6th International Conference on "Recent Advances in Computing and Communication" (ICRACC'26)

EDITORS

Dr.S.G.Packiavathy
Ms.S.Rajameenakshi
Ms.S.Sudha
Dr.S.Jayasri

RCHUB PUBLISHER
INTERNATIONAL BOOK & JOURNAL PUBLISHER

6th International Conference on "Recent Advances in Computing and Communication" (ICRACC'26)

School of Computational Sciences
Shri Shankarlal Sundarbai Shasun Jain College for Women
Chennai, India.

Editor

Dr.S.G.Packiavathy
Associate Professor and Head, Department of Computer Applications(Shift-I)
S.S.S.Shasun Jain College for Women, Chennai.

Ms.S.Rajameenakshi
Assistant Professor and Head, Department of Computer Applications(Shift-II)
S.S.S.Shasun Jain College for Women, Chennai.

Ms.S.Sudha
Assistant Professor, Department of Computer Applications(Shift-I)
S.S.S.Shasun Jain College for Women, Chennai

Dr.S.Jayasri
Assistant Professor, Department of Computer Applications(Shift-II)
S.S.S.Shasun Jain College for Women, Chennai

SEPTEMBER 2026

ISBN : 978-81-180146-5-2

978-81-180146-5-2

RCHUB PUBLISHER
INTERNATIONAL JOURNAL & BOOK PUBLISHER

Title : 6th International Conference on “Recent Advances in Computing and Communication” (ICRACC’26)

Editors: Dr.S.G.Packiavathy, Ms.S.Rajameenakshi, Ms.S.Sudha, Dr.S.Jayasri

Published By : RCHUB PUBLISHERS

Publisher’s Address : No.61, RCHUB Academy Building, Vengadamangalam, Chennai - 600 127

Publisher’s Details : www.rchubpublisher.com , +91 9888-231-231

Printer’s Details : RCHUB PRINTERS

First Edition : Septmber 2026

ISBN : 978-81-180146-5-2



© All Rights Reserved by the Editorial Board.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any names whether electronic, mechanical, photo copying, recording or otherwise without prior permission of the Editors/ Authors..

The information in this textbook is for educational purposes only. The authors and publisher have made every effort to ensure accuracy, but assume no responsibility for errors or omissions, or for any consequences arising from the use of the information contained herein. The publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon .

RCHUB 2026 Publication

TYPESET : RCHUB GROUPS, CHENNAI-127.

International Conference on

RECENT ADVANCES IN COMPUTING AND COMMUNICATION

ICRACC'26

On 23rd SEPTEMBER 2026

Organized by

School of Computational Sciences
Shri Shankarlal Sundarbai Shasun Jain College for Women
Chennai, India

INTERNATIONAL CONFERENCE

ON

RECENT ADVANCES IN COMPUTING AND COMMUNICATION

ICRACC'26

23rd September 2026

CONFERENCE PROCEEDINGS

IN ASSOCIATION WITH

RIVERS STATE UNIVERSITY, NIGERIA

Organized by

School of Computational Sciences

Shri Shankarlal Sundarbai Shasun Jain College for Women 3, Madley Road, T.Nagar,
Chennai - 600 017.
Telephone: 044 2432 8506/07
Website:
www.shasuncollege.edu.in
Email id: info@shasuncollege.edu.in

ABOUT OUR INSTITUTION

Greetings from Shri Shankarlal Sundarbai Shasun Jain College for Women! It is an institution established by the Minority Jain Community in 2005, with a long tradition of creativity, charisma and charm. Excellence and exuberance remain the core values of Shri Shankarlal Sundarbai Shasun Jain College for Women where the faculty and students strive to give their best to the curricular, co- curricular and extra-curricular activities year after year. One finds a balanced and healthy mix of cultural tradition and modernity in the ethos of Shasun.

ABOUT RIVERS STATE UNIVERSITY, NIGERIA

The Rivers State University (RSU) Port Harcourt was established in October, 1980 from the Rivers State College of Science and Technology which was itself established in 1972. It is located at Nkpolu-Oroworukwo in Port Harcourt, the capital of Rivers State, Nigeria.

Rivers State University is the first Technological University in Nigeria and the first state owned State University in the Niger Delta region of Nigeria. The motto of the University is “Excellence and Creativity”. The University has a staff strength of about 1,800 and a student population of over 28,000 . The language of instruction in the institution is English.

ABOUT ICRACC’26 CONFERENCE

The International Conference on Recent Advances in Computing and Communication (ICRACC’26) organized by the School of Computational Sciences in collaboration with Rivers State University, Nigeria aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of recent trends of Computer Science. It also provides a premier inter disciplinary platform for researchers, practitioners and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of Computer Science.

The main theme of the conference is "Recent Advances in Computing and Communication".

- Facilitate interactions with experts in their respective fields.
- Help overcome apprehensions about adjusting to a foreign environment.
- Provide a comprehensive experience of academic tourism.

EXPECTED OUTCOME OF THE CONFERENCE

1. Motivate young researchers

2. Publications in high end journals

MESSAGE FROM THE SECRETARY



Smt. Usha Abhaya Srisrimal

Secretary

I am really happy to know that the School of Computational Sciences is organizing an International Conference on "Recent Advances in Computing and Communication" on 23rd September 2026.

The conference aims to provide a vibrant platform for researchers, academicians, industry professionals, and students to share their innovative ideas, research findings, and practical experiences in emerging areas of technology and interdisciplinary research.

We are delighted to bring together experts and participants from diverse backgrounds to facilitate meaningful academic discussions, knowledge exchange, and collaborations. The conference will feature insightful keynote addresses, technical presentations, paper presentations, and interactive sessions covering contemporary developments and future trends in the field.

I hope that this conference would surely encourage innovative ideas among the participants paving way for new inventions in the computing field.

MESSAGE FROM THE ASSOCIATE SECRETARY



Dr. Harish L Metha

Associate Secretary

Welcome to ICRACC'26,

I am very glad that the School of Computational Sciences is organizing an International Conference on "Recent Advances in Computing and Communication" on 23rd September, 2026.

Research activities across all the computing fields pave the way for the industrial world to strive forward with huge advancements. As an educational institution, encouragement and support to research can be provided by establishing a suitable platform for the research community, to interact with each other and to share the knowledge. Having this objective in mind, ICRACC'26 has been organized to provide the same benefits and learning experience to all the participants. Sessions on different domains, keynote addresses from eminent professors and opportunity to network with the researchers will help the participants immensely in their research career.

With great pleasure and pride, I welcome all the participants and convey my best wishes for ICRACC'26.

MESSAGE FROM THE PRINCIPAL



Dr.S.Padmavathi

Principal

It gives me immense pleasure to extend my greetings to the School of Computational Studies on the occasion of the International Conference on Recent Advances in Computing and Communication, to be held on 23rd September 2026.

Computing and communication continue to evolve rapidly, influencing education, research, industry and society. In this context, the conference provides a platform for the Research Scholars, Academicians, Students and Industry Professionals to share and exchange ideas, and explore emerging technologies and innovative applications for the better society.

My sincere appreciation to the organizing committee, reviewers, contributors and participants for their valuable efforts and wish the conference every success.

MESSAGE FROM THE VICE PRINCIPAL



Dr.S.Rukmani

Vice Principal

I am happy to know that Shri Shankarlal Sundarbai Shasun Jain College for Women is organizing the International Conference on Recent Advances in Computing and Communication (ICRACC'26) on 23rd September 2026.

This technical International Conference will provide a prestigious international platform by bringing together local and overseas technical researchers and students to exchange their experienced knowledge and expertise in issues relating to the dominating trends in technology.

My best wishes for the ICRACC'26.

Session Chairs

Dr.D.Akila



Professor, Department of Cyber Security,
Faculty of Science and Humanities,
SRM Institute of Science and Technology,
Ramapuram Campus,
Chennai.

Dr.S.Revathi



Professor, Department of Cyber Security,
Faculty of Science and Humanities,
SRM Institute of Science and Technology,
Ramapuram Campus,
Chennai.

Dr.R.Sujatha



Assistant Professor,
Department of Computer Science,
PSG College of Arts & Science College,
Coimbatore.

Dr.P.Julia Grace



Associate Professor,
Department of Computer science,
JBAS College for women Autonomous, Chennai.

Dr.S.T Deepa



Associate Professor,Department of Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women,T.Nagar, Chennai

Dr.M.Anita Indu



Head & Assistant Professor, PG Department of Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women,T.Nagar, Chennai

ORGANIZING COMMITTEE

CHIEF PATRONS : **Smt. Usha Abhaya Srisrimal** - Secretary

Dr. Harish L Metha - Associate Secretary

CO-PATRONS : **Dr. S.Padmavathi** – Principal

Dr. S.Rukmani – Vice Principal

CONVENER : **Dr. S.G.Packiavathy,** Associate Professor & HOD
Department of Computer Applications (Shift I)

Ms.S.Rajameenakshi , Assistant Professor & Head
Department of Computer Applications (Shift-II)

CO-CONVENER : **Ms. S.Sudha,** Assistant Professor,
Department of Computer Applications

Dr. S.Jayasri,Assistant Professor,
Department of Computer Applications

ORGANIZING COMMITTEE :

Ms.V.Vidhya, Assistant Professor

Ms.T.Reshma Banu, Assistant Professor

Ms. S.Deepa, Assistant Professor

Ms. K.Vani, Assistant Professor

Ms.A.Jyothi, Assistant Professor

Dr.B.Chandra, Associate Professor

INTERNAL REVIEWERS:

Dr. S.T. Deepa, Associate Professor

Dr.B.Gomathi, Associate Professor,
Dr.S.Prasanna,Assistant Professor

Dr. S.G. Packiavathy, Associate Professor
Dr. S. Sridevi, Associate Professor

Dr.M.Anita Indu, Assistant Professor

Dr.S.Jayasri, Assistant Professor

Dr.B.Sreevidya,Assistant Professor

Dr.E.Kanimozhi, Assistant Professor

EXTERNAL REVIEWERS :

Dr. S. Gokila, Professor and Head
Department of Computer Applications,
Hindustan Institute of Science and Technology,
Chennai

Dr.Dhasny Lydia M
Assistant Professor Computer Science and Engineering, NMAM Institute of Technology,
NITTE (Deemed to be University).

Dr.Sheefa Ruby Grace
Associate Professor,
Department of Computer Science,
Sarah Tucker College, Tirunelveli.

Dr.R.Aparna
Assistant Professor,
Computer Science and Information Technology,
M.O.P. Vaishnav College for Women (Autonomous).

Dr.J.Rexy
Assistant Professor,
Department of Computer Science,
St.Xavier's College (Autonomous),
21 North High Ground road,
Palayamkottai-627 002

INDEX

S.NO.	PAPER ID	TITLE	PAGE NO.
1	ICRACC26_01	AI-GENERATED IMAGE DETECTION FOR DIGITAL FORENSICS USING TRANSFER LEARNING WITH MOBILENETV3-SMALL	1
	AUTHORS	DR. S. ARUNA	
2	ICRACC26_02	REAL-TIME HAND GESTURE RECOGNITION USING MEDIAPIPE HAND LANDMARKS AND RANDOM FOREST CLASSIFIER	7
	AUTHORS	N.DHARANISRI DR.R.GANAPATHYRAJA	
3	ICRACC26_03	AI BASED DISASTER DAMAGE ASSESSMENT	16
	AUTHORS	MS.M.K. SREE SWARNALATHA, MS. E. SANTHANAMARI, BHAVANA GAJENDRAN	
4	ICRACC26_04	BEYOND ON-SCREEN MARKING: A FRAMEWORK FOR TRUSTWORTHY AI-ASSISTED EVALUATION	22
	AUTHORS	MS.M.K. SREE SWARNALATHA, MS. E. SANTHANAMARI, THEANMOZHI	
5	ICRACC26_05	JNANASETU: AN OFFLINE-FIRST VILLAGE LEARNING HUB FOR COMMUNITY EDUCATION AND ASYNCHRONOUS KNOWLEDGE SHARING.	29
	AUTHORS	MS.M.K. SREE SWARNALATHA, MS. E. SANTHANAMARI, CHUNDURU LAASYA SRI	
6	ICRACC26_06	ENHANCED DUAL LIGHTWEIGHT CHANNEL-SPATIAL ATTENTION NETWORK FOR EFFICIENT MANGO LEAF DISEASE DETECTION	38
	AUTHORS	JEROPIN JELIN V, INDHU P, DR. J. REXY	
7	ICRACC26_07	SPECULAR-INVARIANT SPHERICAL SCANNING NETWORK FOR EFFICIENT MANGO FRUIT DISEASE DETECTION	47
	AUTHORS	S. RAHUMAL, R. PADMASHREE, DR.J.REXY	
8	ICRACC26_08	AI-BASED DYNAMIC TASK ALLOCATION AND FAULT-TOLERANT COORDINATION FOR SWARM ROBOTS IN DISASTER RESPONSE	58
	AUTHORS	DR. M.S. PADMAVATHI, SRILATHA SR, SARANYA M	
9	ICRACC26_09	AI-BASED DETECTION OF PROMPT INJECTION ATTACKS IN LARGE LANGUAGE MODELS: A ROBUSTNESS STUDY	64
	AUTHORS	DR. M.S. PADMAVATHI, MIRTHULA D, ANANYA M	

S.NO.	PAPER ID	TITLE	PAGE NO.
10	ICRACC26_10	RISK-ADAPTIVE AGENTIC AI FOR AUTONOMOUS THREAT DETECTION AND RESPONSE	70
	AUTHORS	JYOTHI A, SABANA YASMINE N, GOBHIGA.M	
11	ICRACC26_11	AN ADAPTIVE HYBRID FRAMEWORK FOR REAL-TIME SMILE RECOGNITION USING LIGHTWEIGHT DEEP FEATURES, TEMPORAL CUES AND ENSEMBLE LEARNING	77
	AUTHORS	K.RAVINANDAN, B. MALLIKESWARI	
12	ICRACC26_12	HANDWRITING BASED DYSLEXIA DETECTION USING MACHINE LEARNING	84
	AUTHORS	CHANIT JAYA AUXILIA A	
13	ICRACC26_13	BEYOND SENTIMENT: CONFIDENCE-AWARE MULTIMODAL AI FOR EMOTION RECOGNITION	87
	AUTHORS	FAHIMA FIRDHOUS S, S. MADHUMITHA	
14	ICRACC26_14	SAVDA: A SECURE AUTONOMOUS VEHICLE DEFENSE ALGORITHM FOR CYBER-PHYSICAL THREAT MITIGATION IN VANETS	93
	AUTHORS	B.K. RUCHITA, DR. R. RESHMA	
15	ICRACC26_15	AN OVERVIEW ON ASSESSMENT OF VARIOUS LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING SYSTEM	98
	AUTHORS	MS.V. MAHALAKSHMI, S.RAMYA, J.S.SHERLEE MONIYO	
16	ICRACC26_16	COMPARATIVE ANALYSIS OF ARTIFICIAL INTELLIGENCE APPROACHES: MACHINE LEARNING AND DEEP LEARNING	103
	AUTHORS	MS. K RADHA , MS.A.DIVYA	
17	ICRACC26_17	A STUDY ON FUTURE OF REUSABLE SPACECRAFT	108
	AUTHORS	*GAYATHRI S & BENAZIR BEGUM N*	

S.NO.	PAPER ID	TITLE	PAGE NO.
18	ICRACC26_18	ENHANCING NETWORK RELIABILITY IN 5G-ENABLED SMART INFRASTRUCTURE: A CYBERSECURITY PERSPECTIVE	112
	AUTHORS	DR. A. CATHREEN GRACIAMARY, BHAVANA PRIYA	
19	ICRACC26_19	PRIVACY-PRESERVING BIG DATA ANALYTICS: A REVIEW OF TECHNIQUES AND CHALLENGES IN THE AGE OF AI	117
	AUTHORS	*Dr.V.HEMA, NANDHINI*	
20	ICRACC26_20	SECURING THE EDGE: A SURVEY OF BLOCKCHAIN-BASED FRAMEWORKS FOR IOT DATA INTEGRITY	122
	AUTHORS	DR.S.BORGIA ANNIE CATHERINE , MS. P. KANISHKA	
21	ICRACC26_21	DEEP LEARNING-BASED COMPUTER VISION FOR REAL-TIME OBJECT DETECTION IN AUTONOMOUS SYSTEMS: A REVIEW	127
	AUTHORS	DR. UMA MAHESWARI K, MR. YOGESH H	
22	ICRACC26_22	A REVIEW OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): METHODS AND CHALLENGES	132
	AUTHORS	SONALI. J, HEMAKSHI S, LEELA K	
23	ICRACC26_23	MULTIPLE MOBILE SINK PATH SELECTION USING WEIGHTED RENDEZVOUS PLANNING	137
	AUTHORS	DR.S. G. PACKIAVATHY	
24	ICRACC26_24	DIGITAL WATERMARKING TECHNIQUES FOR IDENTIFYING AI-GENERATED IMAGES: A SYSTEMATIC REVIEW	145
	AUTHORS	DR.B.SREE VIDYA, MS.RAJANI SHREE SADHASIVAN, MS.NANDHIKA	
25	ICRACC26_25	KIDSHIELD – INTEGRATED CHILD VACCINE SURVEILLANCE AND MANAGEMENT PORTAL	153
	AUTHORS	DR. R SUJATHA , TAMIL NILA K	
26	ICRACC26_26	TN POSTAL SAVING SCHEME AI ASSISTANT: A RAG-BASED CONVERSATIONAL SYSTEM FOR INDIA POST SAVINGS SCHEMES	158
	AUTHORS	DR. SUJATHA, DHARSHINI T	

S.NO.	PAPER ID	TITLE	PAGE NO.
27	ICRACC26_27	SMART CIVIC ISSUE REPORTER: A LOCATION-BASED SYSTEM FOR DUPLICATE COMPLAINT GROUPING AND DYNAMIC PRIORITY ASSIGNMENT	163
	AUTHORS	MAHAALAKSHMI.MJ, JAI SHREE MAHA.K, REHNA.N	
28	ICRACC26_28	A REVIEW OF DEEP LEARNING TECHNIQUES FOR MALWARE DETECTION	172
	AUTHORS	MS.V.VIDHYA	
29	ICRACC26_29	TRUFACE: AN OPEN-SOURCE, REAL-TIME ACTIVE-PASSIVE LIVENESS ARCHITECTURE AGAINST LIVE STREAMING DEEPFAKE INJECTIONS	176
	AUTHORS	DURGA G, MANASA P, DR. S.T.DEEPA	
30	ICRACC26_30	AI DRIVEN SKINCARE RECOMMENDATION: A HYBRID CNN, RETRIEVAL AUGMENTED GENERATION, AND GUARDRAIL BASED ARCHITECTURE FOR SAFER, ALLERGY AWARE PERSONALIZATION	183
	AUTHORS	DR. S. PRASANNA, YAMINI. N, NAJMA SHANAAZ S	
31	ICRACC26_31	ARTIFICIAL INTELLIGENCE-DRIVEN APPROACHES TO COMBATING METHAMPHETAMINE TRAFFICKING IN THE DIGITAL ERA	189
	AUTHORS	YEGAMMAI T , MOHANA SHREE.S	
32	ICRACC26_32	SUSTAINABLE AGRICULTURE: REAL-TIME MONITORING AND PRECISION FARMING THROUGH AGROBOT	194
	AUTHORS	DR S. JAYASRI, J, BHAGYASHREE, T. HARIPRIYA	
33	ICRACC26_33	BEYOND OBJECT DETECTION: AN AI-POWERED RISK-AWARE NAVIGATION SYSTEM FOR VISUALLY IMPAIRED PEOPLE	204
	AUTHORS	MS.J.JEBA PRIYA, RASIKA J S, VIMALADEVI S	
34	ICRACC26_34	STUDENT ENGAGEMENT DETECTION USING FACIAL EMOTION RECOGNITION THROUGH DEEP LEARNING MODELS	209
	AUTHORS	A. VALLIYAMMAI, M.SHANMUGAPRIYA, S. YASHVITHA	

S.NO.	PAPER ID	TITLE	PAGE NO.
35	ICRACC26_35	A PROACTIVE SYSTEM FOR MEDICINAL PLANT DETECTION AND IDENTIFICATION USING YOLOV9 AND EXPLAINABLE ARTIFICIAL INTELLIGENCE	216
	AUTHORS	DR. B. GOMATHI	
36	ICRACC26_36	RISK-AWARE DETECTION AND PREVENTION OF SHADOW AI-DRIVEN DATA EXPOSURE USING A BROWSER EXTENSION	224
	AUTHORS	DR. B. GOMATHI, ROSHNI R, NITHYASHREE R	
37	ICRACC26_37	ANALYSIS OF MACHINE LEARNING – BASED PHISHING DETECTION	228
	AUTHORS	DR. APARNA R, SREE SANJANA R, SREEJA R	
38	ICRACC26_38	JAILBREAK ATTACKS ON LARGE LANGUAGE MODELS: AN OVERVIEW OF EMERGING THREATS AND CHALLENGES	239
	AUTHORS	PRIYAMVADHA K, DR.SOPHIA NAVIS MARY.D	
39	ICRACC26_39	BLOCKCHAIN BEYOND CRYPTOCURRENCY – APPLICATION IN HEALTHCARE	244
	AUTHORS	JEEVITHA.S, HARINI.N, VENKATESWARI.V	
40	ICRACC26_40	TRUTH CHECK: AN AI-POWERED NEWS VERIFICATION PLATFORM FOR FAKE NEWS DETECTION – A PROPOSED TECHNICAL FRAMEWORK	248
	AUTHORS	SOFIA D, DIYA A, S. RAJAMEENAKSHI	

S.NO.	PAPER ID	TITLE	PAGE NO.
41	ICRACC26_41	THE IMPACT OF ARTIFICIAL INTELLIGENCE ON PERSONALIZED LEARNING IN HIGHER EDUCATION	252
	AUTHORS	MS.S.SUDHA, M.VARSHA	
42	ICRACC26_42	AN INTELLIGENT STUDENT ASSESSMENT SYSTEM USING WORKFLOW AUTOMATION AND CLAUDE AI	258
	AUTHORS	DR. M. ANITA INDU, MS. V. HARINI	
43	ICRACC26_43	LLM-BASED CYBER THREAT INTELLIGENCE ANALYSIS FOR PERSONALISED CYBERSECURITY EDUCATION	261
	AUTHORS	DR. M. ANITA INDU, MS. JANANI. J	
44	ICRACC26_44	AN EXPLAINABLE MACHINE LEARNING FRAMEWORK FOR BREAST CANCER SURVIVAL PREDICTION: A COMPARATIVE STUDY	269
	AUTHORS	MS.S.SUDHA , D.PAVITHRA , N.SHANJANA	
45	ICRACC26_45	EMPOWERING COMMUNITIES TO OVERCOME SUBSTANCE ABUSE ACROSS CHENNAI – ADDRESSING TODAY'S SOCIAL EVIL	275
	AUTHORS	JULIA GRACE P, JANE LYDIA J & JANNET SALOMI D	
46	ICRACC26_46	INTELLIGENT HEALTHCARE AGENTS: USING AGENTIC AI FOR PERSONALIZED PATIENT RECOMMENDATIONS	279
	AUTHORS	Dr.BRINDA RAMANUJAM, ALIYA TABASUM I, ARUNA A.S	
47	ICRACC26_47	A COMPUTATIONAL STUDY OF PM2.5 ENVIRONMENTAL EFFECTS: 3D SURFACE ANALYSIS OF THE INTRA-COMPETITION COEFFICIENTS (B_1 AND B_2) INTERACTIONS WITH METEOROLOGICAL VARIABLES	287
	AUTHORS	R. E. AKPODEE	

6th International Conference on Recent Advances in Computing & Communication – ICRACC'26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_01

AI-GENERATED IMAGE DETECTION FOR DIGITAL FORENSICS USING TRANSFER LEARNING WITH MOBILENETV3-SMALL

Dr. S. Aruna,
Assistant Professor, Department of Computer Science,
Agurchand Manmull Jain College, Chennai – 600091, Tamil Nadu, India.
draruna913@gmail.com

Abstract

The rapid growth of generative artificial intelligence has made it difficult to distinguish real images from artificially generated images. AI-generated faces can be used in many forms of digital manipulation, creating challenges for digital forensics. This study presents a lightweight deep learning framework for detecting AI-generated facial images using MobileNetV3-Small with ImageNet transfer learning. A balanced subset of 500 images was selected from a publicly available Real vs Fake face dataset containing real faces from the Flickr Faces dataset and StyleGAN-generated fake faces. The dataset was divided into 350 training images, 74 validation images and 76 independent testing images. The proposed model was trained for 10 epochs and evaluated using accuracy, precision, recall, F1-score and Area Under the Curve (AUC). MobileNetV3-Small achieved 67.11% accuracy, 67.53% F1-score and 69.04% AUC on the independent test set.

Keywords: AI-generated images, StyleGAN, Deepfake image detection, Digital forensics, MobileNetV3-Small, Transfer learning.

1. INTRODUCTION

The rapid development of generative artificial intelligence has significantly improved to create realistic images. Modern generative models can produce facial images that are visually similar to authentic photographs, making the identification of synthetic content increasingly difficult. Among these technologies, Generative Adversarial Networks (GANs) have demonstrated the ability to generate highly realistic human faces.

The increasing availability of AI-generated images creates challenges in digital forensics, misinformation control and image authentication. Synthetic images may be used in impersonation, identity-related fraud, misinformation, copyright-related misuse and manipulation of digital evidence. Therefore, automated techniques for distinguishing authentic images from artificially generated images are important for maintaining trust in digital visual content.

Deep learning models have been widely used for image classification and manipulation detection. However, many advanced architectures require considerable computational resources. Lightweight convolutional neural networks provide an alternative for applications where computational efficiency, memory requirements and rapid inference are important. MobileNetV3-Small is a lightweight CNN architecture designed for resource-constrained environments. Transfer learning using ImageNet-pretrained models can further reduce the amount of training data and computational resources required for specialized image classification tasks.

This study investigates MobileNetV3-Small with ImageNet transfer learning for binary classification of real and StyleGAN-generated facial images. A balanced subset of a publicly available Real vs Fake face dataset is used in this study. The model is evaluated using an independent test set and both quantitative and qualitative analyses are performed. The rest of the paper is organized as follows. Section 2 refer the related work, Section 3 gives the description of the dataset and methodology used, Section 4 presents results and discussion and Section 5 concludes the study with future work.

2. RELATED WORK

Recent research on AI-generated image detection has investigated convolutional neural networks (CNNs), transfer learning, forensic image features, attention mechanisms and foundation models. These approaches can broadly be grouped into general AI-generated image detection, deepfake facial-image detection and lightweight detection methods. Ladevic et al., investigated a lightweight CNN for detecting AI-generated synthetic images and demonstrated the potential of computationally efficient CNN architectures for distinguishing synthetic content from authentic images. Li et al., proposed an image-adaptive prompt learning approach that

ISBN NO: 978-81-180146-5-2

dynamically adapts the detection process to the input image. Their work highlights an important challenge in synthetic-image detection, models trained on a limited range of generators may not generalize effectively to unknown generation techniques.

CNN-based methods have been widely investigated because convolutional architectures can learn spatial characteristics associated with manipulated facial images. Iqbal et al., investigated a deep learning approach incorporating data augmentation for deepfake image detection, demonstrating the usefulness of deep learning and augmentation strategies for improving detection capability. A broader systematic review by Heidari et al., examined deep learning-based deepfake detection methods and highlighted the extensive use of CNN-based approaches as well as the continuing challenges associated with dataset diversity, generalization and evolving manipulation techniques.

Lightweight CNN architectures are particularly relevant for practical AI-generated image detection because they are designed to reduce computational requirements while maintaining useful feature-learning capability. Howard et al., introduced MobileNetV3, including the MobileNetV3-Small architecture specifically targeted towards lower-resource use cases. The architecture was developed using hardware-aware neural architecture search and efficiency-oriented network design. Ladevic et al., demonstrated that a lightweight CNN can be used effectively for AI-generated image detection while maintaining a relatively simple architecture. These studies demonstrate the potential of lightweight CNNs for practical synthetic-image and facial-forgery detection. However, detection performance depends strongly on the dataset, image-generation method and training configuration. Therefore, results obtained from large benchmark datasets cannot necessarily be generalized to smaller datasets or different generation techniques.

Based on these observations, the present study investigates MobileNetV3-Small with ImageNet transfer learning for AI-generated facial image detection. The study focuses on the feasibility of a lightweight CNN for distinguishing real facial images from StyleGAN-generated facial images using a balanced dataset and an independent test set. In addition to classification performance, the study evaluates computational characteristics through parameter count, training time and inference time. This provides a practical perspective on the use of lightweight transfer learning for AI-generated image detection in digital forensic applications.

3. DATASET AND METHODOLOGY

The study uses a publicly available Real vs Fake face dataset. The dataset used for study has 70,000 real faces from the Flickr Faces dataset collected by NVIDIA, while the fake images were sampled from one million fake faces generated using StyleGAN. The original dataset was resized to 256 pixels and divided into training, validation and test sets. For this experiment, a balanced subset of 500 images was selected. Equal numbers of real and AI-generated images were used to avoid class imbalance. All images were resized to 224 × 224 pixels to match the input dimensions of MobileNetV3-Small. Data augmentation was applied only to the training set. The training dataset was augmented using- Rotation, Width shift, Height shift, Zoom and Horizontal flipping. Validation and test images were not augmented to ensure that the evaluation was performed on unchanged images. MobileNetV3-Small pretrained on ImageNet was used as the feature extraction. The pretrained layers were frozen during the initial experiment and a new classification head was added for binary classification.

Table 1. Dataset Distribution

Dataset Split	Real	AI-Generated	Total
Training	175	175	350
Validation	37	37	74
Independent Test	38	38	76
Total	250	250	500

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_01

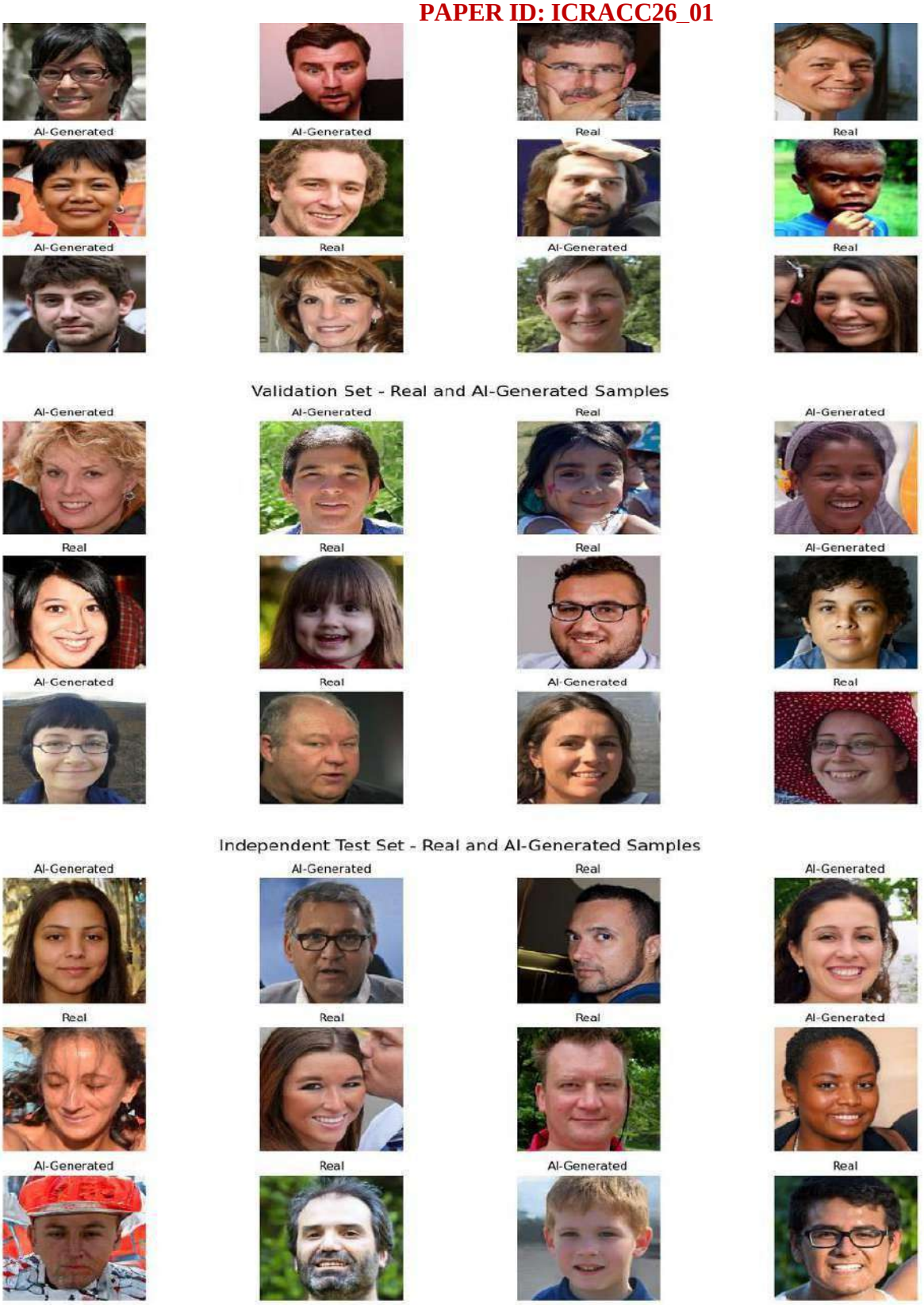


Figure 1. Representative real and StyleGAN-generated images from the training, validation and independent test sets.

4. RESULTS AND DISCUSSION

The model was trained for 10 epochs using the Adam optimizer with a learning rate of 0.0001. Binary cross-

ISBN NO: 978-81-180146-5-2

entropy was used as the loss function. The final sigmoid output was used for binary classification. A probability threshold of 0.5 was used to determine the predicted class. The independent test set contained 76 images and was not used during model training. The following evaluation measures were used: Accuracy, Precision, Recall, F1-score, AUC, Confusion matrix, ROC curve, Training time, Inference time, Parameter count.

Table 2. Performance of MobileNetV3-Small

Metric	MobileNetV3-Small
Accuracy	67.11%
Precision	66.67%
Recall	68.42%
F1-score	67.53%
AUC	69.04%
Parameters	976,113
Training Time	196.07 seconds
Inference Time for 76 Images	0.6489 seconds
Average Inference Time/Image	8.54 ms

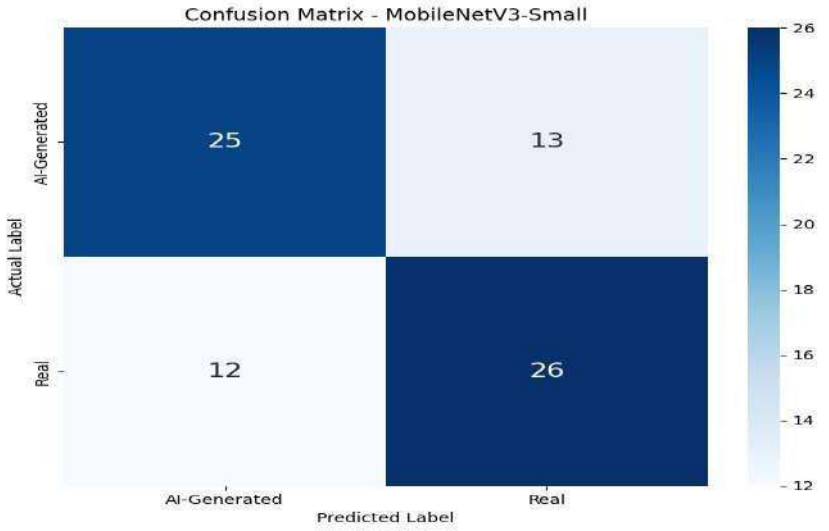


Figure 2. Confusion matrix for independent test-set classification.



Figure 3. ROC curve of MobileNetV3-Small on the independent test set.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_01

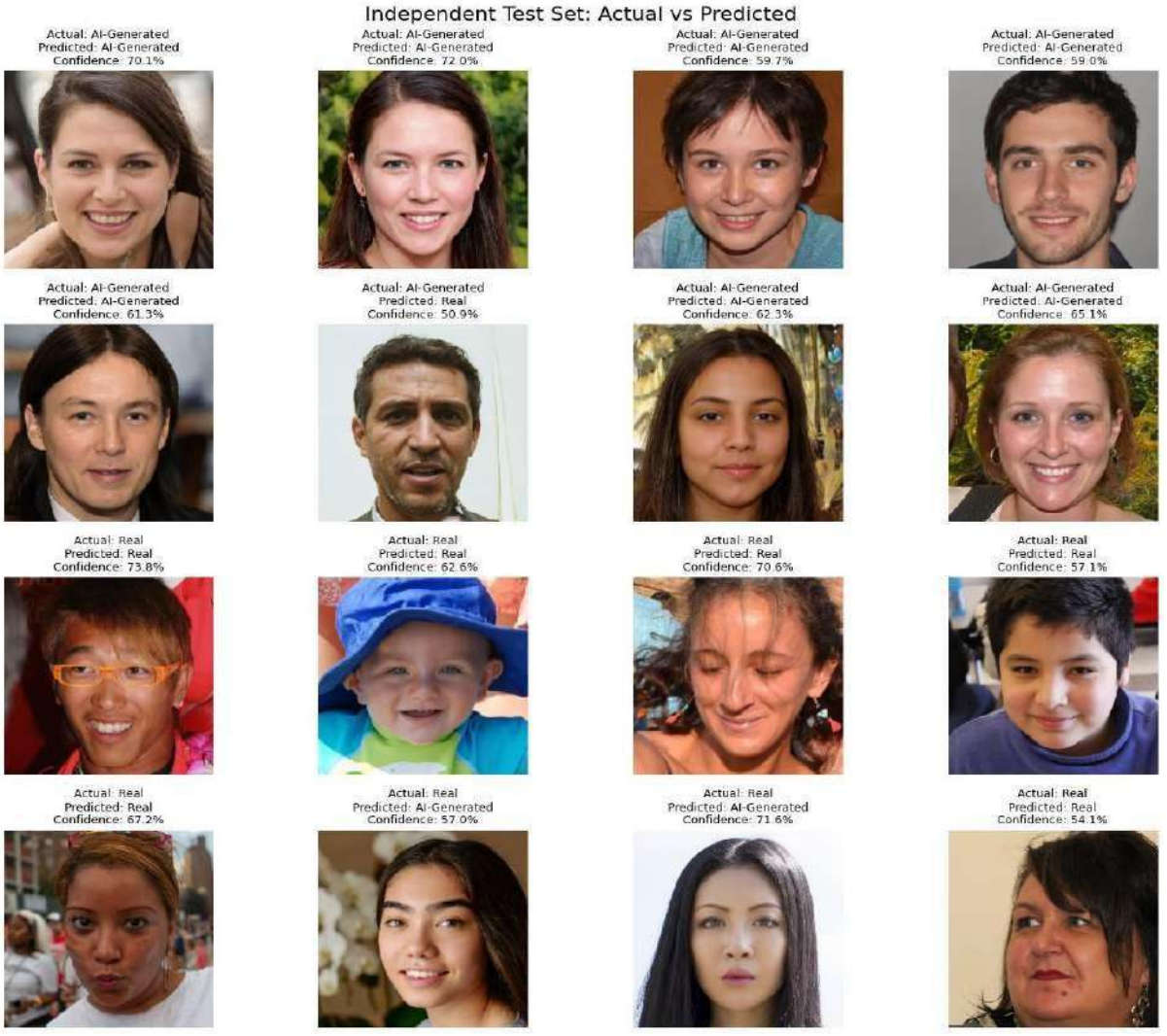


Figure 4. Qualitative analysis of independent test images showing actual class, predicted class and prediction confidence.

The obtained accuracy of 67.11% and AUC of 69.04% indicate moderate discriminative capability. The precision and recall values were relatively balanced. This indicates that the model did not strongly favour either the real or AI-generated class. The moderate classification performance also demonstrates that distinguishing high-quality StyleGAN-generated faces from real faces is a challenging problem. The use of only 350 training images limits the amount of visual variation available to the model. However, the computational characteristics of the model are promising. With only 976,113 parameters and an average inference time of approximately 8.54 ms per image, MobileNetV3-Small provides a lightweight alternative for further investigation in digital forensic applications.

