12	0.863
	(45 -60)	0.44 (significant level of 0.05)	0.010
	+ 60	0.43 (significant level of 0.05)	0.042

There have been significant differences (mean = 0.51, sig. level 0.000) among responses in the level of attitudes toward the use of IoT in the smart home system due to the educational level between the two categories of university and school for the former. Significances were also found between the university and illiterate categories in favor of the university group. The value of the Mean Difference 0.42) was statistically significant at the level of 0.007. There have been also shown a statistical difference (at the significant level of 0.05) for the responses of the study sample members in the level of attitudes toward the use of IoT in the smart home system for 30 to 45-year-old category (mean value of 0.12±0.863), and both the category of 45-60-year-old (mean value of 0.44±0.01) and over 60-year-old category for the benefit of the former (mean value of 0.43±0.042) (see Table VI).

V. DISCUSSION

Results showed that the level of attitudes towards using IoT in the smart home system, according to the estimates of the study sample members, including elderly people and people with disabilities as users and potential users, was significant. The result may be attributed to the fact that smart home and building applications have become a reality that can enhance human lives, including the elderly and people with disabilities [13]. With the advantages of using IoT in the smart home system application, ranging from voice-activated lighting to automated security systems, it provides a lot of convenience and extraordinary safety benefits. Elderly and people with disabilities have complete control over all home appliances and tools from one device, such as a mobile phone or remote control, so there is no need to wander around the house with effort, as well as managing the house from the outside as well [32]. It emphasized the integration of smart technology to ensure the well-being and quality of life for humans and Saudi society [27]. Hence, the positive attitude towards technology and the appropriate individual awareness of the importance of using IoT in applying the smart home system to provide the needs and services of the elderly and people with disabilities easily and conveniently has contributed to the high attitudes towards this modern technology. This result differs from the study of Chang and Nam [13], which showed negative attitudes among the population in the Republic of Korea towards the use of IoT in the smart home by its users or potential users. It also disagrees with that in the study [23], which showed that Malaysian residents are extremely reluctant to adopt IoT in the smart home. Perhaps this difference may be due to factors of time, place, and social and cultural conditions between these two studies and the current study. This result may benefit local and international technology companies in enhancing the local and global market with high-quality products and devices that meet the needs of the elderly and people with disabilities in using IoT in implementing the smart home system. It improves the quality of living and well-being of elderly and disabled individuals and provides special services and care appropriate to their abilities and potential. Thus, it is reflected in national belonging through the promotion of human and social values in society and the right of everyone to live in dignity, and from an economic standpoint, it enhances state revenues and national economic income.

The study examined the impact of demographic factors, including gender, marital status, education, employment status, age, experience, social class, and economic status, on attitudes toward IoT adoption in smart home systems among elderly and disabled individuals, revealing no statistically significant differences based on gender. This may be caused by the similar awareness among the study sample members, regardless of their gender, about the importance of using IoT in the smart home system in daily life and household activities and requirements, and its role in facilitating their lives and living [33]. No statistically significant differences were found due to the marital status variable in favor of the married group[13]. The result may be attributed to the role of spouses in providing great support to one another. Hence, they need to realize the value of this technology in enhancing that role between them as well as its role in relieving pressure and tension in relationships between couples, because it enables them to complete work and tasks easily. There have been significant differences attributed to the employment status variable and in favor of the working group (employees). The result may be attributed to the study sample members' awareness of the importance of this system, "IoT in implementing the smart home system", in facilitating the process of leaving the home and returning to it with all comfort and ease. Also, it plays a role in commuting to work by providing technological devices to serve them, which is reflected in their feelings and desire to use this system with welcome and love, and to encourage others to use it [5]. There were statistically significant differences attributed to the experience variable and in favor of the usage category (use it). The result may be attributed to the awareness of the people who use this system and interact with it of its importance. It includes many additional tools and characteristics that distinguish it from the traditional home, enabling them to deal with and rely on themselves, and making their home life more quality, easy, and convenient. There were statistically significant differences attributed to the category variable and in favor of the category of people with disabilities. The result may be attributed to the fact that individuals with disabilities possess some of the

cognitive and mental abilities that enable them to use this system and deal with it, in contrast to the elderly, who worry and feel afraid of using this system. In addition, most of them suffer from chronic diseases or mental illnesses such as Alzheimer's, which affects the level of their attitudes towards this system and its importance in their lives [28].

There were significant differences attributed to the economic situation variable and in favor of a group above the average. The result may be due to the person's economic factor playing a major role in their desire to own and use such a system. Therefore, people with high income or high economic levels are informed of everything new about this system to become more familiar with it in terms of its benefits, how to deal with it, and the possibility of using it currently or in the future, which affects their attitudes towards this system [7]. There were statistically significant differences due to the educational level variable in favor of the university category. The result may be attributed to the role of higher education in providing individuals with the opportunity to learn new things and know everything about technology and its great role in improving human life. Hence, the study members who received a university education may have gained experience in how to obtain reliable sources about this system, its benefits, how to deal with it, and the possibility of using it currently or in the future. There were statistically significant differences attributed to the age variable, favoring the 30 to 45 age group. This result may be explained by younger individuals' greater willingness to adopt, explore, and engage with technology, as well as their lower levels of apprehension toward its use compared to older age groups [27]. Therefore, their attitudes toward this system and their acquisition of sufficient knowledge about it, its benefits, how to deal with it, and the possibility of using it currently or in the future greatly affect their attitudes toward this system in terms of use or the probability of using it.

This study attempted to collect quantitative data and analyze it descriptively to explore the influence of demographic factors (gender, marital status, educational level, employment status, age, experience, category, economic status) as users or potential users of IoT in the smart home system and attitudes towards it. It also showed how it might affect them when they decide to adopt this system or encourage others to adopt it, and what differences or lack thereof were found based on those demographic factors. This may help in understanding and increasing the awareness of governments and technical companies specialized in smart homes that end consumers do not respond to those services proposed to them in the same way. This result may be due to the difference between them in the demographic factors discussed in the current study, which necessitates expanding the circle of spreading the culture of smart homes in a major selling space, such as international and local premium markets, for example. The influence of factors and circumstances surrounding people, including the elderly and people with disabilities, may affect their attitudes and purchasing attitudes toward this system. It also requires technical commercial companies specialized in this system to develop an organized and thoughtful strategic plan. It includes a large variety of diverse services that are integrated into using IoT in implementing the smart home system to attract the largest possible segments of society, including the elderly and people with disabilities.

VI. Conclusion and Recommendations

This study concluded that elderly individuals and people with disabilities in Saudi Arabia generally exhibit positive attitudes toward the adoption of IoT-based smart home technologies, particularly those who are married, employed, better educated, and from higher-income groups. However, certain demographic factors, such as being unmarried, unemployed, of lower socioeconomic status, or having limited educational attainment, were associated with lower adoption enthusiasm. These findings highlight the potential of IoT to improve quality of life, autonomy, and safety for marginalized groups, but also underscore the presence of critical barriers that must be addressed through targeted interventions. In light of these findings, several actionable recommendations are proposed. Government agencies, particularly the Ministry of Human Resources and Social Development and the Ministry of Communications and Information Technology, should establish inclusive digital policies and launch national awareness campaigns tailored to underserved populations. These should include digital literacy initiatives, financial support schemes, and the integration of smart technologies into broader Vision 2030 strategies to enhance social equity and digital transformation. For technology developers and solution providers, there is a clear imperative to adopt user-centered design approaches that ensure smart home systems are affordable, intuitive, and responsive to the needs of elderly and disabled users. This includes simplifying interfaces, offering Arabic language support, and addressing security and privacy concerns that may hinder adoption. Civil society organizations and charitable institutions are also encouraged to play an active role in bridging the digital divide. Through community outreach and partnerships with public institutions, these organizations can advocate for the practical benefits of IoT and facilitate equitable access to smart technologies among low-income and vulnerable individuals. Lastly, the academic community should pursue in-depth qualitative research to uncover the psychosocial and cultural dimensions that influence IoT adoption. Special attention should be given to rural populations and those with limited technological exposure to develop evidence-based strategies that ensure inclusive innovation. These efforts will contribute significantly to building a sustainable, digitally empowered society in alignment with Saudi Arabia's Vision 2030.

VII. Study's Limitations

The results of this study are limited to the following: The study was limited to investigating the topic of attitudes towards the use of the Internet of Things (IoT) in the smart home system from the point of view of the elderly and people with disabilities as users and potential users. It was also determined to explore the impact of some demographic factors such as (gender, marital status, educational level, employment status, age, experience, class, economic status) on these attitudes. In addition, the study was limited to the opinions of a sample of the elderly (over 60 years old) and people with disabilities (over 15 years old to 60 years old) from Najran region in the south of the Kingdom of

Saudi Arabia, in the period from June to July of the year 2023. Finally, one tool, the questionnaire, was used in the data collection process, so generalizing the results depends on the accuracy in extracting its validity and reliability implications and the seriousness of the respondents in answering them.