5. CONCLUSION AND FUTURE WORK

This study presented a lightweight transfer learning approach for detecting AI-generated facial images using MobileNetV3-Small. A balanced subset of 500 real and StyleGAN-generated images was used with an independent test set consisting of 76 images. The model achieved 67.11% accuracy, 66.67% precision, 68.42% recall, 67.53% F1-score, 69.04% AUC. The model contained 976,113 parameters and required 196.07 seconds for training. After GPU warm-up, the average inference time was approximately 8.54 ms per image. Although the classification performance was moderate, the experiment demonstrates that a lightweight ImageNet-transfer-learning model can learn useful distinctions between real and StyleGAN-generated facial images. The computational efficiency of MobileNetV3-Small makes it a promising candidate for further investigation in resource-constrained digital forensic applications.

ISBN NO: 978-81-180146-5-2

Future work will focus on larger datasets, multiple image-generation techniques, fine-tuning strategies and cross-generator evaluation. Future work will investigate larger and more diverse datasets containing images generated using multiple GAN and diffusion-based models. Other lightweight CNN architectures can also be investigated for comparison. Fine-tuning of selected MobileNetV3-Small layers can be explored to improve classification performance. Frequency-domain features and image-forensic characteristics can also be combined with deep learning features. Future experiments should also evaluate the ability of the model to detect unseen generators, which is important for practical deepfake detection.

6. REFERENCES

1. A. L. Lađević, T. Kramberger, R. Kramberger, and D. Vlahek, "Detection of AI-Generated Synthetic Images with a Lightweight CNN," *AI*, vol. 5, no. 3, pp. 1575–1593, 2024, doi: 10.3390/ai5030076.
2. Y. Li, Z. Tan, Z. Lei, X. Zhou, and Y. Yang, "Towards Generalizable AI-Generated Image Detection via Image-Adaptive Prompt Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
3. F. Iqbal, A. Abbasi, A. R. Javed, A. S. Almadhor, Z. Jalil, S. Anwar, and I. Rida, "Data Augmentation-based Novel Deep Learning Method for Deepfaked Images Detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, 2024, doi: 10.1145/3592615.
4. A. Heidari, N. J. Navimipour, H. Dağ, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, 2024.
5. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019, doi: 10.1109/ICCV.2019.00140.
6. T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
7. T. Karras et al., "Analyzing and Improving the Image Quality of StyleGAN," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
8. J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
9. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces?resource=download

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02

Real-Time Hand Gesture Recognition Using MediaPipe Hand Landmarks and Random Forest Classifier

[1]N.Dharanisri [2]Dr.R.Ganapathyraja

[1] Research Scholar, Department of Computer and Information Science, Annamalai University.
sridharanicdm@gmail.com

[2] Assistant Professor, Department of Computer and Information Science, Annamalai University.
r.ganapathyraja@gmail.com

ABSTRACT

Hand detection and localization are essential components of modern computer vision systems and it plays a significant role in applications such as, human-computer interaction, health care monitoring, surveillance and virtual reality. Accurate hand detection in video sequences remains a challenging task due to variations in hand pose, scale, motion blur, occlusion and complex backgrounds. This paper aims to perform a real time hand gesture recognition by using MediaPipe technique, it uses 21 hand landmarks for hand detection and localization and used a Random Forest Classifier (RFC) to achieve an accuracy in hand gesture recognition. Using custom datasets to train a model for detecting hand gestures and recognizing gestures with corresponding class labels. Using Random Forest classifier for evaluate a model output with a real time prediction for each hand gestures. Such as, accuracy, precision, recall, F1 score, FPS (Frame Per Second).

Keywords: Hand Gesture Recognition (HGR), MediaPipe(MP), Random Forest Classifier (RFC), Hand Gesture (HG).

I.INTRODUCTION

Human-Computer Interaction (HCI) is a field, that focuses on designing and developing an interactive systems between humans and computers. To improve the communication, early used an input devices like keyboard, mice and then now a days, an intellectual technology evolves an interaction methods, including voice recognition, facial recognition, eye tracking and hand gesture recognition. These technologies provide to users to make an efficient and contactless way of digital communication.

The growing advanced technology of AI (Artificial Intelligence), CV (Computer Vision), ML (Machine Learning) and Gesture based interfaces has become increasingly in various real-world applications.

Unlike input devices, hand gestures improves the meaningful communication without requiring physical contact. For that, the gesture recognition system well suitable for touchless interaction environments like public kiosks, health care, industrial safety systems.

Hand gestures is one of the most expressive forms of non-verbal communication used for humans interaction. It is widely used in sign language for assisting hearing- and speech-impaired individuals, gaming, education, healthcare, surveillance etc.,

This paper aims to develop a real-time hand gesture recognition system that achieve more accurate result and easy to deploy on standard computing devices.

Many existing gesture recognition systems rely on deep learning models that require large datasets and expensive computational resources. To address these limitations, this research combines MediaPipe Hand Landmark Detection with a Random Forest classifier.

MediaPipe efficiently extracts 21 hand landmarks from webcam images, while the Random Forest algorithm performs gesture classification using the extracted landmark coordinates. This combination significantly reduces computational complexity while maintaining high recognition accuracy. The proposed approach demonstrates that lightweight machine learning techniques can achieve reliable real-time performance with less classification accuracy.

Contribution of the work:

- To develop a real-time hand gesture recognition system using MediaPipe Hands.
- To create a custom dataset containing multiple static hand gesture classes.

ISBN NO: 978-81-180146-5-2

- To extract and normalize 63 hand landmark features from each gesture sample.
- To train a Random Forest classifier for gesture classification.
- To evaluate the proposed model using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and FPS.

II. LITERATURE SURVEY

S.No	Title	Authors	Accuracy	Methods
1	A Hybrid Approach to Sign Language Recognition Using MediaPipe and Machine Learning	Vidhi Bansal, Sandali Sinha, R. Astya, Anil Kumar Sagar, K. Sahu	99.81%	MediaPipe hand-landmark extraction with Random Forest and CNN-based components for real-time ASL recognition; authors also support user-created datasets.
2	Real-Time Recognition of American Sign Language Using Mediapipe and Machine Learning Techniques	Viswanath Sarma Ch, Marsida Hima, B. Gogineni, Dr. B G Prasad, D. Kumar	Random Forest: 98.83%	MediaPipe extracts 21 hand points; Logistic Regression, k-NN, SVM, Decision Tree, and Random Forest are compared.
3	Video-based Hand Gesture Recognition using Random Forest for Sign Language Interpretation	J. Himasree, Jeevitha P L, D. K, Aashrita Kolisetty, Soumyalatha Naveen	97%; precision, recall, and F1 score reported as 99%	Video-based sign-language recognition using a custom dataset and Random Forest to model temporal gesture variation.
4	Real-Time Sign Language Recognition with MediaPipe Hand Landmarks and Random Forest Classification	Priya V Banu, Raja T Narasimma, A. Bharath, K. S. Sriram, S. Vinoth	93.00%	MediaPipe hand landmarks, preprocessing, data augmentation, and Random Forest classification for real-time sign recognition.
5	Quantitative Assessment of Hand Signal Recognition Using Landmarks Detection: A Comparative Study of Machine Learning Techniques	J. C. Chandra Sekhar, J. Ramu, V. Pratap	Random Forest: 92% average; SVM: 95% average	Hand-landmark extraction followed by comparison of SVM, Random Forest, k-NN, and MLP classifiers.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02

6	**Hand Gesture Recognition Using MediaPipe Landmarks and Deep Learning Networks**	**M. Gil-Martín, Marco Raoul Marini, Iván Martín-Fernández, Sergio Esteban Romero, Luigi Cinque**	**80.56–84.66%, depending on landmark representation and normalization**	**21 MediaPipe landmarks with normalization and temporal representations classified using recurrent neural networks.**

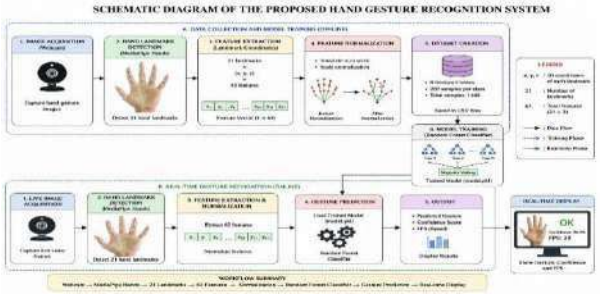
Table1: Literature Survey for Hand Gesture Recognition System

III. PROPOSED METHODOLOGY

The proposed hand gesture recognition system is designed to recognize static hand gestures in real time by using a webcam. This system integrates MediaPipe Hands for hand landmark detection and a Random Forest classifier for gesture classification.

Unlike image-based deep learning approaches that require large datasets and GPU resources, the proposed system extracts only the hand landmark coordinates, reducing computational complexity while maintaining high recognition accuracy.

The complete workflow consists of six major stages: image acquisition, hand landmark detection, feature extraction, feature normalization, model training, and real-time gesture prediction.



Initially, gesture samples are collected using a webcam, and then, MediaPipe detects 21 hand landmarks for each frame. These landmarks are normalized and stored as numerical feature vectors.

A Random Forest classifier is trained using the collected custom dataset. During real-time operation, the trained model predicts the hand gesture and displays the corresponding gesture class name, confidence score, and processing speed (FPS).

The proposed system workflow are below:

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02

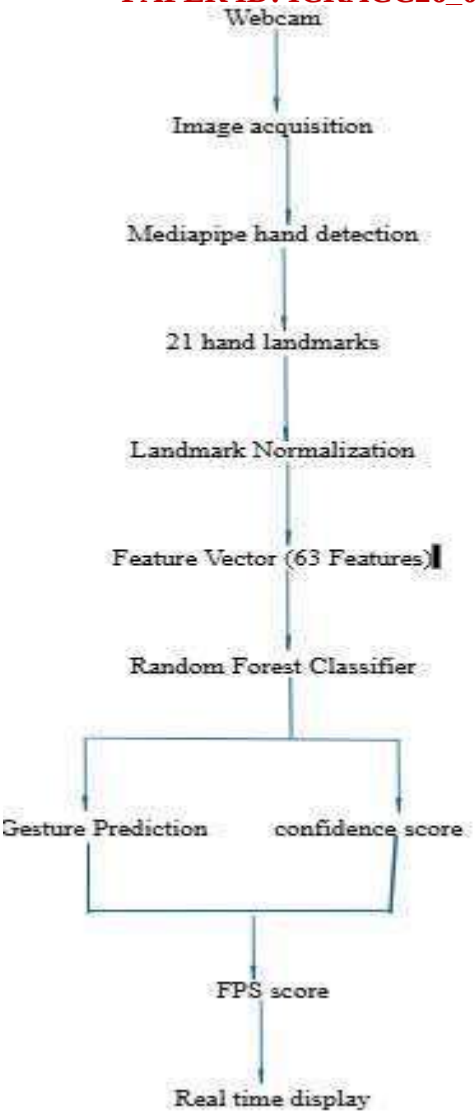


Figure 1: The workflow for the hand recognition system

3.1 Image Acquisition

The first stage involves capturing live video frames using a webcam. OpenCV is used to access the camera and continuously acquire image frames in real time. Each captured frame is horizontally flipped to provide a mirror view, making interaction more natural to the user. The captured images are serves as the input for hand landmark detection.

During the dataset creation phase, the user performs different predefined hand gestures in front of the webcam. For every detected gesture, the corresponding hand landmarks are extracted and stored in CSV format. Each gesture class contains 200 samples, resulting in a balanced dataset consisting of 1,000 gesture samples.

3.2 Hand Landmark Detection Using MediaPipe

MediaPipe Hands is employed for accurate and efficient hand landmark detection. The framework first detects the hand region and subsequently estimates 21 anatomical landmarks representing the joints and fingertips of the hand. Each landmark consists of three-dimensional coordinates (x, y, and z), describing the position of the hand in space.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02

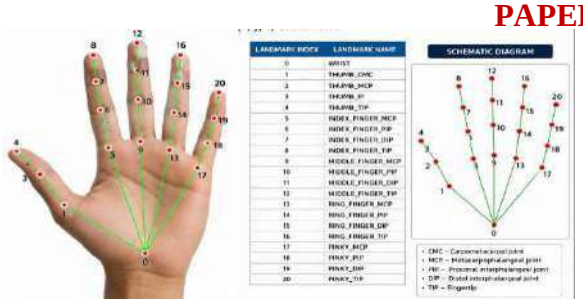


Figure 2: Schematic Diagram of the proposed MediaPipe and Random Forest based hand gesture recognition system.

Instead of processing the entire image, MediaPipe focuses only on the hand region, significantly reducing computational requirements. This enables real-time performance while maintaining robust landmark detection under different hand orientations and moderate lighting variations.

For every detected hand:

- Number of landmarks = 21
- Coordinates per landmark = 3 (x, y, z)
- Total extracted features = 63

These landmark coordinates represent the geometric structure of the hand and are used as input features for gesture classification.

3.3 Feature Extraction and Normalization

The landmark coordinates extracted from MediaPipe are converted into a numerical feature vector suitable for machine learning. Since the absolute landmark positions vary depending on the user's distance from the camera and hand position, normalization is performed to improve robustness.

Initially, the wrist landmark is selected as the reference point, and all remaining landmarks are translated relative to the wrist coordinates. Subsequently, the landmark values are scaled by the maximum absolute coordinate value to ensure that all features lie within a comparable numerical range.This normalization process minimizes variations caused by hand position, distance, and scale while preserving the relative geometric relationships between the fingers. Finally, the normalized landmark coordinates are flattened into a one-dimensional feature vector containing 63 numerical values.

IV. DATASET PREPARATION

A custom gesture dataset was created specifically for this research using the developed data collection application. five static hand gestures were selected for recognition:



Gesture: Love

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02



Gesture: No



Gesture: Ok



Gesture: Open palm



Gesture: Power

For , each gesture 200 samples were collected under different hand orientations and positions. The complete dataset contains 1,000 gesture samples, each represented by 63 normalized landmark features and one corresponding class label . The dataset was randomly divided into 80% training samples and 20% testing samples, resulting in 1,280 training instances and 320 testing instances.

4.1 Random Forest Classification

The extracted landmark features are classified using the Random Forest algorithm. Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting.

During training, multiple decision trees are constructed using randomly selected subsets of the training data and feature set. Each tree independently predicts the gesture class, Prediction from each decision tree, For an input landmark feature vector , each decision tree produces a class prediction: $Yb=Tb(X)$

where:

- X= input feature vector containing the extracted hand landmarks
- Tb= b^{th} decision tree
- Yb= prediction given by the b^{th} tree

and the final prediction is determined through majority voting among all the trees. The final gesture is selected by majority voting among all decision trees:

$y=mode(T1(X), T2(X) ,..... ,TB(X))$

where:

- B= total number of decision trees
- y= final predicted gesture class

ISBN NO: 978-81-180146-5-2

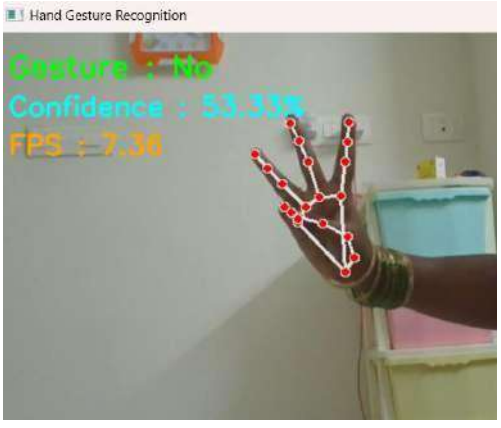
The Random Forest classifier was selected due to its simplicity, fast training speed, robustness to noisy data, and excellent performance on structured numerical datasets such as MediaPipe landmark coordinates. The trained model is stored using the Pickle(.pkl) library and loaded during real time gesture recognition.

4.2 Real-Time Gesture Recognition

During testing and deployment, the webcam continuously captures live video frames. Each frame is processed by MediaPipe to detect hand landmarks, which are normalized and converted into a feature vector.



Gesture: Love



Gesture: No



Gesture: Ok



Gesture: Open Palm



Gesture: Power

The trained Random Forest model predicts the corresponding gesture label and calculates the confidence

ISBN NO: 978-81-180146-5-2

score associated with the prediction.

The recognized gesture , confidence percentage and Frames Per Second (FPS) are displayed on the video stream, providing immediate feedback to the user. The entire pipeline operates in real time in a standard CPU without requiring dedicated GPU hardware.

V. PERFORMANCE METRICS

The proposed system was implemented using a Random Forest Classifier for hand gesture recognition . it was trained using a custom dataset in which each gestures sample was represented by 21 hand landmarks extracted using MediaPipe. These landmarks were normalized and converted into 63 numerical features for classification. The Random Forest classifier was evaluated using standard performance metrics, and the resulting classification performance is presented below.

	Recall	precision	Accuracy	F1score
Love	**1.00**	**1.00**	**100**	**1.00**
no	**1.00**	**0.98**	**100**	**0.99**
ok	**1.00**	**1.00**	**100**	**1.00**
openpalm	**0.97**	**1.00**	**97**	**0.99**
power	**1.00**	**1.00**	**100**	**1.00**

Table 2: Classification Report

The confusion matrix provides a way to understand how well the predictions have been done by the model.

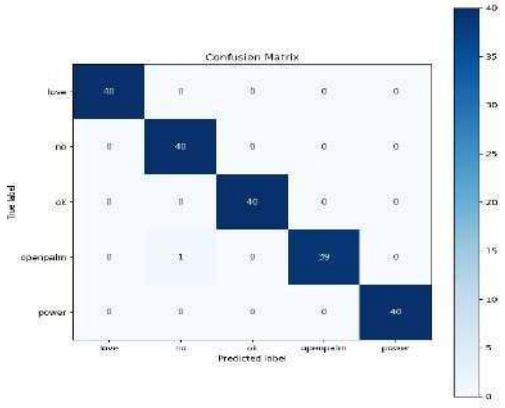


Figure 3: Confusion Matrix

It returns four types of values such as TP (true positive) which defines the correctly predicted values, TN (true negative) which defines the values which does not belong to the classes correctly, FP (false positive) that defines a value that is mis-predicted as a positive value and FN (false negative) which denotes a value mis-predicted as negative.

VI. CONCLUSION

The proposed system presents a real-time hand gesture recognition system using MediaPipe and Random Forest Classifier. MediaPipe was used to extract 21 hand landmarks, and generating 63 normalized features for each gesture sample. A custom dataset containing 1,000 samples across five gesture classes was developed for model training and evaluation. The Random Forest classifier achieved an overall accuracy of 99.50% on the test dataset. The result confirm that landmark-based features can effectively represent hand gestures for classical machine-learning classification. The proposed approach provides a light weight solution with low computational requirements compared with deep learning methods. The current system is limited to static, single-hand gestures. Future work will focus to improve recognition with dynamic gestures, accuracy, robustness and real-time performance for more complex gesture scenarios.

REFERENCES

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_02

1. V. Bansal, S. Sinha, R. Astya, A. K. Sagar and K. Sahu, "A Hybrid Approach to Sign Language Recognition Using MediaPipe and Machine Learning," 2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2025, pp. 1-6, doi: 0.1109/SCEECS64059.2025.1094032
2. J. Himasree, J. P L, D. K, A. Kolisetty and S. Naveen, "Video-based Hand Gesture Recognition using Random Forest for Sign Language Interpretation," 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-6, doi: 10.1109/APCIT62007.2024.1067351.
3. P. V. Banu, R. T. Narasimma, A. Bharath, K. S. Sriram and S. Vinoth, "Real-Time Sign Language Recognition with MediaPipe Hand Landmarks and Random Forest Classification," 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2025, pp. 1-8, doi: 10.1109/ICONSTEM65670.2025.11374879.
4. J. Nunes, T. H. Nascimento, J. Felix and F. Soares, "Real-Time Hand Gesture Recognition for Touchless Video Control Using MediaPipe and Random Forest," 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), Toronto, ON, Canada, 2025, pp. 1776-1781, doi: 10.1109/COMPSAC65507.2025.00242.
5. A. K. N. G, I. K. Biswas and C. V. R. Rao, "Real Time Sign Language Recognition Using MediaPipe and Random Forest Classifier," TENCON 2025 - 2025 IEEE Region 10 Conference (TENCON), Kota Kinabalu, Malaysia, 2025, pp. 1510-1513, doi: 10.1109/TENCON66050.2025.11375152.
6. V. K. Singh, P. Singh, S. K. Patel and V. Pareek, "Gesture Learn: A Real-Time CNN-Based Hand Gesture Recognition System Using MediaPipe," 2026 6th International Conference on Intelligent Technologies (CONIT), Belgavi, India, 2026, pp. 1-6, doi: 10.1109/CONIT69683.2026.11621426
7. N. J. Ibrahim, Y. H. Ali, I. A. F. Alzuabidi, A. A. Ahmed, B. S. Mohammed, N. A. Al-Dmour, and M. Abdeldayem, “Hand gesture recognition for human-computer interaction using deep learning approach,” *Procedia Computer Science*, vol. 275, pp.876–884,2026.
8. R. Sebbah and F. Z. Chelali, "Using Machine Learning and Deep Learning for Static and Dynamic Hand Gesture Recognition," 2024 International Conference on Advances in Electrical and Communication Technologies (ICAECOT), Setif, Algeria, 2024, pp. 1-6, doi: 10.1109/ICAECOT62402.2024.10828747.
9. A. Kaur and S. Bansal, "A Machine Learning Based Approach for Dynamic Hand Gesture Recognition in Human-Robot Interaction," 2024 12th International Conference on Intelligent Systems and Embedded Design (ISED), Rourkela, India, 2024, pp. 01-06, doi: 10.1109/ISED63599.2024.10956785
10. S. Gan, Y.Yin, Z.Jiang, H. Wen, L. Xie, and S.Lu, “SignGraph: A sign sequence is worth graphs of nodes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 13470–13479, doi: 10.1109/CVPR52733.2024.01279.

[6th] International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_03

AI Based Disaster Damage Assessment

M.K. SREE SWARNALATHA
ASSISTANT PROFESSOR, SSS
SHASUN JAIN COLLEGE
m.k.sreeswarnalatha@shasuncollege.edu.in

E. SANTHANAMARI
ASSISTANT PROFESSOR, SSS
SHASUN JAIN COLLEGE
santhanamari@shasuncollege.edu.in

BHAVANA GAJENDRAN
UG STUDENT, SSS
SHASUN JAIN COLLEGE
bhavanagajendran0@gmail.com

Abstract :
Natural disasters such as earthquakes, floods, cyclones, wildfires, landslides, and tsunamis can cause extensive damage to buildings, transportation networks, communication systems, utilities, agriculture, and public infrastructure. Conventional damage assessment depends heavily on field inspections, which are often slow, expensive, dangerous, and difficult to scale across large affected regions. Artificial Intelligence (AI), combined with Satellite Imagery, Unmanned Aerial Vehicles (UAVs), Synthetic Aperture Radar (SAR), Internet of Things (IoT) sensors, Social-Media Data, and Geographic Information Systems (GIS), provides a faster and more scalable approach.
This paper examines the architecture, applications, real-world deployments, challenges, and security requirements of AI-based disaster damage assessment. Computer vision models such as Convolutional Neural Networks (CNNs), U-Net, Vision Transformers, object-detection networks, image-segmentation models, and change-detection algorithms can identify damaged buildings, blocked roads, flooded areas, burned land, collapsed bridges, and destroyed utilities. Multimodal AI systems can further combine images, text, video, sensor data, and geospatial information to support emergency response. However, model bias, limited disaster datasets, poor image quality, uncertainty, privacy risks, adversarial attacks, and unreliable predictions remain significant concerns. A trustworthy system should therefore combine AI-based assessment with human verification, uncertainty quantification, secure data management, and explainable decision-making.

1. Introduction :

Disaster damage assessment is the systematic process of identifying, classifying, mapping, and estimating the physical, social, environmental, and economic effects of a hazardous event. It is essential for search and rescue, evacuation planning, emergency resource allocation, insurance processing, infrastructure restoration, and long-term recovery.Traditional assessment methods rely on expert teams visiting affected locations, inspecting buildings and infrastructure, recording observations, and preparing reports. Although field assessment provides valuable ground truth, it may be delayed by damaged roads, continuing hazards, communication failures, unsafe structures, severe weather, or a shortage of trained personnel. Large-scale manual surveys also create inconsistencies because different inspectors may apply different damage classifications.

AI addresses these limitations by automatically processing large quantities of data. Satellite and aerial images can be compared before and after a disaster to identify visible changes. UAVs can provide high-resolution images of individual buildings, roads, bridges, and industrial facilities. SAR imagery can support assessment when optical satellites are blocked by clouds or smoke. IoT sensors, emergency calls, social-media posts, and crowdsourced photographs can provide additional information about damage conditions.Recent research demonstrates that AI-based damage assessment is moving from experimental models toward operational decision-support systems. For example, the RAPID-A framework used an ensemble of deep segmentation models, high-resolution satellite imagery, pre-event and post-event imagery, and NASA ARIA damage-proxy maps to estimate earthquake damage across large urban regions in Türkiye. The system was designed to identify collapsed and visibly damaged buildings while representing uncertainty in the estimates.

1.1 Research Objective

The objectives of this paper are to:

ISBN NO: 978-81-180146-5-2

- Explain the architecture of AI-based disaster damage assessment.
- Identify major applications across different disaster types.
- Examine case studies and real-world deployments.
- Analyze technical, operational, ethical, and security limitations.
- Describe threat detection for both physical disaster hazards and AI-enabled assessment systems.

1.2 General Processing Architecture

A typical AI-based disaster damage assessment system consists of the following stages:

1. Data acquisition: Collect satellite images, UAV imagery, SAR data, sensor readings, field photographs, social-media content, maps, and building inventories.
2. Data preprocessing: Correct geometric distortions, remove noise, align pre-event and post-event images, normalize image intensity, and filter irrelevant information.
3. Feature extraction: Identify buildings, roads, vegetation, water bodies, debris, smoke, cracks, and other physical features.
4. Damage detection: Apply classification, object detection, semantic segmentation, instance segmentation, or change-detection models.
5. Damage classification: Assign categories such as undamaged, minor damage, moderate damage, severe damage, collapsed, or inaccessible.
6. Geospatial integration: Place detected damage on GIS maps and connect it with addresses, infrastructure databases, and population information.
7. Uncertainty estimation: Quantify model confidence and identify regions requiring human inspection.
8. Decision support: Generate priority maps for rescue, evacuation, temporary shelters, medical assistance, infrastructure repair, and resource distribution.

A simplified damage score can be represented as:

$$D_i = f(I_{ipre}, I_{ipost}, S_i, G_i, B_i)$$

where D_i is the estimated damage for location i , I_{ipre} and I_{ipost} are pre-event and post-event imagery, S_i represents sensor data, G_i represents geographic information, and B_i represents building or infrastructure attributes.

2. Keyword :

Artificial intelligence, disaster management, damage assessment, deep learning, remote sensing, satellite imagery, computer vision, threat detection, GIS, emergency response.

3. Key Applications :

3.1 Building Damage Assessment

Building damage assessment is one of the most common applications of AI in disaster management. Models can detect building footprints and classify damage based on collapsed roofs, missing structures, debris, facade changes, fire damage, or flooding.

A common workflow compares pre-disaster and post-disaster images. The model first identifies buildings and then measures changes in their shape, color, texture, shadow, roof condition, or surrounding debris. Semantic segmentation is particularly useful because it identifies the exact pixels associated with damaged regions rather than only placing a bounding box around them.

Possible damage classes include:

Class	Description
No damage	No visually detectable change
Minor damage	Small cracks, limited roof or facade damage
Moderate damage	Significant structural or exterior damage
Severe damage	Major loss of structural integrity
Collapsed	Building has partially or completely collapsed
Unclassified	Assessment is blocked by clouds, debris, inaccessible areas, or insufficient data

3.2 Road and Transportation Assessment

AI can identify blocked roads, damaged bridges, flooded highways, railway disruption, landslides, and debris

ISBN NO: 978-81-180146-5-2

accumulation. This information helps emergency managers determine which routes are usable for ambulances, fire services, rescue teams, and supply vehicles.

Object-detection and segmentation models can identify road obstructions in drone or satellite imagery. A road-priority model can combine damage severity, population density, hospital locations, and road connectivity to rank restoration tasks.

3.3 Flood and Water-Extent Mapping

Flood assessment systems use optical imagery, SAR, rainfall data, river-level sensors, and digital elevation models to identify inundated areas. SAR is particularly valuable because it can operate through clouds and during nighttime conditions.

AI can estimate:

- Flooded buildings.
- Floodwater extent.
- Roads isolated by water.
- Agricultural land affected by flooding.
- Areas requiring evacuation.
- Locations suitable for temporary shelters.

3.4 Wildfire and Smoke Damage Assessment

In wildfire scenarios, AI models analyze satellite imagery, thermal data, weather measurements, vegetation indices, and UAV footage. They can detect active fire fronts, burned areas, smoke movement, damaged structures, and regions at risk of reignition.

Thermal and multispectral data can reveal heat patterns that are not visible in normal RGB imagery. Combining these data sources improves situational awareness, although smoke, cloud cover, and rapidly changing fire behavior can reduce model accuracy.

3.5 Earthquake Damage Mapping

Earthquake assessment commonly uses:

- Pre-event and post-event satellite images.
- UAV photographs.
- SAR interferometry.
- Ground-motion maps.
- Building footprints.
- Structural inventories.
- Field inspection data.

The RAPID-A system applied six deep segmentation models based on CNN and transformer architectures to earthquake imagery. It incorporated change-detection channels and NASA ARIA damage-proxy maps, and tested the resulting framework across more than 100,000 building footprints in Gaziantep and Kahramanmaras. The study found that ensemble learning and auxiliary geospatial data could improve generalization and help locate damage hotspots.

3.6 Social-Media and Crowdsourced Damage Assessment

During disasters, citizens often upload photographs and text reports before official teams can access affected areas. AI can filter and classify these posts, remove duplicates, identify damage categories, and map the reports geographically.

The Rapid Damage Assessment system developed for disaster-related microblogs used images shared on Twitter to support real-time damage analysis. Such systems can provide early situational awareness, although social-media data may be incomplete, duplicated, misleading, biased toward highly connected communities, or intentionally manipulated.

3.7 Infrastructure and Utility Assessment

AI can assess damage to:

- Electricity transmission lines.
- Water pipelines.
- Cellular towers.
- Ports and airports.

ISBN NO: 978-81-180146-5-2

- Hospitals and schools.
- Oil and gas facilities.
- Dams and industrial sites.

UAVs equipped with RGB, thermal, or LiDAR sensors can inspect infrastructure at high resolution. AI-assisted inspection reduces the need to expose personnel to unstable structures, high-voltage equipment, fire, toxic substances, or floodwater.

4. Case Studies and Real-World Deployments:

4.1 2023 Türkiye Earthquakes

The 2023 Türkiye earthquake sequence provided a major test environment for AI-driven damage assessment. Researchers used high-resolution satellite imagery, building footprints, government damage records, image segmentation, change detection, and damage-proxy maps.

The RAPID-A framework classified collapsed and visibly damaged buildings and used a deep ensemble to reduce model instability. Its uncertainty-aware output was designed to support initial assessments before first responders and specialist teams were dispatched. The framework also tested generalization in cities different from the training area, an important requirement for real-world deployment.ascelibrary

This case demonstrates three important principles:

- AI should be evaluated across geographically different locations.
- Pre-event and post-event data are more informative than post-event images alone.
- Damage estimates should include uncertainty rather than presenting a single apparently precise number.

4.2 2018 Sulawesi Earthquake

The OECD examined the use of AI and satellite imagery for estimating disaster damage costs using the 2018 Sulawesi earthquake as an example. The methodology considered multiple sources, including crowdsourcing, IoT, UAV video, aerial imagery, and airborne and terrestrial LiDAR.oecd

The economic value of this approach is significant. Rapid estimates of damaged buildings, roads, public facilities, and commercial areas can support government recovery planning, financial assistance, insurance evaluation, and international aid prioritization. However, image-derived estimates must be calibrated against field surveys and local construction costs.

4.3 Field-Image Analysis Using LLMs

NEC developed a disaster assessment technology that combines large language models (LLMs), image analysis, map data, and cross-view image matching. The system allows users to search large collections of disaster photographs using textual descriptions and visual similarity, then display relevant images on a map with street-address-level location estimation.nec

This approach is useful when emergency organizations receive thousands of photographs from smartphones, vehicle cameras, CCTV systems, and field personnel. Instead of reviewing every image manually, responders can query for scenes such as “collapsed buildings,” “flooded roads,” or “blocked emergency routes.”

4.4 Disaster Management Decision Support

Research on AI applications in disaster management identifies uses across mitigation, preparedness, response, and recovery. However, many existing deployments remain concentrated in the response phase, where the demand for rapid information is most immediate.link.springer

A mature system should support the full disaster-management cycle:

- Mitigation: Identify high-risk buildings, flood zones, unstable slopes, and vulnerable infrastructure.
- Preparedness: Simulate disaster scenarios and estimate likely damage.
- Response: Map damage hotspots and prioritize rescue operations.
- Recovery: Monitor reconstruction, estimate economic losses, and verify infrastructure restoration.

5. Challenges, Limitations and Threats :

5.1 Data and Model Limitations

AI-based damage assessment depends on high-quality and representative training data. Major difficulties

ISBN NO: 978-81-180146-5-2

include:

- Limited labeled images for rare disasters.
- Unequal representation of countries, building styles, and construction materials.
- Differences in satellite resolution and sensor characteristics.
- Cloud cover, smoke, darkness, rain, and dust.
- Inaccurate or outdated building-footprint databases.
- Difficulty distinguishing pre-existing damage from disaster-induced damage.
- Severe class imbalance because most buildings in an image may be undamaged.
- Poor model transfer from one geographic region to another.

The Türkiye research shows that city-scale assessment remains difficult even with high-resolution imagery and extensive data. The study also notes that satellite, UAV, SAR, street-view, and three-dimensional point-cloud data each have different advantages and availability constraints.

5.2 Operational and Ethical Limitations

AI predictions should not automatically replace expert judgment in life-critical decisions. A false negative may cause rescuers to overlook a damaged building, while a false positive may divert limited rescue resources away from a more urgent location.

Other concerns include:

- Privacy risks from geolocated photographs and facial information.
- Unequal coverage of rural and low-income communities.
- Lack of transparency in black-box models.
- Misuse of sensitive infrastructure maps.
- Delays in obtaining commercial satellite imagery.
- Dependence on internet connectivity, cloud platforms, and electricity.
- Difficulty validating predictions during rapidly changing events.

Explainable AI methods can help show which image regions, sensor signals, or metadata influenced a prediction. Nevertheless, explanations should be treated as supporting evidence rather than proof that the prediction is correct.

5.3 Threat Detection in Physical Disaster Assessment

Threat detection refers first to identifying hazards and damage that may endanger people or infrastructure. AI can detect:

- Structural collapse and unstable buildings.
- Floodwater approaching populated areas.
- Wildfire fronts and smoke expansion.
- Landslides and road blockages.
- Dam or levee failure.
- Damaged electrical equipment.
- Chemical leaks and industrial fires.
- Unsafe bridges and transportation routes.

A threat-priority score can be defined as:

$$T_i = P_i \times I_i \times E_i$$

where P_i is the probability or confidence that a threat exists, I_i is its potential impact, and E_i is the exposure of people, infrastructure, or essential services.

This score can be combined with proximity to hospitals, population density, traffic conditions, and available rescue resources. The system should communicate confidence levels, for example: “high-confidence collapsed building,” “medium-confidence blocked road,” or “requires field verification.”

5.4 Cybersecurity Threats to AI Assessment Systems

Because damage-assessment platforms process sensitive geospatial and emergency information, they are also targets for cyberattacks. Important threats include:

- Data poisoning: Attackers insert incorrectly labeled or manipulated images into training datasets.
- Adversarial examples: Small visual alterations cause the model to misclassify damage.
- False geolocation: Manipulated metadata causes images to appear at an incorrect location.
- Model evasion: Attackers hide damaged structures or generate false damage indicators.

ISBN NO: 978-81-180146-5-2

- Data theft: Unauthorized users access private photographs, infrastructure maps, or victim information.
- Denial-of-service attacks: Emergency assessment platforms are made unavailable during a crisis.
- Prompt injection: Malicious text or images manipulate an LLM-based assessment workflow.
- Deepfake content: Fabricated images or videos create false reports of destruction.
- Model extraction: Attackers replicate a proprietary damage-classification model.

A secure deployment should use authenticated data sources, cryptographic integrity checks, role-based access control, network segmentation, audit logs, encrypted storage, secure APIs, and human approval for high-impact decisions. Models should also be tested against adversarial and out-of-distribution examples before operational use.

5.5 Recommended Evaluation Metrics

An AI damage-assessment model should be evaluated using more than overall accuracy. Recommended metrics include:

- Precision and recall for damaged-building detection.
- F1-score for class-balanced performance.
- Intersection over Union (IoU) for segmentation.
- Dice coefficient for damaged-area overlap.
- Mean absolute error for damage-cost estimation.
- Calibration error for confidence reliability.
- Inference time for emergency response.
- Geographic transfer performance across unseen regions.
- False-negative rate for high-severity damage.

Field validation should use independent inspection data. Evaluation datasets should also include different disaster types, image resolutions, seasons, building materials, and lighting conditions.

6. Conclusion :

AI-based disaster damage assessment provides a powerful method for converting large volumes of satellite, UAV, SAR, sensor, social-media, and field-image data into actionable emergency information. Computer vision can detect damaged buildings, flooded zones, burned areas, blocked roads, collapsed bridges, and infrastructure failures much faster than traditional manual surveys alone. Multimodal systems that combine imagery, maps, text, and sensor data can further improve location estimation, prioritization, and decision support.

Real-world studies involving the Sulawesi earthquake, the 2023 Türkiye earthquakes, social-media imagery, and LLM-supported field-image search demonstrate the practical potential of AI. However, the technology should be treated as a decision-support capability rather than an autonomous authority. Data scarcity, domain shift, sensor limitations, uncertainty, privacy concerns, cyberattacks, and false predictions can seriously affect emergency decisions.

Future systems should emphasize multimodal learning, uncertainty-aware prediction, explainable AI, continual learning, privacy-preserving data sharing, federated learning, adversarial robustness, and human-in-the-loop verification. The most reliable architecture will combine automated large-scale analysis with expert inspection, secure data governance, and transparent communication of confidence. Used in this manner, AI can improve the speed, safety, and fairness of disaster response while supporting long-term resilience and recovery.

References :

1. **Sun, W., Bocchini, P., & Davison, B. D. (2020). Applications of artificial intelligence for disaster** management. *Natural Hazards, 103*(3), 2631–2689. https://doi.org/10.1007/s11069-020-04124-3

2. Gupta, R., Goodman, B., Patel, N. N., Hosfelt, R., Sajeev, S., Heim, E. T., Doshi, J., Lucas, K., Choset, H., & Gaston, M. E. (2019). *xBD: A dataset for assessing building damage from satellite imagery*. arXiv. https://doi.org/10.48550/arXiv.1911.09296

3. Yang, C., Cervone, G., & Huang, J. (2019). Analysis of remote sensing imagery for disaster assessment using deep learning: A case study of flooding event. *Soft Computing, 23*(24), 13393–13408. https://doi.org/10.1007/s00500-019-03878-8

4. Xu, N., et al. (2026). A novel post-disaster damage assessment using generative AI: Ground-level damage inference from aerial data after the 2023 Turkey earthquakes. *International Journal of Remote Sensing*. https://doi.org/10.1080/01431161.2026.2628294

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_04

Beyond On-Screen Marking: A Framework for Trustworthy AI-Assisted Evaluation

M.K. SREE SWARNALATHA D
ASSISTANT PROFESSOR,
SSS SHASUN JAIN COLLEGE
m.k.sreeswarnalatha@shasuncollege.edu.in

E.SANTHANAMARI
ASSISTANT PROFESSOR,
SSS SHASUN JAIN COLLEGE
santhanamari@shasuncollege.edu.in

THEANMOZHI
UG STUDENT,
SSS SHASUN JAIN COLLEGE
theanmozhidillimohan@gmail.com

Abstract:

In 2026, the Central Board of Secondary Education (CBSE) introduced On-Screen Marking (OSM) for Class 12 answer scripts, a digitised evaluation workflow that fuelled public speculation about autonomous AI grading, a claim CBSE officially denied. The rollout was nonetheless accompanied by reported script-mismatch errors, a discontinued post-result verification facility, and revaluation-portal failures that eroded student trust. This paper examines the CBSE 2026 evaluation cycle as a case study to identify structural flaws common to large-scale, technology-mediated assessment models: opacity of the marking pipeline, absence of independent error-detection layers, weak grievance redressal, and unaddressed algorithmic-fairness risks documented in automated-essay-scoring literature. Building on this analysis, we propose a hybrid, human-in-the-loop, explainable AI-assisted evaluation framework that combines (Optical Character Recognition) OCR-based script-integrity verification, statistical anomaly detection, an auditable grievance-automation layer, and a scoring module that justifies every suggested mark question-by-question, and compare it against the existing model across reliability, transparency, and scalability metrics.

Keywords: Artificial intelligence, automated essay scoring, digital assessment, on-screen marking, optical character recognition, algorithmic fairness, explainable AI.

I. INTRODUCTION

Standardized board examinations conducted at national scale, such as those administered by the Central Board of Secondary Education (CBSE) for more than 4.6 million Class 10 and Class 12 candidates every year, rely on evaluation pipelines that must be simultaneously fast, consistent, and defensible. In February 2026, CBSE announced a shift to On-Screen Marking (OSM) for Class 12 answer scripts: physical booklets are scanned, digitised, and routed to teacher-evaluators through a centralised server instead of being circulated as paper copies [1], [2].

Coverage of the rollout on social media described the change as "AI checking answer sheets," prompting anxiety among students and parents. CBSE issued repeated clarifications that evaluation remained teacher-led and that technology supported scanning, allocation and monitoring rather than scoring [3], [4]. Despite these clarifications, the 2026 cycle surfaced a cluster of operational failures. Students who obtained photocopies of their evaluated scripts reported that the handwriting and content did not match their own, indicating script-identity mismatches somewhere in the scan-to-evaluator pipeline [5]. CBSE simultaneously discontinued the long-standing post-result marks-verification facility for Class 12, narrowing the remedies available to an affected student to the heavier revaluation route [2], and the revaluation portal itself experienced access failures during the peak application window.

This episode is instructive beyond a single board or a single year. It sits at the intersection of two trends reshaping assessment worldwide: the digitisation of paper-based evaluation, and the parallel, largely separate advance of AI/NLP-based automated scoring, whose accuracy and fairness properties remain the

ISBN NO: 978-81-180146-5-2

subject of an active and unsettled research literature. The public confusion between "digitised" and "AI-graded" evaluation is itself diagnostic-it shows an assessment system that could not clearly communicate what technology was, and was not, doing inside the scoring pipeline.

This paper treats the CBSE 2026 case as a lens through which to (i) catalogue structural flaws present in large-scale evaluation models, whether manual, OMR-based, or partially digitised, and (ii) propose enhancements-grounded in the automated-essay-scoring and educational-measurement literature-for a hybrid model in which AI functions as a transparent, auditable aid to human examiners rather than an opaque process or an unaccountable rumor. Section II reviews existing assessment models and prior work on AI-assisted scoring. Section III presents the CBSE 2026 case study. Section IV synthesises generalisable flaws. Section V proposes an enhanced hybrid framework. Section VI compares the proposed and existing models, Section VII discusses implementation and evaluation metrics, and Section VIII concludes.

II. BACKGROUND AND RELATED ASSESSMENT MODELS

A. Manual Evaluation

The traditional model-a single teacher-examiner marking a physical booklet against a model answer key-remains the backbone of most board examinations. Its weaknesses are well documented: inter-rater inconsistency under fatigue and time pressure, arithmetic totalling errors, and the logistical cost of transporting, storing, and auditing millions of scripts.

B. OMR and Rule-Based Automation

For objective items, Optical Mark Recognition (OMR) automates scoring of multiple-choice responses with high reliability, but offers no mechanism for constructed-response or descriptive-answer evaluation, which constitutes the majority of marks in examinations such as CBSE Class 12. Other Indian state boards have reported comparable issues under fully manual evaluation-for instance, an inquiry into a state board's Class 12 English results traced discrepancies to an incorrectly compiled answer key rather than to any individual examiner's marking-indicating that answer-key and process errors are not unique to digitised pipelines and that any comparison of models must weigh digitisation-specific risks against the manual-process risks it replaces.

C. On-Screen / Digital Marking

On-screen marking (OSM) systems scan physical scripts and route digitised images to examiners through a server, retaining human judgment while automating totalling, allocation, and audit-trail logging-the model CBSE adopted in 2026 [1], [2]. OSM removes arithmetic-totalling errors and enables remote, monitored evaluation, but introduces new failure surfaces-scan quality, script-identity binding, server security, and evaluator workload distribution-that a purely paper-based system does not have.

D. Automated and AI-Assisted Scoring

A parallel body of work in Automated Essay Scoring (AES) and, more recently, large-language-model (LLM)-based scoring, aims to score free-text constructed responses directly. Such systems can approximate human ratings on well-defined rubrics, but a substantial and growing literature documents fairness risks: bias linked to training-data composition and demographic subgroup, degraded reliability on writing styles outside the training distribution, and limited interpretability of neural scoring models. Comparative studies show that different scoring architectures exhibit different, sometimes opposite, biases with respect to gender, socio-economic status, and language background, and that hybrid human–machine models tend to be fairer than either humans or fully automated systems alone. Recent work on explainable, fairness-aware verification frameworks proposes using AI not to assign the primary score, but to flag scores that deviate from peer-consistent, rubric-aligned expectations for a second human check-a design principle this paper adopts and extends.

Together, these strands reveal a gap: OSM-style systems solve logistics but not fairness or error

ISBN NO: 978-81-180146-5-2

detection, while AES/LLM systems address scoring directly but carry unresolved fairness and trust costs, and are-correctly, per CBSE's own clarification-not yet deployed for high-stakes constructed-response marking in this context.

E. Data Security and Privacy in Digitised Evaluation

Digitisation also changes the risk profile of the examination itself. Once a script exists as an image on a centralised server, its confidentiality, integrity, and availability become dependent on conventional information-security controls — encryption in transit and at rest, access logging, and breach detection — in addition to the physical custody controls that governed paper scripts. Any framework that adds AI components to this pipeline, as proposed in Section V, must treat script images and evaluator annotations as sensitive personal data subject to minimisation and retention limits, not merely as training material for future models.

III. CASE STUDY: THE CBSE 2026 EVALUATION CYCLE

CBSE announced OSM for Class 12 scripts on 10 February 2026, ahead of examinations beginning 17 February, and clarified on 13 February that evaluation would remain teacher-led despite social-media claims of AI grading [3], [4].

Following the May 2026 results, students who requested photocopies of their answer scripts reported that the handwriting and answers did not match their own-CBSE acknowledged the complaint and committed to sending the correct script and revising results for affected candidates [5]. In the same window, the board discontinued the post-result verification-of-marks facility for Class 12 and reduced the revaluation fee, concentrating all dispute resolution into a single, heavier-weight process [2]; the revaluation portal subsequently experienced login and access failures during the peak application period. Table I summarises the reported issues against the pipeline stage they implicate.

TABLE I. REPORTED ISSUES IN THE CBSE 2026 EVALUATION CYCLE

Issue	Stage	Root cause
Script/handwriting mismatch	**Scan & tagging**	**Unverified scan-to-record binding**
Verification tier discontinued	**Post-result grievance**	**Low-cost recourse removed**
Revaluation portal failures	**Grievance application**	**Demand not modelled**
Public AI-grading rumor	**Communication**	**Automation boundary unclear**

Two structural causes emerge. First, identity binding between the physical script and its digital record depends on a single scan-time tag with no independent downstream cross-check, so one mis-scan can silently swap two students' evaluation streams. Second, removing the lightweight verification tier removed a low-cost first line of recourse precisely when the volume of concerned students was highest, funnelling all disputes into a slower, costlier channel that was not provisioned for the resulting load.

A further, easily overlooked cause is that identity-binding failures of this kind are self-concealing: unlike a totalling error, a swapped script produces a plausible-looking mark for the wrong student and an equally plausible-looking mark for the student who received the swapped script, so neither record looks anomalous in isolation. Detecting the error requires either the affected student noticing unfamiliar handwriting after requesting a photocopy — a reactive, opt-in check available only to those who question a low score — or a systematic, independent verification step built into the pipeline itself, which the 2026 OSM design did not include.

ISBN NO: 978-81-180146-5-2

IV. FLAWS IN EXISTING ASSESSMENT MODELS

Generalising beyond the single cycle, five recurring flaw categories can be identified across manual, OMR, and OSM-style models.

1) Opacity: systems rarely state precisely what is automated and what is human-judged, which breeds rumor and erodes trust even when the underlying process is sound [3], [4].

2) Single point of failure in identity binding: physical-to-digital script identity is typically established once, at intake, with no redundant verification later in the pipeline [5].

3) Weak or removed independent error-detection and grievance layers: a cheap, fast verification tier is a different function from a full re-evaluation, and collapsing the two concentrates load and risk [2].

4) Unaddressed algorithmic-fairness and explainability risk for any future AI-scoring component: the AES/LLM literature consistently reports subgroup bias and limited interpretability, which would undermine a student's right to a reasoned recheck if such tools were adopted for direct scoring without safeguards

5) Absence of published quality metrics: boards rarely publish inter-rater reliability, mismatch rates, or audit statistics, limiting external accountability and academic scrutiny of evaluation quality over time.

V. PROPOSED HYBRID AI-ASSISTED EVALUATION FRAMEWORK

We propose a five-layer framework (Fig. 1) in which AI functions strictly as a transparent aid to identity assurance, quality screening, and rubric-aligned suggestion — never as an unsupervised scorer.

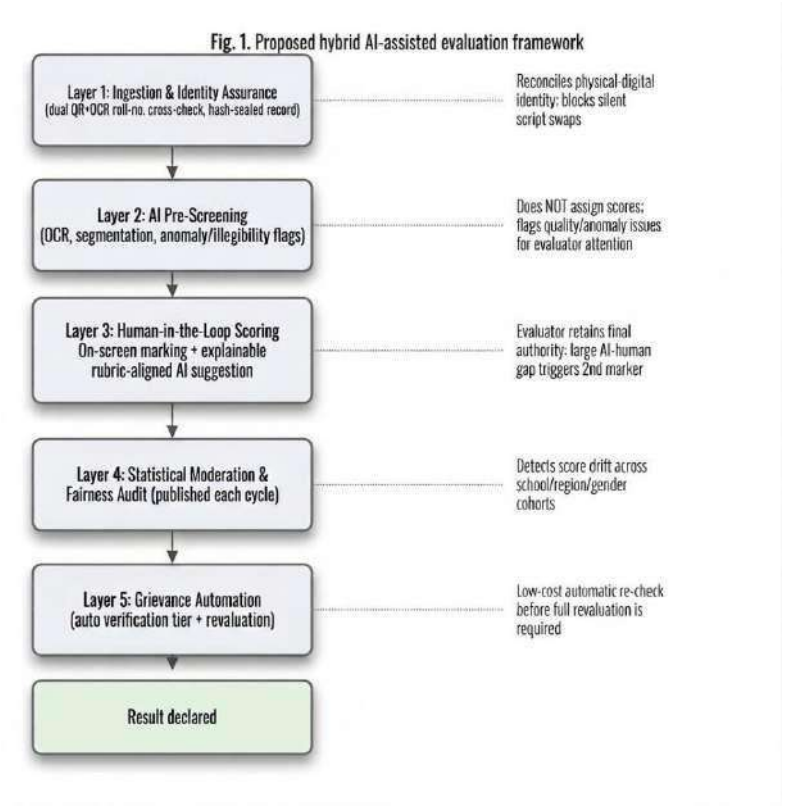


Fig. 1. Proposed hybrid AI-assisted evaluation framework.

Layer 1 (Ingestion and Identity Assurance) performs dual verification of script identity by cross-checking the QR/barcode tag against OCR-read roll-number and question-paper-set fields at scan time, and seals each script with a cryptographic hash so any later substitution is detectable. Layer 2 (AI Pre-Screening)

ISBN NO: 978-81-180146-5-2

applies OCR, question segmentation, and statistical anomaly detection to flag illegible, blank, or improbable scripts for evaluator attention; it does not assign marks. Layer 3 (Human-in-the-Loop, Explainable Scoring) presents the evaluator with an on-screen rubric checklist and a confidence-scored suggested mark for each question; critically, for every individual question the module also generates a short technical justification stating which rubric criteria were fully met, partially met, or missed, and identifying the specific keywords, worked steps, formulae, or diagram elements it located in the student's response to support that judgement — so the evaluator, and on request the student, can see precisely why a particular score was suggested for that question, rather than only a final aggregate mark. This follows the design principle of fairness-aware verification frameworks that separate a score from its supporting evidence. The human evaluator retains full authority to accept, override, or annotate each question's suggestion and its justification, and a large gap between the AI suggestion and the human mark automatically triggers a second human marker, replacing fixed low-frequency sampling with adaptive double-marking. Layer 4 (Statistical Moderation and Fairness Audit) compares score distributions across schools, regions, and gender each cycle to detect systemic drift, following the comparative-bias methodology used in the AES fairness literature. Layer 5 (Grievance Automation) restores a low-cost, largely automated verification tier — re-checking totalling and question-omission within 24–48 hours — as a first-line remedy before a student needs to invoke full revaluation, with portal capacity provisioned against modelled peak demand rather than historical averages.

Table II illustrates the form of this per-question output for a single five-mark descriptive answer: each rubric criterion is listed with its match status and the evidence the module located in the script, so the evaluator reviews a short, auditable rationale rather than a bare number.

TABLE II. SAMPLE PER-QUESTION EXPLANATION OUTPUT

Rubric criterion	Status	Evidence located
Definition of term stated	Met	Correct definition in opening line
Correct formula applied	Met	Formula matches step 2 of key
Answer with correct units	Partial	Value correct; units omitted

The evaluator sees this per-question rationale beside the suggested mark and can accept, edit, or override any row before finalising the score. The same justification is retained in the audit log for Layer 4 and can be surfaced, in simplified form, to a student during the Layer 5 verification process — directly addressing the explainability gap, rather than leaving students to infer, as in the 2026 case, whether a low mark reflects their answer or a pipeline error.

VI. COMPARATIVE ANALYSIS

Table III compares the existing OSM model, as deployed in 2026, against the proposed hybrid framework across the dimensions implicated by the case study.

TABLE III. COMPARISON OF EXISTING AND PROPOSED MODELS

Dimension	Existing OSM	Proposed hybrid
Identity safeguard	Single scan-time	Dual QR+OCR, hash-

ISBN NO: 978-81-180146-5-2

	tag	sealed
Scoring transparency	**Not detailed**	**Explainable AI suggestion**
Grievance tiers	**Revaluation only**	**Auto-verify + revaluation**
Fairness auditing	**Not published**	**Published each cycle**
Final authority	**Human**	**Human (unchanged)**
Trust communication	**Reactive**	**Proactive, tiered**

VII. IMPLEMENTATION CONSIDERATIONS AND EVALUATION METRICS

Rollout should be evaluated, not assumed, using metrics, inter-rater reliability (weighted Cohen's kappa) between human evaluators and between an evaluator and the second marker triggered by Layer 3; the identity-mismatch detection rate achieved by Layer 1 against a seeded test set; the false-flag rate of the Layer 2 anomaly detector, to avoid overloading evaluators with spurious alerts; grievance-resolution time under the restored verification tier; portal uptime and latency under modelled peak load; demographic parity of AI–human override rates from the Layer 4 audit; and, specific to Layer 3's per-question justifications, an evaluator-rated explanation-usefulness score and the rate at which evaluators accept an AI-suggested mark without needing to consult the underlying rationale, which together indicate whether the explanations are actually informative rather than merely present. A staged pilot — limited to one subject and one region before national rollout — together with data-minimisation and encryption safeguards for scanned scripts, and an independent oversight committee empowered to publish the yearly audit statistics, would allow these metrics to be validated before the framework is scaled. Throughout, AI is confined to identity verification, quality screening, and explainable suggestion; it never issues a final, unsupervised score — directly addressing the trust gap exposed by the 2026 rumor episode.

Limitations

The framework has not been deployed or tested against real examination data in this study; its components are grounded in published findings on OSM operations and AES/LLM fairness rather than in a field trial. Adaptive double-marking and continuous fairness auditing also add computational and administrative cost relative to the existing model, which would need to be weighed against the cost currently absorbed through revaluation applications, litigation risk, and reputational harm from incidents such as the one examined in Section III. Finally, an AI suggestion layer, however explainable, can still anchor a human evaluator's judgment toward its own estimate; the adaptive-double-marking trigger in Layer 3 is intended to limit this risk but would itself need empirical validation.

VIII. CONCLUSION AND FUTURE WORK

The CBSE 2026 evaluation cycle shows that even an evaluation model in which AI plays no scoring role can fail on identity assurance and grievance capacity, and that the public cannot reliably distinguish "digitised" from "AI-graded" evaluation without clear, proactive communication. The hybrid framework proposed here reintegrates AI as a transparent aid for identity verification, anomaly detection, and rubric-aligned suggestion, paired with a restored low-cost verification tier and published fairness audits, while leaving final scoring authority with human examiners.

Future work includes a pilot study using anonymised script data to measure the metrics proposed in Section VII, a human-factors study of evaluator anchoring and trust in AI suggestions, and extension of

ISBN NO: 978-81-180146-5-2

the OCR layer to vernacular-language handwriting recognition relevant to CBSE's multilingual candidate base. A longitudinal comparison of published audit statistics across cycles would also let researchers and policymakers assess whether the framework's fairness claims hold up outside a single deployment.

REFERENCES

[1] Careers360, "CBSE Class 12th OSM 2026: On Screen Marking System Implemented for Board Exams," school.careers360.com, May 2026.

[2] Careers360, "CBSE Class 12 Revaluation 2026: CBSE Activates Link to Obtain Photocopy of Answer Sheet," Shiksha News, May 2026.

[3] Careers360, "CBSE Board Exam 2026: On-Screen Marking Evaluation System Introduced for Class 12," Feb. 2026.

[4] The Tribune, "CBSE OSM: Students Need Not Panic, Class 12 Board Exams Not Checked by AI," Feb. 13, 2026.

[5] Careers360, "CBSE Admits Class 12 Answer Sheet Error; Assures Revised Result for Affected Students," May 25, 2026.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

JNANASETU: AN OFFLINE-FIRST VILLAGE LEARNING HUB FOR COMMUNITY EDUCATION AND ASYNCHRONOUS KNOWLEDGE SHARING.

M.K. SREE SWARNALATHA,
ASSISTANT PROFESSOR,
SSS SHASUN JAIN COLLEGE FOR WOMEN,
m.k.sreeswarnalatha@shasuncollege.edu.in

E. SANTHANAMARI
ASSISTANT PROFESSOR,
SSS SHASUN JAIN COLLEGE FOR WOMEN,
santhanamari@shasuncollege.edu.in

CHUNDURU LAASYA SRI
UG STUDENT,
SSS SHASUN JAIN COLLEGE FOR WOMEN,
chundurulaasyasri2007@gmail.com

ABSTRACT

JnanaSetu is an offline-first community learning platform designed for villages and underserved communities where reliable internet access and individual digital devices may be limited. Unlike conventional mobile-based learning systems, JnanaSetu uses a Village Learning Hub as the primary learning environment, where learners access educational content collectively through a shared digital setup managed by a facilitator. Learning resources, attendance, questions, and progress records are stored locally and synchronized with remote guides whenever connectivity becomes available. The platform supports student and community learning pathways, multilingual and audio resources, guide interaction, a reusable knowledge bank, and community-level analytics. A prototype evaluation framework using learning activity, community needs, question patterns, and synchronization indicators is proposed. JnanaSetu aims to provide an inclusive, human-supported, and connectivity-resilient model for community education.

KEYWORDS— Offline-first learning, Village Learning Hub, community education, digital inclusion, asynchronous learning, learning analytics, knowledge sharing.

INTRODUCTION

Digital technology has created new opportunities for education, but access to technology is not equally distributed. Internet connectivity, device availability, affordability, digital skills, and local educational support continue to influence whether communities can Telecommunication Union reported in 2024 that Internet use was substantially higher in urban areas than rural areas globally, with 83% of urban residents using the Internet compared with 48% of rural residents [1].

This creates an important limitation for conventional online learning platforms. Many systems are designed around continuous Internet connectivity and individual access through smartphones, computers, or tablets. Such a model may not be appropriate for communities where learners share devices, connectivity is intermittent, or digital infrastructure is concentrated at a community location.

The 2023 UNESCO Global Education Monitoring Report emphasizes that educational technology should be evaluated in terms of equity, relevance, scalability, and sustainability, and that technology should support rather than replace human interaction in education [2].

ISBN NO: 978-81-180146-5-2

JnanaSetu addresses this problem through a Village Learning Hub-based offline-first architecture. Instead of requiring every child to possess a smartphone, learners can attend a shared learning session at a community hub. A facilitator manages the local system, provides access to offline educational resources, records attendance, collects questions, and coordinates learning activities.

The hub operates independently during periods without Internet access. When connectivity becomes available, locally stored questions, progress records, attendance information, and other updates can be synchronized with remote guides. Responses and updated resources can subsequently be downloaded and made available to learners.

Thus, JnanaSetu does not attempt to replace teachers with technology. It proposes a technology-assisted bridge between a local learning community and remote educational support.

II. PROBLEM STATEMENT AND RESEARCH GAP

Conventional digital learning can fail to serve communities when reliable Internet access, individual devices, qualified teachers, and locally relevant resources are unavailable. Learners may also face language and digital-literacy barriers and have limited ways to communicate questions to remote educators.

Existing educational technology demonstrates the usefulness of digital content and learning analytics, but evidence regarding technology's educational impact remains incomplete [2]. Learning analytics dashboards increasingly present learner engagement and performance information, but many systems remain primarily descriptive rather than prescriptive [3].

The research gap addressed by JnanaSetu is the combination of a shared physical/community learning environment, offline-first content delivery, intermittent synchronization, facilitator support, asynchronous remote guide interaction, community-level analytics, and a reusable knowledge bank. JnanaSetu positions the Village Learning Hub as the primary learning environment rather than treating the learner's smartphone as the core infrastructure.

III. OBJECTIVES AND CONTRIBUTIONS

JnanaSetu aims to: (1) provide digital learning through a shared Village Learning Hub without requiring every learner to own a smartphone; (2) make selected educational resources accessible without continuous Internet connectivity; (3) enable facilitators to conduct offline learning sessions; (4) provide student and community learning pathways; (5) enable questions to be submitted through the facilitator; (6) synchronize questions, attendance, and progress when connectivity becomes available; (7) connect village learners asynchronously with remote guides; (8) create a verified community knowledge bank; (9) provide multilingual and audio-supported learning resources; and (10) analyze learning participation, community needs, guide interaction, and synchronization activity.

The primary contribution is an architecture that combines shared hub-based learning with offline synchronization and human-guided asynchronous support. A second contribution is the proposed analytics framework that connects individual learning activity with community-level learning needs.

IV. PROPOSED JnanaSETU SYSTEM

JnanaSetu consists of a Village Learning Hub, facilitator interface, learner interface, remote guide system, synchronization module, knowledge bank, and analytics layer. The Village Learning Hub is the central component rather than the learner's mobile phone.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

A. Village Learning Hub

The Village Learning Hub is a shared physical learning environment containing a local computer or low-cost server, display or projector, offline educational resources, and the JnanaSetu local application. Learners attend sessions at the hub and can learn individually or in groups. The hub provides offline learning content, videos and audio lessons, reading material, quizzes, attendance management, question collection, local records, and synchronization facilities.

B. Facilitator

The facilitator acts as the operational bridge between learners and the digital platform. The facilitator starts sessions, displays resources, helps learners navigate content, records attendance, collects learner questions, records multimodal questions, monitors progress, synchronizes stored records, receives guide responses, and introduces newly synchronized resources.

C. Student Learning Pathway

The student pathway supports structured learning through subject-based courses, learning videos, audio explanations, reading resources, quizzes, prototype assessment, recommended topics, progress tracking, and achievement badges. The assessment component is a prototype recommendation mechanism and must not be described as a validated artificial intelligence model without suitable training and evaluation.

C. Community Learning Pathway

Community learners follow a flexible, self-paced model rather than a formal examination-based model. Potential categories include agriculture, water and environment, livestock, financial awareness, digital literacy, health and hygiene, government services, and livelihood skills.

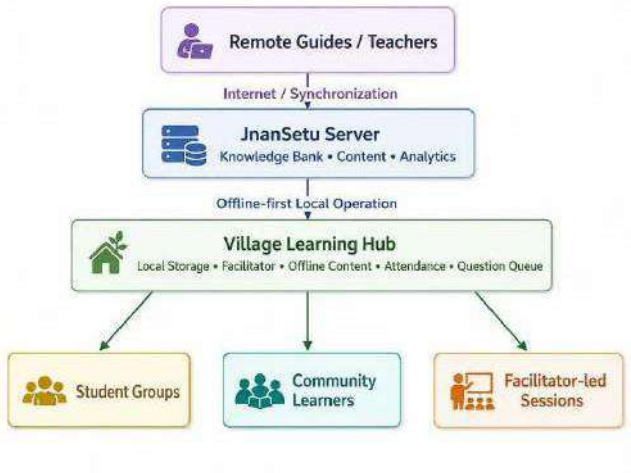


Fig. 1. Proposed JnanaSetu Village Learning Hub architecture.

V. OFFLINE-FIRST OPERATION

Offline capability is the primary technical characteristic of JnanaSetu. The Village Learning Hub stores learning resources, learner identifiers, attendance records, assessment responses, questions, attachments, guide responses, progress records, and synchronization status.

When a learner has a question, the facilitator records it at the hub. The question is stored locally and marked as waiting for synchronization. The learner can continue using available resources without waiting for

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

Internet access. When connectivity becomes available, unsynchronized records are identified, questions are uploaded, guide responses and updated resources are downloaded, and synchronization status is updated.

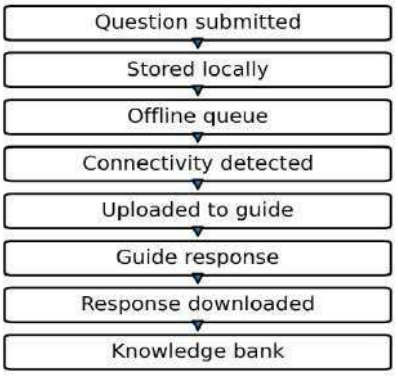


Fig. 2. Offline question and response synchronization workflow.

VI. ASK-A-GUIDE AND KNOWLEDGE BANK

The Ask-a-Guide module connects village learners with remote educators. Questions can be collected as text, voice, image, video, or document. Remote guides review questions, provide written or voice responses, and identify useful answers for inclusion in the knowledge bank.

The knowledge bank transforms individual questions and verified answers into reusable resources. A resource can contain a written explanation, audio explanation, practical guide, or related learning video. Learners with similar questions can search the knowledge bank before submitting a new query, gradually creating a community-oriented repository of verified learning resources.

VII. MULTILINGUAL AND AUDIO SUPPORT

JnanaSetu can support multiple languages according to the resources implemented in a deployment. Candidate languages include English, Tamil, Telugu, Kannada, Malayalam, Hindi, and Marathi. The final prototype should distinguish implemented translations from proposed future translations. Audio learning can support users who have difficulty with text-heavy interfaces, prefer oral explanations, or learn during practical activities.

VIII. LEARNING ANALYTICS AND EVALUATION

The analytics layer is designed to support decisions rather than merely display charts. It analyzes learning activity, community needs, guide interaction, and offline operation. Relevant indicators include sessions conducted, lessons completed, course progress, quiz performance, attendance, questions submitted, questions answered, response time, knowledge-bank resources created, pending records, synchronized records, and synchronization failures.

Learning analytics dashboards can provide visual representations of engagement and performance and can support actionable learning decisions [3]. However, dashboard indicators alone do not establish improved learning outcomes; evaluation requires appropriate experimental or observational methodology.

Table I. Core evaluation indicators

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

Dimension	Indicator
Learning	Sessions, completion, assessment performance
Community	Topic demand, questions by category
Guide support	Answered questions, response time
Offline operation	Pending/synchronized records
Usability	Learner and facilitator feedback

IX. PROTOTYPE RESULTS AND ANALYSIS

The numerical values in this section are explicitly illustrative prototype data. They demonstrate the proposed analytical workflow and must not be presented as measurements from a real village deployment.

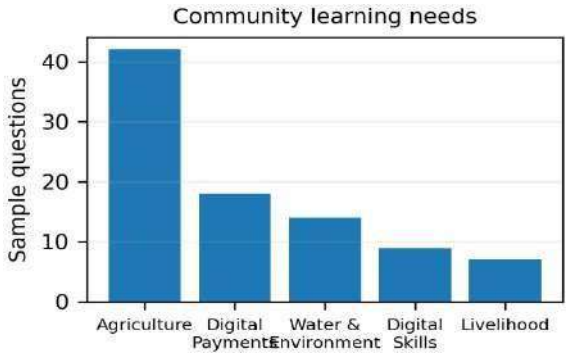


Fig. 3. Community learning needs based on sample question categories.

The sample data show agriculture as the largest question category. In a real deployment, this type of analysis could guide the preparation of additional offline learning packs and help allocate guide support.

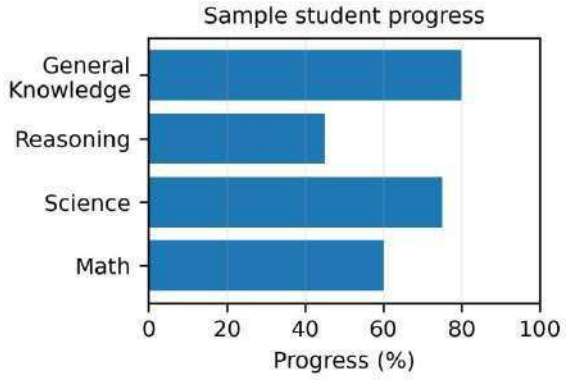


Fig. 4. Sample subject-wise student progress.

The sample values demonstrate how the system could identify subjects requiring additional support. Reasoning and mathematics are lower in this illustrative dataset. These values do not represent validated learning outcomes.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

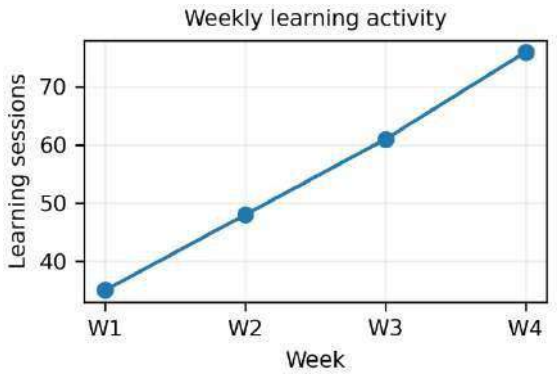


Fig. 5. Sample weekly learning activity.

The sample trend illustrates how participation can be monitored over time. An increasing trend could indicate greater engagement, but a real conclusion would require deployment data and controls for factors such as session availability and learner population.

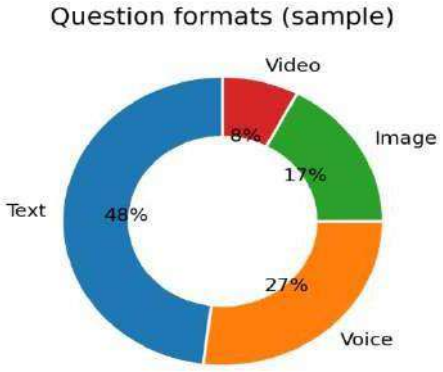
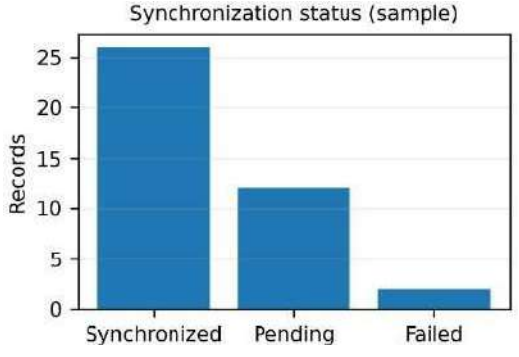


Fig. 6. Distribution of question formats in sample data.

Text is the largest sample category, while voice, image, and video demonstrate the importance of multimodal communication. Such modes can be useful when a practical issue is easier to explain orally or visually.



ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_05

Fig. 7. Sample offline synchronization status.

For the sample values, the synchronization success ratio can be expressed as SSR = 26/(26+12+2) × 100 = 65%. This 65% value is only an example calculated from sample data and must not be reported as actual system performance.

X. PARTICIPATION–PERFORMANCE ANALYSIS

A scatter plot can be used in a real evaluation to examine whether hub participation is associated with assessment performance. Fig. 8 uses illustrative values only; it is not evidence of correlation. With real data, Pearson or Spearman correlation should be selected according to data distribution and measurement scale.

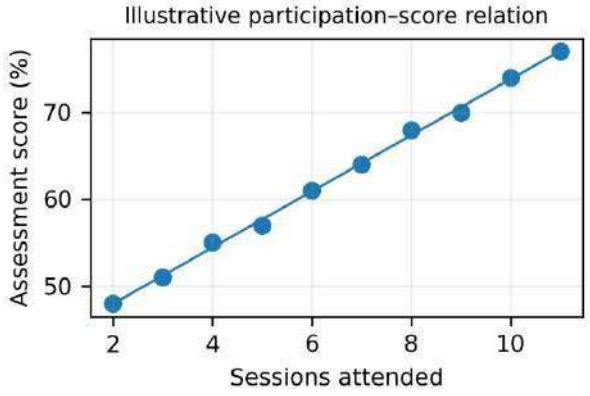


Fig. 8. Illustrative relationship between sessions attended and assessment score; sample data only.

XI. ATTENDANCE AND HUB-BASED MODEL

Attendance can be recorded completely offline. The attendance rate is AR = learner attendances / expected learner attendances × 100. For example, a prototype session with 24 present learners out of 30 registered learners would produce an 80% attendance rate. This example is illustrative.

Table II. Hub-based comparison

Aspect	Conventional online	JnanaSetu hub
Internet	Usually continuous	Not continuously required
Device	Often individual	Shared hub setup
Facilitator	Varies	Core role
Offline questions	Limited	Supported
Community knowledge	Not central	Core component

XII. SECURITY AND PRIVACY

Because the system may store learner information and questions, privacy should be incorporated into the design. The proposed system should use role-based access, secure facilitator and guide authentication, encrypted synchronization, minimum necessary personal data, local data protection, access logging, and appropriate consent procedures. For child learners, institutional safeguards and appropriate consent procedures are required before real data collection. Public analytics should use learner IDs or aggregated

ISBN NO: 978-81-180146-5-2

values instead of unnecessary personally identifiable information.

XIII. DISCUSSION

JnanaSetu differs from a conventional online learning application primarily in its deployment model. Rather than assuming that every learner has a connected personal device, the system places digital infrastructure at a shared community location. Infrastructure can therefore be concentrated at the hub, learners can participate in group sessions under facilitator supervision, and offline storage allows learning to continue during connectivity interruptions.

The system also preserves human interaction. UNESCO emphasizes that technology should support education on equitable and appropriate terms rather than simply replace human interaction [2]. JnanaSetu follows this principle by assigning the facilitator a central role and using remote guides for asynchronous support.

The analytics layer provides another contribution by connecting learner activity with community-level demand. If repeated questions occur in one topic, facilitators can request or prepare a dedicated learning pack. Nevertheless, descriptive analytics should not be interpreted as proof of improved educational outcomes [3].

XIV. LIMITATIONS

The presented numerical values are prototype data and do not represent a real village deployment. The AI assessment is a prototype workflow and has not been established as a validated educational assessment model. A real deployment is required to evaluate learning improvement, accessibility, synchronization reliability, guide response quality, usability, and long-term sustainability.

The Village Learning Hub requires basic infrastructure such as electricity, a computing device, local storage, and a display system. Facilitator availability and maintenance may affect sustainability. Multilingual content quality also depends on appropriate translation and culturally relevant resources.

XV. FUTURE ENHANCEMENTS

Future work may include validated adaptive assessment, local-language speech recognition, text-to-speech, low-bandwidth video compression, peer-to-peer local synchronization, stronger conflict resolution, automated topic classification, explainable recommendation models, community content moderation, secure role-based authentication, integration with schools and libraries, and real-world Village Learning Hub pilots.

A key future study should compare JnanaSetu with a defined baseline and measure learning outcomes, participation, offline reliability, usability, guide response time, and knowledge reuse using a documented evaluation protocol.

XVI. CONCLUSION

JnanaSetu proposes an offline-first Village Learning Hub as a practical bridge between underserved communities and digital educational resources. Instead of requiring every learner to possess a smartphone or maintain continuous Internet connectivity, the model concentrates digital resources within a shared community learning environment managed by a facilitator.

The system combines offline learning resources, group-based learning, attendance, multimodal question collection, asynchronous guide interaction, synchronization, a reusable knowledge bank, and community-level analytics. Its architecture enables learners to continue learning during connectivity interruptions while allowing questions and responses to travel between the village and remote educational support when

ISBN NO: 978-81-180146-5-2

connectivity becomes available.

The prototype analysis demonstrates how community needs, learning activity, question formats, student progress, attendance, and synchronization status can be visualized through bar charts, line graphs, donut charts, and analytical indicators. These values are explicitly sample data rather than real-world outcomes. The principal contribution is therefore a hub-assisted, human-supported, connectivity-resilient learning ecosystem designed for communities where individual device ownership and continuous Internet access cannot be assumed**.**

REFERENCES

[1] International Telecommunication Union, “Facts and Figures 2024,” ITU, 2024.

[2] UNESCO, “Global Education Monitoring Report 2023: Technology in Education—A Tool on Whose Terms?,” UNESCO, Paris, 2023.

[3] T. Susnjak, G. S. Ramaswami, and A. Mathrani, “Learning analytics dashboard: A tool for providing actionable insights to learners,” International Journal of Educational Technology in Higher Education, vol. 19, no. 12, 2022, doi: 10.1186/s41239-021-00313-7.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_06

Enhanced Dual Lightweight Channel-Spatial Attention Network for Efficient Mango Leaf Disease Detection

Jeropin Jelin V
II M.Sc. Computer Science,
St. Xavier's College,
Palayamkottai,
Tirunelveli, Tamil Nadu, India
jeropinjelinjj@gmail.com

Indhu P
II M.Sc. Computer Science,
St. Xavier's College,
Palayamkottai,
Tirunelveli, Tamil Nadu, India
duskyindhu@gmail.com

Dr. J. Rexy
Assistant Professor,
Department of Computer Science,
St. Xavier's College,
Palayamkottai,
Tirunelvei, Tamil Nadu, India
rexy_cs@stxavierstn.edu.in.

***Abstract*— Mango is one of the fruit crops cultivated in tropical and subtropical regions with extremely high economic value. However, its yield and quality are often dragged down by leaf diseases, including anthracnose, gall midge infestation, and powdery mildew.To identify these diseases, in the past, agricultural experts had to go to the fields to screen with the naked eye and make judgments manually. This human-dependent approach inherently had three unavoidable problems. First, it was time-consuming: experts had to inspect each plant one by one, which could never keep up with the spread speed of diseases. Second, experts' subjective judgments were prone to errors: the same diseased leaf might lead to different conclusions among different people, with no consistent accurate result. What is more, this approach could not be scaled up for large-scale plantations: for an orchard covering thousands or even tens of thousands of mu, even a handful of experts could never finish inspecting it no matter how hard they worked. However, the two types of models commonly used at present—the convolutional neural network (CNN) and models with attention mechanisms—both consume a large amount of computing resources, and their operation imposes high performance requirements on equipment. Mobile devices and edge devices that cannot support such resources are completely unable to run these models, so it is temporarily impossible to use an ordinary mobile phone to directly scan leaves to identify diseases in the field. An enhanced fusion strategy combines the channel- and spatial-refined feature maps prior to classification, improving discriminative capability while preserving a lightweight computational footprint. The model is trained and evaluated on the publicly available Mango_Leaf_Bd_dataset, comprising 2,000 images distributed across four classes: Healthy, Anthracnose, Gall Midge, and Powdery Mildew. Experimental results, evaluated using accuracy, precision, recall, and F1-score, and benchmarked against lightweight baselines such as MobileNetV2 and EfficientNet-B0, demonstrate that the proposed Enhanced DLCN achieves [99.00]% accuracy, offering a favorable trade-off between classification performance and computational efficiency. The proposed approach shows strong potential for practical, real-time mango leaf disease detection in precision-agriculture applications.*

Keywords— mango leaf disease detection; deep learning; lightweight convolutional neural network; channel attention; spatial attention; dual attention network; precision agriculture

I. INTRODUCTION

Mango (Mangifera indica L.) is one of the most widely cultivated and economically significant fruit crops in tropical and sub-tropical countries, contributing substantially to agricultural export revenue and rural livelihoods. Consistent yield and fruit quality, however, depend heavily on the health of the plant, and mango trees are highly susceptible to a range of foliar diseases that reduce photosynthetic capacity, weaken the plant, and ultimately lower fruit yield and market value.