ACKNOWLEDGMENT

The researchers would like to thank and appreciate the study participants for their efforts in responding to the study tool objectively. We thank Ajloun National University for their support.

ETHICAL APPROVAL

This scientific research carries ethical approval from the Standing Committee for Scientific Research Ethics in the Research Deanship of Najran University, Saudi Arabia. Reference No.: 010118-022124-DS. Accredited as a local committee by the National Committee for Bioethics at King Abdulaziz City for Science and Technology (HAPO-11-N-102).

INFORMED CONSENT STATEMENT

Informed consent was obtained from all study participants.

REFERENCES

[1] S. Gøthesen, M. Haddara, and K. N. Kumar, "Empowering homes with intelligence: An investigation of smart home technology adoption and usage," Internet of Things, vol. 24, p. 100944, 2023/12/01/ 2023.

[2] Z. Almusaylim and N. Jhanjhi, "A review on smart home present state and challenges: linked to context-awareness internet of things (IoT)," Wireless Networks, vol. 25, 08/01 2019.

[3] C. Wilson, T. Hargreaves, and R. Hauxwell-Baldwin, "Smart homes and their users: A systematic analysis and key challenges," Personal and Ubiquitous Computing, vol. 19, pp. 463-476, 02/01 2014.

[4] C. Stolojescu-Crisan, C. Crisan, and B.-P. Butunoi, "An IoT-Based Smart Home Automation System," Sensors, vol. 21, p. 3784, 2021.

[5] S. Leo Willyanto, L. Resmana, and T. Kevin, "Smart Home System Using Internet of Things," Journal of Information and Communication Convergence Engineering, vol. 16, pp. 60-65, 2018.

[6] J. Rosslin, R. Robles, and T.-H. Kim, "A Review on Security in Smart Home Development," International Journal of Advanced Science and Technology, vol. 15, 03/01 2010.

[7] M. Hubert, M. Blut, C. Brock, R. Zhang, V. Koch, and R. Riedl, "The influence of acceptance and adoption drivers on smart home usage," European Journal of Marketing, vol. 53, 09/21 2018.

[8] A. Eltaleb, M. Khaleel, S. Abulifa, D. Abdelkareem, A. Abodwair, and M. Elsharif, "Smart Home Sensor Systems: Advancements and Applications," vol. 1, pp. 120-128, 08/06 2023.

[9] A. Coskun, G. Kaner, and İ. Bostan, "Is Smart Home a Necessity or a Fantasy for the Mainstream User? A Study on Users' Expectations of Smart Household Appliances," International Journal of Design, vol. 12, 04/30 2018.

[10] C. Patience and T. Apaokueze, "The Impact of Smart Home Technologies on Energy Efficiency, Cost Savings, and Environmental Benefits," Journal of Energy Engineering and Thermodynamics, vol. 4, pp. 21-32, 07/16 2024.

[11] S. Chang and K. Nam, "Exploring the Sustainable Values of Smart Homes to Strengthen Adoption," Buildings, vol. 12, p. 1919, 2022.

[12] J. Bennett, O. Rokas, and L. Chen, "Healthcare in the Smart Home: A Study of Past, Present and Future," Sustainability (Switzerland), vol. 9, 05/17 2017.

[13] S. Chang and K. Nam, "Smart Home Adoption: The Impact of User Characteristics and Differences in Perception of Benefits," Buildings, vol. 11, p. 393, 09/03 2021.

[14] N. Aggarwal and S. Singh, "Embracing the Future by Unlocking the Door to Smart Living: A Systematic Review of Consumers Adoption and Acceptance of Smart Home Devices," NMIMS Management Review, vol. 32, 02/11 2025.

[15] M. Shajalal, A. Boden, G. Stevens, D. Du, and D.-R. Kern, "Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments," in Explainable Artificial Intelligence, Cham, 2024, pp. 418-440.

[16] J. Rey-Jouanchicot, A. Bottaro, E. Campo, J.-L. Bouraoui, N. Vigouroux, and F. Vella, Leveraging Large Language Models for enhanced personalised user experience in Smart Homes, 2024.

[17] A. Vrančić, H. Zadravec, and T. Orehovački, "The Role of Smart Homes in Providing Care for Older Adults: A Systematic Literature Review from 2010 to 2023," Smart Cities, vol. 7, pp. 1502-1550, 2024.

[18] H. A. Alharbi, K. K. Alharbi, and C. A. U. Hassan, "Enhancing Elderly Fall Detection through IoT-Enabled Smart Flooring and AI for Independent Living Sustainability," Sustainability, vol. 15, p. 15695, 2023.

[19] A. J. Perez, F. Siddiqui, S. Zeadally, and D. Lane, "A review of IoT systems to enable independence for the elderly and disabled individuals," Internet of Things, vol. 21, p. 100653, 2023/04/01/ 2023.

[20] M. Raed, I. Huseyinov, G. Ozdemir, I. Kotenko, and E. Doynikova, "An IoT-Based Smart Home for Elderly Suffering from Dementia," ed, 2024, pp. 362-371.

[21] F. Shu and J. Shu, "An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box," Scientific Reports, vol. 11, p. 2471, 01/28 2021.

[22] K. De Miguel, A. Brunete, M. Hernando, and E. Gambao, "Home Camera-Based Fall Detection System for the Elderly," Sensors, vol. 17, p. 2864, 2017.

[23] Z. Rasyidah, A. Hariati, M. Rosadah, and M. Maryanti, "Perceptions on smart home concept among the millennials in Johor," IOP Conference Series: Materials Science and Engineering, vol. 849, p. 012055, 05/30 2020.

[24] E. Bhatnagar and D. Nim, "Impact of Housekeeping Services and Practices on Customer Satisfaction and Repeat Business," Prabandhan: Indian Journal of Management, vol. 12, p. 46, 08/31 2019.

[25] S. He, Y. Lei, X. Zhan, C. Zhang, J. Ye, R. Abu-Salma, et al., Privacy Perspectives and Practices of Chinese Smart Home Product Teams, 2025.

[26] T. Magara and Y. Zhou, "Security and Privacy Concerns in the Adoption of IoT Smart Homes: A User-Centric Analysis," American Journal of Information Science and Technology, vol. 8, pp. 1-14, 03/19 2024.

[27] M. Alasmari, "The Attitudes of Public-School Teachers towards E-learning in Saudi Arabia," Arab World English Journal, pp. 245-257, 01/24 2022.

[28] K. Maswadi, N. Ghani, and S. Hamid, "Factors influencing the elderly's behavioural intention to use smart home technologies in Saudi Arabia," PLOS ONE, vol. 17, p. e0272525, 08/30 2022.

[29] M. Torroba Diaz, A. Bajo-Sanjuan, Á. M. Callejón Gil, A. Rosales-Pérez, and L. López Marfil, "Environmental behavior of university students," International Journal of Sustainability in Higher Education, vol. 24, pp. 1489-1506, 2023.

[30] S. Shah, M. Aljawarneh, and S. Akbar, "Assistive Living in IoT-Based Smart Home Systems for Aging in Place: A Review," VFAST Transactions on Software Engineering, vol. 13, pp. 83-98, 05/13 2025.

[31] A. A. Alshdadi, "Evaluation of IoT-Based Smart Home Assistance for Elderly People Using Robot," Electronics, vol. 12, p. 2627, 2023.

[32] R. Al-Mutawa and F. Eassa, A Smart Home System based on Internet of Things, 2020.

[33] N. X. Tou, Y. H. Kee, K. Koh, M. Camiré, and J. Y. Chow, "Singapore teachers' attitudes towards the use of information and communication technologies in physical education," European Physical Education Review, vol. 26, pp. 481-494, 05/01 2020.