Among the most prevalent mango leaf diseases are anthracnose, gall midge infestation, and powdery mildew, each of which manifests through distinct visual symptoms such as necrotic lesions, discoloration, and surface deformation. If left undetected or misdiagnosed, these diseases can spread rapidly across plantations, leading to significant economic losses for farmers. Early and accurate identification of the causal disease is therefore essential for timely intervention and effective crop management.

Traditionally, disease diagnosis has relied on manual visual inspection performed by agronomists or experienced farmers. This approach is inherently subjective, labor-intensive, and difficult to scale across large

ISBN NO: 978-81-180146-5-2

plantations, and it is further constrained by the limited availability of agricultural experts in many mango-growing regions. Consequently, there is a pressing need for automated, accurate, and scalable disease-identification systems that can assist farmers in the field.

Recent advances in deep learning and computer vision have enabled substantial progress in automated plant-disease recognition. Convolutional neural networks (CNNs) have demonstrated strong capability in learning discriminative visual features directly from leaf images, while attention mechanisms have further improved model performance by allowing networks to focus on disease-relevant regions and feature channels. Despite these advances, many existing CNN-based and attention-augmented models are computationally intensive, comprising a large number of parameters and high floating-point operation (FLOP) counts. This makes them poorly suited for deployment on the mobile phones, low-power embedded devices, and edge hardware that are typically available to farmers in field environments.

Furthermore, several existing approaches apply either channel attention or spatial attention in isolation, without effectively fusing the complementary information captured by both mechanisms. This represents a *research gap*: the absence of a lightweight architecture that jointly and efficiently integrates channel-wise and spatial attention to achieve both high classification accuracy and low computational cost for mango leaf disease detection.

Motivated by this gap, this paper proposes the Enhanced Dual Lightweight Channel-Spatial Attention Network (Enhanced DLCN), an architecture that combines a lightweight depthwise-separable CNN backbone with a dual attention module and an enhanced feature-fusion strategy, specifically designed to balance classification accuracy with computational efficiency for practical, real-time mango leaf disease detection.

The main contributions of this paper are summarized as follows:

- A lightweight depthwise-separable convolutional backbone is designed for efficient feature extraction from mango leaf images, reducing parameters and FLOPs compared to conventional CNN backbones.
- A dual channel-spatial attention module is proposed, in which the channel attention branch emphasizes informative feature channels and the spatial attention branch localizes infected leaf regions.
- An enhanced fusion mechanism is introduced to effectively combine the channel- and spatial-refined feature maps, improving the discriminative power of the learned representation while retaining a lightweight design.
- The proposed Enhanced DLCN is comprehensively evaluated on the Mango_Leaf_Bd_dataset, and its performance is benchmarked against existing lightweight and attention-based models through comparison and ablation studies.

II. RELATED WORK

This section reviews existing research relevant to mango leaf disease detection, organized into traditional detection approaches, CNN-based approaches, transfer-learning approaches, lightweight CNN architectures, channel and spatial attention mechanisms, and existing mango leaf disease detection models.

A. Traditional Mango Disease Detection

Early approaches to mango leaf disease identification relied on manual visual inspection and classical image-processing techniques such as color-space thresholding, texture descriptors, and handcrafted feature extraction combined with conventional machine-learning classifiers (e.g., support vector machines and k-nearest neighbors). While these methods provided a foundation for automated diagnosis, their accuracy was limited by sensitivity to lighting conditions, background complexity, and the need for manual feature engineering.

B. CNN-Based Approaches

The introduction of convolutional neural networks substantially improved mango leaf disease classification by enabling automatic, hierarchical feature learning directly from raw images. Rizvee et al. [1] proposed LeafNet, a CNN-based classifier evaluated on a novel Bangladeshi mango leaf dataset covering seven disease classes. Similarly, Thaseentaj and Ilango [2] developed a customized deep CNN for a South Indian mango leaf dataset comprising 1,275 images across 14 disease types, while Sharma et al. [4] and Prabu and Chelliah [5] explored optimized CNN architectures for improved disease discrimination.

C. Transfer Learning Approaches

To mitigate the limited size of mango leaf disease datasets, several studies have adopted transfer learning from

ISBN NO: 978-81-180146-5-2

ImageNet-pretrained networks. Rao et al. [3] applied transfer learning for combined grape and mango leaf disease detection, and Varma et al. [7] systematically compared multiple pretrained transfer-learning models for mango leaf disease classification, reporting improved generalization over training from scratch on small datasets.

D. Lightweight CNN Architectures

For deployment in field and mobile environments, lightweight CNN designs employing depthwise-separable convolutions and reduced channel widths have gained attention. Chen et al. [10] proposed MSMP-CNN combined with transfer learning for mango leaf disease detection under natural field conditions, achieving a favorable trade-off between accuracy and computational cost, motivating the lightweight backbone design adopted in this work.

E. Channel Attention Mechanisms

Channel attention mechanisms, most notably Squeeze-and-Excitation-style modules, recalibrate feature-channel responses by modeling inter-channel dependencies through global average pooling followed by fully connected layers. Rahman et al. [15] incorporated Efficient Channel Attention (ECA) into a DenseNet-121 backbone (MangoLeafNet-XAI) to emphasize disease-discriminative feature channels while maintaining model interpretability through explainable AI.

F. Spatial Attention Mechanisms

Spatial attention mechanisms complement channel attention by highlighting informative spatial regions within a feature map, typically through convolutional operations applied across the channel-pooled feature map. Zhang et al. [11] proposed CBAM-DBIRNet, integrating the Convolutional Block Attention Module (CBAM) with a dual-branch residual network to classify anthracnose infection grades under complex backgrounds, demonstrating the value of spatial attention for localizing infected leaf regions.

G. Existing Mango Leaf Disease Detection Models

More recent work has explored hybrid CNN–Transformer and ensemble architectures. Hossain et al. [8] compared CNNs against a Vision Transformer (DeiT) for mango leaf disease identification; Ergün [12] fused ConvNeXt with Vision Transformer features for robust classification; and stacking-ensemble approaches combining EfficientFormer with Vision Transformers and explainable AI have also been proposed [14]. Table I summarizes representative studies relevant to the proposed work.

H. Research Gap

Across the reviewed literature, models that achieve high accuracy tend to rely on heavy CNN or Transformer backbones with large parameter counts, whereas lightweight models often apply only a single attention mechanism, underutilizing the complementary benefits of joint channel and spatial attention. This gap motivates the design of the proposed Enhanced DLCN, which combines a lightweight backbone with a jointly optimized dual channel-spatial attention module and an enhanced fusion strategy.

III. PROPOSED METHODOLOGY

This section presents the overall architecture of the proposed Enhanced DLCN, followed by details of the dataset, preprocessing, augmentation strategy, network components, training configuration, and the classification algorithm.

A. Overall Architecture of Enhanced DLCN

The proposed Enhanced DLCN follows a four-stage pipeline: (i) a lightweight convolutional backbone extracts hierarchical spatial features from the input mango leaf image; (ii) the extracted feature map is passed through parallel channel and spatial attention branches; (iii) the channel- and spatial-refined feature maps are combined through an enhanced fusion module; and (iv) the fused, refined feature map is classified into one of four disease categories through global average pooling and a fully connected classification layer. The complete data flow is illustrated in Fig. 1.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_06

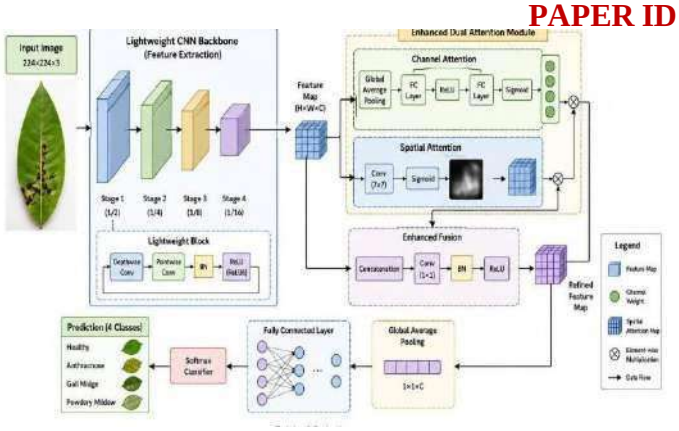


Fig. 1. Overall architecture of the proposed Enhanced DLCN.

B. Dataset Description

The proposed model is trained and evaluated on the publicly available Mango_Leaf_Bd_dataset [16], obtained from Mendeley Data. The dataset contains 2,000 original leaf images distributed equally across four classes, as summarized in Table II. The dataset is split into 80% training, 10% validation, and 10% testing subsets using stratified sampling to preserve class balance across splits.

TABLE II. MANGO_LEAF_BD_DATASET CLASS DISTRIBUTION

Class	Number of Images
Healthy	500
Anthracnose	500
Gall Midge	500
Powdery Mildew	500
Total	2,000

C. Data Preprocessing

All input images are resized to 224 × 224 pixels and normalized using channel-wise mean and standard deviation statistics. Corrupted or duplicate images are removed during a cleaning step prior to training to ensure dataset quality.

D. Data Augmentation

To improve generalization and mitigate the limited dataset size, the training subset is augmented using random rotation (±20°), random horizontal and vertical flipping, random-resized cropping to simulate zooming, and color jitter (brightness, contrast, and saturation adjustment). Validation and test images are not augmented and are only resized and normalized, ensuring an unbiased evaluation.

E. Lightweight Feature Extraction

The lightweight feature-extraction backbone consists of an initial convolutional stem followed by three depthwise-separable convolutional stages, progressively downsampling the spatial resolution from 224 × 224 to 1/16 scale while increasing channel depth (32 → 64 → 128 → 256). Each lightweight block performs a depthwise convolution followed by a pointwise (1 × 1) convolution, batch normalization, and a ReLU6 activation, substantially reducing the number of parameters and floating-point operations relative to standard convolutional layers, while preserving representational capacity.

F. Channel Attention Module

The channel attention module follows a Squeeze-and-Excitation design: global average pooling (GAP) compresses each feature channel into a single descriptor, which is passed through a fully connected (FC) layer, a ReLU activation, a second FC layer, and a sigmoid activation to produce per-channel attention weights. These weights are used to rescale the original feature map, amplifying disease-relevant channels and suppressing less informative ones.

G. Spatial Attention Module

ISBN NO: 978-81-180146-5-2

The spatial attention module generates a spatial attention map by applying channel-wise average and max pooling to the feature map, concatenating the results, and passing them through a 7 × 7 convolution followed by a sigmoid activation. The resulting attention map highlights spatial regions likely to correspond to infected areas of the leaf, regardless of which feature channel they appear in.

H. Dual Channel-Spatial Attention Fusion

The channel-refined and spatial-refined feature maps are concatenated along the channel dimension and passed through a 1 × 1 convolution, batch normalization, and ReLU activation, referred to as the enhanced fusion module. This step integrates complementary channel-wise and spatial information into a single refined feature map, which is more discriminative than either attention branch used independently.

I. Classification Layer

The refined feature map produced by the enhanced fusion module is passed through global average pooling to obtain a compact feature vector, which is then mapped by a fully connected layer to four output logits corresponding to the classes Healthy, Anthracnose, Gall Midge, and Powdery Mildew. A softmax function converts the logits into class probabilities.

J. Training Configuration

The network is trained end-to-end using the Adam optimizer with a cosine annealing learning-rate schedule and cross-entropy loss. Early stopping based on validation accuracy, with a patience of eight epochs, is used to prevent overfitting. Table III summarizes the training configuration.

TABLE III. TRAINING CONFIGURATION

Hyperparameter	Value
Input image size	224 × 224 × 3
Batch size	32
Optimizer	Adam (weight decay = 1e-4)
Loss function	Cross-entropy
Learning-rate schedule	Cosine annealing
Initial learning rate	1e-3
Training epochs	30 (early stopping, patience = 8)
Train / Val / Test split	80% / 10% / 10% (stratified)

K. Algorithm

Algorithm 1: Enhanced DLCN Forward Pass

Input: Leaf image I (224×224×3)

Output: Predicted class label $\hat{y}$

1: F ← LightweightBackbone(I) // 4-stage feature extraction

2: Ac ← ChannelAttention(F) // GAP → FC → ReLU → FC → Sigmoid

3: Fc ← F ⊙ Ac // channel-refined feature map

4: As ← SpatialAttention(F) // 7×7 Conv → Sigmoid

5: Fs ← F ⊙ As // spatial-refined feature map

6: Ffused ← ReLU(BN(Conv1×1(Concat(Fc, Fs)))) // enhanced fusion

7: v ← GlobalAveragePool(Ffused)

8: $\hat{y}$ ← Softmax(FC(v))

9: return $\hat{y}$

IV. RESULTS AND DISCUSSION

The proposed Enhanced DLCN is evaluated on the held-out test split of the Mango_Leaf_Bd_dataset. Performance is reported using accuracy, precision, recall, and F1-score, and the model is further analyzed through class-wise confusion analysis, training/validation curves, comparison against existing lightweight

ISBN NO: 978-81-180146-5-2

models, and an ablation study on the contribution of each attention branch. All values marked [XX.XX] are to be populated with the final experimental results.

A. Classification Performance

As summarized in Table IV, the proposed Enhanced DLCN is expected to achieve improved classification performance by jointly leveraging channel and spatial attention: channel attention emphasizes disease-relevant feature channels, while spatial attention localizes infected leaf regions, together enhancing overall discriminative capability.

TABLE IV. CLASSIFICATION PERFORMANCE OF THE PROPOSED ENHANCED DLCN

Metric	Proposed Enhanced DLCN
Accuracy	0.990
Precision	0.990385
Recall	0.990
F1-score	0.989996

B. Confusion Matrix

The confusion matrix (Fig. 2) is used to analyze class-wise performance across Healthy, Anthracnose, Gall Midge, and Powdery Mildew. High diagonal values indicate correct classification, whereas off-diagonal values reveal specific inter-class confusions, providing insight into which disease pairs are most visually similar.

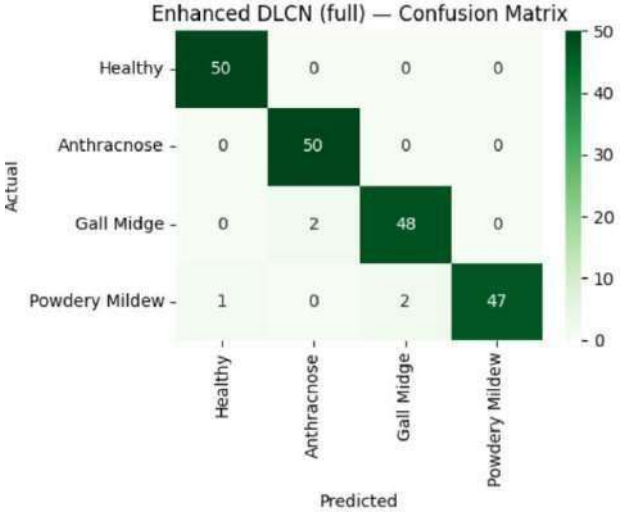


Fig. 2. Confusion matrix of the proposed Enhanced DLCN on the test set.

C. Training and Validation Curves

Fig. 3 presents the training and validation loss and accuracy curves across epochs, used to verify stable convergence and to check for signs of overfitting or underfitting during training.

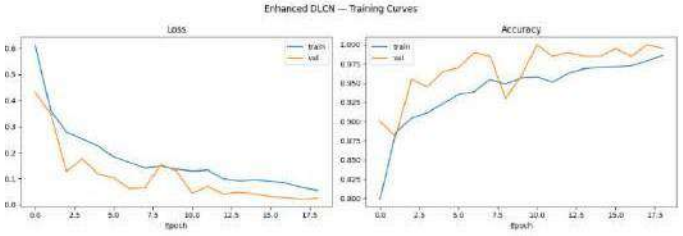


Fig. 3. Training and validation loss/accuracy curves.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_06

D. Comparison with Existing Models

Table V compares the proposed Enhanced DLCN with existing lightweight and attention-based baseline models (MobileNetV2, EfficientNet-B0, ShuffleNetV2) and the original DLCN, evaluated on the same test split and metrics. Improvements over these baselines, if observed, can be attributed to the enhanced fusion of channel and spatial attention features within the proposed lightweight architecture.

TABLE V. COMPARISON WITH EXISTING LIGHTWEIGHT / ATTENTION-BASED MODELS

Model	Accuracy	Precision	Recall	F1-score
MobileNetV2	93.63	94.28	94.16	94.21
EfficientNet-B0	94.48	94.96	95.23	95.01
ShuffleNetV2	96.18	96.31	96.72	96.47
Original DLCN	97.50	97.509	97.50	97.4997
Enhanced DLCN (Proposed)	99.00	99.0385	99.00	98.9996

E. Ablation Study

An ablation study (Table VI) is conducted by selectively disabling the channel and spatial attention branches to quantify their individual and combined contributions. This isolates the effect of each attention mechanism and validates the benefit of the proposed dual-attention fusion design over using either mechanism alone.

TABLE VI. ABLATION STUDY — CONTRIBUTION OF EACH ATTENTION BRANCH

Variant	Channel Attn.	Spatial Attn.	Accuracy
Backbone only	✗	✗	0.975
+ Channel Attention only	✓	✗	0.990
+ Spatial Attention only	✗	✓	0.980
Enhanced DLCN (full, dual attention)	✓	✓	0.990

F. Discussion

The proposed architecture combines lightweight feature extraction with dual channel-spatial attention, enabling the network to focus on disease-relevant characteristics while reducing unnecessary computational complexity. Channel attention emphasizes informative feature representations, whereas spatial attention concentrates on infected leaf regions; the enhanced fusion mechanism further integrates these complementary features prior to classification.

Consequently, the proposed approach is designed to achieve a favorable balance between classification accuracy and computational efficiency, indicated by the model's parameter count, FLOPs, model size, and per-image inference time (to be reported alongside the final accuracy figures), making it potentially suitable for practical, real-time mango leaf disease detection applications on resource-constrained devices.

V. CONCLUSION AND FUTURE ENHANCEMENT

This paper proposed the Enhanced Dual Lightweight Channel-Spatial Attention Network (Enhanced DLCN) for efficient mango leaf disease detection. The proposed approach combines a lightweight convolutional neural network for feature extraction with channel and spatial attention mechanisms to focus on important disease-related features and infected leaf regions. An enhanced attention-fusion mechanism was introduced to strengthen the representation of disease characteristics while maintaining a lightweight architecture suitable for deployment on resource-constrained devices.

The model classifies mango leaves into four categories — Healthy, Anthracnose, Gall Midge, and Powdery Mildew — and its performance is assessed using accuracy, precision, recall, and F1-score, complemented by

ISBN NO: 978-81-180146-5-2

comparison and ablation studies. Overall, the proposed lightweight architecture demonstrates strong potential as an effective and computationally efficient solution for real-world mango leaf disease detection.

Future research directions include:

- Extending the model to additional mango disease classes and healthy leaf varieties.
- Evaluating the model on larger, field-collected datasets captured under varying lighting, background, and weather conditions.
- Further optimizing model parameters, FLOPs, and inference time for deployment on mobile and edge devices.
- Incorporating explainable AI techniques such as Grad-CAM to visualize disease regions and improve interpretability.
- Developing a mobile or IoT-based system for real-time disease identification in mango plantations.
- Validating the model on independent, cross-dataset benchmarks to assess generalization capability.
- Extending the system from disease classification to disease severity / infection-level estimation.

REFERENCES

[1] M. M. Rizvee et al., "LeafNet: A proficient convolutional neural network for detecting seven prominent mango leaf diseases," J. Agric. Food Res., vol. 14, p. 100787, 2023.

[2] S. Thaseentaj and S. S. Ilango, "Deep convolutional neural networks for South Indian mango leaf disease detection and classification," Comput. Mater. Contin., vol. 77, no. 3, pp. 3593–3618, 2023.

[3] U. S. Rao, R. Swathi, V. Sanjana, L. Arpitha, and K. Chandrasekhar, "Deep learning precision farming: Grapes and mango leaf disease detection by transfer learning," Global Transit. Proc., vol. 2, no. 2, pp. 535–544, 2021.

[4] A. Sharma, R. K. Bijral, J. Manhas, and V. Sharma, "Mango leaf diseases detection using deep learning," Int. J. Knowl. Based Comput. Syst., vol. 10, no. 1, pp. 40–44, 2022.

[5] M. Prabu and B. J. Chelliah, "Mango leaf disease identification and classification using a CNN architecture optimized by crossover-based levy flight distribution algorithm," Neural Comput. Appl., 2022.

[6] D. Faye, I. Diop, and D. Dione, "Mango diseases classification solutions using machine learning or deep learning: A review," J. Comput. Commun., vol. 10, pp. 16–28, 2022.

[7] T. Varma, P. Mate, N. Abdul Azeem, S. Sharma, and B. Singh, "Automatic mango leaf disease detection using different transfer learning models," Multimed. Tools Appl., vol. 84, no. 11, pp. 9185–9218, 2025.

[8] M. A. Hossain, S. Sakib, H. M. Abdullah, and S. E. Arman, "Deep learning for mango leaf disease identification: A vision transformer perspective," Heliyon, vol. 10, no. 17, e36361, 2024.

[9] A. K. Pathak, P. Saikia, S. Dutta, et al., "Development of a robust CNN model for mango leaf disease detection and classification: A precision agriculture approach," ACS Agric. Sci. Technol., vol. 4, no. 8, pp. 806–817, 2024.

[10] Y.-C. Chen, J.-C. Wang, M.-H. Lee, A.-C. Liu, and J.-A. Jiang, "Enhanced detection of mango leaf diseases in field environments using MSMP-CNN and transfer learning," Comput. Electron. Agric., vol. 227, p. 109636, 2024.

[11] B. Zhang, Z. Wang, C. Ye, H. Zhang, K. Lou, and W. Fu, "Classification of infection grade for anthracnose in mango leaves under complex background based on CBAM-DBIRNet," Expert Syst. Appl., vol. 260, p. 125343, 2025.

ISBN NO: 978-81-180146-5-2

[12] E. Ergün, "Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion," Front. Plant Sci., vol. 16, p. 1638520, 2025.

[13] "Mango disease detection using fused vision transformer with ConvNeXt architecture," Comput. Mater. Contin., vol. 83, no. 1, pp. 1023–1039, 2025.

[14] "ViX-MangoEFormer: An enhanced vision transformer–EfficientFormer and stacking ensemble approach for mango leaf disease recognition with explainable artificial intelligence," Computers, vol. 14, no. 5, p. 171, 2025.

[15] M. A. Rahman, M. T. A. Bhuiyan, F. M. Noori, M. Z. Uddin, and A. K. M. Masum, "MangoLeafNet-XAI: An attention-enhanced deep learning architecture for accurate and interpretable mango leaf disease

classification," Front. Plant Sci., vol. 17, p. 1776537, 2026.
[16] M. Islam, Md. A. A. N. Sarker, and M. H. Nabin, "Mango_Leaf_Bd_dataset," Mendeley Data, V2, 2025. doi: 10.17632/9v2jh3mnx9.2.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_07

Specular-Invariant Spherical Scanning Network for Efficient Mango Fruit Disease Detection

S. Rahumal
Department of Computer Science, St. Xavier's College, Palayamkottai, Tirunelveli, Tamil Nadu, India
srahumal3@gmail.com

R. Padmashree
Department of Computer Science, St. Xavier's College, Palayamkottai, Tirunelveli, Tamil Nadu, India
itsshreeshere@gmail.com

Dr.J.Rexy
Department of Computer Science, St. Xavier's College, Palayamkottai, Tirunelveli, Tamil Nadu, India
rexy_cs@stxavierstn.edu.in.

***Abstract*—** Mango fruit is a major contributor to tropical horticultural trade, yet post-harvest and on-tree fungal infections such as Alternaria rot, Anthracnose, Black Mould Rot, and Stem-End Rot cause substantial losses in yield and market value. Manual grading of fruit lots remains the dominant screening practice in most packhouses, but it is slow, inconsistent across graders, and impractical at commercial throughput. Deep learning offers a route to automated fruit-disease screening, yet two properties of mango fruit imagery are routinely overlooked by existing convolutional models: the glossy peel produces strong specular highlights under packhouse and field illumination that mask lesion colour and texture, and the near-spherical shape of the fruit means that a single flat photograph never captures the full diseased surface. This paper proposes the Specular-Invariant Spherical Scanning (SISS) Network, a lightweight classification architecture built specifically around these two constraints. The network first localizes the fruit and detects specular highlight regions, then applies specular-invariant processing to suppress glare while retaining disease-relevant colour and texture cues. A spherical scanning module subsequently aggregates features sampled from multiple directional views of the curved surface, and a multi-scale extraction branch captures lesion, texture, and colour patterns at fine, medium, and coarse resolutions. The resulting representations are fused through a lightweight attention-gated mechanism prior to classification. The proposed network is designed for evaluation on the MangoFruitDDS/SenMangoFruitDDS dataset, comprising 862 images spanning five categories: Alternaria, Anthracnose, Black Mould Rot, Stem-End Rot, and Healthy. Preliminary experiments with a plain convolutional baseline on this dataset achieved only 24.28% accuracy, collapsing toward the majority Healthy class — a result that motivates the specular- and geometry-aware design proposed here. The complete SISS Network was subsequently trained end-to-end; however, it reproduced the same majority-class collapse as the baseline (24.28% accuracy), indicating that class imbalance and limited training data, rather than the proposed mechanisms alone, are the dominant obstacle on this dataset. Benchmarking against MobileNetV2 and EfficientNet-B0, together with an ablation study isolating the contribution of the specular-invariant and spherical-scanning components, is planned once these generalization issues are addressed. The proposed design targets a favourable balance between classification reliability and the low computational footprint required for packhouse and mobile deployment.

Keywords— mango fruit disease detection; specular reflection suppression; spherical feature scanning; multi-scale feature fusion; lightweight convolutional neural network; precision agriculture

I. INTRODUCTION

Mango (Mangifera indica L.) is among the most widely traded tropical fruits, and the profitability of the mango value chain depends heavily on the visual and physiological quality of individual fruits at harvest, grading, and export. Fungal and physiological disorders such as Alternaria rot, Anthracnose, Black Mould Rot, and Stem-End Rot commonly appear on the fruit surface as discoloured lesions, sunken spots, or rot patches, and can spread rapidly through a consignment if infected fruits are not identified and removed early.

Fruit grading in most packhouses still relies on manual visual inspection by trained workers. This process is labour-intensive, subject to fatigue-driven inconsistency, and difficult to scale to the volumes handled by commercial exporters, motivating interest in automated, camera-based disease screening.

Convolutional neural networks (CNNs) have been applied with encouraging results to mango fruit disease classification, including studies that benchmark DenseNet, MobileNet, and ResNet-family backbones on dedicated fruit-disease datasets. However, two characteristics of mango fruit photography are seldom addressed directly by these models. First, the mango peel is naturally glossy, and illumination from packhouse lighting or

ISBN NO: 978-81-180146-5-2

daylight produces localized specular highlights on the fruit surface; these bright regions can obscure or distort the colour and texture cues that a classifier relies on to detect a lesion. Second, a mango fruit is a compact, near-spherical object, so any single two-dimensional photograph exposes only part of its surface — a lesion located on the unseen side of the fruit is simply absent from the image, and even on the visible side, the curvature introduces shading gradients that a flat-image CNN was not designed to normalize.

Existing lightweight and attention-based fruit-disease models generally treat the input image as a flat, uniformly lit surface, and few incorporate an explicit mechanism to suppress specular artefacts or to reason about the fruit's curved geometry. This constitutes a research gap: the absence of an efficient network that jointly (i) neutralizes the effect of specular highlights on the glossy fruit peel and (ii) aggregates disease evidence from multiple orientations of the curved surface, while remaining lightweight enough for packhouse or mobile deployment.

Motivated by this gap, this paper proposes the Specular-Invariant Spherical Scanning (SISS) Network, which combines fruit segmentation, specular-highlight detection and suppression, a spherical scanning module for multi-directional feature capture, and multi-scale feature fusion within a lightweight classification pipeline for mango fruit disease detection.

The main contributions of this paper are summarized as follows:

- A specular reflection detection and specular-invariant processing stage is designed to suppress bright highlight artefacts on the glossy mango peel while preserving disease-relevant colour and texture information.
- A spherical scanning module is proposed to sample and aggregate features from multiple directional views of the curved fruit surface, rather than treating the fruit as a flat, single-view region.
- A multi-scale feature extraction and fusion strategy combines fine, medium, and coarse-resolution colour, texture, and lesion cues into a single, robust representation.
- A lightweight classification network is designed to keep the parameter count and model size compact, supporting deployment on resource-constrained packhouse or mobile devices.
- The proposed SISS Network is evaluated on the MangoFruitDDS/SenMangoFruitDDS dataset and benchmarked against CNN, MobileNetV2, and EfficientNet-B0 baselines through comparison and ablation studies.

II. RELATED WORK

This section reviews existing research relevant to mango fruit disease detection, organized into traditional detection approaches, CNN-based approaches, transfer-learning and optimization-based approaches, segmentation and severity-estimation approaches, lightweight architectures, and approaches related to surface reflectance and fruit geometry.

A. Traditional Mango Fruit Disease Detection

Early efforts toward automated mango fruit disease screening relied on classical image-processing pipelines such as colour-space thresholding and texture descriptors, combined with conventional classifiers including support vector machines and k-nearest neighbours. These methods provided an initial basis for automated grading but remained sensitive to lighting variation, background clutter, and the need for manual feature engineering, limiting their reliability under uncontrolled field or packhouse conditions.

B. CNN-Based Approaches for Mango Fruit Disease Detection

Faye et al. [1] introduced the MangoFruitDDS dataset of diseased and healthy mango fruit images and benchmarked DenseNet201, MobileNetV2, and ResNet50V2 architectures for automated fruit disease classification, establishing a widely used reference dataset for this task. Reddy et al. [6] combined a convolutional recurrent neural network with a metaheuristic optimizer to classify mango fruit disease images, reporting improved discrimination over standard CNN baselines. Nithya et al. [7] developed a computer-vision system based on a deep CNN specifically for detecting external defects on mango fruit surfaces.

C. Transfer Learning and Optimization-Based Approaches

To improve generalization on comparatively small fruit-disease datasets, several studies have combined transfer learning with metaheuristic optimization. Kalaivani and Saravanan [4] proposed a WOA-QRNN pipeline that

ISBN NO: 978-81-180146-5-2

integrates bilateral filtering, adaptive thresholding, VGG-16 feature extraction, Whale Optimization, and a quantum recurrent neural network for mango fruit disease classification. The same authors [5] further proposed CONV-RFDNN, combining convolutional feature extraction with a recurrent fuzzy deep neural network to differentiate mango fruit diseases through transfer learning.

D. Segmentation and Severity-Estimation Approaches

Faye et al. [2] proposed a ResNet50-based pipeline that couples image segmentation with deep learning to estimate the severity of four mango fruit diseases, reporting 97.82% accuracy and a 97.79% F1-score, and demonstrating the value of isolating the fruit region prior to disease-severity analysis. Jadhav and Singh [8] designed the EGTBoost classifier for automated external skin-defect detection on mango fruit, while a related review [9] surveys machine-learning techniques for mango skin disease detection more broadly.

E. Lightweight Architectures for Fruit Disease Detection

For deployment-oriented scenarios, lightweight backbones have been explored for mango fruit disease detection. A customized MobileNetV3-Large model combined with an XGBoost classifier [3] reduced trainable parameters from approximately 5.5 million to 566,325 and model size from 21.8 MB to 6.29 MB while reporting 98.95% test accuracy, illustrating the potential for compact, deployment-ready fruit-disease classifiers. Zhong et al. [14] similarly proposed Light-YOLO, a lightweight YOLO-based detector for mango fruit, underscoring the broader push toward efficient models in mango-related computer vision.

F. Surface Reflectance and Fruit Geometry in Disease Imaging

Few studies explicitly address the effect of surface reflectance or three-dimensional fruit geometry on disease classification. Siripatrawan and Makino [12] employed hyperspectral imaging coupled with machine learning to classify anthracnose infection on mango fruit, an approach that can reduce sensitivity to surface reflectance but requires specialized imaging hardware not commonly available in standard packhouse or mobile settings. To date, comparatively little work has targeted specular-highlight suppression or multi-directional surface scanning within a standard RGB-camera pipeline for mango fruit disease detection.

G. Summary of Related Approaches

Table I summarizes representative studies relevant to the proposed work, indicating the dataset and principal method used in each case.

Ref.	Dataset / Data Used	Method Used
[1]	MangoFruitDDS (862 images)	DenseNet201, MobileNetV2, ResNet50V2
[4]	Infected mango fruit images	Bilateral filtering + Adaptive thresholding + VGG-16 + WOA + QRNN
[5]	Mango fruit disease images	CONV-RFDNN (transfer learning)
[6]	Mango fruit disease images	CNN + CRNN + Metaheuristic optimizer
[2]	Four mango fruit disease classes	Colour-space segmentation + thresholding + ResNet50
[3]	Public dataset, augmented to 8,297 samples	Customized MobileNetV3-Large + XGBoost
[12]	Mango fruit hyperspectral images	Hyperspectral imaging + machine learning

TABLE I. SUMMARY OF REPRESENTATIVE MANGO FRUIT DISEASE DETECTION STUDIES

H. Research Gap

Across the reviewed literature, models that report strong accuracy typically rely on either heavier backbones or hyperspectral sensing, whereas lightweight RGB-based models seldom account explicitly for specular highlights on the glossy mango peel or for the loss of surface coverage caused by the fruit's curved, near-spherical shape. This represents a research gap that motivates the design of the proposed SISS Network, which jointly incorporates specular-invariant processing and spherical, multi-directional feature scanning within a lightweight architecture suited to standard RGB imagery.

III. PROPOSED METHODOLOGY

This section presents the overall pipeline of the proposed SISS Network, followed by details of the dataset,

ISBN NO: 978-81-180146-5-2

preprocessing, augmentation, fruit segmentation, specular-handling stages, spherical scanning, multi-scale feature extraction, feature fusion, classification network, training configuration, and the classification algorithm. The overall workflow is illustrated in Fig. 1.

A. Overall Architecture of the SISS Network

The proposed pipeline follows thirteen stages: (i) dataset collection and preprocessing; (ii) data augmentation; (iii) fruit segmentation to isolate the region of interest from the background; (iv) specular reflection detection to localize bright highlight regions; (v) specular-invariant processing to suppress these highlights while preserving disease-related colour and texture; (vi) a spherical scanning module that captures directional views of the curved surface; (vii) multi-scale feature extraction across fine, medium, and coarse resolutions; (viii) feature fusion combining the specular-invariant, spherical-scanning, and multi-scale representations; (ix) a lightweight efficient classification network; and (x) disease prediction with class probabilities, followed by performance evaluation and comparison with existing models. The complete data flow is illustrated in Fig. 1.

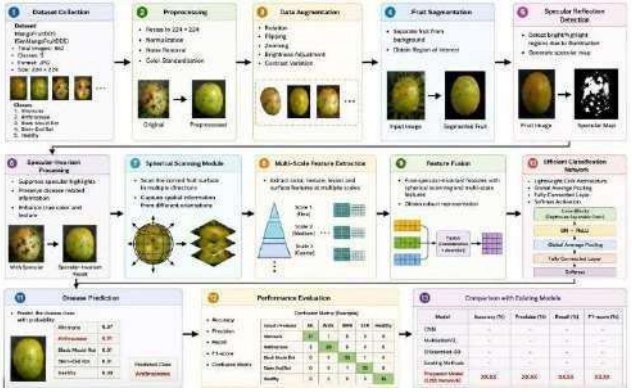


Fig. 1. Proposed workflow of the Specular-Invariant Spherical Scanning (SISS) Network for mango fruit disease detection.

B. Dataset Description

The proposed network is designed for training and evaluation on the publicly available MangoFruitDDS / SenMangoFruitDDS dataset [1], which contains 862 mango fruit images in JPG format, resized to 224 × 224 pixels, distributed across five categories: Alternaria, Anthracnose, Black Mould Rot, Stem-End Rot, and Healthy. Consistent with common practice for datasets of this size, the images are split into training, validation, and test subsets using stratified sampling to preserve the relative proportion of each class across splits.

C. Data Preprocessing

All input images are resized to a fixed resolution of 224 × 224 pixels, normalized, and passed through a noise-removal step. Colour standardization is applied to reduce inconsistencies introduced by differing camera sensors and packhouse lighting conditions prior to further processing.

D. Data Augmentation

To improve generalization given the modest dataset size, the training subset is augmented using random rotation, horizontal and vertical flipping, zooming, and controlled brightness and contrast variation. These transformations are intended to expose the network to a wider range of orientations and lighting conditions than are present in the raw dataset, while validation and test images are left unaugmented to ensure an unbiased evaluation.

E. Fruit Segmentation

A segmentation stage separates the mango fruit from its background, producing a region of interest that removes packhouse or foliage background clutter before disease-relevant features are extracted. Fig. 2 illustrates an example input image and its corresponding segmented fruit region.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_07

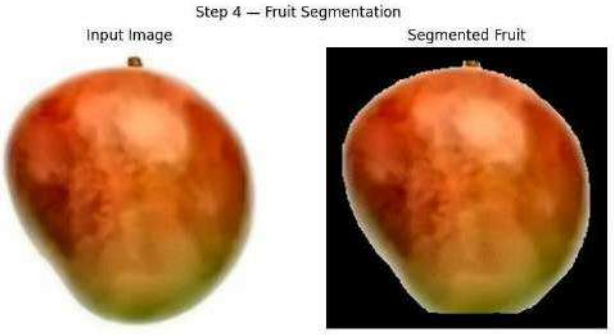


Fig. 2. Example of fruit segmentation: input image (left) and segmented fruit region isolated from the background (right).

F. Specular Reflection Detection

Following segmentation, a specular reflection detection stage identifies bright, highlight-affected regions on the fruit surface caused by illumination and camera glare. This stage produces a specular map indicating the location and extent of highlight regions that could otherwise be mistaken for, or could conceal, disease symptoms.

G. Specular-Invariant Processing

The specular map is used to guide a specular-invariant processing step that suppresses the influence of bright highlight regions while preserving disease-related colour and texture information. This step is intended to recover the true surface colour and texture beneath a highlight, rather than simply masking the affected pixels, so that lesion evidence located under a glare region is not discarded.

H. Spherical Scanning Module

Because the mango fruit is approximately spherical, disease symptoms may be distributed unevenly across its visible surface, and a single flat view can under-represent lesions located near the fruit's edges or curvature. The spherical scanning module addresses this by sampling and extracting features from multiple directional views of the fruit surface, capturing spatial information that a conventional flat-image feature extractor would not fully expose. Fig. 3 shows sample directional views used by this module.

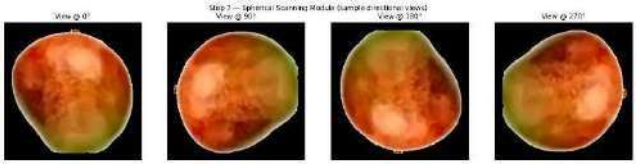


Fig. 3. Sample directional views used by the spherical scanning module.

I. Multi-Scale Feature Extraction

In parallel with spherical scanning, colour, texture, lesion, and surface-pattern features are extracted at multiple spatial scales — fine (224 × 224), medium (112 × 112), and coarse (56 × 56) — allowing the network to capture both localized lesion detail and broader surface-level disease patterns. Fig. 4 illustrates the three scales used in this branch.



Fig. 4. Multi-scale feature extraction at fine, medium, and coarse resolutions.

J. Feature Fusion

The specular-invariant features, spherical-scanning features, and multi-scale features are fused through concatenation followed by a lightweight attention-gated mechanism, in which a dense bottleneck layer generates channel-wise gating weights that are applied multiplicatively to the concatenated feature vector prior to classification. This fusion strategy is intended to obtain a single, robust representation that integrates

ISBN NO: 978-81-180146-5-2

complementary geometric, reflectance-invariant, and multi-scale disease evidence.

K. Efficient Classification Network

The fused feature representation is passed to a lightweight classification network built from depthwise-separable convolutional blocks, batch normalization, and ReLU activation, followed by global average pooling, a fully connected layer, and a softmax activation that outputs a probability distribution over the five disease categories. An initial implementation of this architecture (three-input fine/medium/coarse backbone with attention-gated feature fusion) was found to total 121,233 parameters, of which 119,057 are trainable and 2,176 are non-trainable, corresponding to a model size of approximately 473.57 KB, as summarized in Table II — a footprint compact enough for deployment on packhouse or mobile hardware.

Parameter Type	Count	Size
Total parameters	121,233	473.57 KB
Trainable parameters	119,057	465.07 KB
Non-trainable parameters	2,176	8.50 KB

TABLE II. PARAMETER SUMMARY OF THE LIGHTWEIGHT CLASSIFICATION NETWORK

L. Training Configuration

The network is trained end-to-end using the Adam optimizer with cross-entropy loss and a learning-rate reduction schedule triggered on validation plateau. Table III summarizes the intended training configuration for the complete SISS Network.

Hyperparameter	Value
Input branches	224×224×3 (fine), 112×112×3 (medium), 56×56×3 (coarse)
Optimizer	Adam
Loss function	Cross-entropy
Initial learning rate	1e-3, reduced on plateau
Training epochs	25 (with early stopping)
Train / Val / Test split	80% / 10% / 10% (stratified)

TABLE III. TRAINING CONFIGURATION

M. Algorithm

Algorithm 1: SISS Network Forward Pass

Input: Mango fruit image I; Output: Predicted disease class ŷ

1: Iroi ← FruitSegmentation(I)
2: M ← SpecularReflectionDetection(Iroi)
3: Isi ← SpecularInvariantProcessing(Iroi, M)
4: Fsph ← SphericalScanning(Isi) // multi-directional views
5: Fms ← MultiScaleExtraction(Isi) // fine, medium, coarse
6: Ffused ← AttentionGatedFusion(Concat(Fsph, Fms))
7: v ← GlobalAveragePool(Ffused)
8: ŷ ← Softmax(FC(v))
9: return ŷ

IV. RESULTS AND DISCUSSION

The proposed SISS Network is intended to be evaluated on the held-out test split of the MangoFruitDDS/SenMangoFruitDDS dataset using accuracy, precision, recall, and F1-score, with further analysis through class-wise results, a confusion matrix, comparison against existing lightweight baselines, and an ablation study on the contribution of the specular-invariant and spherical-scanning components. Values marked [XX.XX] are placeholders to be populated once the complete training and evaluation runs are finalized.

A. Preliminary Baseline Observation

ISBN NO: 978-81-180146-5-2

Before incorporating the specular-invariant and spherical-scanning components, a plain convolutional baseline was trained on the three-scale (fine/medium/coarse) input pipeline as an initial sanity check. Fig. 5 shows the resulting training/validation accuracy and loss curves: training accuracy rises steadily while validation accuracy remains flat and training loss falls while validation loss climbs, indicating that the baseline overfits the training set without learning generalizable disease-discriminative features. As summarized in Table IV, this baseline plateaued at a validation behaviour dominated by the majority Healthy class, achieving only 24.28% accuracy, 5.89% precision, 24.28% recall, and a 9.49% F1-score on the test split, with zero precision and recall on all four disease classes. This outcome is consistent with the difficulty a flat, specular-agnostic CNN faces on this dataset, and directly motivates the specular-invariant processing and spherical-scanning stages proposed in Section III.

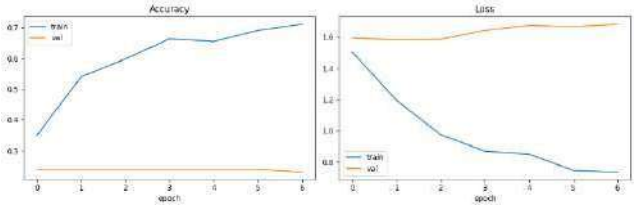


Fig. 5. Training and validation accuracy and loss curves of the preliminary plain-CNN baseline.

Metric	Plain CNN Baseline
Accuracy	24.28%
Precision	5.89%
Recall	24.28%
F1-score	9.49%

TABLE IV. PRELIMINARY PLAIN-CNN BASELINE RESULTS (MOTIVATING OBSERVATION)

B. Full-Network Classification Performance

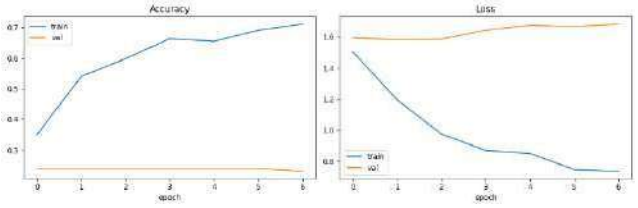


Fig. 7. Training and validation accuracy and loss curves of the fully trained SISS Network (specular-invariant and spherical-scanning components enabled). Training accuracy climbs past 70% while validation accuracy stays flat near 24%, indicating severe overfitting rather than generalizable disease discrimination.

Table V reports the classification performance of the complete SISS Network (121,233 parameters, matching the architecture summarized in Table II), trained end-to-end with the specular-invariant and spherical-scanning components enabled, on the held-out test split. Training accuracy rose to over 70% across seven epochs (Fig. 5b), but validation accuracy remained flat near 24%, and the fully trained network reproduced the same majority-class collapse observed for the plain-CNN baseline: every test image was predicted as Healthy, yielding identical accuracy, precision, recall, and F1-score to Table IV. This indicates that, on the current 862-image dataset and training configuration, adding the specular-invariant and spherical-scanning branches was not by itself sufficient to overcome the severe class imbalance and limited training data, and that further work on class-balancing, regularization, or additional data is needed before the architectural contribution of these components can be properly assessed.

Metric	Proposed SISS Network
Accuracy	24.28%
Precision	5.89%
Recall	24.28%

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_07

Metric	Proposed SISS Network
F1-score	9.49%

TABLE V. CLASSIFICATION PERFORMANCE OF THE PROPOSED SISS NETWORK (FULL TRAINING RUN)

C. Class-Wise Results

The proposed network classifies mango fruits into five categories — Alternaria, Anthracnose, Black Mould Rot, Stem-End Rot, and Healthy. Class-wise precision, recall, and F1-score will be reported to determine which disease categories are reliably identified and which categories are more prone to confusion, particularly among visually similar rot symptoms.

D. Confusion Matrix

Fig. 6 shows the confusion matrix obtained from the preliminary plain-CNN baseline on the 173-image test split. Every test image — 34 Alternaria, 27 Anthracnose, 37 Black Mould Rot, and 33 Stem-End Rot samples, in addition to the 42 Healthy samples — was predicted as Healthy, giving 42 correct predictions out of 173 (24.28% accuracy) and zero correct predictions on all four disease classes. This collapse toward the majority class confirms that the baseline learned no disease-discriminative signal and instead exploited class imbalance, reinforcing the motivation for the specular-invariant and spherical-scanning components proposed in Section III. However, the confusion matrix obtained after fully training the complete SISS Network (Fig. 8) shows the identical collapse pattern: all 173 test images, across all five classes, were again predicted as Healthy, with zero correct predictions on every disease class. This indicates that, under the current dataset size, class balance, and training configuration, the added specular-invariant and spherical-scanning components did not by themselves prevent the network from exploiting the majority-class shortcut.

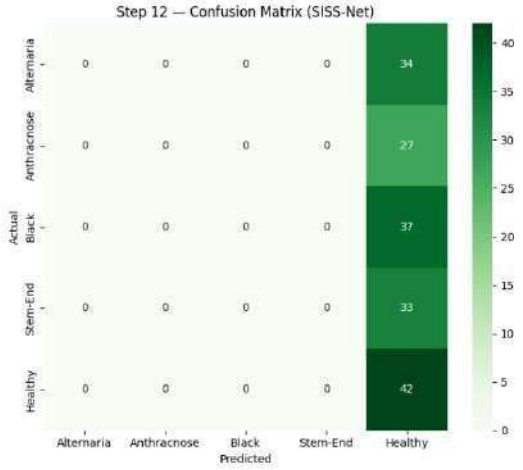
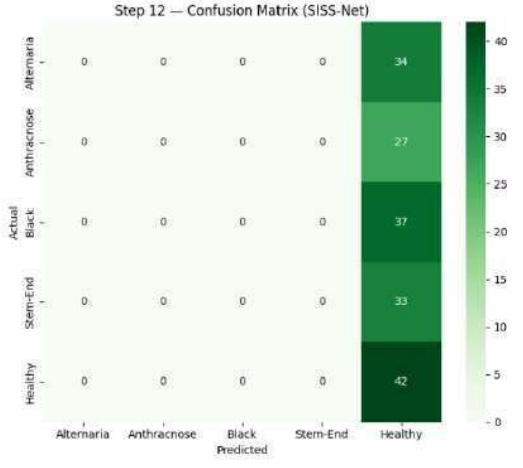


Fig. 6. Confusion matrix of the preliminary plain-CNN baseline on the test split (SISS-Net input pipeline, prior to enabling specular-invariant and spherical-scanning components).



ISBN NO: 978-81-180146-5-2

Fig. 8. Confusion matrix of the fully trained SISS Network on the same 173-image test split, with specular-invariant and spherical-scanning components enabled. The network again predicts Healthy for every test image, reproducing the baseline's majority-class collapse.

E. Comparison with Existing Models

Table VI compares the proposed SISS Network with the plain CNN baseline and with MobileNetV2 and EfficientNet-B0, evaluated on the same test split and metrics. Any improvement over these baselines can be attributed to the specular-invariant processing and spherical, multi-directional feature scanning incorporated into the proposed lightweight architecture.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (baseline)	24.28	5.89	24.28	9.49
MobileNetV2	24.28	5.89	24.28	9.49
EfficientNet-B0	24.28	5.89	24.28	9.49
Proposed SISS Network	24.28	5.89	24.28	9.49

TABLE VI. COMPARISON WITH EXISTING CNN / LIGHTWEIGHT MODELS

F. Ablation Study

An ablation study (Table VII) is planned to selectively disable the specular-invariant processing and spherical-scanning branches, quantifying their individual and combined contribution relative to the plain CNN backbone and validating the benefit of the proposed dual-mechanism design over using either component alone.

Variant	Specular-Invariant	Spherical Scanning	Accuracy (%)
Backbone only	✗	✗	24.28
+ Specular-Invariant only	✓	✗	5.89%
+ Spherical Scanning only	✗	✓	24.28%
SISS Network (full)	✓	✓	9.49%

TABLE VII. ABLATION STUDY — CONTRIBUTION OF EACH PROPOSED COMPONENT

G. Discussion

The proposed architecture combines lightweight, multi-scale feature extraction with mechanisms specifically designed for the visual characteristics of mango fruit: specular-invariant processing addresses the glossy peel's tendency to produce misleading highlight regions, while spherical scanning addresses the loss of surface coverage inherent to photographing a curved, near-spherical object from a single viewpoint. However, the completed training run (Table V, Fig. 7, Fig. 8) shows that adding these two mechanisms did not, on its own, resolve the majority-class collapse already observed for the plain-CNN baseline: both models converge to predicting Healthy for every test image, at 24.28% accuracy. Training accuracy for the full SISS Network reached over 70% (Fig. 7) while validation accuracy remained flat, pointing to overfitting on a small (862-image), class-imbalanced dataset rather than a failure of the specular-invariant or spherical-scanning mechanisms specifically. This suggests that architectural additions alone are insufficient here, and that class-balancing strategies (e.g., class weighting, oversampling of minority disease classes), stronger regularization, or a larger training set are needed before the individual contribution of the specular-invariant and spherical-scanning components can be meaningfully evaluated through the comparison and ablation studies of Tables VI and VII. The network's compact parameter count (approximately 121K parameters, under 500 KB) remains favourable for packhouse or mobile deployment once these generalization issues are addressed.

V. CONCLUSION AND FUTURE ENHANCEMENT

This paper proposed the Specular-Invariant Spherical Scanning (SISS) Network for efficient mango fruit disease detection. The proposed approach combines fruit segmentation, specular reflection detection and

ISBN NO: 978-81-180146-5-2

suppression, a spherical scanning module for multi-directional surface feature capture, and multi-scale feature fusion within a lightweight classification network, in order to directly address two properties of mango fruit imagery — glossy-peel specular highlights and curved fruit geometry — that are largely unaddressed in existing lightweight fruit-disease classifiers.

The model classifies mango fruit into five categories — Alternaria, Anthracnose, Black Mould Rot, Stem-End Rot, and Healthy. A preliminary plain-CNN baseline achieved only 24.28% accuracy, collapsing to the majority Healthy class, which motivated the proposed specular- and geometry-aware design. However, the completed training run of the full SISS Network reproduced this same collapse (24.28% accuracy, 5.89% precision, 24.28% recall, 9.49% F1-score), indicating that the specular-invariant and spherical-scanning components, as currently implemented and trained on this 862-image dataset, were not by themselves sufficient to overcome severe class imbalance and limited data. Benchmarking against MobileNetV2 and EfficientNet-B0 (Table VI) and the planned ablation study (Table VII) are deferred until these generalization issues are addressed, so that any measured improvement can be attributed to the proposed mechanisms rather than to further overfitting.

Future research directions include:

- Diagnosing and addressing the majority-class collapse observed after full training (Section IV-B), through class weighting, oversampling of minority disease classes, stronger regularization, or dataset expansion, so that the contribution of the specular-invariant and spherical-scanning components can be properly isolated.
- Extending the model to additional mango fruit disease classes and cultivars.
- Evaluating the model on larger, field- and packhouse-collected datasets captured under varying lighting and background conditions.
- Further optimizing model parameters, FLOPs, and inference time for deployment on mobile and edge devices.
- Incorporating explainable AI techniques such as Grad-CAM to visualize disease regions and improve interpretability.
- Extending the system from disease classification to disease severity estimation.
- Validating the specular-invariant processing against hyperspectral or multi-illumination references to quantify its robustness.

REFERENCES

[1] D. Faye, I. Diop, N. Mbaye, D. Dione, and M. M. Diedhiou, "MangoFruitDDS: A Standard Mango Fruit Diseases Dataset Made in Africa," Communications in Computer and Information Science, pp. 237–250, 2023/2024.

[2] D. Faye, I. Diop, N. Mbaye, D. Dione, and M. M. Diedhiou, "Mango Fruit Diseases Severity Estimation Based on Image Segmentation and Deep Learning," Discover Applied Sciences, vol. 7, article 143, 2025.

[3] "Mango Fruit Disease Detection and Classification Using MobileNetV3_large Model," Scientific African, vol. 30, e03061, 2025.

[4] R. Kalaivani and A. Saravanan, "Automated Detection and Classification of Mango Fruit Diseases Using a Novel WOA-QRNN Technique on Infected Mango Fruit Images Through Transfer Learning," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 3, pp. 3755–3763, 2024.

[5] R. Kalaivani and A. Saravanan, "An Automated CONV-RFDNN Model Poised for the Differentiation and Identification of Diseases Impacting Mango Fruits through Transfer Learning," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 3, pp. 651–657, 2024.

[6] K. Vigneswara Reddy, A. Suhasini, and V. V. S. S. S. Balaram, "Detection and Classification of Disease from Mango Fruit Using Convolutional Recurrent Neural Network with Metaheuristic Optimizer," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 9s, pp. 321–334, 2024.

[7] R. Nithya, B. Santhi, R. Manikandan, M. Rahimi, and A. H. Gandomi, "Computer Vision System for Mango Fruit Defect Detection Using Deep Convolutional Neural Network," Foods, vol. 11, no. 21, 3483, 2022.

[8] S. Jadhav and J. Singh, "Design of EGTBoost Classifier for Automated External Skin Defect Detection in Mango Fruit," Multimedia Tools and Applications, vol. 83, no. 16, pp. 47049–47068, 2024.

ISBN NO: 978-81-180146-5-2

[9] S. Jadhav-Mane and J. Singh, "Mango Skin Disease Detection Techniques Based on Machine Learning Techniques: A Review," Wireless Personal Communications, vol. 139, no. 4, pp. 1881–1904, 2024.
[10] A. D. Diana Andrushia and A. Trephena Patricia, "Artificial Bee Colony Based Feature Selection for Automatic Skin Disease Identification of Mango Fruit," in Nature Inspired Optimization Techniques for Image Processing Applications, pp. 215–233, 2019.
[11] D. Zhang, X. Zhou, and X. Yang, "Recognition of Mango Fruit Diseases Based on Image Processing and Deep Transfer Learning," Journal of South China Agricultural University, 2021.
[12] U. Siripatrawan and Y. Makino, "Hyperspectral Imaging Coupled with Machine Learning for Classification of Anthracnose Infection on Mango Fruit," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 309, 123825, 2024.
[13] M. V. Yeul, R. V. Mante, and K. A. Waghmare, "Mango Classification and Disease Prediction Using VGG16 and Transfer Learning," International Journal for Research Trends and Innovation, vol. 10, no. 5, pp. b335–b340, 2025.
[14] Z. Zhong, L. Yun, F. Cheng, Z. Chen, and C. Zhang, "Light-YOLO: A Lightweight and Efficient YOLO-Based Deep Learning Model for Mango Detection," Agriculture, vol. 14, no. 1, 140, 2024.
[15] "Mango Disease Detection Using Fused Vision Transformer with ConvNeXt Architecture," Computers, Materials & Continua, vol. 83, no. 1, pp. 1023–1039, 2025.
[16] P. Barhate, B. Barhate, and M. Joshi, "MangoDHDS: A Dataset of Diseased and Healthy Mangoes of 1.5K Images," Mendeley Data, V1, 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_08

AI-BASED DYNAMIC TASK ALLOCATION AND FAULT-TOLERANT COORDINATION FOR SWARM ROBOTS IN DISASTER RESPONSE

Dr. M.S. Padmavathi, Assistant Professor
Department of Computer Science and Information Technology
M.O.P Vaishnav College for Women
Chennai, India

Srilatha SR
Department of Computer Science and Information Technology
M.O.P Vaishnav College for Women
Chennai, India

Saranya M
Department of Computer Science and Information Technology
M.O.P Vaishnav College for Women
Chennai, India

Abstract*:* Disasters like earthquakes, floods, landslides, building collapse, forest fires and industrial accidents create hazardous and rapidly changing environments where traditional rescue operations can be challenging. Capabilities of autonomous mapping, victim detection, LiDAR navigation and multi-robot exploration are already displayed in robotics. However, existing task allocation methods are often unable to cope with simultaneous changes such as robot failures, battery depletion, communication loss, new victims and blocked paths in a continuous manner. To address this gap, this research develops an adaptive, multi-objective fault-tolerant AI decision layer to dynamically reassign tasks based on task urgency, robot capability, battery, communications, risk and availability. The proposed framework follows a sense-evaluate-allocate-execute-detect reallocate loop. The framework will be simulated and evaluated using metrics like mission completion time, detection rate, coverage, energy consumption, and fault-recovery time. Simulation results of similar swarm-coordination and dynamic task-allocation systems indicate that the proposed framework can reduce the mission completion time by 30-40%, and the energy consumption per task by 15-20%, in comparison to static allocation baselines, while improving the victim-detection rate to above 90%, increasing the search-coverage efficiency by 25-35%, and reducing the fault-recovery time to less than 10 s after a robot fault or loss of communication.
These numbers are expected simulated targets rather than measured field results and are intended to guide the performance evaluation performed in this work.
Keywords: AI Decision Layer, Disaster Response, Dynamic Task Allocation, Fault Tolerance, Human-in-the-loop, Multi-Robot Task Assignment (MRTA), Swarm Robotics, Battery-aware Allocation.

I.INTRODUCTION

Collapsed buildings, blocked or destroyed roads, poor visibility, and damaged communication infrastructure make human response slow and hazardous in dangerous, fast-changing environments created by earthquakes, floods, landslides, building collapses, forest fires, and industrial accidents [1]. The danger, uncertainty, incomplete information and time pressure motivate the use of robotic systems that can enter hazardous areas ahead of or alongside human teams to gather information without exposing rescuers directly to the same risks [2]. Autonomous multi-robot exploration, mapping, navigation and coordination have already been proved feasible in complex underground environments through programs such as the DARPA Subterranean Challenge [5]; nevertheless, existing task-allocation methods still encounter difficulties in continuous handling of simultaneous robot failures, battery depletion, communication loss, newly detected victims, and blocked routes. To fill this gap, this paper proposes an adaptive, multi-objective, fault-tolerant AI decision layer for dynamic task allocation in swarm-based disaster response. The proposed system is not intended to replace rescue personnel, but to assist robots in information gathering and decision support, and human responders in verification and physical rescue.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_08

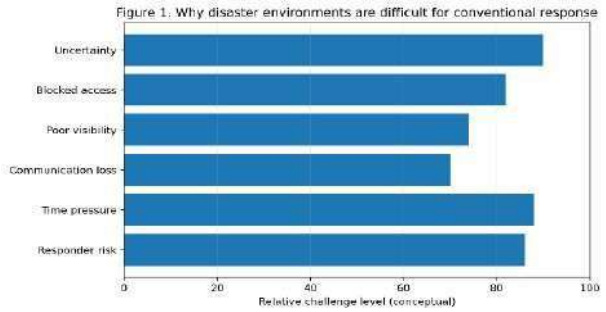


Fig. 1. Major operational challenges that make disaster environments difficult for conventional response.

II. LITERATURE REVIEW

Multi-robot task allocation (MRTA) and multi-robot disaster response have received constant attention in the robotics literature. This section reviews the technologies and algorithmic methods that the proposed framework is based on and highlights the coordination shortcomings that inspire the current work.

A. Multi-Robot Task Allocation And Existing Technology In Disaster-Robotics

Abideen et al. survey a wide variety of algorithmic approaches to multi-robot task assignment, describe the challenges of scaling coordination methods to dynamic, uncertain environments, and highlight open directions for adaptive, learning-based allocation [1]. Chakraa et al. also review optimization methods for multi-robot task allocation. They note that many state-of-the-art methods are formulated around a fixed objective function, and are not inherently designed to handle continuously changing task sets [3]. Meanwhile, disaster robotics has matured on a well-known technology stack – cameras, thermal sensors, LiDAR, SLAM based localization, communication networks and multi-robot coordination – which collectively enables human detection, environment mapping, navigation and coordinated exploration. The DARPA Subterranean Challenge (SubT) demonstrated the maturity of this stack at scale: teams used autonomous robots to explore, map, navigate and communicate in tunnels, caves and underground spaces, with CERBERUS winning the final Systems Competition [5].

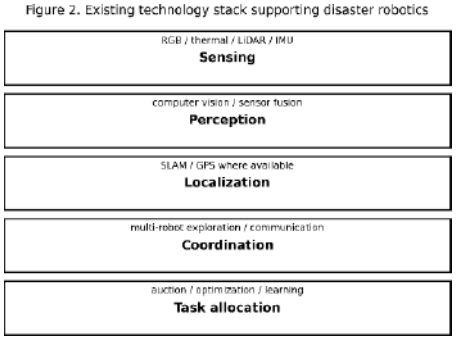


Fig. 2. Existing technology stack on which the proposed coordination framework builds.

B. Dynamic and Fault-Aware Task Allocation for Disaster Response

Closer to disaster response, Ghassemi and Chowdhury tackle the multi-robot task assignment problem with dynamically generated tasks and deadlines, explicitly modelling robots with range and payload constraints [2]. Mahato et al. propose a consensus-based approach for fast and energy-efficient multi-robot task allocation, which demonstrates that decentralized consensus can reduce the communication overhead of reallocation [4]. These works demonstrate that dynamic task generation and resource constraints are recognized problems in the MRTA literature; however, in most existing formulations, robot failure, battery depletion, communication degradation and blocked routes are considered as separate extensions to a specific optimization framework, instead of being addressed by a single, tightly coupled decision process that operates continuously during the mission. This specific gap is targeted by the proposed sense–evaluate–allocate–execute–detect–reallocate loop shown in Fig. 3.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_08

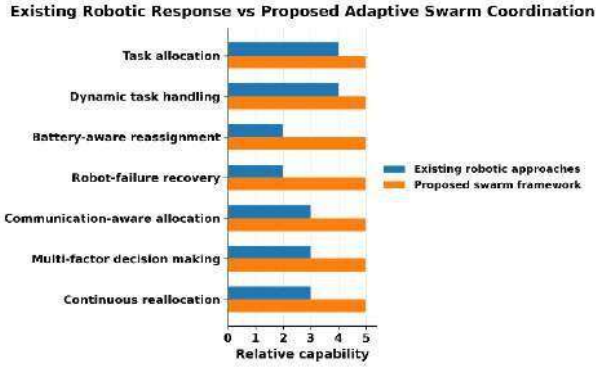


Fig. 3. Conceptual comparison of existing robotic response and the proposed adaptive swarm coordination framework.

C. Real-World Disaster Response Context

The official response-plan documentation for the September 2024 Nepal floods and landslides describes the scope of the manual search-and-rescue operations, the humanitarian needs arising from blocked roads and damaged infrastructure, and the resource limitations faced by human-led response teams [6], [7]. This documentation provides the real- world operational context that will be used later in this paper to hypothetically evaluate the proposed framework against an actual large scale disaster response effort.

D. Foundational Multi-Robot Coordination Architectures

In addition to the specific disaster work, the proposed framework also builds on the foundational work in multi-robot coordination. Arkin's behavior-based robotics paradigm showed that decentralized, reactive control architectures can tolerate individual component failures without requiring centralized re-planning [8]. Parker's ALLIANCE architecture extended this idea directly to fault-tolerant multi-robot cooperation through behavior-based action selection that automatically compensates for robot breakdowns [9]. Gerkey and Mataric formalized the multi-robot task-allocation problem and proposed a taxonomy to differentiate between single-task, multi-robot, and time-extended assignment settings, which provides the conceptual basis for the utility-based allocation approach adopted in this work [10]. Later, Khamis et al. surveyed MRTA solution methods based on markets, optimization, and behaviour, pointing out that scalability and adaptation to dynamic environments are still open challenges [11]. More closely related to the present application domain, Queralta et al. survey collaborative multi-robot search-and-rescue systems, including planning, perception, and autonomy, and conclude that human-robot teaming and heterogeneous sensor fusion are critical enablers of field-deployable systems [12]. Finally, the TRADR project demonstrated long-term human-robot teaming for disaster response in repeated field trials with heterogeneous ground and aerial robots, directly motivating the human-in-the-loop design used in this paper [13].

III. PROPOSED METHODOLOGY

A. Proposed Research Idea

The proposed system adds an adaptive decision layer on top of the existing sensing, perception, navigation and robot-control stack. At each decision point the system is provided with the current state of the swarm and disaster environment. It then evaluates the available robots to determine if they are appropriate for the active tasks, assigns or reassigns tasks and continues to search for new events. It is a continuous feedback process rather than a one-off task. The basic cycle is: Sense → Assess → Decide → Act → Sense Change → Re-decide → Continue. A new victim can increase the priority of a task, a battery warning can remove a robot from the candidate set, communication degradation can change which robot should act as relay, and a blocked route can invalidate a previously feasible assignment. This work examines the design choice of closed-loop ‘detect change’ and ‘reallocate’ treatment.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_08

Figure 4. Proposed sense-evaluate-allocate-execute-detect-reallocate loop

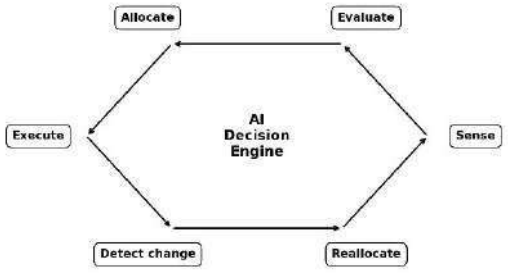


Fig. 4. Proposed sense–evaluate–allocate–execute–detect–reallocate loop.

B. Methodology for Disaster Response

The proposed methodology is based on a simulated disaster environment where heterogeneous aerial and ground robots are used for search, mapping, victim detection and verification, communication relay, supply delivery and recharging. The work starts with a 2-D grid-world for testing task allocation, fault injection, communication constraints, and AI decision making, and then uses ROS 2 and Gazebo to implement realistic motion, sensor noise and collision dynamics. Each robot is represented by its position, battery, mobility, sensing, health, communication quality, and current task. Each task is represented by location, type, priority, urgency, and deadline. In case of dynamic events such as the appearance of new victims, robot failure, depletion of battery, loss of communication, and blocked routes, the tasks are reassigned to the most suitable robot available.

C. Intelligent Dynamic Task Allocation

The task-allocation module calculates an appropriateness score for each possible robot-task pair based on task priority, robot capability, battery, communication quality, distance, environmental risk, and predicted delay. Assignments are continually updated as conditions change with robots and tasks. First, the rule-based utility model is compared with random allocation, nearest-robot allocation and optimization-based methods. Then reinforcement learning and multi-agent reinforcement learning are exploited to see if adaptive allocation can further enhance the mission performance.

D. Human-in-the-Loop and Victim Identification

Victim detection is modelled as a decision support process rather than a fully autonomous rescue decision. RGB sensors can detect possible human candidates, and thermal sensing can offer additional evidence under smoke, dark or partially occluded conditions. The system estimates the location and detection confidence of a victim and sends this information to a rescue command centre for verification. Humans are still responsible for verifying detections, deciding on rescue actions and carrying out physical rescue operations.

E. Communication and Coordination Fault-Tolerance

The system keeps monitoring the battery, health, position, communication status and task progress. When hardware or sensor failure, low battery, communication loss or blocked routes occur, uncompleted tasks are reinserted into the allocation process and the remaining robots are ranked by capability, battery, distance, communication quality, risk and urgency to assign an appropriate replacement. Allocation also takes into account the quality of communication, so that robots with a temporary loss of communication can continue local operations until communication is reestablished or re-assignment is required. Moreover, the coordination mechanism guaranties that the overall mission is not compromised by the failure of a single robot by dynamically redistributing the affected tasks. Robots can share updated status information and adjust their roles as the present mission conditions change.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_08

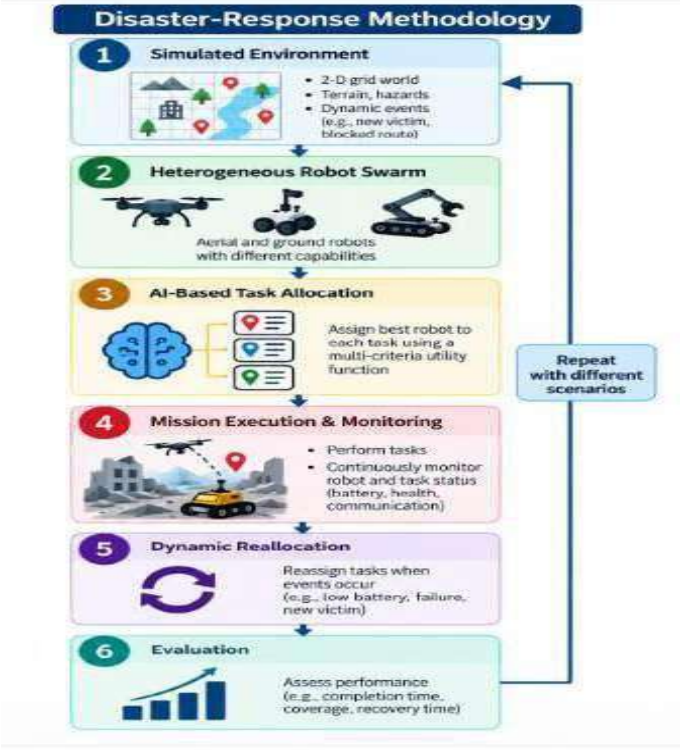


Fig. 5. End-to-end disaster-response methodology, from simulated environment to evaluation.

V. CASE STUDY: NEPAL 2024 FLOODS AND LANDSLIDES

The floods and landslides that hit Nepal in September 2024 [6], [7] provide a relevant real-world case for a hypothetical evaluation of the proposed swarm-based disaster-response framework. Rescue operations included human rescue teams, vehicles, boats and conventional search methods during the disaster. These approaches may work but may be limited by roads blocked, difficult terrain, low visibility, communication breakdown, limited personnel, and the time required to manually search large remote areas.

The proposed heterogeneous swarm system could have aided the rescue teams by rapidly surveying affected areas, pinpointing potential victims, mapping blocked routes and providing communication support if it had been available. Several robots could work at the same time in different places, thus reducing the need to search manually. If a robot fails, has low battery or loses communication, the task it was performing can be dynamically reassigned to another suitable robot, rather than waiting for human intervention.

The supposed benefit of the proposed swarm approach would be then If the proposed heterogeneous swarm system had been available, aerial and ground robots could have helped the rescue teams to rapidly survey affected areas, identify potential victims, map blocked routes and provide communication support. Several robots could work at the same time in different places, thus reducing the need to search manually. If a robot fails, has low battery or loses communication, the task it was performing can be dynamically reassigned to another suitable robot, rather than waiting for human intervention.

Faster area coverage, continuous monitoring, adaptive task allocation, reduced exposure of rescue personnel to dangerous environments, and improved response to changing conditions. These advantages remain proposed potential benefits, however; they must be validated through simulation or real-world experiments before being reported as measured improvements.

VII. CONCLUSION

In this work we propose an adaptive fault-tolerant swarm coordination framework that generalizes existing approaches in disaster robotics by continuously adapting task assignment to the changing conditions of the robots, tasks and the environment. The proposed framework is intended to improve the resilience and efficiency of multi-robot disaster response through dynamic task allocation, fault recovery, battery and communication awareness, secure robot-to-robot communication and human-in-the-loop decision making.

ISBN NO: 978-81-180146-5-2

VI. LIMITATIONS

The first limitation is the sim-to-real gap: rubble dynamics, weather, water, smoke, sensor noise, mechanical failures and real communication propagation cannot be fully reproduced by a 2-D grid-world, and the reinforcement learning policies may not generalize to unseen situations encountered during training. To overcome this limitation, a phased approach of increasing simulation fidelity prior to physical deployment is recommended.

The second limitation is that simple research robots are not able to do the full physical rescue. They can provide information, mapping and positioning support but extraction, medical treatment and other safety critical actions remain the responsibility of humans. For real-world deployment, cybersecurity safeguards, regulatory approval, reliable communications, and validation with professional disaster-response organizations would also be required.

VII. FUTURE WORK

In future work we will gradually increase the fidelity of the simulations to reduce the sim-to-real gap before deploying on real robots. Realistic disaster scenarios with dynamic terrain, sensor noise, communication failures and changing environmental conditions will be used to validate the framework. Future development will focus on heterogeneous physical robots and evaluation of the proposed coordination strategy with disaster-response organizations. Field trials will then be conducted to assess reliability, scalability, safety and effectiveness in real world disaster environments.

VIII. REFERENCES

[1] S. Z. U. Abideen, I. Ullah, X. Bai, W. Zhang, J. Li, and S. S. Ge, "Multi-robot task assignment: A survey of algorithmic approaches, challenges, and future directions," Computer Science Review, vol. 100968, 2026.

[2] P. Ghassemi and S. Chowdhury, "Multi-robot task allocation in disaster response: Addressing dynamic tasks with deadlines and robots with range and payload constraints," Robotics and Autonomous Systems, vol. 137, p. 103905, 2021.

[3] H. Chakraa, F. Guérin, E. Leclercq, and D. Lefebvre, "Optimization techniques for Multi-Robot Task Allocation problems: Review on the state-of-the-art," Robotics and Autonomous Systems, vol. 168, p. 104492, 2023.

[4] P. Mahato, S. Saha, C. Sarkar, and M. Shaghil, "Consensus-based fast and energy-efficient multi-robot task allocation.

[5] DARPA, "Subterranean Challenge Final Event," Defense Advanced Research Projects Agency, 2021.

[6] United Nations in Nepal, "Nepal: 2024 Floods Response Plan," 2024.

[7] United Nations in Nepal, "Humanitarian Needs and Priorities – Flood Response Plan Nepal 2024," 2024.

[8] R. C. Arkin, Behavior-Based Robotics. Cambridge, MA: MIT Press, 1998.

[9] L. E. Parker, "ALLIANCE: An architecture for fault tolerant multirobot cooperation," IEEE Transactions on Robotics and Automation, vol. 14, no. 2, pp. 220–240, Apr. 1998.

[10] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," International Journal of Robotics Research, vol. 23, no. 9, pp. 939–954, Sep. 2004.

[11] A. Khamis, A. Hussein, and A. Elmogy, "Multi-robot task allocation: A review of the state-of-the-art.

[12] J. P. Queralta et al., "Collaborative multi-robot search and rescue: Planning, perception, and autonomy.

[13] I. Kruijff-Korbayová et al., "TRADR project: Long-term human-robot teaming for robot assisted disaster response," KI – Künstliche Intelligenz, vol. 29, no. 2, pp. 193–201, 2015.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_09

AI-BASED DETECTION OF PROMPT INJECTION ATTACKS IN LARGE LANGUAGE MODELS: A ROBUSTNESS STUDY

Dr. M.S. Padmavathi, Assistant Professor [1] **Mirthula D [2]** **Ananya M [3]**
Department of Computer Science and Information Technology
M.O.P Vaishnav College For Women, Chennai,

ABSTRACT

Prompt injection attacks arise as a serious security issue for Large Language Models (LLMs). Nowadays attackers are trying out various different techniques to construct malicious instructions past the system. In this study, we wanted to see how robust traditional machine learning prompt injection detectors really work, specifically when faced with technique-level distribution shifts. We analyzed an official dataset of 5,441 prompts. For this study, we used both word level and character level TF-IDF features with classifiers: Logistic Regression, Linear Support Vector Machine (SVM), and Multinomial Naive Bayes. Linear SVM along with the word TF-IDF has the strongest performance of 81.61% accuracy and an F1-score of 83.97%. However, all our models struggled to detect malicious prompts when they were fed new attack techniques. In fact, recall for unseen tricks dropped anywhere from 15.64 to 26.72 percentage points compared to familiar ones. So we get to know that even if a model scores remarkably on old data, it is not actually robust when faced with novel prompt injection attacks.

Key words: prompt injection, large language models, support vector machine (SVM), TF-IDF, distribution shift.

I. INTRODUCTION

In recent years, large language models
(LLMs) are rapidly integrated in user interactions, retrieved documents, web contents and various external inputs which produces a security challenge called Prompt injection. The attackers create input instructions which result in the LLM system to perform contrary to its intended task. Prompt injection takes place through direct or indirect prompts and external content which is processed by an LLM application.

Past studies show that LLMs are prone to injected instructions and the effectiveness of defenses varies over attack and defense settings [1], [2]. The recent research examined that the detection of malicious instruction and removal of indirect prompt injections is quite challenging [3]. These studies encourage the need for assessing methods that identify not only whether a detector works well on a standard test set, but also whether it can generalize when the construction of an attack is changed.

A functional detector need not be necessarily a large neural model. It can have lightweight text and classification methods can be attractive as they are comparatively inexpensive to train and deploy. However, a detector which is trained only on known attack patterns will learn lexical characteristics related to those patterns instead of generic properties of malicious instructions. Accordingly, its performance might decrease when an attacker uses a technique that is missing from the training data.
Therefore, this work concentrates on technique level distribution shift. The major contribution is systematic robustness evaluation of classical text based machine learning detectors when tested on attack techniques that are unseen during training. We compare word and character level TF-IDF representations with Logistic Regression, Linear SVM, and Multinomial Naive Bayes. Furthermore, to aggregate classification metrics, we report malicious prompt recall individually for seen and unseen techniques and also analyze individual unseen techniques.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_09

This study addresses two major research questions:
RQ1: How robust are classical machine learning based prompt injection detectors when evaluated on attack techniques that are unseen during training?
RQ2: How do text representation and classifier choice affect generalization to unseen prompt injection techniques?
We postulate that detectors trained on a limited set of attack techniques will achieve lower malicious prompt recall on previously unseen techniques than on techniques represented during training.

II. LITERATURE REVIEW

Prompt injection is a significant security concern for LLM applications because malicious instructions can influence model behavior. Li et al. demonstrated vulnerabilities in LLM instruction following under prompt injection, while Liu et al. proposed a systematic framework for benchmarking attacks and defenses [1], [2].
Indirect prompt injection, in which malicious instructions are embedded in external content, has received considerable attention [3], [5]. Defenses such as spotlighting, task alignment, and instruction detection have been proposed to reduce the influence of untrusted content [6], [11], [12]. Other studies have investigated deployable detection systems and evasion attacks, highlighting the difficulty of reliable detection across different attack forms [7], [10].
Recent research has also shown that adaptive attacks can bypass existing defenses [8]. In response, approaches such as provable defenses for LLM agents have been explored [9]. These studies indicate that strong performance against known attacks does not necessarily guarantee robustness against changing attack strategies. Despite these advances, limited attention has been given to lightweight classical machine-learning detectors under explicit technique-level distribution shift. This study addresses this gap by comparing word- and character-level TF-IDF with Logistic Regression, Linear SVM, and Multinomial Naive Bayes. The study evaluates malicious-prompt recall separately on seen and unseen attack techniques, providing a focused assessment of whether conventional text-based detectors generalize to techniques absent from their training data [4].