Enhancing Dendritic Cell Algorithm by Integration with Multi-Layer Perceptron for Anomaly Detection

Yousra Abudaqqa, Zulaiha Ali Othman, Azuraliza Abu Bakar
Research Center for Artificial Intelligence Technology-Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia

Abstract*—Anomaly detection is crucial in a variety of areas, with the Dendritic Cell Algorithm (DCA) being one of the most used artificial immune systems (AIS) and introduced for binary classification of data. Both traditional and current perspectives on classification in DCA have primarily been threshold-based methods. Such approaches are limited in important ways, including inflexibility, manual tuning, and not being context-aware. The latest improvements in literature have provided adaptive dynamic threshold mechanisms that allow the system to adjust the sensitivity of the threshold using some statistical data of real-time observations. Although this is progress, the systems proposed are still based on rules, which have traditionally struggled with the more complex, higher-dimensional and nonlinear nature of data. This is common in most complex anomaly detection tasks today. In this study, we propose an improved DCA-MLP framework that uses a Multi-layer Perceptron (MLP) classifier replacing the thresholding phase. The MLP allows the DCA to learn from data context adaptively through a context-sensitive learning mechanism that can also change with the data distribution as it evolves, eliminating the need to robotically calibrate based on static or heuristic thresholds. The framework was tested thoroughly on fourteen benchmark datasets, and performance was evaluated against standard DCA in terms of accuracy, sensitivity and specificity measures. The performance results revealed considerable enhancements in DCA-MLP's performance: 12%–50% improvements in accuracy (increasing accuracy to 93%–99%), 46% improvements in sensitivity (sensitivity as 98%), and 39% improvements in specificity. This shows that DCA-MLP is better adaptable, with learning capacity and robustness - a paradigm shift away from thresholds or threshold-based systems to an intelligent self-adjusting anomaly detection classification scheme.

Keywords*—Dendritic Cell Algorithm (DCA); anomaly threshold; Multi-Layer Perceptron (MLP); anomaly detection

I. Introduction

Anomaly detection, or the process of identifying patterns that are identified as the unusual portion of normal behavior, is important across several domains, including cybersecurity, healthcare, finance, and industrial systems. Anomalies such as fraudulent financial transactions, malignant tumors in a medical image, and cyberattacks on network traffic, all signal important implications, transforming raw data into actionable insights [7, 24, 33, 37]. Traditional classification algorithms often falter on dynamic, high-dimensional data; so researchers have resorted to diagnostic and biologically inspired solutions, such as artificial immune systems (AISs).AISs emulate the human immune system's ability to distinguish self from non-self, offering unique advantages in adaptability and unsupervised learning [1]. Among AIS-based approaches, the Dendritic Cell Algorithm (DCA)[2]stands out for its success in anomaly detection, leveraging the biological "danger theory" to classify threats by monitoring antigen behavior through metrics like the Multi-Context Antigen Value (MCAV) [3].

Despite its effectiveness, the DCA faces significant limitations. Its reliance on fixed, empirically derived thresholds for anomaly classification introduces rigidity, as static parameters fail to adapt to evolving data patterns or outliers [4]. For instance, threshold-setting methods like try-and-test experiments or class distribution analysis depend heavily on historical data and expert input, limiting scalability and real-world applicability [4]. These constraints are particularly evident in complex tasks such as network intrusion detection or time-series analysis, where dynamic environments demand adaptive solutions [5].

To address these challenges, this study proposes a novel integration of the DCA with a Multi-Layer Perceptron (MLP), replacing the static threshold mechanism with a dynamic, data-driven classification framework. Inspired by the immune system's interaction between dendritic cells (DCs) and T-cells (TCs), our approach enhances the DCA's decision-making process by embedding an MLP to learn complex relationships within antigen signals. Unlike traditional threshold-based methods, the MLP adaptively adjusts classification boundaries, improving robustness against outliers and reducing false alarms. This integration leverages the MLP's capacity to handle non-linear patterns while preserving the DCA's unsupervised learning strengths.

The contributions of this work are threefold:

1) Improved accuracy: The MLP refines [6] anomaly classification through supervised learning on fused antigen signals (e.g., danger, safe, and co-stimulatory signals), achieving accuracy improvements of 12% to 50% across benchmark datasets.

2) Dynamic adaptability: By eliminating fixed thresholds, the DCA-MLP framework dynamically adjusts to new data patterns, enhancing resilience in evolving environments.

3) Scalability: Non-Negative Matrix Factorization (NMF) streamlines signal extraction, enabling efficient processing of high-dimensional data without compromising detection performance.

Experimental validation on fourteen datasets proves the framework's superiority over traditional DCA, with sensitivity

and specificity gains of up to 46% and 39%, respectively. These advancements position the DCA-MLP as a scalable, adaptive solution for complex anomaly detection tasks, bridging the gap between immune-inspired algorithms and modern machine learning.

The rest of this study is organized as follows: Section II reviews related work on AIS and DCA enhancements. Section III details the proposed DCA-MLP architecture, while Section IV presents the experiment setup. Section V details the experimental results. Section VI presents the discussion of the study. Finally in Section VII, the study concludes with implications and future research directions.

II. Related Work

Anomaly detection plays a crucial role in various fields, including network security, healthcare, and industrial systems, where identifying unusual patterns or behaviors is essential for maintaining system integrity and performance [7]. While traditional methods such as (MLP, KNN, SVM and Decision Tree) have been effective in specific contexts, they often face challenges when dealing with the complexity and dynamic nature of modern datasets [8]. Advanced algorithms like the Dendritic Cell Algorithm (DCA) have appeared as promising solutions to these challenges, offering improved capabilities for solving anomaly detection problems [4]. The DCA, introduced by [2] operates through four key phases [9], as illustrated in Fig. 1.

In the first phase**,** pre-processing, the algorithm performs two critical tasks: feature reduction and signal categorization. Feature reduction selects the most relevant attributes from the dataset, and these attributes are then categorized into three signal types: safe, danger, and Pathogen-Associated Molecular Patterns (PAMP). The most popular methods for feature reduction include Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF). While PCA is widely used, it can obscure data interpretation by transforming the dataset into components that may lose connection to the original features [10]. In contrast, NMF excels by identifying weakly correlated or uncorrelated factors, revealing hidden patterns, and transforming data into a reduced yet interpretable space [11]. This approach avoids overfitting, improves prediction accuracy, and effectively manages data sparsity. Furthermore, NMF operates without strict statistical assumptions, making it highly adaptable for large-scale datasets. By maintaining essential data characteristics, NMF simplifies computational demands related to time and storage while ensuring robust model performance.

In the second phase, detection**,** the DCA generates a signal database by combining the input signals with antigens, resulting in cumulative output signals. The third phase, context assessment, evaluates the context of antigens using the cumulative signals. If a Dendritic Cell (DC) collects more Mature Dendritic Cells (mDC) than Semi-Mature Dendritic Cells (smDC), the antigen is labelled as anomalous (1); otherwise, it is classified as normal (0) [12]. Finally, during the classification phase**,** the Mature Context Antigen Value (MCAV) is calculated for each antigen to assess the likelihood of an anomaly. The MCAV is determined by dividing the number of times an antigen appears in the mature context by the total number of antigen presentations. This value is then compared to a predefined anomaly threshold, classifying antigens with higher MCAV values as anomaly.

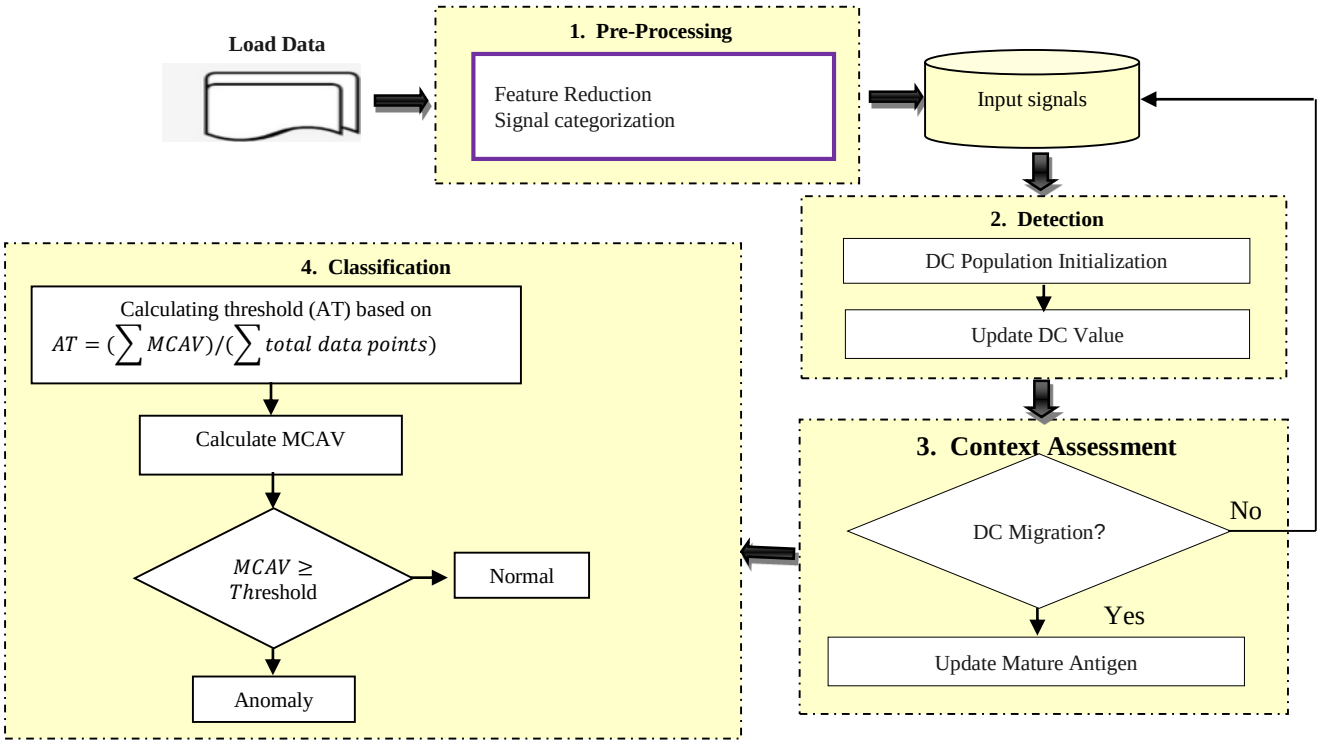


Fig. 1. The standard architecture of the Dendritic Cell Algorithm (DCA) diagram.