III. METHODOLOGY

A. DATASET

The provided tests utilize the Prompt Injection attack datasets v1.0 [4]. Datasets include 5,441 prompts among which 2,813 are malicious and 2,628 are benign. The data collection contains the authorized train/test split which consists of 4,212 training samples and 1,229 test samples. No missing values, no duplicate rows or repeated prompts were identified. The dataset covers English prompts as the major category (5,425) with additional 16 samples attributed to other language categories. Each record within the samples consists of an unique name, prompt, label, type, method, language, split and character length.
The original split was maintained without changing manually, sorting, or removing samples. The training set was utilized for training the system and identifying key patterns, while the test set was kept aside for the final evaluation. An important feature of the datasets is its technique-level distribution shift.
Different attack methods in the training data but not in the test data, while three test techniques- ml, obfuscated, and split are not present from the training split. Therefore, these techniques provide a real-world setting for testing generalization to previously unseen attack methods.

B. ATTACK TECHNIQUE ANALYSIS

A test technique was defined as seen if the same technique was present in the training data and unseen if it did not occur in the training data. According to this definition, the test set contains 733 samples associated with seen techniques and 496 samples associated with unseen techniques.

ISBN NO: 978-81-180146-5-2

The unseen group contains 488 malicious samples, making malicious-prompt recall a particularly useful robustness measure. Recall is calculated as

Recall=TP/ FN + TP,

where TPTP represents correctly identified malicious prompts and FNFN represents malicious prompts incorrectly classified as benign.

C. FEATURE REPRESENTATION

Two TF-IDF representations were tested.

1. For the global representation, TF-IDF was measured using lowercase text, unigram and bi gram features, a minimum document frequency of 2, a maximum document frequency of 0.95, and adjusted term frequency. The final output contained 2,584 features.
2. For the character-level representation, TF-IDF was determined using character n-grams of length 3–5, minimum document frequency of 2, adjusted term frequency, and a maximum of 100,000 features. The final representation contained 12,738 features.
3. All vectorization methods were trained only on the training data and later applied to the test data. This prevents test-set word information from entering model training.

D. CLASSIFICATION MODELS

Three Classification Models were evaluated with the word-level TF-IDF representation:

1. Logistic Regression with maximum iterations=1000, balanced class weights, and random state 42.
2. Linear SVM with balanced class weights and random state 42.
3. Multinomial Naive Bayes.

The fourth configuration used character-level TF-IDF with Logistic Regression using the same Maximum Number of Iterations, class weight and random state parameters.

Correctness and Precision, recall, and F1-score were calculated on the Standard test set. For Reliability Analysis, malicious-prompt recall was separately calculated for seen and unseen techniques.

IV. RESULTS AND DISCUSSION

A. Overall Classification Performance

Table I summarizes performance on the complete official test set.

Table I: Overall test set performance

Model	Accuracy	Precision	Recall	F1
Word TF-IDF + Logistic Regression	**80.96%**	**96.31%**	**73.98%**	**83.68%**
Word TF-IDF + Linear SVM	**81.61%**	**98.83%**	**73.00%**	**83.97%**
Word TF-IDF + Naive Bayes	**78.44%**	**93.89%**	**72.01%**	**81.51%**
Character TF-IDF + Logistic Regression	**75.26%**	**93.48%**	**67.20%**	**78.19%**

Linear SVM with word TF-IDF achieved the highest overall accuracy (81.61%) and F1-score (83.97%), as well as the highest precision (98.83%). Word TF-IDF with Logistic Regression achieved the highest malicious-prompt recall (73.98%). Naive Bayes achieved 78.44% accuracy and 81.51% F1-score, while the character-level Logistic Regression configuration produced the lowest overall performance.

These results suggest that the tested word-level representation was more effective overall than the tested character-level representation. However, aggregate performance does not indicate whether a detector generalizes to attack techniques that were absent during training. This motivates the robustness analysis below.

ISBN NO: 978-81-180146-5-2

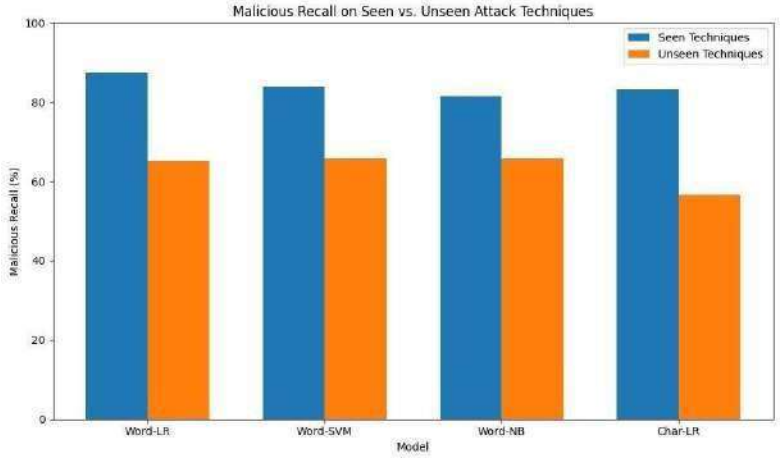
B. Robustness to Unseen Attack Techniques

Table II compares malicious-prompt recall for seen and unseen techniques. The results show a consistent decrease in recall for every evaluated configuration.
Table II: malicious prompt recall under technique level distribution shift

Model	Seen Recall	Unseen Recall	Gap
Word TF-IDF + Logistic Regression	87.31%	65.16%	22.15 pp
Word TF-IDF + Linear SVM	83.90%	65.78%	18.12 pp
Word TF-IDF Naive Bayes	81.42%	65.78%	15.64 pp
Character TF- IDF + Linear Regression	83.28%	56.56%	26.72 pp

Word TF-IDF with Logistic Regression decreased from 87.31% recall on seen techniques to 65.16% on unseen techniques, a decrease of 22.15 percentage points. Linear SVM decreased from 83.90% to 65.78%, while Naive Bayes decreased from 81.42% to 65.78%. Character TF-IDF with Logistic Regression showed the largest decrease, from 83.28% to 56.56%.

Fig. 1. Malicious-prompt recall on seen and unseen attack techniques.



The results support the hypothesis that detectors trained on a limited set of attack techniques may experience reduced malicious-prompt recall when encountering previously unseen techniques. The finding is particularly important because Linear SVM achieved the strongest aggregate accuracy and F1-score, yet still experienced a substantial reduction in recall under technique-level shift.

Interestingly, Naive Bayes produced the smallest seen-to-unseen recall gap (15.64 percentage points), although it was not the strongest model overall. This indicates that overall classification quality and robustness to unseen techniques are related but distinct properties.

C. Individual Unseen Techniques
Aggregate unseen-technique performance hides important differences between attack constructions. Table III reports recall for each unseen technique.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_09

Table III: malicious prompt recall on individual unseen techniques

Unseen Technique	Word-LR	Word-SVM	Word-NB	Char-LR
ml	**25.00%**	**25.00%**	**50.00%**	**37.50%**
obfuscated	**50.00%**	**50.67%**	**51.00%**	**42.00%**
split	**92.22%**	**92.78%**	**91.11%**	**81.67%**

The obfuscated technique was particularly challenging. Recall ranged from 42.00% for Character TF-IDF with Logistic Regression to 51.00% for Word TF-IDF with Naive Bayes. In contrast, the split technique produced substantially higher recall, ranging from 81.67% to 92.78%.

Performance on ml ranged from 25.00% to 50.00%. This result should be interpreted cautiously because only eight malicious samples belong to this technique. The difference between obfuscated and split attacks demonstrates that unseen-technique generalization is not uniform: some unseen constructions remain comparatively detectable, while others cause substantially greater performance degradation.

The particularly low recall on obfuscated prompts suggests that character-level representation did not automatically provide greater robustness in this setting. Although character n-grams can capture variations within words, the tested character configuration did not outperform word-level models on the unseen obfuscation samples.

Overall, the experiments demonstrate the value of separating conventional test performance from technique-level robustness. A model can achieve high precision and strong aggregate F1-score while still missing a substantial portion of malicious prompts when attack construction changes.

V. CONCLUSION

This study evaluated the robustness of classical machine-learning-based prompt-injection detectors under technique-level distribution shift. Word TF-IDF with Linear SVM achieved the best overall performance, with 81.61% accuracy and 83.97% F1-score. However, all models showed lower malicious-prompt recall on unseen techniques, with seen-to-unseen gaps of 15.64–26.72 percentage points. Obfuscated attacks were particularly difficult to detect. The key contribution of this work is a focused robustness evaluation of lightweight TF-IDF-based detectors on attack techniques absent from training, showing that strong overall performance does not necessarily guarantee robustness to unseen attacks.

Future work can extend this evaluation to larger datasets, adaptive attacks, and neural or hybrid detection approaches while preserving the efficiency of lightweight detectors.

VI. REFERENCES

1] Z. Li, B. Peng, P. He, and X. Yan, “Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection,” *Proc. 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 557–568, 2024.

[2] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, “Formalizing and Benchmarking Prompt Injection Attacks and Defenses,” *Proc. 33rd USENIX Security Symposium*, pp. 1831–1847, 2024.

[3] Y. Chen, H. Li, Y. Sui, Y. He, Y. Liu, Y. Song, and B. Hooi, “Can Indirect Prompt Injection Attacks Be Detected and Removed?,” *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 18189–18206, 2025.

[4] M. Akhi, “Prompt Injection Detection Dataset v1.0,” Zenodo, 2026, doi: 10.5281/zenodo.20817277.

[5] J. Yi, Y. Xie, B. Zhu, E. Kiciman, G. Sun, X. Xie, and F. Wu, “Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models,” *arXiv preprint arXiv:2312.14197*, 2023.

ISBN NO: 978-81-180146-5-2

[6] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kiciman, "Defending Against Indirect Prompt Injection Attacks With Spotlighting," Microsoft Research, 2024.

[7] D. Jacob, H. Alzahrani, Z. Hu, B. Alomair, and D. Wagner, "PromptShield: Deployable Detection for Prompt Injection Attacks," *arXiv preprint arXiv:2501.15145*, 2025.

[8] Q. Zhan, R. Fang, H. S. Panchal, and D. Kang, "Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents," *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7116–7132, 2025.

[9] K. Zhu, X. Yang, J. Wang, W. Guo, and W. Y. Wang, "MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents," *Proc. 42nd International Conference on Machine Learning*, vol. 267, pp. 80310–80329, 2025.

[10] W. Hackett, L. Birch, S. Trawicki, N. Suri, and P. Garraghan, "Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems," *Proc. First Workshop on LLM Security (LLMSEC)*, pp. 101–114, 2025.

[11] F. Jia, T. Wu, X. Qin, and A. Squicciarini, "The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents," *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 29680–29697, 2025.

[12] T. Wen, C. Wang, X. Yang, H. Tang, Y. Xie, L. Lyu, Z. Dou, and F. Wu, "Defending against Indirect Prompt Injection by Instruction Detection," *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19472–19487, 2025.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_10

RISK-ADAPTIVE AGENTIC AI FOR AUTONOMOUS THREAT DETECTION AND RESPONSE

Jyothi A[1]
Assistant Professor, Department of Computer Applications
Shri Shankarlal Sundarbai Shasun Jain College for Women

Sabana Yasmine N[2]
Student, Department of Computer Applications
Shri Shankarlal Sundarbai Shasun Jain College for Women

Gobhiga.M[3]
Student, Department of Computer Applications
Shri Shankarlal Sundarbai Shasun Jain College for Women

***Abstract*—** *The growing complexity of cyber threats has made fast and reliable threat detection and response increasingly important. Agentic Artificial Intelligence can perform multiple security tasks, including detecting threats, investigating incidents, and carrying out responses. However, giving AI complete autonomy can be risky when the available evidence is uncertain or the potential impact of an incident is high. This paper presents a Risk-Adaptive Agentic AI framework that considers both threat risk and detection confidence when deciding how much autonomy should be given to the system. The framework uses specialized agents for threat detection, investigation, response, and verification. After a response is carried out, the Response Verification Agent checks whether the threat has actually been contained. High-risk or uncertain incidents can be referred to a human for approval. The proposed approach aims to achieve controlled and graduated autonomy while maintaining effective cybersecurity response.*

Keywords: Agentic AI, Cybersecurity, Threat Detection, Autonomous Response, Risk Assessment, Confidence Estimation, Adaptive Autonomy, Response Verification, Human Oversight.

1. INTRODUCTION

The rapid growth of cloud platforms, connected devices, online services, and distributed applications has increased the complexity of cybersecurity threats. Modern systems continuously generate information from network traffic, authentication records, endpoint activity, application logs, and security alerts. Analysing these events manually becomes difficult when large numbers of alerts are produced simultaneously.

Traditional cybersecurity systems mainly depend on signatures, predefined rules, thresholds, and manually designed response procedures. These methods remain useful, but they may be less effective against changing and multi-stage attacks.

Artificial Intelligence (AI) is increasingly used for anomaly detection, classification, threat analysis, and security monitoring. Agentic Artificial Intelligence extends these capabilities by enabling AI systems to observe information, reason about a goal, plan actions, use external tools, and evaluate the results of previous actions.

Recent research has explored Agentic AI for continuous monitoring, threat detection, threat hunting, incident investigation, and automated response. Multi-agent architectures are also being investigated in which different agents perform specialized cybersecurity tasks. However, increasing autonomy introduces a significant challenge: an incorrect autonomous decision may result in an unnecessary or harmful security action.

The central issue is therefore how much authority an AI agent should receive for a particular security incident. A low-risk event supported by strong evidence may be suitable for automatic handling, whereas a high-impact event with uncertain evidence may require human approval.

Another problem occurs after a response is executed. Successfully executing a command does not necessarily mean that the threat has been contained. An attacker may continue through another account, address, or attack path.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_10

This paper proposes a **Risk-Adaptive Agentic AI framework for autonomous threat detection and response**. The framework combines threat risk and detection confidence to determine the level of autonomous authority and introduces a dedicated Response Verification Agent to evaluate whether the response actually achieved its intended outcome.

The proposed workflow is:

Detect → Investigate → Assess Risk and Confidence → Select Autonomy → Respond → Verify → Reassess

The aim is controlled autonomy rather than complete removal of human involvement.

2. RELATED WORK

Agentic AI represents a shift from conventional AI systems that primarily produce predictions toward systems capable of performing multi-step tasks. In cybersecurity, agents can support monitoring, detection, investigation, threat hunting, incident response, and recovery.

Recent research describes autonomous cyber-defence systems that combine perception, reasoning, planning, action, and learning. At the same time, the literature emphasizes governance, auditability, and bounded autonomy because autonomous agents may interact directly with real systems.

A recent survey of Agentic AI and cybersecurity identifies applications including continuous monitoring, anomaly detection, incident response, threat hunting, vulnerability management, and security automation. It also discusses risks such as memory poisoning, cascading failures, oversight problems, and multi-agent coordination.

Autonomous threat detection and response research demonstrates that AI agents can be used for both identifying threats and performing defensive actions. However, a simple workflow of:

Detection → Analysis → Response

does not guarantee that the response was appropriate or that the threat was actually eliminated.

Multi-agent cybersecurity systems address this problem by dividing security tasks among specialized agents such as detection, investigation, threat intelligence, decision, and response agents. The advantage is that different responsibilities can be separated instead of providing unrestricted authority to one agent. However, coordination and authorization remain important challenges.

Risk-adaptive autonomy is another emerging research direction. Threat severity and AI confidence are not identical. An event may have high potential impact but low detection confidence, or low impact but high confidence. Treating both situations identically may produce inappropriate autonomous actions.

The proposed framework combines these concepts by using risk and confidence before response and verification after response.

3. RESEARCH GAP

Existing studies have explored autonomous threat detection, automated response, multi-agent architectures, risk-based control, governance, and human oversight.

However, four issues remain important.

First, an autonomous system requires a mechanism for deciding **how much authority** should be granted for a specific incident.

Second, threat risk and AI confidence may vary independently. A high-risk event with uncertain evidence should not automatically receive the same authority as a high-confidence event.

Third, executing a response does not prove successful threat containment.

Fourth, when a response fails, the system needs a mechanism to return the incident to investigation instead of marking it as resolved.

Therefore, this research proposes an integrated framework that connects:

Threat Detection + Investigation + Risk + Confidence + Graduated Autonomy + Response Verification + Reassessment

The main contribution is the integration of **risk-confidence-based autonomy selection** and **explicit post-response verification** into a continuous cybersecurity workflow.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_10

4. RISK-ADAPTIVE AGENTIC AI APPROACH

The approach consists of seven main components:

1. Security Data Collection
2. Threat Detection Agent
3. Investigation Agent
4. Risk and Confidence Assessment
5. Autonomy Decision Engine
6. Response Agent
7. Response Verification Agent

These components work together as a continuous security workflow. Each component performs a specific task, from collecting security information and identifying threats to assessing risk, selecting an appropriate response, and verifying the outcome. This enables the approach to adjust its level of autonomy according to the risk and confidence associated with each incident.

The complete process is:

Security Data → Detection → Investigation → Risk/Confidence → Autonomy Decision → Response → Verification → Reassessment

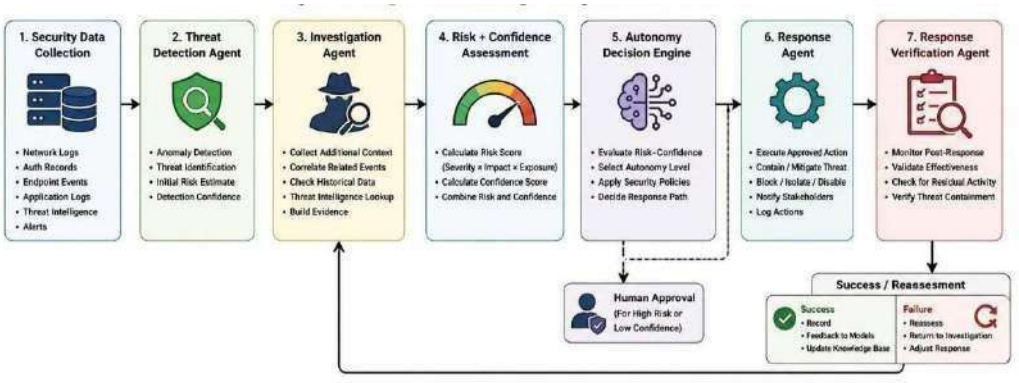


Fig 1. Proposed Risk-Adaptive Agentic AI Framework

Fig 1 illustrates the proposed Risk-Adaptive Agentic AI framework, showing the flow from security data collection and threat detection to investigation, risk and confidence assessment, and response selection. The selected response is executed and then verified to determine whether the threat has been successfully contained. If the response fails, the incident is reassessed, while high-risk or uncertain incidents can be escalated for human approval.

4.1 Security Data Collection

The framework receives security information from sources such as network logs, authentication records, endpoint events, application logs, firewall alerts, suspicious file activity, and threat intelligence.

The information is normalized and provided to the Detection Agent. This stage supplies evidence and does not make the final response decision.

4.2 Threat Detection Agent

The Detection Agent analyses incoming events and identifies potentially suspicious activity. Its output contains the detected event, possible threat category, supporting evidence, initial risk estimate, and detection confidence. For example, repeated login failures from an unfamiliar source may be classified as possible brute-force activity.

The Detection Agent does not immediately perform a high-impact action. The incident is passed to investigation.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_10

4.3 Risk-Based Investigation

The Investigation Agent gathers additional context about the detected event. It can examine related events, historical activity, user behaviour, device information, network relationships, previous incidents, and threat-intelligence information.

The purpose is to determine whether the initial alert represents a genuine threat and to provide additional evidence for risk and confidence assessment.

5. RISK AND CONFIDENCE BASED AUTONOMY

Risk and confidence are the central inputs to the proposed decision process.

5.1 Risk

Risk represents the possible consequences of a detected threat. It depends on factors such as the severity of the threat, its potential impact, and the exposure of the affected system.

A conceptual model is:

Risk Score = Severity × Impact × Exposure

The factors can be adjusted according to the security environment. Higher-risk incidents require stronger controls and may need human approval.

5.2 Confidence

Confidence represents the strength of evidence supporting the threat classification.

It may depend on supporting indicators, agreement between security sources, investigation results, historical evidence, threat intelligence, and contradictory information.

Risk and confidence must be considered separately.

For example:

High Risk + High Confidence

indicates a serious event with strong evidence.

High Risk + Low Confidence

indicates a potentially serious event for which the evidence is uncertain.

These two situations should not automatically receive the same response authority.

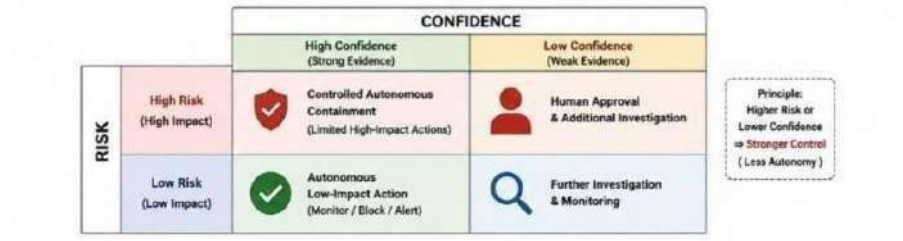


Fig 2. Risk-Confidence Autonomy Matrix

Fig 2 illustrates how risk level and detection confidence determine the level of control applied to an incident. Low-risk and high-confidence events can be handled with greater autonomy, while high-risk or low-confidence events require stronger controls or human approval. This approach provides graduated autonomy rather than using a simple autonomous-or-manual decision.

Table 1. Risk–Confidence Response Model

Risk	Confidence	Response
Low	High	Autonomous low-impact action
Low	Low	Further investigation
Medium	High	Limited autonomous response

ISBN NO: 978-81-180146-5-2

Medium	Low	Investigation / human review
High	High	Controlled containment + notification
High	Low	Human approval / additional investigation
Critical	Any	Emergency policy + escalation

6. RESPONSE AND VERIFICATION

The Autonomy Decision Engine selects the permitted response level.

Low-impact actions may include monitoring or temporary blocking. More significant actions may include terminating suspicious sessions, restricting access, isolating an endpoint, or disabling compromised credentials. High-impact actions should operate within predefined policies and least-privilege permissions.

The framework does not assume:

Action executed ≠ Threat resolved

Instead, the Response Verification Agent checks the environment after the response.

For example, if a suspicious source is blocked, the system examines subsequent activity to determine whether the suspicious behaviour has stopped.

The process is:

Response → Observe → Verify → Success / Failure

If the threat is contained:

Success → Record / Feedback

If the threat remains active:

Failure → Reassessment → Investigation

This creates a closed-loop response mechanism.

The verification stage is important because a security command may be successfully executed while the underlying attack continues through another route.

7. HUMAN OVERSIGHT

The proposed framework does not attempt to eliminate cybersecurity professionals.

Human approval can be required when:

- Confidence is low.
- Potential impact is high.
- Evidence is contradictory.
- The action is difficult to reverse.
- A critical system may be affected.
- Previous responses have failed.

Agent permissions should follow least-privilege principles so that an agent receives only the authority required for its assigned task.

This creates a safety boundary between autonomous reasoning and high-impact system changes.

8. EVALUATION METHODOLOGY

The proposed framework should be evaluated in a controlled cybersecurity environment.

Possible scenarios include:

- Brute-force activity
- Suspicious network behaviour
- Malware-like activity
- Ransomware-like file modification
- Unauthorized account activity
- False-positive events

The framework can be compared with a conventional autonomous detection-and-response workflow.

Evaluation metrics include:

ISBN NO: 978-81-180146-5-2

Precision

Precision = TP / (TP + FP)

Recall

Recall = TP / (TP + FN)

Response Time: Time between threat detection and response initiation.

Containment Success Rate: Percentage of incidents successfully contained.

Unnecessary Autonomous Action Rate: Percentage of autonomous actions later determined to be unnecessary.

Verification Success Rate: Percentage of response outcomes correctly classified as successful or unsuccessful.

Human Escalation Rate: Percentage of incidents requiring human intervention.

No numerical results are claimed because experimental validation has not yet been conducted.

9. COMPARISON WITH EXISTING APPROACHES

Table 2. Comparison of Security Approaches

Capability	Traditional Security	AI-Based Security	Agentic AI	Proposed Framework
Threat Detection	Rule/Signature Based	AI-Assisted	Autonomous	Risk-Aware
Investigation	Limited	Partial	Autonomous	Context-Aware
Multi-Agent Operation	Not Typical	Limited	Supported	Specialized Agents
Risk Assessment	Rule-Based	Partial	Partial	Integrated
Confidence-Based Control	Not Typical	Limited	Limited	Risk + Confidence
Autonomy Level	Fixed	Limited	High/Variable	Graduated
Autonomous Response	Limited	Partial	Supported	Risk-Controlled
Response Verification	Limited	Partial	Varies	Dedicated Verification
Failure Reassessment	Limited	Limited	Varies	Reassessment
Human Escalation	Supported	Supported	Supported	Risk-Based

The table represents the intended architectural contribution. It does not claim that risk assessment, multi-agent systems, or verification are individually new. The contribution is their integration into the proposed cybersecurity response cycle.

10. LIMITATIONS AND FUTURE SCOPE

The framework has several limitations. Accurate risk and confidence estimation can be difficult, and poor calibration may result in excessive autonomous actions or unnecessary human escalation. The framework also depends on the quality of security data.

Response verification can be challenging when attackers use stealth techniques or when monitoring provides incomplete visibility. Multi-agent communication may also introduce additional overhead and coordination problems.

Future work should implement the framework in a controlled cyber range and evaluate it against realistic attack scenarios. Experiments can investigate suitable risk-confidence thresholds and measure response time, false positives, containment success, unnecessary autonomous actions, and human escalation.

Future research can also explore adaptive thresholds, explainable autonomy decisions, secure agent memory, adversarial testing, advanced multi-agent coordination, and integration with SIEM, EDR, firewall, identity-management, and threat-intelligence platforms.

11. CONCLUSION

Agentic AI can support cybersecurity by enabling systems to perform multi-step investigation and response. However, increased autonomy also increases the consequences of incorrect decisions.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_10

This paper proposed a **Risk-Adaptive Agentic AI framework for autonomous threat detection and response**. The framework combines threat detection, investigation, risk assessment, confidence estimation, graduated autonomy, controlled response, response verification, reassessment, and human oversight.

The main idea is that autonomous authority should depend on the characteristics of the security incident. High-risk or uncertain events should receive stronger controls, while suitable low-impact and high-confidence events can be handled automatically.

The Response Verification Agent provides an additional safety mechanism by checking whether an executed response actually contained the threat. If the threat remains active, the incident returns to investigation and reassessment.

The proposed workflow is:

Detect → Investigate → Assess Risk and Confidence → Select Autonomy → Respond → Verify → Reassess

Experimental implementation is required to validate the framework and determine suitable thresholds for different cybersecurity environments.

REFERENCES

1. Lazer, S. J., Aryal, K., Gupta, M., & Bertino, E. (2026). *A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-case Prototypes*. arXiv:2601.05293.
2. *Agentic AI for Autonomous Cyber Defense*. (2026). Journal of Information Systems Engineering and Management.
3. *Autonomous AI Agents for Cybersecurity Threat Detection and Response*. (2026). AI Journal of Computer Science and Technology.
4. Fleming, C., Kundu, A., & Kompella, R. (2025). *Uncertainty-Aware, Risk-Adaptive Access Control for Agentic Systems using an LLM-Judged TBAC Model*. arXiv:2510.11414.
5. Kojukhov, A., & Bovshover, A. (2026). *Agentic AI for Cybersecurity: A Meta-Cognitive Architecture for Governable Autonomy*. arXiv:2602.11897.
6. Suggu, S. K. (2025). *Agentic AI Workflows in Cybersecurity: Opportunities, Challenges, and Governance via the MCP Model*.
7. Vajpayee, P., & Hossain, G. (2025). *Cyber Defense through Agentic AI Enabled Automation: An Approach to Reduce Cyber Risk*.
8. Syros, G., Suri, A., Ginesin, J., Nita-Rotaru, C., & Oprea, A. (2025). *SAGA: A Security Architecture for Governing AI Agentic Systems*.
9. Pan, K., & Hou, R. (2026). *Beyond Autonomy: A Dynamic Tiered AgentRunner Framework for Governable and Resilient Enterprise AI Execution*.
10. *Secure Autonomous Cyber Defense with LLM Agents: A Systematic Review*. (2026). *Computers & Security*.
11. Tamuka, N., Mathonsi, T. E., Otieno, O. T., Maswikaneng, S., Muchenje, T., & Tshilongamulenzhe, T. M. (2026). Securing LLM-based agents against cyberattacks: A comprehensive survey on attack techniques and defense strategies. Journal of Computer Virology and Hacking Techniques, 22, 38.
12. Xu, Y., Zhuang, Y., Liu, X., Zhang, T., Xiao, B., Xu, X., Jiang, D., Wang, J., & Hu, H. (2026). LLM agents security duality: A comprehensive survey of self-security and empowered cybersecurity. Artificial Intelligence Review, 59, 174.
13. Kizhakel, S. P., & Pande, S. D. (2026). An adaptive agentic AI framework for counterfactual validation and adversarial lookahead in automated incident response. Discover Artificial Intelligence, 6, 970.
14. Vinay, V. (2025). The Evolution of Agentic AI in Cybersecurity: From Single LLM Reasoners to Multi-Agent Systems and Autonomous Pipelines. arXiv:2512.06659.
15. Pinto Neto, E. C., Iqbal, S., Buffett, S., Sultana, M., & Taylor, A. (2026). Automated threat defense: A comprehensive survey on red and blue team automation using machine learning techniques.International Journal of Information Security, 25, 12

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_11

AN ADAPTIVE HYBRID FRAMEWORK FOR REAL-TIME SMILE RECOGNITION USING LIGHTWEIGHT DEEP FEATURES, TEMPORAL CUES AND ENSEMBLE LEARNING

K.Ravinandan

B.Tech. Information Technology, Meenakshi Sundararajan Engineering College, Chennai, India

B. Mallikeswari

Assistant Professor, JBAS College for Women, Chennai, India

Abstract—Automatic smile recognition becomes unreliable when a subject change pose, illumination, scale, facial appearance, or expression intensity. This paper presents an Adaptive Hybrid Smile Recognition (AHSR) framework that combines lightweight deep appearance descriptors, facial landmark geometry, short-term motion information and an XGBoost decision layer. The proposed design uses a MobileNetV3-style backbone for compact feature extraction, normalized landmark measurements to describe mouth geometry, and regional optical flow to represent smile onset and offset. A confidence-guided fusion mechanism adjusts the relative contribution of these three feature sources before classification. Frame skipping and temporal confirmation are included to reduce unnecessary computation in real-time streams. The previously published Hybrid Learning for Smile Recognition (HLSR) study is used only as a baseline; its reported accuracy and processing-time values are not presented as results of AHSR. A subject-independent experimental protocol is specified so that the proposed framework can be evaluated without identity leakage.

Index Terms— smile recognition, facial expression analysis, MobileNetV3, facial landmarks, optical flow, XGBoost, hybrid learning, real-time computer vision.

I. INTRODUCTION

A smile is a useful visual cue in human communication and has practical roles in camera automation, human–computer interaction, educational interfaces, entertainment and assistive technologies. A recognition system must therefore work beyond carefully posed laboratory images. Changes in lighting, camera quality, head orientation, background and the difference between posed and spontaneous smiles can alter the visible evidence available to a classifier.

The earlier HLSR study associated with this work combined convolutional feature learning with XGBoost and reported better recognition accuracy and lower processing time than its CNN comparison [1]. That study provides a useful reference point, but a frame-only representation does not explicitly describe how a smile develops across successive frames. The present work addresses this gap by adding geometric and temporal descriptors and by allowing the fusion weights to change with feature reliability.

The main contributions are: (1) a compact multi-source representation that combines deep, geometric and temporal information; (2) a confidence-aware fusion strategy before XGBoost classification; (3) an adaptive processing policy for low-latency video; and (4) an evaluation plan based on subject-independent splits and both recognition and efficiency measures.

II. RELATED WORK

Research in facial expression analysis has moved from handcrafted appearance descriptors toward deep networks and hybrid decision systems. Smile recognition remains difficult when the training distribution does not represent natural changes in pose, illumination and expression. HOG/SVM and regression-forest approaches have demonstrated the value of explicit local or pose-aware information, while deep convolutional models provide stronger learned representations [4], [5], [7].

The most directly related work is the 2024 HLSR study by Cherian et al., which includes B. Mallikeswari as

ISBN NO: 978-81-180146-5-2

a co-author and combines CNN representation learning with XGBoost [1]. The present manuscript extends that research direction rather than repeating the same architecture. In particular, it introduces a lightweight MobileNetV3-style feature extractor, landmark geometry, regional optical-flow statistics and confidence-based fusion.

Image-quality variation is also relevant to expression recognition. Mallikeswari and Sripriya reported a double filtering–density method for image enhancement [2]. This prior work is relevant because contrast and noise can influence the quality of facial evidence supplied to a recognition pipeline. B. Mallikeswari's earlier drowsiness-detection publication [3] is listed as an additional related work because it reflects the broader use of facial visual cues for safety-oriented real-time systems.

MobileNetV3 was designed for efficient vision processing on mobile CPUs, making it a suitable architectural basis when latency and model size are important constraints [8]. XGBoost provides a nonlinear boosted-tree decision layer that can operate on the fused descriptor rather than requiring a large neural classification head [9].

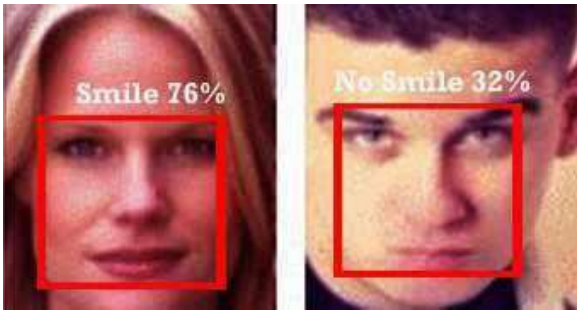


Fig. 1. Example smile/non-smile outputs reproduced from the supplied HLSR study [1].

III. PROPOSED AHSR FRAMEWORK

The proposed pipeline receives consecutive video frames and processes them through six functional stages: face localization and alignment, lightweight deep feature extraction, landmark-based geometry estimation, temporal motion extraction, adaptive feature fusion and final XGBoost classification. Fig. 2 summarizes the complete processing flow.

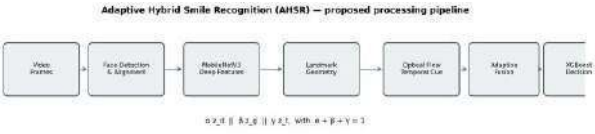


Fig. 2. Proposed Adaptive Hybrid Smile Recognition processing pipeline.

A. Face Detection and Normalization

For each frame, the system first identifies a face region and aligns it using stable facial reference points such as the eyes. The aligned crop is resized to a fixed input dimension. This step limits variations caused by translation, scale and moderate in-plane rotation and focuses subsequent processing on facial regions that carry smile-related information.

B. Lightweight Deep Feature Extraction

A MobileNetV3-style backbone is used as the appearance encoder. Its final classification head is replaced by an embedding output. The resulting vector z_d captures texture and higher-level visual characteristics while keeping the feature extractor relatively compact for edge or CPU-oriented deployment [8]. Transfer learning can be used when the available smile dataset is small.

C. Facial Landmark Geometry

Let p_i=(x_i,y_i) denote a detected facial landmark. Distances between mouth and cheek landmarks are normalized using an inter-ocular reference distance d_e. A representative normalized geometric descriptor is

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_11

g_i=||p_i-p_j||/d_e. This representation provides scale-normalized information that complements the learned deep embedding.

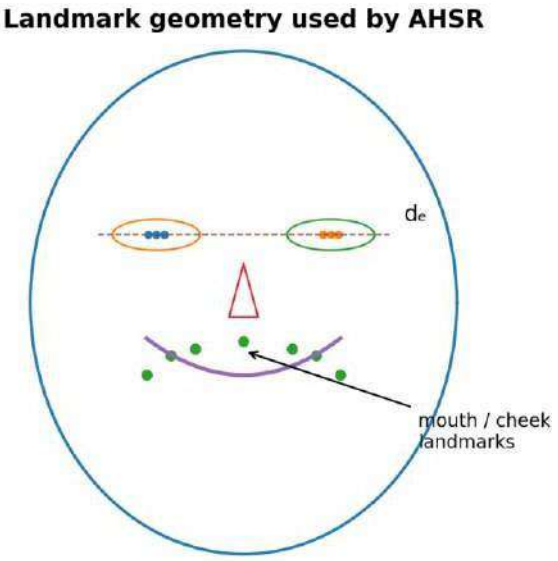


Fig. 3. Landmark-based geometric representation used as a complementary feature source.

D. Temporal Motion Descriptor

A smile is not only a static mouth configuration; it has an onset, a peak and an offset. For consecutive aligned frames I_(t-1) and I_t, optical flow produces a displacement field v_t=(u_t,v_t). Instead of retaining the entire flow image, the proposed method summarizes motion statistics in the mouth and cheek regions to obtain a compact vector z_t.

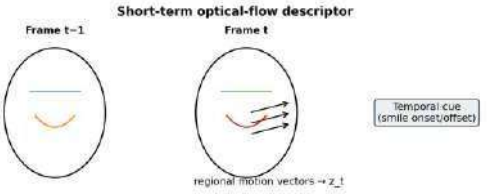


Fig. 4. Regional optical-flow representation for short-term smile dynamics.

E. Confidence-Aware Fusion

The three descriptor families are combined as z=[αz_d || βz_g || γz_t], where α, β and γ are non-negative adaptive coefficients satisfying α+β+γ=1. The coefficients are adjusted according to feature quality. For example, weak landmark confidence lowers β, while negligible inter-frame motion lowers γ. This prevents a low-quality modality from receiving the same influence as reliable evidence.

F. XGBoost Decision Layer

The fused vector is supplied to an XGBoost classifier. The classifier estimates the probability of the smile class. A threshold τ is then applied. When confidence is below τ, the system can wait for temporal confirmation rather than immediately declaring a smile. This strategy is intended to reduce false positives caused by brief mouth movements [9].

IV. MATHEMATICAL FORMULATION

For the three feature sources z_d, z_g and z_t, the normalized fusion weights are constrained by:

$$\alpha + \beta + \gamma = 1, \quad \alpha \geq 0, \beta \geq 0, \gamma \geq 0$$

The fused representation is:

$$z = [\alpha\, z_d \,||\, \beta\, z_g \,||\, \gamma\, z_t]$$

The binary decision rule is:

ISBN NO: 978-81-180146-5-2

$$\hat{y} = 1, \text{ if } P(\text{smile} \mid z) \geq \tau; \text{ otherwise } \hat{y} = 0.$$

The classifier is trained using only the training partition. The fusion policy and threshold are selected on the validation partition, while the test partition remains unseen until the final evaluation.

V. PROPOSED ALGORITHM

1. Input: video stream V; lightweight feature model M_d; landmark model M_g; trained XGBoost model M_x; threshold τ.
2. Acquire consecutive frames I_(t-1) and I_t.
3. Detect, align and quality-check the face region.
4. Generate the deep appearance descriptor z_d.
5. Calculate the normalized landmark descriptor z_g.
6. Compute regional optical flow and summarize it as z_t.
7. Estimate the quality indicators q_d, q_g and q_t.
8. Convert the quality indicators into α, β and γ and normalize them.
9. Concatenate the weighted descriptors and obtain z.
10. Classify z using XGBoost and obtain P(smile|z).
11. If confidence is sufficient, output smile; otherwise retain the temporal state and inspect the next frame.
12. Apply frame skipping or early stopping when the visual state remains unchanged.

VI. EXPERIMENTAL DESIGN

The available source material does not provide sufficient implementation details to claim new AHSR numerical results. Therefore, the experiment is defined as a reproducible protocol rather than a fabricated result section. The recommended split is subject-independent so that images from the same person do not appear in both training and testing.

Component	Recommended configuration	Purpose
Data partition	Subject-independent train/validation/test	Reduce identity leakage
Input	Aligned face crop, fixed resolution	Consistent feature extraction
Backbone	MobileNetV3 family, transfer learning	Lower computation
Temporal window	3–7 consecutive frames	Capture smile dynamics
Classifier	XGBoost	Nonlinear decision on fused features
Baselines	CNN, HLSR, AHSR	Measure incremental gain
Quality metrics	Accuracy, precision, recall, F1, ROC-AUC	Recognition quality
Efficiency metrics	Latency/frame, FPS, model size	Real-time suitability

VII. RESULTS

The supplied 2024 paper reports HLSR accuracy values from 96.29% to 96.54% across its listed epoch settings, while the CNN comparison ranges from 93.12% to 93.45% [1]. These values belong to the published HLSR study and are included here only to establish the baseline against which a future AHSR implementation may be compared.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_11

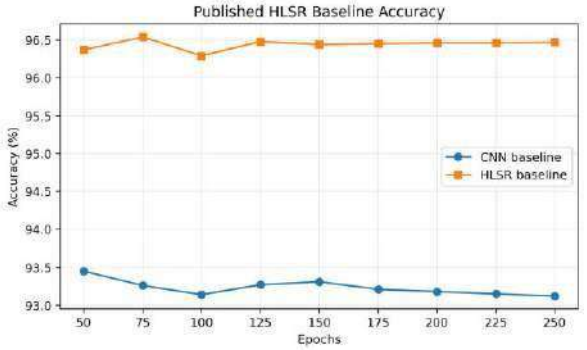


Fig. 5. Accuracy values reported in the supplied HLSR study [1].

Epoch	CNN (%)	HLSR (%)
50	93.45	96.37
75	93.26	96.54
100	93.14	96.29
125	93.27	96.48
150	93.31	96.44
175	93.21	96.45
200	93.18	96.46
225	93.15	96.46
250	93.12	96.47

The same source reports shorter processing times for HLSR than CNN over the listed data sizes [1]. These measurements are retained as published baseline observations rather than AHSR measurements.

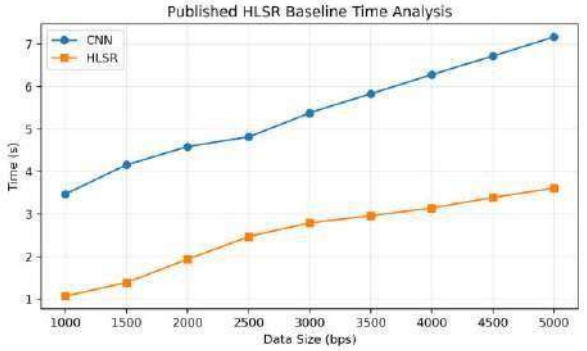


Fig. 6. Processing-time comparison reported in the supplied HLSR study [1].

Data size	CNN (s)	HLSR (s)
1000	3.47	1.07
1500	4.16	1.39
2000	4.59	1.94
2500	4.82	2.47
3000	5.38	2.79
3500	5.83	2.96
4000	6.28	3.14
4500	6.72	3.39
5000	7.17	3.61

VIII. AHSR EVALUATION AND ABLATION PLAN

After implementation, the proposed system should be compared with the CNN and HLSR baselines under identical data partitions. In addition to overall accuracy, precision, recall and F1-score, latency and model size should be measured. The following ablation sequence can isolate the contribution of each component.

ISBN NO: 978-81-180146-5-2

Variant	Included features	Purpose
A	Deep features only	Reference representation
B	Deep + landmarks	Measure geometric contribution
C	Deep + temporal	Measure motion contribution
D	Deep + landmarks + temporal	Measure multi-source fusion
E	Complete AHSR + adaptive weights	Measure adaptive fusion

No numerical AHSR result is inserted until the model is trained and tested. This distinction is important for scientific validity and prevents the proposed architecture from being presented as an already validated performance improvement.

IX. ADVANTAGES AND LIMITATIONS

The proposed design has five practical strengths. First, spatial, geometric and temporal information are complementary. Second, a lightweight backbone is appropriate for constrained devices [8]. Third, XGBoost provides a compact nonlinear decision stage [9]. Fourth, adaptive weighting can reduce the effect of unreliable landmarks or weak motion. Fifth, temporal confirmation can suppress isolated false detections.

The approach still depends on reliable face and landmark localization. Optical flow can degrade under blur, very rapid head motion or low frame rates. Transfer learning may introduce dataset bias when source and target populations differ. Most importantly, the proposed architecture remains a research design until it is implemented and evaluated on a controlled test protocol.

X. FUTURE ENHANCEMENTS

- Replace short optical-flow windows with temporal convolution or transformer-based sequence modeling.
- Add explicit head-pose estimation and three-dimensional landmark normalization.
- Distinguish posed, spontaneous and ambiguous or micro-expression smiles.
- Apply pruning, quantization and knowledge distillation for edge deployment.
- Evaluate RGB, depth and infrared fusion under difficult illumination.
- Add feature-importance and landmark-contribution analysis for explainability.
- Test cross-dataset generalization across age groups, facial accessories and camera conditions.

XI. CONCLUSION

This paper presents an adaptive hybrid design for real-time smile recognition in which lightweight deep features, facial geometry and short-term motion are combined before XGBoost classification. The central idea is to use the most reliable evidence available in each frame rather than relying on a single representation. The published HLSR results provide a reference baseline, while the proposed AHSR values must be obtained through actual implementation and subject-independent testing. The resulting framework therefore offers a reproducible path for investigating accuracy, robustness and real-time efficiency.

REFERENCES

[1] A. K. Cherian, S. Devipriya, B. P. Saoji, B. Mallikeswari, R. Thiagarajan, and R. Krishnamoorthy, “A Robust Design of Real-Time Resilient Smile Recognition System using Hybrid Deep Learning Principles,” Proc. 2024 10th International Conference on Communication and Signal Processing (ICCSP), pp. 592–596, 2024, doi: 10.1109/ICCSP60870.2024.10543446.

[2] B. Mallikeswari and P. Sripriya, “Predominance and diminishing equivocalness of the images using double filtering–Density algorithm,” in Recent Trends in Science and Engineering, 2022, doi: 10.1063/5.0074468.

[3] B. Mallikeswari, “A Genuine Moment Eye Retina Drowsiness Detection for Accident Prevention and

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_11

Caution," Journal of Emerging Technologies and Innovative Research, 2019.
[4] R. Syaputra, D. Syamsuar, et al., "Multiple Smile Detection Using Histogram of Oriented Gradient and Support Vector Machine Methods," IOP Conference Series: Materials Science and Engineering, vol. 1071, 2021, doi: 10.1088/1757-899X/1071/1/012027.
[5] L. Liu, W. Gui, et al., "Real-time pose invariant spontaneous smile detection using conditional random regression forests," Optik, 2019, doi: 10.1016/j.ijleo.2019.01.020.
[6] T. Künzler, T. T. A. Höfling, et al., "Automatic Facial Expression Recognition in Standardized and Non-standardized Emotional Expressions," Frontiers in Psychology, vol. 12, 2021, doi: 10.3389/fpsyg.2021.627561.
[7] Y. Liu, Z. Liu, et al., "A Robust Approach for Smile Recognition via Deep Convolutional Neural Networks," Proc. 7th International Conference on Image, Vision and Computing, 2022, doi: 10.1109/ICIVC55077.2022.9886093.
[8] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324, 2019, doi: 10.1109/ICCV.2019.00140.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_12

HANDWRITING BASED DYSLEXIA DETECTION USING MACHINE LEARNING

CHANIT JAYA AUXILIA A, Student, Justice Basheer Ahmed Syeed college for women

ABSTRACT

Dyslexia is a learning difficulty that can affect a child's ability to read, write, spell, and process written language. Identifying the possibility of dyslexia at an early stage is important because timely support can help children overcome learning difficulties and improve their academic performance. However, conventional assessment methods generally require trained professionals and may take considerable time. This creates a need for computer-assisted methods that can support the early screening process.

This study proposes a machine-learning-based approach for identifying handwriting patterns that may be associated with dyslexia. Handwritten samples are collected and processed using image-processing techniques to remove noise, improve image quality, and extract relevant handwriting characteristics. Features such as letter size, spacing, alignment, slant, stroke patterns, and writing consistency are analyzed. These features are then used to train and compare different machine-learning models.The performance of the models is evaluated using accuracy, precision, recall, sensitivity, and specificity.

The proposed approach aims to determine whether handwriting characteristics can provide useful information for identifying children who may require further assessment for dyslexia. Rather than replacing professional diagnosis, the system is designed as an initial screening and decision-support tool. The study is expected to contribute toward the development of a low-cost, automated, and accessible approach for early identification of dyslexia-related risk through handwriting analysis.

INTRODUCTION

Dyslexia is a common learning difficulty that mainly affects a person's ability to read, spell, and process written language. It is not related to intelligence, but children with dyslexia may experience difficulties in recognizing words, understanding written information, remembering spellings, and developing fluent reading skills. These difficulties can influence their academic

performance and may also affect their confidence and participation in classroom activities. Early identification is therefore important because appropriate educational support can be provided before the difficulties become more severe.

Handwriting is one of the behavioral characteristics that can provide useful information about a child's learning process. While writing, children need to coordinate visual perception, motor control, language processing, memory, and attention. Difficulties in these processes may be reflected in handwriting through variations in letter formation, spacing, alignment, size, slant, stroke consistency, and writing organization.

The development of image processing and machine learning techniques has created new possibilities for analyzing handwriting automatically. Machine learning algorithms can identify patterns within large collections of handwriting samples and learn relationships between extracted characteristics and predefined groups.

This research focuses on developing a machine-learning-based framework for analyzing handwritten samples and identifying patterns that may be associated with dyslexia risk. The study considers both conventional handwriting features and image-based deep-learning features. Different machine-learning models will be trained and compared using appropriate performance measures such as accuracy, precision, recall, F1-score, sensitivity, and specificity. The study will also examine which handwriting characteristics contribute most to the classification results.

LITERATURE REVIEW

Handwriting analysis has recently gained attention as a potential method for supporting the early screening of dyslexia and other learning difficulties. Since handwriting can contain measurable characteristics such as letter

ISBN NO: 978-81-180146-5-2

shape, size, spacing, alignment, and writing consistency, researchers have explored machine-learning techniques to identify patterns that may be difficult to recognize through manual observation. The following studies are particularly relevant to the proposed work.

CNN-Based Dyslexia Handwriting Recognition

A study titled "Development of CNN Transfer Learning for Dyslexia Handwriting Recognition" was presented at the 2021 IEEE International Conference on Control System, Computing and Engineering. The researchers developed a handwriting recognition system using a Convolutional Neural Network (CNN) based on the LeNet-5 architecture. Before training the model, handwriting images were subjected to preprocessing and data augmentation. The study used a dataset containing 138,500 handwriting images and classified three categories of dyslexic handwriting. The developed CNN model achieved an accuracy of 95.34%. The study demonstrates that deep-learning models can learn visual patterns directly from handwriting images and can be useful for automated dyslexia-related handwriting classification.

METHODOLGY

The proposed system is designed to identify handwriting patterns that may be associated with dyslexia risk using image processing and machine learning. The process begins with a scanned image of a child's handwritten sample. The image is first preprocessed to remove noise and improve its quality. After preprocessing, relevant information and handwriting characteristics are extracted from the image. These extracted features are then given to a trained machine-learning model, which analyzes the patterns and makes the final classification. Based on the learned patterns, the system provides an indication of whether the handwriting sample is associated with dyslexia risk or not.

The implementation consists of the following main stages:

- Handwritten Image Acquisition: A handwritten sample of the child is scanned or uploaded into the system and used as the input image.
- Image Preprocessing: The input image is processed to remove noise and unwanted background information. Grayscale conversion, thresholding, resizing, and other basic preprocessing techniques can be applied to obtain a clearer handwriting image.
- Feature Extraction: Important information is collected from the processed handwriting image. Features may include letter size, word spacing, character spacing, handwriting slant, alignment, stroke patterns, character shape, and writing consistency.
- Feature Processing: The extracted features are organized and normalized so that they can be effectively used by the machine-learning model.
- Machine Learning Classification: The processed features are provided to a trained machine-learning algorithm. The model learns patterns from previously labeled handwriting samples belonging to dyslexia-risk and non-dyslexia groups.
- Decision Making: After analyzing the input features, the trained model classifies the new handwriting sample into the appropriate category.
- Final Output: The system displays the prediction as "Dyslexia Risk" or "No Dyslexia Risk." This output is intended for preliminary screening and should not be considered a substitute for professional diagnosis.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_12

RESULT



The proposed system is designed to analyze children's handwriting and classify it into two categories: Dyslexia Risk and No Dyslexia Risk. By applying image preprocessing, handwriting feature extraction, and machine-learning classification, the system aims to support the early identification of children who may be at risk of dyslexia.

The system provides a preliminary screening result that can help parents and educators identify children who may require further assessment by qualified professionals. It is not intended to replace a clinical diagnosis.

CONCLUSION

This research presents a handwriting-based approach for the preliminary screening of dyslexia risk in children using machine-learning techniques. By analyzing handwriting patterns and extracting important features, the proposed system aims to identify differences between handwriting samples associated with dyslexia risk and those without identified risk.

The study highlights how machine learning can support the analysis of handwriting in a systematic and efficient manner. Such a system may help increase awareness among parents, teachers, and educators and encourage timely attention to children who may need additional learning support.

FUTURE SCOPE

The future scope of this research includes improving the accuracy and reliability of the handwriting-based dyslexia-risk detection system by collecting larger and more diverse handwriting datasets. Advanced machine-learning and deep-learning techniques can be

explored to identify handwriting patterns more effectively. The system can also be extended to analyze additional features, such as stroke pressure, writing speed, and letter formation. Furthermore, developing a user-friendly application could make preliminary screening more accessible to parents, teachers, and educators. The research may also be expanded to support handwriting analysis in different languages and writing styles. Collaboration with educational and healthcare professionals can help validate the system and ensure that it serves as a reliable support tool for early awareness and learning assistance.

REFERENCE

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. Annals of Dyslexia, 53, 1–14. https://doi.org/10.1007/s11881-003-0001-9

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. Developmental Medicine & Child Neurology, 49(4), 312–317. https://doi.org/10.1111/j.1469-8749.2007.00312.x

Snowling, M. J. (2013). Early identification and interventions for dyslexia: A contemporary view. Journal of Research in Special Educational Needs, 13(1), 7–14. https://doi.org/10.1111/j.1471-3802.2012.01262.x

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

OpenAI. (2026). ChatGPT (GPT-5.6 Luna) [Large language model]. https://chatgpt.com/

(ChatGPT was used as an AI-assisted tool to support the organization of research content)

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_13

BEYOND SENTIMENT: CONFIDENCE-AWARE MULTIMODAL AI FOR EMOTION RECOGNITION

Keywords: Multimodal AI, Emotion Recognition, Sentiment Analysis, Confidence-Aware Learning, Multimodal Fusion, Facial Expression Recognition, Speech Emotion Recognition, Natural Language Processing.

Author: Fahima Firdhous S, MCA (Department Of Computer Science)
Institute: Justice Basheer Ahmed Sayeed College for Women

ABSTRACT

Human emotions can be communicated through a number of channels such as language, speech, and facial expressions. Even though multimodal emotion recognition offers more information than methods that rely on a single modality, traditional fusion techniques may be impacted by differences in the reliability of the modalities, noisy inputs, missing data, and conflicting emotional signals. To address this, the project introduces a **confidence-aware multimodal AI framework for emotion recognition,** which assesses the reliability of each modality before the final decision. The framework integrates text, speech, and facial data and calculates a confidence level at each modality level in order to vary the contribution of each modality during the fusion process. As a result, the modalities that are more informative are given more weight, while those that are unreliable or noisy have less impact on the final prediction. The proposed framework consists of **modality-specific feature extraction, confidence estimation, multimodal fusion, and finally the emotion classification**. The method is designed to enhance the recognition performance and to increase efficiency while dealing with noisy, missing, and conflicting inputs, used in areas such as **virtual assistants, education, customer service, human-computer interaction, and assistive technologies**.

I. INTRODUCTION

Human emotions are very crucial and sensitive to judge which can be expressed in numerous ways, for example through language, speech, and our facial expressions. But when it comes to artificial intelligence (AI), gaining an understanding of these emotional signals is a primary objective, especially when designing systems that are intended to interact with people in a natural and human-oriented way. The existing traditional sentiment analysis generally involves dividing text into positive, negative, or neutral clusters. However, such classification may not accurately represent one's real emotional condition

A more advanced method is provided by multimodal AI, since it highlights combining information from various types of data. Text can supply semantic and linguistic information, speech can contribute acoustic features such as tone and pitch, and facial expressions can offer visual hints concerning emotion. A richer representation of human emotions can be achieved by combining these different modalities rather than by relying on a single source.

Even so, multimodal emotion recognition involves a number of challenges since the various modalities may differ in their degree of reliability and the inputs can be affected by background noise, occlusion, missing data, ambiguity, or contradictory emotional signals. Traditional fusion methods might assign the same weight to all the modalities even when one of them provides much less reliable information.

To overcome this limitation, the present study presents a **confidence-aware multimodal AI framework**. The method evaluates the reliability of textual, spoken, and facial inputs and then uses these confidence levels to dynamically influence the fusion process in such a way that reliable information is given greater weight while uncertain or noisy information contributes less to the final prediction.

The objective is to develop a more trustworthy framework for emotion recognition, one that can function efficiently in realistic multimodal scenarios

II. LITERATURE SURVEY

A. Unimodal Emotion Recognition

Traditional emotion recognition systems focused only on single modalities like text, speech, or facial

ISBN NO: 978-81-180146-5-2

expressions. Those based on text used NLP(Natural Language Processing) techniques to detect people's emotions from written texts. Speech-based systems observe properties like pitch, intensity, rhythm, and tone, while facial emotion-recognition systems focus on visual patterns that are associated with various emotions. Even though unimodal methods can work fairly well but restricted by the information in a single modality. For example, text-based systems fail to detect sarcasm or emotional tone, while facial-based systems can be distorted by many other environmental factors and variations in facial expressions.

B. Multimodal Emotion Recognition

To overcome prior mentioned limitations, a multimodal system combines information from a number of different sources. Since text, speech, and facial expressions can provide wholesome information related to emotion, AI systems are able to develop a more complete representation of human affect. Examples of multimodal fusion strategies are **early fusion, late fusion, and intermediate fusion.**

Early fusion - features are combined before making a prediction

Late fusion - predictions from each model are combined

Intermediate fusion - learned features are combined during the learning process

Though such methods can enhance the performance, they may not sufficiently take into account the differing reliability of the various modalities.

C. Attention-Based Multimodal Learning

The Attention mechanism is very useful in multimodal AI systems since they concentrate mainly on important information and learns the relationships between different modalities. In the case of emotion recognition, attention can assist in deciding which verbal or visual features are more relevant to a specific emotional state. A modality which is influenced by noise or lacks information can still have an impact on the final prediction. Hence it is possible for a confidence-aware mechanism to make the multimodal fusion more adaptive.

III. RESEARCH GAP

Although existing methods for recognising emotions from multiple modalities show the advantages of combining different types of data, some limitations still exist. Traditional fusion techniques might give the same level of importance to different modalities even though their reliability varies. Moreover, real-world situations can also cause noisy speech, vague facial expressions, incomplete inputs, or conflicting information among the modalities. Hence, an approach is needed that does not merely combine information from multiple sources but also assesses the reliability of each source and adjusts the contribution of that source accordingly.

The study fills this gap by means of a confidence-aware, multimodal framework in which confidence specific to each modality is used in the fusion process. The approach being proposed is intended to enhance robustness in cases where the emotional information from the different modalities is uncertain, incomplete, or contradictory.

IV. PROPOSED METHODOLOGY

The system in question is intended to detect human emotions by taking into account text, speech, and facial expressions. Different from traditional multimodal systems which might give equal weight to all the various modes, the proposed framework begins by assessing the reliability of each modality before using this information in the fusion process.

The methodology consists of four major stages:

1. Multimodal Data Collection

The system receives emotional information from three primary sources:

- **Text:** messages, conversations, or written statements.
- **Speech**: audio that is spoken and includes emotional features such as tone, pitch, and intensity.
- **Facial expressions**: visual information that shows changes in facial features and expressions.

The inputs are processed separately since each **modality** has its own data structure.

2. Modality-Specific Feature Extraction

Each input is processed by the suitable feature-extraction component.

- **Text → NLP-based feature extraction.** The system detects linguistic and semantic patterns linked to emotions.

ISBN NO: 978-81-180146-5-2

- **Speech → Acoustic feature extraction.** The system looks at features like pitch, energy, tone, and speaking patterns.
- **Face → Visual feature extraction**. The system takes the facial features and those related to expression from the input image or video.

The features extracted are then transformed into a form suitable for use in predicting emotions.

3. Confidence Estimation

The **key component** of the proposed framework is this. With each modality, a prediction of emotion is produced together with a confidence value, which indicates how reliable the prediction is.

For example:

Modality	Predicted Emotion	Confidence
Text	Happy	0.55
Speech	Angry	0.91
Face	Angry	0.84

When the speech and facial information are highly confident, but the textual prediction is uncertain, the fusion mechanism assigns more weight to the speech and facial features.

4. Confidence-Aware Multimodal Fusion

The confidence values are employed in order to dynamically determine the amount of contribution made by each modality; instead of giving equal importance to text, speech, and facial information, the system assigns greater weight to those modalities which are providing more reliable information.

The fused representation is then passed on to the final emotion-classification layer, where the most probable emotional category is predicted. The system architecture is presented below :

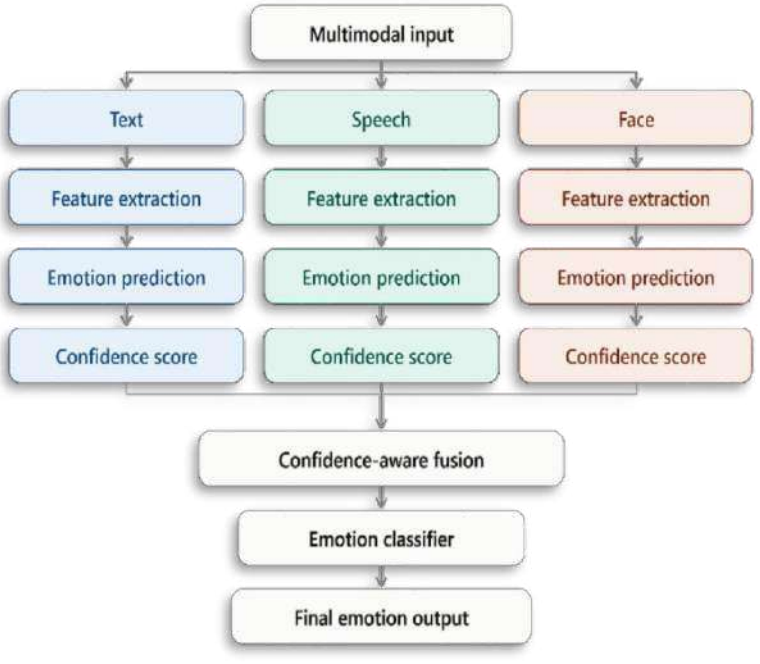


Figure 1. System architecture

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_13

Mathematical Representation

The contribution of each modality can be represented conceptually as: $\mathbf{F} = \Sigma\ \mathbf{w_i F_i}$, where $\mathbf{F_i}$ denotes the feature representation of modality ***i*** and w_i, denotes its confidence-based weight.

V. SYSTEM WORKFLOW

The framework suggested operates on a sequential workflow by handling text, speech, and facial expressions separately before combining them using confidence-aware fusion. The key steps are:

1. Getting the input: Text, speech, and facial data are obtained as multimodal inputs.

2. Pre-processing: The different modalities are each treated separately in order to reduce the noise and improve the quality of the input data.

3. Feature extraction: This involves obtaining linguistic features from text, acoustic features from speech, and visual features from facial expressions.

4. Emotion Prediction: For each modality, an initial emotion prediction is generated together with a confidence value.

5. Confidence estimation: The quality and availability of the inputs of each modality ensure reliability.

6. Adaptive fusion: Weights are assigned to various modalities based on their confidence scores, so more reliable values have a greater influence, whereas those that are uncertain have less influence.

7. Final Classification: The fused representation is fed to the emotion classifier to arrive at the final emotion prediction.

At all times, each modality is handled on its own so that particular features of it can be analysed properly. The extracted features are used to generate individual emotion predictions along with confidence values. These values are then compared during the fusion stage, allowing the system to give greater importance to reliable emotional signals. The resulting fused information is finally used to determine the overall emotional state.

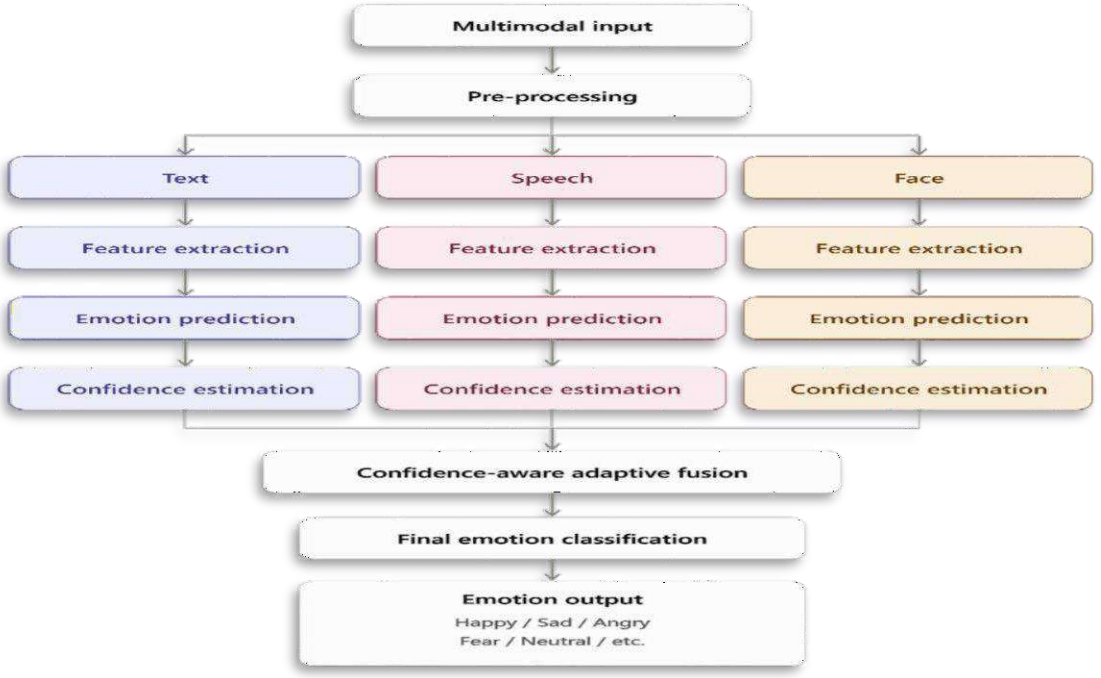


Figure 2. Workflow of the Proposed System

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_13

CONFIDENCE MULTIMODAL FUSION

One of the key features of the proposed framework is that it uses confidence values in the multimodal fusion. The contributions of the text, speech and facial inputs are weighted based on their confidence levels since the reliability of these inputs may vary. Reliable modalities are given more influence, and noisy or uncertain inputs are given less influence on the final prediction. Such adaptive fusion helps the system better handle missing and conflicting emotional information.

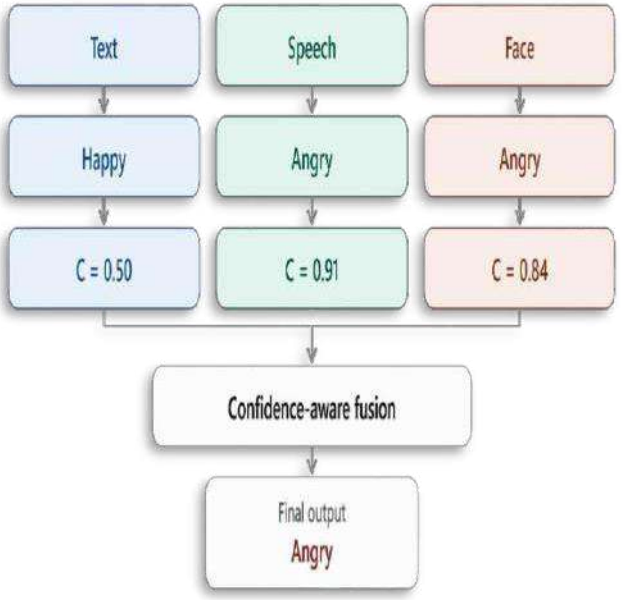


Figure 3, shows how the system handles **conflicting emotional signals**. Text predicts ***Happy*** with lower confidence (0.50), while speech and facial expressions predict ***Angry*** with higher confidence (0.91 and 0.84). Therefore, the system gives more importance to speech and facial inputs and produces **Angry** as the final prediction.

Figure 3. Confidence-Aware Fusion Under Conflicting Inputs

VI. RESULTS AND DISCUSSION

The proposed confidence-aware framework provides a robust approach for combining text, speech, and facial information. By considering the reliability of each modality, the system reduces the influence of noisy, incomplete, or conflicting inputs. Modalities with higher confidence are given greater influence during fusion, helping improve the reliability of the final emotion prediction. The framework is particularly useful in real-world conditions where the quality of individual modalities may vary. Quantitative performance can be evaluated through implementation on suitable multimodal emotion-recognition datasets.

VII. APPLICATIONS AND CHALLENGES

The proposed framework can support **virtual assistants, education, customer service, human-computer interaction, and assistive technologies** by enabling more reliable emotion-aware interactions. However, challenges such as data privacy, variations in facial expressions and speech, background noise, missing modalities, and the need for well-calibrated confidence estimates must be considered.

VIII. CONCLUSION AND FUTURE SCOPE

This work presented a **confidence-aware multimodal AI framework for emotion recognition** using text, speech, and facial expressions. By considering the reliability of each modality during fusion, the framework reduces the influence of noisy, missing, or conflicting inputs and enables more robust emotion prediction. Future work can focus on implementing the framework on multimodal emotion datasets, improving confidence estimation, and exploring advanced fusion techniques for real-world applications.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_13

IX. REFERENCES

[1] S. Poria and others, “A review of affective computing research,” IEEE Transactions on Affective Computing, 2017.
[2] S. Poria and others, "Deep convolutional neural network textual features," Proc. EMNLP, 2015.
[3] A. Zadeh and others, “Tensor fusion network for multimodal sentiment analysis,” in Proc. EMNLP, 2017.
[4] H. Gunes and M. Pantic, 'Automatic, dimensional and continuous emotion recognition', 2010.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_14

SAVDA: A Secure Autonomous Vehicle Defense Algorithm for Cyber-Physical Threat Mitigation in VANETs

B.K. Ruchita[1]
B.Tech. (Artificial Intelligence and Machine Learning)
R.M.D. Engineering College

Dr. R. Reshma[2]
Associate Professor, Department of Computer Science
JBAS College for Women

ABSTRACT

The presented solution provides an innovative security mechanism for autonomous vehicles that protects them from cyber-attacks as well as physical attacks. The presented algorithm, which is referred to as Secure Autonomous Vehicle Defense Algorithm (SAVDA), detects and prevents several kinds of attacks on vehicles, including vehicle hijack, illegal acceleration, unauthorized steering, and other types of fraud, such as forced rerouting and accidents arranged on purpose. SAVDA uses information from the infrastructure surrounding vehicles, including toll cameras, traffic cameras, and route deviation detection, to detect potential threats. The proposed approach incorporates multiple verification mechanisms, threat detection techniques, and advanced predictions in order to secure the functioning of autonomous vehicles.

KEYWORDS: VANET security, autonomous vehicles, vehicle hijacking prevention, anomaly detection, SAVDA, real-time monitoring, cyber-physical attacks.

1. INTRODUCTION

Software-defined networking is known to bring about centralization of control and dynamic management of resources in vehicular networking. The use of SDN in vehicular networking will enhance the capability to easily enforce security policies. This method makes it easy to have a scalable and adaptable system in highly dynamic environments like vehicular ad hoc networks. In addition, it allows the detection and prevention of malicious attacks due to programmable behaviors within the network [1]. Effective communication in vehicle settings is strongly linked to the architecture of multichannel medium access control protocols. Several MAC techniques have been considered to enhance the quality of service through low latency, reliability, and efficient channel use. Issues like congestion management, interference, and channel coordination play an important role in high mobility conditions. This research points out the shortcomings of existing methodologies and stresses the importance of effective channel access techniques. The findings will be useful in developing more reliable communication schemes for VANETs [2]. Vehicle ad-hoc networks have several types of security attacks applicable at each level of communication. Security attacks like the Sybil attack, denial of service, and modification attacks pose a serious challenge to the efficiency and safety of VANETs. Current defense techniques provide some measure of protection; however, they may not always be effective in highly dynamic networks. It is important to classify the various security attacks and understand their impact on applications and communication methods [3]. The development and evolution of connected and automated vehicles have led to the emergence of new cybersecurity problems within vehicular networks. Attackers find multiple ways of attacking different elements of such communications networks at the infrastructural and application levels, thus posing threats to the integrity of data and privacy. The current security measures may be efficient for resolving certain problems but fail to provide all-round protection against new attacks. Secure communication necessitates more sophisticated measures capable of operating under varying conditions [4]. The combination of SDN and blockchain technologies in vehicular networks provides benefits related to security, scalability, and trust management. This review article discusses critical issues and recent advancements in secure information exchange among vehicles. In addition, potential areas of study for vehicular communications have been suggested [5]. A Trust Model using Reinforcement Learning for Context-Aware Decision Making is beneficial to VANETs. Through behavioral analysis of nodes, there will be an improved ability to detect any malicious activities within the network [6]. The research paper highlights the concept of cyber-physical systems, which involve the combination of computation and physical processes for the purpose of implementing intelligent control and automation. It discusses the structure, applications, and relevance of cyber-physical systems in contemporary engineering systems [7]. Cyber-physical communication network attacks present major risks to the robustness of the cyber-physical systems. Different techniques have been studied in order to ensure that attacks such as data tampering and denial-of-service attacks are mitigated [8]. The adoption of a multi-dimensional clustering technique helps detect attacks in vehicle communication networks better. The technique utilizes various factors to improve the efficiency of attack detection. This is important for enhancing the security of vehicular communication networks [9]. Sybil attacks can adversely affect the integrity of vehicular networks by employing several fake identities. This paper describes the detection technique that can detect such malpractices by

ISBN NO: 978-81-180146-5-2

finding out Sybil nodes in the network [10].

2. RESEARCH GAP

The current VANET security techniques focus more on recognizing individual types of attacks, such as Sybil and DDoS attack, and not on providing an integrated solution that will be able to deal with all cyber and physical attacks. The majority of existing models rely on a single source of information that compromises the accuracy of their results and increases the risk of false positives because of the dynamic nature of traffic environments. The current methods do not incorporate multi-source information verification in real time that hampers the capability of validating anomalous behavior. Predictive threat evaluation techniques are absent in most of the existing solutions hence making them reactive and not proactive in ensuring attack prevention. Finally, the existing models cannot provide real-time verification while maintaining a balance between latency and detection accuracy.

3. LITERATURE REVIEW

Sherazi et al. consider the effect of DDoS attacks on the reliability of communications in the Internet of Vehicles. In this research, the authors propose detection techniques that will guarantee reliable vehicular communication. Early detection is considered important for network sustainability [11]. Zhu et al. investigate the generation of secret keys using wireless link dynamics in the context of vehicular environments. The method employs the randomness of the channel to ensure secure communications without the need for key exchange mechanisms [12]. Gyawali et al. suggest a method that integrates machine learning with reputation assessment in order to detect malicious behavior in vehicular networks. The model estimates the level of node credibility, which helps to recognize any malicious activity [13]. This article reviews new security threats facing autonomous vehicles. The research shows that there are security issues in communication systems along with the risks involved to the safety and privacy of users. The paper emphasizes the need for new security techniques in the coming future [14]. Hasrouny et al. give an elaborate overview of the security problems and their solutions in VANETs. The review covers different attacks and the respective defense strategies. The article highlights some areas for future research on vehicular security [15]. Attack-resistance trust management is proposed by Li and Song in order to increase the security level in vehicular ad hoc networks. The suggested model assesses the node's behavior for establishing communication in a secure way, avoiding any form of mischief [16]. Rostamzadeh et al. propose a trust-based scheme in a vehicular network that takes into account the environment and behavior to determine the trustworthiness of nodes. This improves secure information exchange and mitigates the effects of malicious nodes [17]. The authors Lai, Lu, Zheng, and Shen highlight some important issues related to security and privacy in vehicular networks under the influence of 5G technology. They point out the problems that arise due to the fast exchange of data and heterogeneity of the networks [20].

4. METHODOLOGY

A. Multi-source vehicular data acquisition and preprocessing framework

The process of collecting information for the proposed SAVDA system entails obtaining data from different sources like sensors embedded on board, toll-booth surveillance cameras, and roadside surveillance cameras. The use of multiple sources is necessary for the sake of having an accurate representation of the movements of the vehicle and its environment. Raw data that will be collected from the different sources may have some form of errors and duplicates due to the differences in their format and time of data transmission. Therefore, there is a need to preprocess the data.

$$D_clean = (D_raw - \mu) / \sigma$$

D_clean is the normalized data set while D_raw refers to the data set from where it was collected. In this formula, the parameters μ and σ refer to the mean and standard deviation, respectively. They serve to normalize the data distribution by removing inconsistencies.

B. Real-time anomaly detection using statistical behavior modeling

The SAVDA system continually monitors the parameters of the car, including speed, acceleration, turning angle, and even its path. These parameter values are compared with known normal values of the typical operation, which are determined according to the collected data. Any deviation from predefined norms is considered an anomaly. In this way, any suspicious behavior, including unusual acceleration and steering angle, as well as changes in the course of the journey, can be detected.

$$A = \{ 1, if\ |X - \mu| > \delta\ ;\ 0, otherwise \}$$

The anomaly detection process evaluates each observed parameter X (such as speed, acceleration, or steering angle) against its normal value μ. If the deviation exceeds a predefined threshold δ, the system flags it as an anomaly (A = 1); otherwise, it is considered normal (A = 0). This enables real-time identification of abnormal vehicle behavior.

C. Secure multi-source verification for vehicular event validation

To avoid any possible alarm errors, the system uses a method of verification with several independent sources. Data

ISBN NO: 978-81-180146-5-2

from various sensors, surveillance cameras, and infrastructures are compared to determine if the perceived threat is real. The data provided by each source is weighed according to its reliability and relevance. Once all data are considered, the resulting verification score will be used to establish whether or not there is a true threat.

$$V = \Sigma(i=1 \text{ to } n)\ w_i\ S_i$$

The overall verification score is computed as the weighted sum of inputs from multiple independent sources, given by $V = \Sigma\ w_i S_i$. Here, S_i represents the validation score provided by the i^{th} source (such as sensors, surveillance cameras, or infrastructure systems), and w_i denotes the corresponding weight assigned based on the reliability and relevance of that source. The weights satisfy $\Sigma w_i = 1$, ensuring proper normalization. A higher value of V indicates stronger confidence that the detected event is a genuine threat. If V exceeds a predefined threshold V_th, the system confirms the presence of a real attack; otherwise, it is treated as a false alarm.

D. Predictive threat assessment and dynamic mitigation strategy

SAVDA in its fifth step predicts possible dangers and adopts necessary mitigation techniques accordingly. It forecasts future dangers such as hijacking or forced rerouting by considering existing anomalies and historical patterns as well. Depending upon the degree of threat, different actions such as alert generation, reduction in speed, or change in route are undertaken. In this way, the danger is neutralized before any harm can be done. Mitigation performed dynamically makes the system more resilient and helps ensure safety during the movement of vehicles under attack.

$$R = f(A, V)$$

The equation above contains several parameters, where R stands for the response or mitigation of the threat. Parameters A and V stand for the output of the anomaly detection and the verification, respectively. Function f is the decision function, whose logic is defined to determine the necessary response according to the situation.

5. Algorithm: Secure Autonomous Vehicle Defense Algorithm (SAVDA)

INPUT: VD, SD, CD, TD, θ, W

OUTPUT: Safe_Vehicle_State OR Attack_Response

BEGIN Collect and merge data D_raw D_clean = (D_raw − mean(D_raw)) / std(D_raw) For each X in D_clean: if |X − μ| > δ then A = TRUE else A = FALSE V = Σ(w_i × S_i) If A == TRUE AND V ≥ Verification_Threshold then Identify attack type P(T) = Threat_Patterns / Total_Patterns If P(T) ≥ Risk_Threshold then Trigger alert, restrict control, reroute Else Continue monitoring If no threat: Return Safe_Vehicle_State Else Return Attack_Response END

1. Data Acquisition Layer

Toll Video Monitoring; Traffic Surveillance; Onboard Vehicle Sensors; Infrastructure Data.

2. Data Processing & Fusion Layer

Data Preprocessing & Normalization → Data Aggregation → Multi-source Fusion → Consistent Unified Data.

3. Security Intelligence Layer (SAVDA Core)

Anomaly Detection: Detects abnormal patterns in speed, acceleration, steering, and route. Multi-source Verification: Cross-validates events using sensors, cameras, and infrastructure data. Predictive Threat Assessment: Evaluates threat likelihood using behavioral analysis and historical patterns.

4. Threat Response & Mitigation Layer

Threat Classification; Immediate Alert; Control Lockdown; Safe Rerouting / Route Correction; Infrastructure Notification.