The MCAV plays a pivotal role in the DCA, representing the ratio of how often an antigen appears in a "mature context" compared to the total number of antigen presentations[13]. In the DCA, dendritic cells exist in two states**:** Mature Dendritic Cells (mDC)**,** which detect danger or anomalies, and Semi-Mature Dendritic Cells (smDC)**,** which reflect normal or non-threatening behavior. The MCAV measures how frequently an antigen is associated with dangerous contexts (mDC) versus normal ones (smDC), aiding in determining whether the antigen is more likely to be abnormal or normal. A higher MCAV indicates a higher likelihood of the antigen being anomalous, while a lower MCAV suggests it is normal. The MCAV is calculated as Eq. (1):

$$\text{MCAV} = (\text{Mature})/(\text{Semi Mature} + \text{Mature}) \quad (1)$$

The importance of the MCAV lies in its adaptive nature, allowing the DCA to assess anomalies based on the environment rather than static rules [14]. This flexibility makes the DCA highly capable of handling unfamiliar data points by evaluating how often an antigen appears in mature contexts, improving the algorithm's ability to classify new or unseen data accurately. Furthermore, combining the MCAV with a flexible anomaly threshold allows the DCA to adapt to different domains, such as network intrusion detection, spam filtering, and fault diagnosis, where normal and abnormal patterns vary significantly.

Anomaly Threshold (AT) is a default value used to differentiate between normal and abnormal antigens by comparing it to the MCAV of an antigen. If the MCAV exceeds the threshold, the antigen is considered abnormal or an anomaly. Currently, there are three primary strategies for determining the AT in DCA: 1) trial-and-error experimentation, 2) class distribution between normal and abnormal groups and 3) the average MCAV method [10]. The class distribution method is based on the ratio of normal to abnormal classes, requiring both classes to be balanced to generate a suitable threshold value, as shown in Eq. (2). However, achieving this balance is challenging, especially since anomalies are often rare and can create a significant gap between the two classes.

$$AT = (\sum MCAV)/(\sum total\ data\ points) \quad (2)$$

The anomaly threshold (AT) in standard DCA implementations has been applied across various domains with different strategies. In [15], the authors has enhanced the traditional DCA by introducing a migration threshold strategy that adapts to various attack scenarios, applied in network intrusion detection on the UNSW-NB15 dataset, achieving an anomaly value of 76.69% using the min MCAV approach. In [16], the authors also focused on network intrusion, using the NSL-KDD dataset to classify traffic as normal or anomalous based on immune-like danger signals. This approach identified DoS attacks through a class and min MCAV distribution strategy, providing scalability in dynamic network environments. Similarly, [17] employed a try-and-test approach in conjunction with min MCAV within a fuzzy logic system on the KDD99 dataset, enhancing adaptability to varied data types.

In [7], the authors adopted a different method, using class distribution to determine AT by assessing the ratio of abnormal to normal traffic in the NSL-KDD dataset, achieving a 0.60% anomaly threshold. In [5], the authors applied min MCAV across general classification domains with datasets such as Sonar and GCD, adjusting the threshold based on dataset-specific characteristics. Finally, [10] utilized min MCAV for anomaly detection in medical diagnosis, using datasets like WBC and LDR, focusing on metrics such as Mean Correlation Activity Value to identify anomalies dynamically. This variety of approaches underscores the flexibility of DCA in different application areas, with each study refining AT to suit specific datasets and anomaly detection needs. Some investigations have focused on the DCA classification phase, aiming to refine and enhance its effectiveness. More precisely, researchers have explored ways to improve the decision-making process by optimizing how the algorithm classifies data points as either normal or anomalous. The traditional classification approach in the DCA relies on fixed thresholds based on the multi-context antigen value (MCAV) [16], which can be limited in applications such as network intrusion detection [13] and medical anomaly detection [17]. The core of this process revolves around the calculation of the Mature Context Antigen Value, handling complex or extreme data patterns. A critical stage for identifying anomalous patterns across different domains (MCAV) [8], which determines whether a pattern is a normal presentation. When the MCAV exceeds a predefined threshold, the pattern is classified as anomalous (depicted or anomalous. The MCAV is the ratio of antigen encounters in a mature context to the total number of antigens by the decision boundary in the figure). Otherwise, it is classified as normal [18]. This classification method, though effective in many cases, highlights the challenge of setting an appropriate threshold, which directly impacts the detection accuracy of the algorithm.

Several techniques have been proposed to establish the anomaly threshold for MCAV, each with its own limitations [19]. For instance, the try-and-test method involves iterative testing to find the optimal threshold, though this process is time-consuming and highly reliant on expert knowledge. Other approaches, such as the minimum MCAV threshold introduced by[7] and adaptive thresholds based on cumulative sum (CUSUM)[10], have also been explored. While these methods offer systematic approaches to threshold determination, they deal with outliers and data changes over time, which can limit their effectiveness for longer-term anomaly detection. However, the challenge remains that these methods still depend on historical data for initial threshold values may not be reliable in rapidly evolving datasets [10, 16, 20].

An effective method for enhancing the traditional threshold methods in the DCA is to adopt the Multi-Layer Perceptron (MLP) algorithm into the DCA. Unlike the standard method of comparing MCAV values to a predefined threshold for anomaly detection, the MLP algorithm learns from the data, adjusting its parameters to better manage outliers and extreme values. This adaptability enables the MLP-DCA combination to have greater accuracy and success in detecting and classifying anomalies. By leveraging the machine learning capabilities of MLP, this integration enhances the overall process, making it more resilient against unusual data points that might otherwise challenge a simple MCAV-based threshold approach.

Integrating the DCA with an MLP algorithm presents several key advantages. To begin with, it enhances system robustness by giving the MLP the ability to better deal with extreme values

and outliers. This ensures that the algorithm maintains reliable performance even when faced with unexpected data. Secondly, the learning capabilities in the MLP improve the ability to detect and classify anomalies by allowing the model to recognize complex patterns and associations in the data assignments. The MLP refines its understanding over time, it produces more precise and reliable results. Moreover, the adaptability of the MLP enables it to respond to evolving data patterns, making it a flexible and sustainable solution for anomaly detection.

This combination overcomes the limitations of traditional threshold-based approaches by introducing adaptive learning from data, making it a valuable advancement in the field of anomaly detection.

III. The Proposed Method

This study proposes an enhanced anomaly detection framework known as MLP-DCA**,** which integrates the Dendritic Cell Algorithm (DCA) with a Multi-Layer Perceptron (MLP) classifier to overcome critical limitations of traditional threshold-based decision-making in DCA. Two primary limitations of the standard DCA model motivate this integration:

1) Its reliance on a fixed threshold for anomaly classification, which leads to poor adaptability when handling datasets with imbalanced class distributions **or** dynamic patterns.

2) The manual effort and expert knowledge required to tune this threshold, making it impractical for real-world, large-scale, or evolving data environments.

To address these challenges, the proposed MLP-DCA model retains the original architecture of the standard DCA, which includes four primary phases: signal preprocessing, antigen processing, signal integration**,** and classification (see Fig. 1). The first three phases are preserved to leverage the biologically inspired behavior and contextual data fusion capabilities of DCA, which have proven effective in prior studies [11].

In the preprocessing phase, the model introduces Non-Negative Matrix Factorization (NMF) for automated feature transformation and signal categorization**.** This enhancement allows the model to handle high-dimensional datasets by reducing complexity and focusing on the most informative features. NMF serves as a dimensionality reduction technique, helping to categorize input attributes into the necessary signal types (PAMP, danger, and safe) without manual mapping.

The major innovation lies in the classification phase, where the traditional MCAV-based anomaly threshold (AT) is replaced with an MLP classifier (see Fig. 2). Instead of relying on static decision rules, the MLP dynamically learns non-linear and high-level feature relationships from the processed signals and antigen information. This adaptive learning approach enables the system to generalize better across diverse datasets, respond to class imbalance, and improve classification robustness On the whole, by combining context-aware processing capabilities of DCA with the learning power of MLP, we proposed model achieves a more flexible and scalable anomaly detection system. The MLP-DCA architecture in addition to completely removing threshold tuning, these powerful neurons improve the accuracy, sensitivity and specificity across many benchmark datasets, demonstrated in the experimental section.

As depicted in Fig. 2, the classification process begins the same as a standard DCA which is by calculating the MCAV [Eq. (1)]. The calculated MCAV values (acting as the independent variable) are fed into the MLP that classifies the antigens, as shown in Eq. (3):

$$f(MCAV) = (\sum_{i=1}^{m} w_i \,.\, MCAV_i) + b \qquad (3)$$

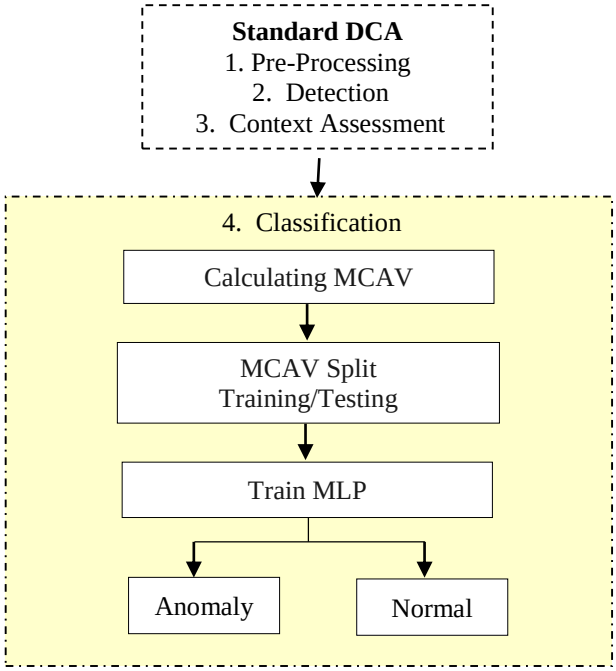


Fig. 2. Proposed method MLP_DCA.

where:

- $MCAV$: Represents the mean of context activation vectors
- derived from the inputs or data points.
- w_i: Weights assigned to each $MCAV_i$ during training.
- b: Bias term.
- m: Total number of MCAV inputs.
- The MLP leverages this transformation to predict the class of each antigen (e.g., normal or abnormal) based on learned parameters.

Fig. 3 illustrates the pseudocode for integrating the Dendritic Cell Algorithm (DCA) with a Multi-Layer Perceptron (MLP) during the classification phase. This integration comprises three primary steps:

1) Step 4.1 Calculation of MCAV values: Initially, for each antigen generated in prior stages, the Mature Context Antigen Value (MCAV) is computed using Equation (1). These MCAVs are then stored in the MCAV_values []. array for subsequent classification.

2) Step 4.2 Training the MLP model: After the MCAV values are recorded, and their labels normal or abnormal, we

can use those as inputs to train the MLP. The MLP contains three hidden layer sizes of 10, 10, and 5. Initial weights and biases are initialized random. Then the dataset is split into training and test data. The model trains through many epochs, and in each epoch all the samples will go through forward propagation in the layers of the MLP. The activations of the neurons are calculated using Eq. (3). The loss of the model is determined through cross entropy loss, and gradients are calculated by backpropagation for the weights and biases and updating them in many epochs until convergence.

3) Step 4.3 Prediction using the trained MLP: Finally, for each test sample, forward propagation is executed using the trained model to compute activations. The output activation is determined by Eq. (4):

$$MCAV(l) = f(W(l).MCAV(l-1)b(l) \qquad (4)$$

where, l indicates the output layer. A threshold of 0.5 is applied to classify the sample: outputs exceeding 0.5 are labeled as abnormal, while those below are labeled as normal.

This method uses the DCA's ability to generate MCAVs, and then it will be used as input to the MLP. The MLP is trained to classify the data into normal or abnormal categories, enhancing the overall classification performance by combining the strengths of both algorithms.

Phase 4: Classification
#Step 4.1: calculate MCAV values
1. For each antigen: // where antigen is the output of previous steps
2. Compute MCAV (antigen) using Equation 1
3. Append the computed MCAV (antigen) to the array MCAV_values [].

#Step 4.2: Feed MCAV Values into MLP for Classification
4. Train the MLP model as follow:
5. Input: MCAV_values []and their corresponding predefined labels (normal/abnormal).
6. Output: A trained MLP model with updated weights and biases.
7. Randomly initialize weights and biases for all layers based on the architecture (hidden layer sizes: 10, 10, 5).
8. Split the data into training and testing sets: X_{train}, X_{test}
9. Repeat the following steps until convergence:
10. For each epoch
11. For each training sample in X_{train}:
Perform forward propagation through the layers: Input Layer → Hidden Layers → Output Layer.
Compute activation for each neuron using equation (3)
Compute the loss between the predicted output and the actual label using across-entropy loss
Backpropagate the error using the chain rule to compute gradients of the loss with respect to the weight and biases.
Update weights and biases.
12 Store Final Weights and Biases:

#Step 4.3: Use trained MLP for prediction
13. For each test sample in X_{test}:
14. Perform forward propagation the trained model: Input Layer → Hidden Layers → Output Layer.
15. Compute activations for the output layer using
$$MCAV(l) = f(W(l).MCAV(l-1) + b(l))$$
16. Classify each test sample based on the output
If output > 05 → Assign ABNORMAL.
Else → Assign NORMAL.

Fig. 3. The pseudo-code for the proposed DCA_MLP algorithm.

The reason for the connection of the MLP to the DCA is that the MLP is a very strong learning procedure that can learn many complex, non-linear relationships from the data; without relying on any fixed decision rules. As a type of artificial neural network, the MLP classifier builds models through multiple interconnected layers of neurons, allowing it to capture intricate patterns between input features—such as the Mean Context Antigen Value (MCAV)and corresponding class labels (e.g., normal or anomalous).

During training, this MLP will repeatedly adjust its internal variables (weights and biases) to minimize overall classification error, in the hopes of maximizing performance across multiple types and dimensions of data. Unlike threshold methods, the dynamic nature of the learning process allows the model to learn and adjust to a number of data distributions with no need for human threshold adjustment. By combining MLP with the biologically inspired architecture of DCA—which is retained through its four-phase structure, the proposed MLP-DCA hybrid model enhances anomaly detection by preserving the contextual fusion strengths of DCA while improving classification robustness, generalization, and adaptability through machine learning.

IV. The Experimental Setup

As previously discussed, the standard Dendritic Cell Algorithm (DCA) [20] suffers from key limitations, particularly its reliance on manually defined, static thresholds for classification. This section details the experimental procedures undertaken to investigate these limitations and to validate the proposed enhancement of integrating a Multi-Layer Perceptron (MLP) into the DCA classification phase. The following hypotheses were formulated:

- H1: The performance of the standard DCA is significantly influenced by using static, predefined threshold values, particularly in datasets with varying class distributions.
- H2: Replacing the static threshold mechanism with an adaptive classifier, such as an MLP, improves classification performance by dynamically learning from the data.

To validate these hypotheses, we designed a two-part experimental framework. All experiments across both parts were conducted using a consistent set of fourteen benchmark datasets, listed in Table I, which span multiple domains including medical diagnosis, credit scoring, email classification, and network intrusion detection. These datasets were obtained from the UCI Machine Learning Repository[21], as well as widely used intrusion detection benchmarks—NSL-KDD [22] and UNSW-NB15[23].

1) Threshold sensitivity in the standard DCA model: To explore H1, we looked at the responsiveness of the standard DCA algorithm to different threshold values. For demonstrative purposes, the Wisconsin Breast Cancer (WBC) dataset was used under two distinct scenarios:

- Scenario A: Original dataset with 100% of the instances are anomalous.

- Scenario B: Modified version containing only 10% anomaly instances.

2) In each case, the DCA's decision threshold was manually tuned to maximize classification accuracy. The findings showed significant variability in performance depending on configuration of threshold which validated our assertion that the model was responsive and furthermore, not robust to changing class distributions. Evaluating the Proposed DCA-MLP Model.

To test H2**,** we implemented the DCA-MLP model by integrating the Multi-Layer Perceptron classifier into the DCA's classification stage. Unlike static thresholding, the MLP enables adaptive and data-driven decision-making by learning complex, nonlinear relationships in the data.

For a comprehensive evaluation, we applied the DCA-MLP model to all fourteen benchmark datasets mentioned in Table I. These datasets were selected to ensure diversity in feature dimensionality, sample size, and application domain. The model's performance was compared to the standard DCA using metrics including accuracy, sensitivity, and specificity, demonstrating the robustness and generalizability of the proposed approach across different types of data distributions and class imbalances.

To assess the performance of the proposed DCA-MLP model, the following evaluation metrics were used:

- Accuracy (ACC): Overall classification correctness.
- Sensitivity (SNS): The ability to correctly identify abnormal (positive) instances.
- Specificity (SPS): The ability to correctly identify normal (negative) instances.
- False Discovery Rate (FDR): The proportion of abnormal predictions that are normal.

TABLE I. DATASET DESCRIPTION

Dataset	Source	Feature	Record	Target class	Class count
Sonar (sonar)	UCI	60	208	2	Normal (111), Abnormal (97)
Wisconsin Breast Cancer (WBC)	UCI	9	699	2	Normal (458), Abnormal (241)
Wisconsin Diagnostic Breast Cancer (WDBC)	UCI	30	569	2	Malignant (212 cases) or benign (357 cases)
Pima Indians Diabetes (PID)	UCI	8	768	2	500 (non-diabetic), 268 (diabetic).
Indian Liver Patient Dataset (ILPD)	UCI	10	583	2	**416** liver disease and **167** without liver disease
Horse Colic(horse)	UCI	28	368	2	**Survival (**204), **non-survival (**164)
German-Credit (GC)	UCI	20	1000		**Good** 700, **Bad 300**
Red-Win-quality(win)	UCI	13	6497		**Low quality (63, High quality (217))**
Ionosphere (ionosphere)	UCI	35	351	2	Normal (225), Abnormal (126)
Statlog (Heart)	UCI	16	270		**(105)** normal, **(165) abnormal**
Spambase (SP)	UCI	57	4601		**Spam (**1,813**), not spam (**2,788**)**
BUPA liver disorder (LDR)	UCI	7	345	2	Liver disorder present (145), Not present (200)
UNSW-NB15 (UNSW-NB15)	IDS	42	2007	2	Normal (243), Abnormal (1764)
NSL-KDD (NSL_KDD)	IDS	41	3577	2	Normal (1577), Abnormal (2000)

High values of ACC, SNS, and SPS indicate good performance, while a lower FDR reflects fewer false alarms. These metrics were computed as follows [see Eq. (5) to Eq. (8)] [24]:

$$ACC = TP + TN/(TP + TN + FP + FN) \quad (5)$$

$$SNS = TP/(TP + FN) \quad (6)$$

$$SPS = TN/(TN + FP \quad (7)$$

$$\text{FDR} = FP/(TP + FP) \quad (8)$$

To ensure statistical reliability, each experiment was repeated 30 times, and mean values of each metric were reported. The statistical significance of performance improvements (Δ) between the standard DCA and the proposed DCA-MLP was evaluated using p-values from independent t-tests.

3) Parameter configuration and preprocessing: In all the experiments we conducted, the setup for the Dendritic Cell Algorithm (DCA) included a population of 100 dendritic cells (DCs), and each iteration selected 10 DCs to sample the antigen vector. The migration threshold for each DC was set to 10 iterations. Signal classification employed a feature-based approach- the smallest standard deviation represented the PAMP and safe signals, while the largest standard deviation represented the n-NS signal. standard deviation was selected during preprocessing to prioritize informative features.

For the Multi-Layer Perceptron (MLP) classifier, we applied standard training approaches; adaptive learning rate, early stopping to prevent overfitting, and K-fold cross validation in order to ensure robust model evaluation. These parameters were not only chosen based on best practices in literature but were also found to significantly influence detection performance. For example, increasing the amount of DCs can increase the number

of antigen samples, but the cost of the computation may be increased. Likewise, if the migration threshold is lowered, it will improve the speed of classification but may decrease decision-making accuracy in noisy data. Within the MLP, we found that adaptive learning rate and early stopping had the most significant impacts, effectively allowing the MLP to converge without overfitting. While the proposed model demonstrated stability over a range of parameter values, a sensitivity analysis revealed that signal selection and learning rate had the most effect on classification accuracy and generalization capability.

Comparative Analysis of Benchmark Literature Finally, we validated the proposed DCA-MLP model by comparing its performance to established benchmark literature methods, including GA-DCA, dDCA, NMF, standard DCA, and COID-DCA, on several datasets (Sonar, NSL-KDD, UNSW-NB15, HC, GC, LR). Key metrics (Accuracy, Sensitivity, Specificity) were compared to confirm the superiority of the DCA-MLP model. In addition, Receiver Operating Characteristic (ROC) curve analysis was conducted to further highlight the performance robustness of the proposed method.

V. Experimental Results

A. Threshold Sensitivity in the Standard DCA Model

The traditional Dendritic Cell Algorithm (DCA) relies on a predefined threshold during the classification phase. While this threshold allows for the differentiation between normal and anomalous instances, its effectiveness is dataset dependent. Particularly in imbalanced datasets, a static threshold leads to suboptimal performance and reduced sensitivity.

As shown in Table II, reducing anomaly proportion to 10% results in a performance drop of 5% in accuracy. This demonstrates the sensitivity of standard DCA to class distribution, which compromises its generalizability across.

B. Performance of the Proposed DCA-MLP Model

To overcome the above limitations, we propose an enhanced DCA model by integrating a Multi-Layer Perceptron (MLP) classifier, replacing the static threshold with dynamic learning. We evaluated the proposed DCA-MLP on fourteen datasets, comparing its performance with the standard DCA using Accuracy (ACC), Sensitivity (SNS), and Specificity (SPS), alongside statistical significance (p-values).

As shown in Table III, DCA-MLP consistently outperforms the standard DCA across all datasets, with statistically significant improvements ($p < 0.001$) in all key metrics.

C. Comparative Evaluation with Benchmark Literature

To further validate the proposed model, we compared it with recent state-of-the-art DCA extensions and hybrid methods, as shown in Table IV.

TABLE II. Effect of Threshold on Anomaly Detection Accuracy in DCA

Dataset	Total data		Threshold value	Result
WBC	Normal	Anomaly		Accuracy
Original (100% anomaly)	458	241	0.65	0.98
Reduced (10% anomaly)	241	40	0.26	0.93

TABLE III. The Results of Comprehensive Evaluation of Performance Metrics (ACC, SNS & SPS) with Statistical Significance (P-Values) for the Standard DCA and the Proposed DCA-MLP Across Datasets

Datasets	Accuracy (%)				Sensitivity (%)				Specificity (%)			
	DCA	DCA-MLP	Δ	pval	DCA	DCA-MLP	Δ	pval	DCA	DCA-MLP	Δ	pval
Sonar	0.88	0.99	0.11	2.26E-07	0.59	0.98	0.39	0.0004	0.78	0.98	0.2	0.000726
WBC	0.93	0.98	0.05	2.26E-07	0.90	0.98	0.08	0.0004	0.90	0.96	0.06	0.000726
WDBC	0.88	0.94	0.06	2.26E-07	0.39	0.86	0.47	0.0004	0.91	0.99	0.08	0.000726
PID	0.61	0.95	0.34	2.26E-07	0.39	0.87	0.48	0.0004	0.91	0.99	0.08	0.000726
ILPD	0.69	0.92	0.23	2.26E-07	0.57	0.77	0.20	0.0004	0.47	0.98	0.51	0.000726
horse	0.75	0.98	0.23	2.26E-07	0.95	0.99	0.04	0.0004	0.10	0.96	0.86	0.000726
GC	0.75	0.97	0.22	2.26E-07	0.95	0.99	0.04	0.0004	0.27	0.91	0.64	0.000726
win	0.70	0.91	0.21	2.26E-07	0.75	0.91	0.16	0.0004	0.42	0.91	0.49	0.000726
Ionosphere	0.75	0.95	0.20	2.26E-07	0.97	0.98	0.01	0.0004	0.88	0.9	0.02	0.000726
Heart	0.72	0.92	0.20	2.26E-07	0.25	0.85	0.60	0.0004	0.87	0.97	0.10	0.000726
SP	0.88	0.94	0.06	2.26E-07	0.84	0.90	0.15	0.0004	0.90	0.98	0.32	0.000726
LDR	0.67	0.97	0.30	2.26E-07	0.51	0.97	0.46	0.0004	0.79	0.95	0.16	0.000726
UNSW-NB15	0.86	0.98	0.12	2.26E-07	0.86	0.98	0.12	0.0004	0.60	0.99	0.39	0.000726
NSL-KDD	0.75	0.94	0.19	2.26E-07	0.82	0.84	0.02	0.0004	0.78		0.20	0.000726

TABLE IV. RESULT COMPARISON BETWEEN THE PROPOSED MODEL AND STATE-OF-THE-ART APPROACHES

Ref.	Method	Dataset	Results in percentage		
			Acc.	Sen.	Spe.
[5]	GA-DCA	Sonar	83.4	-	64.2
[7]	dDCA	NSL-KDD	93.29	-	88.93
		UNSW-NB15	97.25	-	95.01
[15]	DCA	UNSW-NB15	79.8	-	-
[25]	COID-DCA	HC	87.77	91.66	82.23
		GC	87.90	83.94	89.66
		LR	82.45	80.00	83.78
Proposed Model	MLP_DCA	Sonar	**0.99**	**0.98**	**0.98**
		NSL-KDD	**0.94**	**0.84**	**0.98**
		UNSW-NB15	**0.98**	**0.98**	**0.99**
		HC	**0.98**	**0.99**	**0.96**
		GC	**0.97**	**0.99**	**0.91**

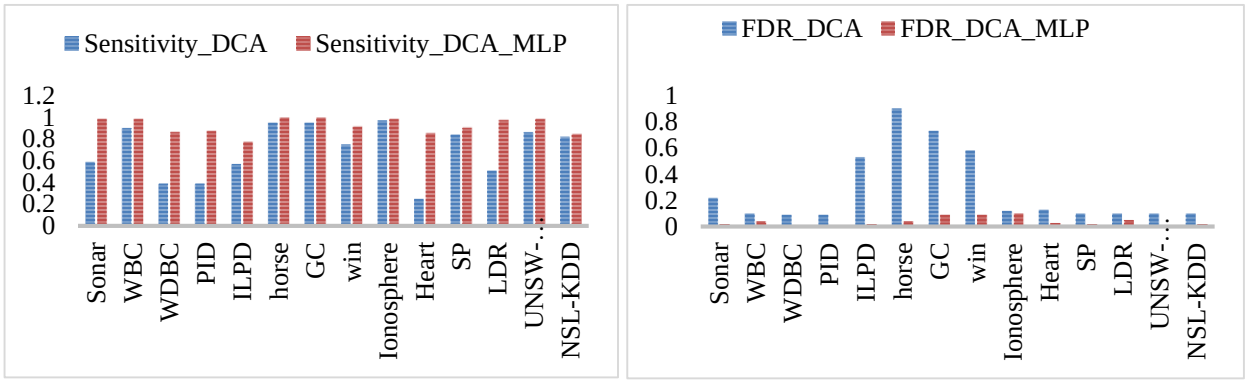
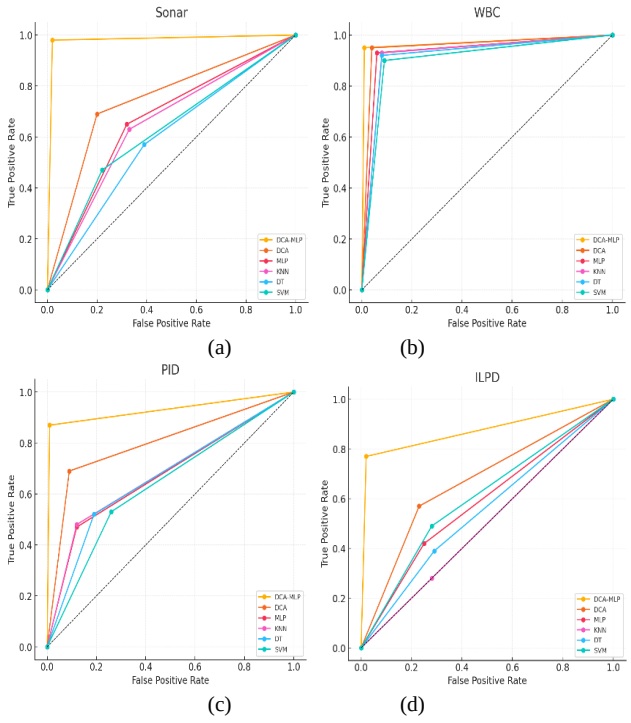


Fig. 4. The range between Sensitivity (SNS) and False Detection Rate (FDR) for DCA-MLP and DCA in benchmark datasets.

As illustrated in Fig. 4**,** the results are summarized in terms of SNS and FDR, which are critical performance metrics for anomaly detection models. The increased distance between the two metrics for the DCA-MLP model demonstrates its improved ability to clearly distinguish between normal and abnormal data groups. This larger gap indicates that the DCA-MLP not only increases the detection of actual anomalies but also appropriately limits false positives, leading to more accurate and dependable classification results. This capability is essential in high-stakes domains like cybersecurity, healthcare, or fraud detection, where accuracy and reducing false alarms are crucial.

Across all fourteen datasets, the proposed MLP-DCA model consistently achieved the highest Area Under the Curve (AUC) values when compared to other DCA variants and traditional classifiers, including stand-alone MLP, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree. As illustrated in Fig. 5, this robust and consistent performance highlights the superior generalization capability of the MLP-DCA model. Furthermore, the statistically significant improvement over both the standard DCA and conventional classifiers underscores the effectiveness of integrating adaptive learning through MLP into the biologically inspired DCA framework.



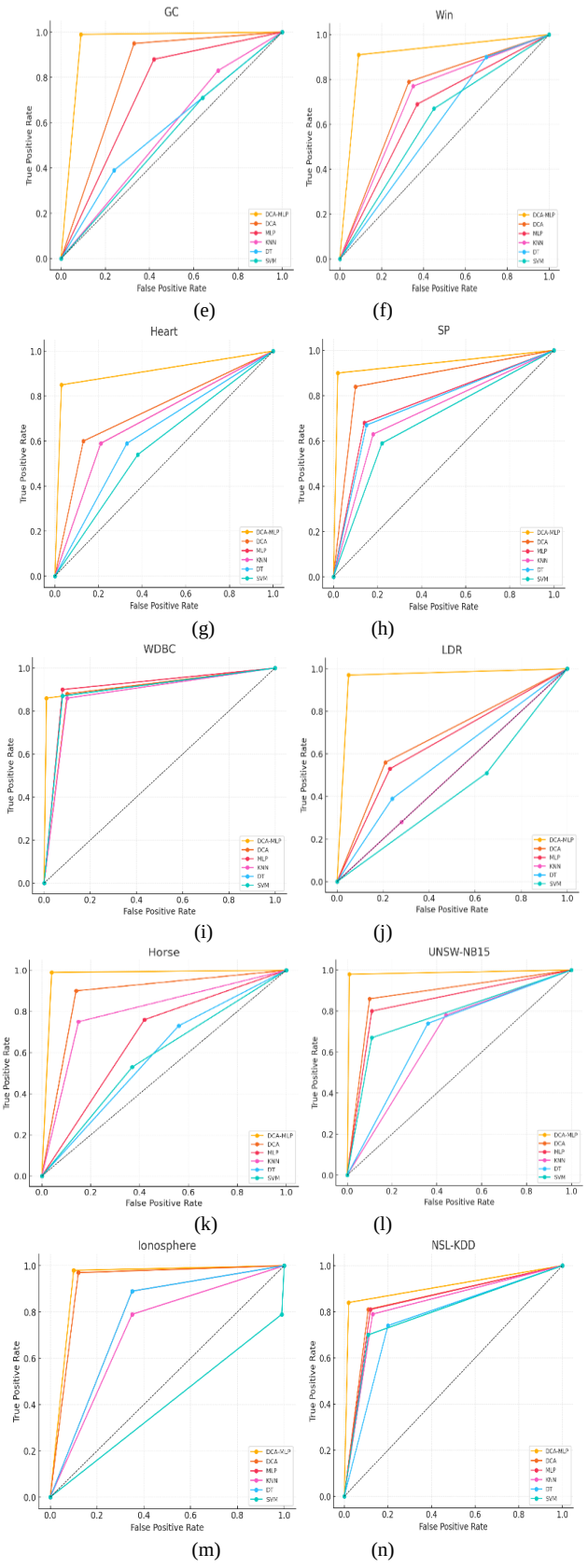


Fig. 5. Illustrates the comparative classification performance of the various DCA expansions and traditional classifiers in terms of ROC analysis across multiple datasets.

VI. Discussion

The experimental results provide ample evidence for the effectiveness and robustness of the proposed DCA-MLP model against the traditional Dendritic Cell Algorithm (DCA) model. One of the major limitations of a standard DCA is that it relies on a fixed threshold value during classification. As the results show, the fixed threshold is very sensitive to dataset specific distributions, especially with class imbalance (which is typically the case with anomaly detection tasks). The static nature of the threshold often results in decreased detection accuracy and increased false negatives when applied to datasets with evolving or uneven data distributions.

Integrating a Multi-Layer Perceptron (MLP) into the DCA framework, the model overcomes this threshold sensitivity and gains the ability to learn and generalize complex decision boundaries. The MLP enables dynamic classification by continuously adjusting to the patterns within the input data. This adaptability significantly improves classification performance across a wide variety of datasets. For example, substantial improvements were observed in both sensitivity and specificity on datasets like Sonar, PID, and NSL-KDD, which are known for their variability and complexity.

Another important aspect highlighted by the results is the consistent superiority of the DCA-MLP model across different domains and data types, including medical (WBC, WDBC), cybersecurity (NSL-KDD, UNSW-NB15), and signal processing (Sonar). This indicates that the proposed hybrid model is not only effective for a single category of data but is also generalizable and robust across diverse anomaly detection scenarios. The model's capacity to handle both structured and unstructured features contributes to its high adaptability and makes it suitable for real-world applications where data distributions are not always known in advance.

Moreover, the comparison with recent state-of-the-art approaches further supports the DCA-MLP model's effectiveness. When evaluated against various DCA variants—such as GA-DCA, COID-DCA, and dDCA—as well as conventional classifiers, the DCA-MLP consistently achieved higher accuracy and demonstrated better sensitivity-specificity trade-offs. These improvements are statistically significant, with p-values well below the 0.01 threshold, indicating that the observed performance gains are not due to random variation but rather a result of the model's enhanced learning capabilities.

The model also shows strong capability in minimizing false positives, a critical factor in anomaly detection systems. The analysis of sensitivity versus false detection rate (FDR), as shown in the experimental figures, illustrates a wider performance gap for the DCA-MLP model. This means it not only detects true anomalies more accurately but also avoids misclassifying normal instances as abnormal, thereby improving the reliability and precision of the system. Such performance is particularly crucial in high-stakes applications such as fraud detection, network intrusion prevention, and medical diagnosis, where incorrect predictions can lead to severe consequences.

VII. Conclusion

This research presents an enhancement of the Dendritic Cell Algorithm (DCA) aimed at improving anomaly detection

performance. The new approach utilizes a Multi-Layer Perceptron (MLP) embedded into the DCA framework, thereby eliminating the threshold-based classification reliant exclusively on the multi context antigen value (MCAV). This innovation solves the fundamental problems of fixed thresholds, which generally perform badly when measured with extreme values and bad from the perspective of typically improving the accuracy of anomaly detection. By leveraging the adaptive learning capabilities of the MLP, the enhanced MLP-DCA model demonstrates increased robustness and adaptability. Experimental results conducted on fourteen benchmark datasets validate that the MLP-DCA model significantly outperforms the conventional DCA, showing notable improvements in sensitivity, specificity, and overall accuracy.

These updates suggest that there is a good reason to believe that combining the strong principles of anomaly detection in the DCA with the extremely capable pattern detection and recognition advantages of the MLP. However, this study also has certain limitations**.** First, while the current model maintains consistent performance across many datasets, its effectiveness will vary when used on large-scale or highly imbalanced real-world data where rare anomalies are often prevalent. Second, the model's performance depends on hyperparameter configurations in both the DCA and MLP components, which may require manual tuning or additional optimization for each dataset. To add, the current model has not yet taken streaming or online data into account so there is no direct application of it towards real-time anomaly detection. As a result, the MLP-DCA model may have significant advantages for complex classification tasks, specifically for anomaly detection, which is characterized by non-linear data distributions or imbalances related to classes. Future research could explore integrating the Dendritic Cell Algorithm with advanced deep learning architectures or reinforcement learning-based decision modules. Hybrid approaches have the capacity to further improve performance on the class of complex datasets, which can be found in time-series anomaly detection, streaming data environments, and image-based pattern identification applications.

REFERENCES

[1] D. Wang, Y. Liang, H. Dong, C. Tan, Z. Xiao, and S. Liu, "Innate immune memory and its application to artificial immune systems," The Journal of Supercomputing, vol. 78, no. 9, pp. 11680-11701, 2022.

[2] J. Greensmith, "The dendritic cell algorithm," Citeseer, 2007.

[3] R. Pinto, G. Gonçalves, J. Delsing, and E. Tovar, "Incremental dendritic cell algorithm for intrusion detection in cyber-physical production systems," in Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 3, 2021: Springer, pp. 664-680.

[4] C. A. Winanto, P. Putra, D. P. Rini, O. Arsalan, S. K. Ningrum, and M. Q. Rizqie, "A Comparative Analysis of Snort and Dendritic Cell Algorithm in Intrusion Detection Systems," in 2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2024: IEEE, pp. 282-287.

[5] D. Zhang, Y. Zhang, and Y. Liang, "A genetic algorithm-based dendritic cell algorithm for input signal generation," Applied Intelligence, vol. 53, no. 22, pp. 27571-27588, 2023.

[6] M. F. M. Mohsin, A. A. Bakar, and A. R. Hamdan, "An Adaptive Anomaly Threshold in Artificial Dendrite Cell Algorithm."

[7] D. Limon-Cantu and V. Alarcon-Aquino, "Multiresolution dendritic cell algorithm for network anomaly detection," PeerJ Computer Science, vol. 7, p. e749, 2021.

[8] L. An et al., "Challenges, tasks, and opportunities in modeling agent-based complex systems," Ecological Modelling, vol. 457, p. 109685, 2021.

[9] B. Boudoua, M. Roche, M. Teisseire, and A. Tran, "EpiDCA: Adaptation and implementation of a danger theory algorithm for event-based epidemiological surveillance," Computers and Electronics in Agriculture, vol. 229, p. 109693, 2025.

[10] M. F. M. Mohsin, A. A. Bakar, A. R. Hamdan, and M. H. A. Wahab, "An improved artificial dendrite cell algorithm for abnormal signal detection," Journal of Information and Communication Technology, vol. 17, no. 1, pp. 33-54, 2018.

[11] M. Belhadj, F. Cherif, and M. Cheriet, "NMF-DCA: An efficient dendritic cell algorithm based on non-negative matrix factorization," International Journal of Computing and Digital Systems, vol. 10, pp. 575-583, 2021.

[12] Z. C. Dagdia, "A scalable and distributed dendritic cell algorithm for big data classification," Swarm and Evolutionary Computation, vol. 50, p. 100432, 2019.

[13] M. F. M. Mohsin, A. A. Bakar, A. R. Hamdan, M. Sahani, and Z. M. Ali, "Dengue Outbreak Detection Model Using Artificial Immune System: A Malaysian Case Study," Journal of Information and Communication Technology, vol. 22, no. 3, pp. 399-419, 2023.

[14] C. Pinto, R. Pinto, and G. Gonçalves, "Towards bio-inspired anomaly detection using the cursory dendritic cell algorithm," Algorithms, vol. 15, no. 1, p. 1, 2021.

[15] E. Farzadnia, H. Shirazi, and A. Nowroozi, "A new intrusion detection system using the improved dendritic cell algorithm," The Computer Journal, vol. 64, no. 8, pp. 1193-1214, 2021.

[16] K. Arora and S. Mahajan, "Detecting denial-of-service attack using dendritic cell approach," in Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI 2020, 2021: Springer, pp. 509-516.

[17] N. Elisa, J. Li, Z. Zuo, and L. Yang, "Dendritic cell algorithm with fuzzy inference system for input signal generation," in Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK, 2019: Springer, pp. 203-214.

[18] N. Elisa, L. Yang, F. Chao, N. Naik, and T. Boongoen, "A secure and privacy-preserving e-government framework using blockchain and artificial immunity," IEEE Access, vol. 11, pp. 8773-8789, 2023.

[19] M. F. M. Mohsin, A. A. Bakar, and A. R. Hamdan, "Integration of the dendritic cell algorithm with K-means clustering," Int. J. Adv. Soft Compu. Appl, vol. 6, no. 3, 2014.

[20] J. Deng, D. Wang, J. Gu, C. Chen, and C. Xie, "NK-DCHS: An Adaptive Hybrid Immune Model for Imbalanced Anomaly Detection," Expert Systems with Applications, p. 128704, 2025.

[21] A. Frank and A. Asuncion, "UCI machine learning repository, url= http://archive. ics. uci. edu/ml," Accessed: June, vol. 59, pp. 80-85, 2013.

[22] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE symposium on computational intelligence for security and defense applications, 2009: Ieee, pp. 1-6.

[23] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 military communications and information systems conference (MilCIS), 2015: IEEE, pp. 1-6.

[24] F. Gu, J. Greensmith, R. Oates, and U. Aickelin, "Pca 4 dca: The application of principal component analysis to the dendritic cell algorithm," arXiv preprint arXiv:1004.3460, 2010.

[25] Z. Chelly Dagdia and Z. Elouedi, "A hybrid fuzzy maintained classification method based on dendritic cells," Journal of Classification, vol. 37, no. 1, pp. 18-41, 2020.