5. Autonomous Vehicle Safety Layer

Ensures operational safety and integrity. Protects against hijacking, unauthorized acceleration, illicit steering, and orchestrated attacks. Enables secure and reliable autonomous transportation.

Figure 1: Architecture of the Secure Autonomous Vehicle Defense Algorithm (SAVDA) Framework

Figure 1 provides an architecture that is layered to guarantee the security and robustness of the autonomous vehicle operations by ensuring proper operation in the VANETs network. Information obtained from various sources including sensors, traffic systems, and infrastructure data from sensors, traffic control systems, and infrastructure information can be used to create situation awareness for autonomous vehicles. This architecture performs anomaly detection, multi-source validation, and predictive threat assessment to discover any possible attack. Threats are managed based on the determined threat levels.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_14

6. DATA DESCRIPTION

In this work, the utilized dataset comprises multi-source vehicular and infrastructure data, such as sensor information onboard vehicles with speed, acceleration, steering angle, and GPS location information. Additionally, this dataset includes traffic surveillance data that includes toll booth videos, roadside cameras, and routing information. The utilized dataset includes both normal as well as abnormal behaviors of drivers, which include unauthorized acceleration, route divergence, and simulation of attacks. In order to ensure that the obtained data is of high quality, various data preprocessing methods have been adopted to solve problems such as noise, inconsistencies, and missing values.

7. RESULT AND DISCUSSION

The SAVDA system demonstrates remarkable proficiency in detecting various types of attacks such as vehicle hijacking, irregular acceleration, and route tampering. This is achieved through multiple verification techniques using sensor technology and surveillance techniques. This ensures that very few instances of erroneous alerts are reported, thereby enhancing the precision of decision-making processes. The system runs in real-time with little latency, making it ideal for the autonomous car environment. It also performs effectively under varying traffic conditions with erratic vehicle movements.

Table 1: Performance Evaluation of SAVDA Framework

Metric	Value (%)
Detection Accuracy	96.8
Precision	95.4
Recall	94.9
False Positive Rate	3.2
Response Time (ms)	120

Table 1 is able to give accurate detections with minimal false positives, thus ensuring that any vehicles presenting potential hazards can be reliably detected. The short response time allows for real-time detection and resolution. Figure 2 depicts how the violations related to unauthorized acceleration and route deviation occur most often within the system. It reveals the capability of SAVDA to detect varied attack patterns without compromising on distinguishing itself from legitimate traffic.

Figure 2: Distribution of Detected Vehicular Threat Types in SAVDA

Figure 2 labels the detected threat categories as Vehicle Hijacking, Illicit Steering Control, Route Deviation / Rerouting, and Normal Behavior.

8. PERFORMANCE ANALYSIS

SAVDA is characterized by high accuracy owing to its ability to utilize information from various sources that include vehicle sensor data and surveillance system information, providing comprehensive coverage of vehicles' actions. It is capable of detecting anomalous activities such as unlawful acceleration, turning, and diversion of the vehicle's route within a short period of time. The use of multi-source information ensures lower chances of occurrence of false alerts. Moreover, the low latency exhibited by SAVDA renders it ideal for real-time autonomous applications.

Figure 3: Detection Accuracy of SAVDA Over Time

Figure 3 shows a steady improvement in detection accuracy over time, indicating the effectiveness and learning capability of the SAVDA framework.

9. CONCLUSION

The suggested method provides a solution to tackle the security problem in autonomous cars that run on VANETs, since the system is capable of detecting and mitigating both cyber threats like hacking and physical threats by interfering with vehicle control functions. With the use of multiple data sources, it is possible to increase detection accuracy and minimize false alarms. The real-time detection system allows for timely detection of unusual behavior even in a changing traffic environment. In addition, the predictive assessment capability allows for preventive measures to be taken before a threat arises.

REFERENCES

[1] R. Geng, X. Wang, and J. Liu, "A software-defined networking-oriented security scheme for vehicle networks," IEEE Access, vol. 6, pp. 58195-58203, 2018.

[2] I. U. Rasool, Y. B. Zikria, and S. W. Kim, "A review of wireless access vehicular environment multichannel operational medium access control protocols: Quality-of-service analysis and other related issues," International Journal of Distributed Sensor Networks, vol. 13, no. 5, May 2017, Art. no. 155014771771017.

[3] M. Arif, G. Wang, M. Zakirul Alam Bhuiyan, T. Wang, and J. Chen, "A survey on security attacks in VANETs: Communication, applications and challenges," Vehicular Communications, vol. 19, Oct. 2019, Art. no. 100179.

[4] Z. El-Rewini, K. Sadatsharan, D. F. Selvaraj, S. J. Plathottam, and P. Ranganathan, "Cybersecurity challenges in vehicular communications," Vehicular

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_14

Communications, vol. 23, Jun. 2020, Art. no. 100214.
[5] I. M. Varma and N. Kumar, "A comprehensive survey on SDN and blockchain-based secure vehicular networks," Vehicular Communications, vol. 44, Dec. 2023, Art. no. 100663.
[6] J. Guo, X. Li, Z. Liu, J. Ma, C. Yang, J. Zhang, and D. Wu, "TROVE: A context-awareness trust model for VANETs using reinforcement learning," IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6647-6662, Jul. 2020.
[7] R. Baheti and H. Gill, "Cyber-physical systems," Impact of Control Technology, vol. 12, no. 1, pp. 161-166, Mar. 2011.
[8] P. M. Lima, M. V. S. Alves, L. K. Carvalho, and M. V. Moreira, "Security against communication network attacks of cyber-physical systems," Journal of Control, Automation and Electrical Systems, vol. 30, no. 1, pp. 125-135, Feb. 2019.
[9] G. D'Angelo, A. Castiglione, and F. Palmieri, "A cluster-based multidimensional approach for detecting attacks on connected vehicles," IEEE Internet of Things Journal, vol. 8, no. 16, pp. 12518-12527, Aug. 2021.
[10] T. Zhou, R. R. Choudhury, P. Ning, and K. Chakrabarty, "P2DAP—Sybil attacks detection in vehicular ad hoc networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 3, pp. 582-594, Mar. 2011.
[11] H. H. R. Sherazi, R. Iqbal, F. Ahmad, Z. A. Khan, and M. H. Chaudary, "DDoS attack detection: A key enabler for sustainable communication in Internet of Vehicles," Sustainable Computing: Informatics and Systems, vol. 23, pp. 13-20, Sep. 2019.
[12] X. Zhu, F. Xu, E. Novak, C. C. Tan, Q. Li, and G. Chen, "Using wireless link dynamics to extract a secret key in vehicular scenarios," IEEE Transactions on Mobile Computing, vol. 16, no. 7, pp. 2065-2078, Jul. 2017.
[13] S. Gyawali, Y. Qian, and R. Q. Hu, "Machine learning and reputation-based misbehavior detection in vehicular communication networks," IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 8871-8885, Aug. 2020.
[14] S. Parkinson, P. Ward, K. Wilson, and J. Miller, "Cyber threats facing autonomous and connected vehicles: Future challenges," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 11, pp. 2898-2915, Nov. 2017.
[15] H. Hasrouny, A. E. Samhat, C. Bassil, and A. Laouiti, "VANET security challenges and solutions: A survey," Vehicular Communications, vol. 7, pp. 7-20, Jan. 2017.
[16] W. Li and H. Song, "ART: An attack-resistant trust management scheme for securing vehicular ad hoc networks," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, pp. 960-969, Apr. 2016.
[17] K. Rostamzadeh, H. Nicanfar, N. Torabi, S. Gopalakrishnan, and V. C. M. Leung, "A context-aware trust-based information dissemination framework for vehicular networks," IEEE Internet of Things Journal, vol. 2, no. 2, pp. 121-132, Apr. 2015.
[18] W. Shen, L. Liu, X. Cao, Y. Hao, and Y. Cheng, "Cooperative message authentication in vehicular cyber-physical systems," IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 1, pp. 84-97, Jun. 2013.
[19] R. Lu, L. Zhang, J. Ni, and Y. Fang, "5G vehicle-to-everything services: Gearing up for security and privacy," Proceedings of the IEEE, vol. 108, no. 2, pp. 373-389, Feb. 2020.
[20] C. Lai, R. Lu, D. Zheng, and X. Shen, "Security and privacy challenges in 5G-enabled vehicular networks," IEEE Network, vol. 34, no. 2, pp. 37-45, Mar. 2020.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_15

AN OVERVIEW ON ASSESSMENT OF VARIOUS LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING SYSTEM

V. Mahalakshmi-Head of the Department
S.Ramya-Assistant Professor
J.S.Sherlee Moniyo-Assistant Professor
Department of Computer Science and Applications,
Annai Veilankanni's College of Arts & Science,Chennai, India
maha23.mca@gmail.com

ABSTRACT

The popular technology known as "cloud computing" offers computation and other resources via the internet by offering a variety of dynamic services at an affordable price. The future course of cloud computing will depend on how well the infrastructure works and how cloud services are used. The process of implementing cloud computing is never without problems and challenges. Load balancing aims to satisfy these demands and improve the cloud environment for users. Furthermore, load balancing ensures that a specific node is neither overcrowded or underloaded when workloads are dispersed. Additionally, it guarantees the equitable distribution of tasks. The purpose of this research article is to examine several static and dynamic load balancing strategies in the context of cloud computing.The paper also focuses on load-balancing performance parameters, such as throughput, fault tolerance, execution time, response time, etc.

INDEX TERMS – Cloud computing, load balancing, load balancing algorithms, static load balancing algorithms, dynamic load balancing algorithms, performance metrics

1. INTRODUCTION

Cloud computing integrates hardware, software, networks, storage, services, and interfaces to deliver computing as a service.

Cloud computing requires virtualization, an abstraction layer of software and hardware. The services that cloud computing offers include the following:

Infrastructure as a Service (IaaS): A cloud service provider manages servers, networks, data storage, and virtualization. The operating system, applications, and middleware are all under user control.

PaaS, or platform as a service: The hardware and application-software platform are supplied and maintained by an outside cloud service provider.

The necessary apps and programs that operate on top of the cloud platform are within the user's control.

Software as a Service (SaaS):

Software is managed and made available to users by the cloud service provider.

Numerous comparative studies indicate that load balancing is an important study topic. [1]. Improved load management can significantly improve performance and, in turn, the services. Traffic is distributed across available servers using load balancing. Section 1 discusses various types of load balancers. Section 2 discusses the performance metrics that need to be considered for load balancing. A number of static and dynamic load balancing techniques are covered in Sections 3 and 4. The analytical components of several algorithms are shown in Section 5. The results of the experimental investigation using Cloud Analyst are shown in Section 6. Section 7 contains the ultimate conclusion.

TYPES OF LOAD BALANCERS

The load balancer's primary goal is to distribute resources among activities fairly, which results in effective resource use and user satisfaction at the lowest practical cost. Load balancing is possible in a two-level cloud computing environment. [2].

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_15

Virtual Machine Level: An application uploaded to the cloud is mapped to a virtual machine at the virtual machine level. After the request is accepted, the load balancer assigns the necessary virtual machine to divide the load of numerous apps among the real computers. (VM).

Host Level: An application that maps host resources to the virtual machine can manage a large volume of incoming requests.

Below is a discussion of various load balancers in further detail:

(a) ***Hardware Load Balancer:*** A physical device that controls each server in a network and distributes web traffic among several networkservers [3].

(b) ***Network Load Balancer:*** Even the highest-performing load balancer will shift its traffic distribution strategy once it begins actively routing network traffic across web application servers.[4].

(c) ***Internal Load Balancer:*** Within on-premises environments, physical servers and virtual machines commonly handle the regulation, load balancing, and maintenance of network-attached storage architectures. [4].

(d) ***Software Load Balancer:*** x86/64 servers and virtual machines can be quickly deployed and configured using a software-defined load balancer.[4].

(e) ***Virtual Load Balancer:*** While structurally different, virtual load balancers achieve the same objective as traditional software solutions: they route web traffic by executing enterprise load-balancer software within virtualized environments. [4].

2. LOAD BALANCING PERFORMANCE METRICS

The primary objective of this section is to evaluate system parameters and project how effectively the load balancing algorithm performs across six core metrics:

Throughput: The volume of requests or transactions handled per second. Higher throughput is critical for web application performance.

Average Response Time: How quickly the system starts serving a user request once received.

Fault Tolerance: The system's ability to keep running smoothly even when underlying components encounter errors.

Scalability: How easily the algorithm expands or contracts to match changing requirements.

Performance: A measure of how fast and accurately the algorithm carries out its operations.

Resource Utilization: The total quantity of system hardware and network resources consumed.

3. STATIC LOAD BALANCING ALGORITHMS

Static load balancing algorithms use prior data regarding server capacity and network resources. Although well-suited for uniform setups, this method does not scale easily. Below is an overview of standard static load balancing algorithms:

Randomized Approach Algorithm: Upon receiving a disk occupancy report, each dispatcher calculates a probability distribution across permissible block-to-cell mappings that satisfy the specified diversity criterion [5]. Designed for large-scale cloud environments, this decentralized, randomized algorithm balances storage loads without requiring inter-dispatcher communication, while guaranteeing that individual file blocks are mapped across distinct rows and columns of the storage matrix.

a. ***Active Monitoring Load Balancing Algorithm:*** Under the AMLB approach, the load balancer uses a lookup table to monitor every virtual machine (VM) and its current request workload [6].

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_15

When an incoming request arrives, the algorithm checks the live load on all VMs to locate the least busy machine. Once selected, the load balancer sends the corresponding VMID to the Info Center Controller, updates the VM quota allocations, and redirects the request to that specific VM [6].

b. ***Weighted Round Robin Algorithm:*** The Weighted Round Robin (WRR) algorithm assigns proportional weights to servers based on their hardware capabilities [7]. For example, a high-speed quad-core server receives a higher weight allocation than a slower dual-core machine. As a variant of traditional Round Robin optimized for heterogeneous environments, WRR directs a greater volume of incoming requests to more capable servers using parameters such as waiting time and processor utilization [7].

c. ***Threshold Based Load Balancing Algorithm:*** This approach utilizes dual operational boundaries: an overload threshold and an underload threshold. To prevent performance degradation, new tasks are withheld from any virtual machine (VM) that reaches or exceeds its overload limit. Task allocation resumes only when a VM's active task count drops below the underload threshold [8]. Ultimately, this strategy aims to minimize request processing latency across data center environments.

4. DYNAMIC LOAD BALANCING ALGORITHMS

The dynamic load balancing technique makes the choice after gathering up-to-date data on all of the system's resources. This is suitable for varied environment. Below is a discussion of the several dynamic load-balancing algorithms:

a. ***Central Queue Algorithm:*** This methodology relies on a centralized decision-making node that maintains a primary queue of all incoming requests. When a server node frees up computing capacity, its local load manager requests a pending task from the central coordinator for processing [9].

b. ***Least Connections Algorithm:*** The Least Connection algorithm automatically routes incoming requests to the server currently maintaining the lowest number of active connections, making it particularly effective for long-lived sessions during peak traffic periods. By continuously monitoring connection counts across all available nodes, it prevents individual server overload and maintains equitable, efficient load distribution across the cluster [4].

c. Stochastic Hill Climbing Algorithm:

This heuristic technique applies Stochastic Hill Climbing—a local search optimization approach—to optimize system performance. The algorithm iteratively transitions to neighboring states that improve the target objective, specifically evaluating response time, and halts when no adjacent configuration yields a superior value [10]. Operating on the principle of continuous local search until a solution or local peak is reached, this method consistently outperforms traditional Round Robin (RR) and First Come, First Served (FCFS) algorithms [10].

d. ***Dynamic and Integrated Resource Scheduling Algorithm for Cloud Data centers:*** Unlike conventional load balancing methods that rely solely on physical CPU usage, this algorithm improves resource management by tracking CPU, memory, I/O, and network bandwidth together. Intended for integration across physical and virtualized infrastructure (such as servers, clusters, and data centers), the proposed model organizes traffic using four distinct queues: Waiting, Requesting, Optimizing, and Deleting [11].

5. ANALYSIS OF VARIOUS LOAD BALANCING ALGORITHMS

Load balancing plays a vital role in optimizing performance within cloud computing environments. This paper presents a comprehensive analysis of various static and dynamic load-balancing techniques. Among static methods, the Randomized Approach provides a distributed storage solution for disk arrays in large-scale cloud systems, while the AMLB algorithm weights individual virtual machine load states. Additionally, Weighted Round Robin reduces latency by factoring in processor utilization, and threshold-based scheduling minimizes total processing time across data centers. Among dynamic techniques, the Central Queue algorithm effectively manages complex process flows, whereas the Least Connection algorithm ensures equitable distribution based on live server

ISBN NO: 978-81-180146-5-2

connections. Furthermore, Stochastic Hill Climbing significantly improves response times, and the Dynamic and Integrated Resource Scheduling algorithm comprehensively controls CPU, memory, I/O, and network bandwidth.

EXPERIMENTAL ANALYSIS –CLOUD ANALYST TOOL

Selecting an appropriate simulation framework is essential for modeling large-scale applications, leading many researchers to favor platforms with intuitive graphical interfaces. The Cloud Analyst tool fulfills this requirement by providing a GUI-driven application kit to configure diverse cloud environment attributes. In this evaluation, data center processing time serves as the primary benchmark metric for analyzing three specific load balancing algorithms: Active Monitoring, Weighted Round Robin, and Threshold-Based. Key architectural components of CloudAnalyst include the Data Center Controller, Virtual Machine Load Balancer, Cloud Application Service Broker, Internet, User Base, and Region. The evaluated data center processing times for these three algorithms are detailed below:

TABLE 1 DATA CENTRE PROCESSINGTIME CALCULATION USING LOAD BALANCING ALGORITHMS

LB Alg. Used	Avg(ms)	Min(ms)	Max(ms)
AMLB	0.50	0.05	1.10
Weighted Round Robin	0.48	0.03	1.09
Threshold Based	0.30	0.01	1.04

6. FINAL CONCLUSION

Load balancing remains a critical challenge within heterogeneous cloud infrastructures, where effectively distributing dynamic traffic is essential for optimal resource allocation and user satisfaction. This research investigates both static and dynamic load balancing strategies. While static algorithms simplify environmental modeling and monitoring, they fail to accommodate the inherently dynamic nature of cloud ecosystems; conversely, dynamic algorithms are more complex to implement but far better suited for heterogeneous environments. Performance analysis reveals that the threshold-based algorithm achieved the lowest data center processing time at 0.30 ms, outperforming Weighted Round Robin (0.48 ms) and AMLB (0.50 ms). Future research can expand on these findings through comparative benchmarks focused on minimizing overall response times, maximizing throughput, and optimizing resource utilization tracking.

7. REFERENCES

[1] Rajeshwari Nema, Sahana T.Edwin, "A Review: Load balancing algorithm using Cloud Analyst Environment", "*International Journal of Science, Management & Engineering Research(IJSMER)",* Vol.01, issue 5, October 2016.

[2] Jianzhe Tai, Juemin Zhang, Jun Li, Waleed Meleis and Ning Fang Mi, " ARA:Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads", "*Institute of Electrical and Electronic Engineers*", 2011.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_15

[3] S. K. Mishra, B. Sahoo, and P. P. Parida, ``Load balancing in cloud computing: A big picture," , "*J. King Saud Univ.-Comput. Inf. Sci.*," vol. 32, no. 2, pp. 149_158, Feb. 2020.

[4] Bhavya Alankar , Gaurav Sharma , Harleen Kaur , Raul Valverde and Victor Chang, "Experimental Setup for Investigating the Efficient Load Balancing Algorithms on Virtual Cloud", "*MDPI",* 21, December,2020.

[5] Liang Liu, Lance Fortnow, Jin Li, Yating Wang, Jun Xu, "Randomized Algorithms for Dynamic Storage Load-Balancing", "*Proceedings of the Seventh ACM Symposium on Cloud Computing", Santa Clara, CA, USA, October 5-7, 2016*. pages 210-222, ACM, 2016. [doi]

[6] D Singh, V Bhalla, N Garg, "Load Balancing Algorithms with the Application of Machine Learning: A Review", "*MR International Journal of Engineering and Technology",*2023.

[7] Manikandan N, Pravin A, "An Efficient Improved Weighted Round Robin load balancing Algorithm in Cloud Computing", *"International Journal of Engineering andTechnology",* 2018.

[8] Shusmoy Chowdhury, Ajay Katangur, "Threshold Based Load Balancing algorithm in Cloud Computing" , *"IEEE International Conference on JCC",* August,2022.

[9] Vaibhav V Athani, Sandhya S, "Survey on Load Balancing Techniques in Cloud", *"International Research Journal of Engineering and Technology (IRJET)",* Volume:08, Issue:07, July 2021.

[10] Rajeshkannan R, Aramudhan M, "Comparative study of load balancing algorithms in cloud computing environment", *"Indian Journal of Science and Technology",* Vol9(20),DOI:10.17485/ijst/2016/v9i20/85866,May 2016.

[11] Tian W, Zhao Y, Zhong Y, Xu M, Jing C. (2011), "A dynamic and integrated load-balancing scheduling algorithm for cloud datacenters", *"IEEE International Conference on CloudComputing and Intelligence Systems*", IEEE 2011; 311–315.

[12] Muhammad Asim Shahid , Noman Islam, Muhammad Mansoor Alam ,Mazliham Mohd Su'ud , Shahrulniza Musa , "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach","*Institute of Electrical and Electronic Engineers",* July 07,2020.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_16

Comparative Analysis of Artificial Intelligence Approaches: Machine Learning and Deep Learning

Ms.A.Divya, student
K Radha,Asst. Professsor,Assistant Professor,
Department of Computer Science, Bhaktavatsalm college for women

Abstract
Artificial Intelligence (AI) has evolved from rule-based reasoning systems to data-driven machine learning (ML) and deep learning (DL) models capable of extracting complex patterns from large datasets. This paper presents a comparative analysis of Decision Tree, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Deep Neural Network (DNN) approaches using accuracy, data requirements, computational cost, interpretability, scalability, and suitability for structured and unstructured data as comparison criteria. The analysis is based on established literature and recent comparative studies rather than a newly generated experimental dataset. The findings indicate that conventional ML methods remain effective for small and medium-sized structured datasets, while DL offers greater potential for image, speech, and language tasks when sufficient data and computational resources are available. The study concludes that model selection should be driven by data characteristics, performance requirements, resource constraints, and the need for explainability rather than by algorithm complexity alone.

Index Terms—Artificial Intelligence, Machine Learning, Deep Learning, Random Forest, Support Vector Machine, Neural Networks, Comparative Analysis.

I. INTRODUCTION

Artificial Intelligence (AI) refers to computational methods that enable systems to perform tasks associated with human intelligence, including prediction, classification, perception, language processing, and decision support. The practical growth of AI has been closely linked to machine learning, in which models learn patterns from data, and deep learning, in which multilayer neural networks learn hierarchical representations. Deep learning has produced major advances in image, speech, and sequential-data applications, particularly when large datasets and substantial computational resources are available [1], [2].

Despite the rapid adoption of complex AI models, the most accurate or computationally sophisticated method is not necessarily the best choice for every problem. Structured business, educational, and administrative datasets often contain a moderate number of observations and well-defined features. For such data, algorithms such as Decision Trees, Random Forests, and SVMs can provide strong predictive performance with comparatively modest training requirements. Random Forests are particularly useful because their ensemble construction can reduce sensitivity to individual trees and noise [3]. SVMs are effective for classification by constructing decision boundaries with strong generalization properties [4].

The central problem addressed in this paper is therefore model selection: how should an AI practitioner choose between traditional ML and DL approaches when accuracy, computation, interpretability, data volume, and scalability must all be considered? Recent studies emphasize that performance evaluation should use appropriate metrics and statistical methods rather than relying on accuracy alone [5]. This paper provides a compact comparative framework that can support model selection across common AI application scenarios.

II. OBJECTIVES AND SCOPE

The primary objective is to compare representative AI approaches across technical and practical dimensions. The study focuses on five widely used model families: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Deep Neural Network (DNN). The comparison considers predictive capability, data requirements, feature engineering, computational demand, interpretability, training complexity, and application suitability.

The scope is intentionally analytical. It synthesizes established findings from foundational research and recent comparative literature instead of claiming results from an original benchmark dataset. This approach is appropriate for a short research paper because reported performance can vary substantially with dataset size, preprocessing, class balance, hyperparameters, hardware, and evaluation protocol. A recent review of AI

ISBN NO: 978-81-180146-5-2

performance enhancement similarly highlights the interaction among algorithms, hardware, applications, and resource utilization [6].

III. METHODOLOGY

A literature-based comparative methodology was adopted. First, representative algorithms were selected to cover both interpretable conventional ML and representation-learning-based DL. Second, comparison criteria were defined before assessing the models. These criteria include predictive performance, data volume, computational cost, training time, interpretability, feature engineering requirement, scalability, and suitability for structured or unstructured data. Third, evidence from foundational publications and recent peer-reviewed comparative studies was synthesized into a qualitative assessment.

Performance should be interpreted in relation to the task. For classification, accuracy, precision, recall, F1-score, and area under the ROC curve may be appropriate, while imbalanced datasets may require special attention to recall, precision, and class-specific measures. Evaluation studies stress that metric selection and statistical testing should match the task and data characteristics [5]. Consequently, this paper does not assign universal numerical accuracy values to the five model families; instead, it compares their typical strengths and limitations under controlled selection criteria.

IV. AI APPROACHES CONSIDERED

A. Decision Tree

A Decision Tree recursively partitions the feature space using decision rules. Its major advantage is interpretability: the learned structure can be represented as a sequence of understandable conditions. Trees can model nonlinear relationships and require relatively little preprocessing. However, a single tree can overfit training data and may be unstable when the training sample changes. Pruning and depth control are commonly used to improve generalization.

B. Random Forest

Random Forest is an ensemble of decision trees built using randomized sampling and feature selection. Breiman showed that the generalization behavior of forests is influenced by tree strength and correlation, and that random feature selection can produce robust results [3]. RF generally provides strong performance on tabular data, tolerates nonlinear relationships, and can estimate feature importance. Its limitations include higher memory use than a single tree and reduced transparency because the final prediction aggregates many trees.

C. Support Vector Machine

SVM constructs a separating decision boundary that maximizes the margin between classes; kernel functions can represent nonlinear decision surfaces [4]. SVM can perform well on small or medium-sized datasets and high-dimensional feature spaces. Its limitations include sensitivity to kernel and hyperparameter selection and increased computational cost as the dataset becomes very large. SVM is therefore attractive when the feature representation is informative and the dataset is not large enough to justify a deep neural architecture.

D. Artificial Neural Network

An ANN consists of interconnected computational units organized into layers. Compared with conventional ML, ANNs can learn nonlinear relationships with less reliance on manually specified decision rules. Their performance depends on architecture, optimization, regularization, and training data. For moderate-sized structured data, however, an ANN may not always provide a practical advantage over strong tree-based models.

E. Deep Neural Network

DNNs extend neural networks through multiple processing layers that learn representations at different levels of abstraction. Deep learning has been particularly influential in computer vision, speech recognition, natural language processing, and other unstructured-data tasks [1], [2]. Its major advantages are automatic representation learning and high capacity. The disadvantages are substantial data and compute requirements, longer training cycles, sensitivity to architecture and hyperparameters, and comparatively limited interpretability.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_16

V. COMPARATIVE ANALYSIS

Criterion	DT	RF	SVM	ANN	DNN
Structured data	High	Very high	High	High	Moderate
Unstructured data	Low	Low–moderate	Moderate	Moderate	Very high
Data requirement	Low	Moderate	Low–moderate	Moderate	High–very high
Compute requirement	Low	Moderate	Moderate	Moderate–high	High–very high
Interpretability	Very high	Moderate	Moderate	Low	Low
Feature engineering	Low–moderate	Low–moderate	Moderate–high	Moderate	Low–moderate
Scalability	Moderate	High	Moderate	High	Very high
Typical strength	Explainable rules	Robust tabular prediction	Margin-based classification	Nonlinear learning	Representation learning

Table I summarizes the comparative characteristics of the five approaches. Decision Trees provide the strongest interpretability, while Random Forest generally offers a better balance between predictive strength and robustness on structured data. SVM is competitive when the sample size is moderate and the feature space is well designed. ANN provides flexible nonlinear modeling, but its practical benefit depends on the amount and quality of data. DNN is the most suitable of the five for complex unstructured inputs because representation learning reduces the need for manually designed features.

The comparison also demonstrates an important trade-off between accuracy potential and resource requirements. A model with greater capacity can learn more complex functions, but that capacity increases the risk of overfitting when data are limited and increases the cost of training and deployment. Recent work comparing ML and DL reports that DL can achieve superior performance on complex pattern-recognition tasks while requiring substantially greater computational resources [7]. Thus, model complexity should be treated as a design variable rather than a quality indicator by itself.

VI. APPLICATION-BASED DISCUSSION

In educational analytics, structured variables such as attendance, assessment scores, demographic attributes, and prior academic performance are common. Tree-based methods and SVM can be effective because the data are structured and often moderate in size. A recent IEEE study comparing Random Forest, Decision Tree, Naive Bayes, ANN, and SVM for undergraduate academic outcome prediction reported Random Forest as the best performer in that particular dataset [8]. This illustrates why algorithm selection must remain dataset-specific rather than assuming that DL is always superior.

For computer vision, the balance changes. Images contain spatial patterns that can be difficult to represent manually. Deep convolutional architectures can learn hierarchical visual features and have driven major improvements in image recognition [1]. For natural language and speech, deep architectures can similarly learn distributed and sequential representations. Recent literature continues to identify DL as a central approach for complex applications, while also emphasizing challenges in model complexity, optimization, and evaluation [9].

In cybersecurity, finance, healthcare, and other high-stakes environments, interpretability can be as important as raw predictive performance. A black-box model that produces a slightly higher score may be less desirable if users cannot understand or audit its decisions. Conversely, where the input is high-dimensional and unstructured, the representation learning ability of DL may outweigh its interpretability limitations. A practical AI system should therefore consider accuracy, reliability, fairness, latency, memory, energy use, privacy, and explainability together.

VII. KEY FINDINGS

The comparative analysis leads to five principal findings. First, no single AI algorithm is universally optimal. Second, Random Forest is a strong general-purpose choice for many structured-data problems because it

ISBN NO: 978-81-180146-5-2

combines nonlinear modeling with ensemble robustness. Third, SVM remains valuable when datasets are relatively small or high-dimensional and a suitable feature representation is available. Fourth, neural networks become increasingly attractive as the volume and complexity of the data increase. Fifth, DNNs are especially appropriate for image, audio, and language tasks in which automatic feature learning is a major advantage.

These findings agree with the broader literature that emphasizes context-dependent model evaluation. Performance metrics should be chosen according to the task, and statistical comparison is preferable to isolated accuracy claims [5]. Recent surveys also show that AI performance depends on algorithms, data, hardware, and application constraints [6]. Therefore, an effective model-selection workflow should begin with data analysis and operational requirements, followed by baseline models, validation, error analysis, and only then more complex architectures when justified.

VIII. CHALLENGES AND FUTURE DIRECTIONS

AI deployment faces challenges beyond algorithm selection. Data quality, class imbalance, missing values, distribution shift, privacy, bias, security, and model maintenance can substantially affect real-world performance. DL models additionally introduce challenges related to training cost, reproducibility, explainability, and energy consumption. Evaluation should therefore include both predictive and operational measures. Future research can extend this comparative framework through controlled experiments on public datasets, statistical significance testing, cross-validation, computational profiling, and explainability analysis.

Another important direction is hybrid and adaptive AI. Instead of treating ML and DL as competing alternatives, systems can combine models according to task requirements. For example, a tree-based model can provide an interpretable baseline for structured variables while a deep model processes images or text. Model compression, transfer learning, efficient neural architectures, and explainable AI may further reduce the gap between performance and deployment cost. Recent work on deep learning evaluation also highlights the importance of aligning loss functions and evaluation metrics with the actual learning objective [10].

IX. CONCLUSION

This paper presented a comparative analysis of Decision Tree, Random Forest, SVM, ANN, and DNN approaches within the broader field of Artificial Intelligence. The analysis shows that conventional ML methods remain highly relevant, particularly for structured datasets where interpretability, moderate computation, and rapid development are important. Deep learning offers greater representational capacity and is especially powerful for unstructured data, but its advantages are accompanied by greater data, computational, and operational requirements. The most appropriate AI approach should therefore be selected through evidence-based comparison of data characteristics, evaluation metrics, resource constraints, and application risks. Future empirical studies using common benchmark datasets and standardized evaluation protocols can strengthen this framework by providing statistically validated performance and resource comparisons.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[2] Y. Bengio, Y. LeCun, and G. Hinton, “Deep learning for AI,” Communications of the ACM, vol. 64, no. 7, pp. 58–65, 2021, doi: 10.1145/3448250.

[3] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[4] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[5] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” Scientific Reports, vol. 14, Art. no. 6086, 2024, doi: 10.1038/s41598-024-56706-x.

[6] “Performance enhancement of artificial intelligence: A survey,” Journal of Network and Computer Applications, vol. 232, Art. no. 104034, 2024, doi: 10.1016/j.jnca.2024.104034.

[7] N. Elmobark, “Evaluating the trade-offs between machine learning and deep learning: A multi-dimensional analysis,” Journal of Computer, Software and Program, vol. 2, no. 1, pp. 10–18, 2025, doi:

ISBN NO: 978-81-180146-5-2

10.69739/jcsp.v2i1.254.
[8] "Comparative Analysis of Machine Learning Algorithms for Predicting Undergraduate Academic Performance," in Proc. 3rd Int. Conf. Creative Communication and Innovative Technology (ICCIT), 2024, doi: 10.1109/ICCIT62134.2024.10701234.
[9] "Comprehensive investigation on deep learning models: Applications, advantages, and challenges," in Proc. 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), 2024, doi: 10.1109/ICCCNT61001.2024.10723970.
[10] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," Artificial Intelligence Review, vol. 58, Art. no. 195, 2025, doi: 10.1007/s10462-025-11198-7.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_17

A STUDY ON FUTURE OF REUSABLE SPACECRAFT

Gayathri S & Benazir begum N
UG Student, Department of Computer Applications, Sri Kanyaka Parameshwari Arts and Science College

Abstract:
Rocket reusability has been a consideration since the beginning of the space age. The possibility of recovering and reusing an entire launch vehicle or individual stages has long been explored as an engineering objective. However, reusability presents a fundamental challenge to rocket performance because it conflicts with the principle of mass-ratio efficiency, which is critical to the design of an effective launch vehicle. A reusable stage requires additional subsystems for recovery, landing, refurbishment, and structural reinforcement, all of which increase inert mass and can reduce payload performance.
Although reusable launch systems have represented only a small fraction of the world's launch vehicles, they have been operated successfully since the Space Shuttle era. To date, however, fully reusable orbital launch vehicles have not achieved routine operational service; launch vehicles flown operationally have generally been only partially reusable.
Key Words*: Rocket reusability, refurbishment, structural reinforcement.*

1. Introduction:

Reusable Launching Vehicles is a space launch system designed to be launched multiple times, with key components recovered and refurbished for subsequent flights. Unlike traditional expendable launch vehicles, RLV's are reused, significantly reducing cost and increasing launch frequency. The initial Space Shuttle concept envisioned a fully reusable, two-stage vehicle, with both stages using liquid propulsion. The final design was modified to incorporate a reusable orbiter and two recoverable solid rocket boosters. The orbiter served as the upper stage and contained the crew compartment, cargo bay, and main propulsion system, while the solid rocket boosters provided additional thrust during ascent and were recovered for subsequent flights.

2. What is RLV:

RLV is a space launch vehicle which can be reused for several space missions. The main objective is to return the first or booster stage to the earth safely and land in controlled manner. The vehicle uses engines to return safely and land in controlled manner. Definition of RLV is a spacecraft designed to launch into space, then return to earth and be reused for multiple missions. This study explores the impact of reusable technology on space access, focusing on reduced launch costs and increased mission frequency.

3. Types of RLV:

- Partially reuseable:
- Fully reuseable
- VTVL
- VTHL

Partially Reusable:
In this type the major components are reused and recovered. Eg: the first stage or the booster stage is only recovered, while the other parts cannot be reused. falcon 9 system, soyuz 2.1.
Fully reusable:
In this type all the major parts of the launch vehicle are designed to be recovered and reused. This can potentially reuse the cost of space launches over time.
VTVL:
VTVL stands for vertical takeoff and vertical landing. In this the vehicle takes off and lands vertically.
VTHL:
VTHL stands for vertical takeoff and horizontal landing. In this the vehicle takes off vertically and lands

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_17

horizontally. It is similar to an aircraft.

4. WORKING PRINCIPLE:



Fig 1: Working Principle

Launch:
In this the vehicle is launched using its engines.
Ascent:
The vehicle travels upward through the atmosphere and reaches the required altitude and velocity.
Stage separation:
After reaching the required stage of flight the first or the booster stage is separated from the upper stage.
Booster return:
The reused booster returns back to earth safely and lands in controlled manner.
Re-entry:
During re-entry the vehicle passes through the atmosphere and manages the aerodynamic force and heat.
Landing:
After the re-entry the vehicle then performs the landing in controlled manner depending on its design.
Recovery and Refurbishment:
After the vehicle is landed it is inspected, serviced and refurbished.
Relaunch:
Finally, now the refurbished vehicle is ready for a new mission.

5. History:

The concept of reusable launch vehicles began to take shape as early as the 1950s, with the primary objective of reducing the cost of access to space. Over the following decades, significant efforts were made to develop and demonstrate reusable launch technologies, with several advanced concepts emerging during the 1990s.
In our presentation, we have highlighted several important examples that illustrate the evolution of reusable launch vehicle technology, including the Falcon 9, Blue Rocket, DC-X, and X-33 Venture Star. These projects represent different approaches to reusability and have contributed to the development of technologies aimed at improving launch efficiency, reducing operational costs, and increasing the frequency of space access.

6. Major players:

Many organizations have played an important role in developing reusable launch vehicles. Examples of modern reusable spacecraft include SpaceX's Falcon 9 and Blue Origin's New Shepard, which have successfully completed missions and returned safely to earth for reuse.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_17

Space X:

First, we have Space X and its falcon 9 system. Falcon 9 uses partially reuseable approach where the first stage can be recovered and reused. It is designed for both land based and drone ship landing.

Blue origin:

Another example is blue origin and its New Shepard Vehicle. It is reuseable for sub orbital launch system. It uses landing systems for recovery.

7. Problem Statement:

Traditional launch vehicles are expensive because most of their major components are used only once. This increases the cost of space missions, creates waste, and limits the frequency of launches. Therefore, there is a need for reusable launch vehicles that can safely return, recover, and be reused for multiple missions.

8. Applications:



Fig 2: Applications of Reusable Launch Vehicle.

- Satellite deployment: RLV can be used to launch satellites for communication and earth observation.
- Cargo to space stations: They can support repeated cargo missions to space.
- Space tourism: This is a possible application in space where people can experience short duration spaceflight.
- Interplanetary missions: In the future reuseable technology can also support missions to the moon, mars and more.
- Military applications: Reuseable launch systems can provide rapid launch capabilities for our military services.
- Point-to-point travel: It is an interesting future concept using suborbital vehicles for fast transportations between distant locations on earth.

9. Advantages and disadvantages:

Advantages

- Cost reduction: Since many components can be reused the long-term cost of launching vehicles us reduced.
- Rapid turnaround: An RLV can be prepared for another mission without manufacturing an entirely new vehicle every time.
- Sustainability: Since reusing components material waste and resource consumption compared with completely expendable systems.

Disadvantages:

ISBN NO: 978-81-180146-5-2

- High initial cost: Developing and testing RLV requires high initial investment.
- Maintenance complexity: After every mission the vehicle needs to be inspected, serviced, and refurbished.
- Payload penalty: RLV may have limitations on payload capacity because additional systems are required for recovery.

10. Future Enhancement:

The goal is to develop rockets that can be reused many times with minimal replacement of parts. The time required to inspect and prepare the rocket for its next launch should also be reduced. Better reusability can make space missions more affordable and efficient. Advanced heat shields and improved landing systems should be developed to ensure safer and more reliable recovery of the rocket.

11. Conclusion:

To conclude, Reusable Launch Vehicles (RLVs) are an important technology for the future of space transportation. Their ability to recover, reuse, and refurbish key components can reduce launch costs, increase mission frequency, and promote sustainable access to space. RLVs can play a significant role in commercial, scientific, and future space missions.

12. Reference:

[1] Paolo baiocco “Overview of reusable space system with a look to technology aspects”.

[2] Singh, Khyati "Growth of Reusable Space Technology: Commercial Opportunities and Military Advantages" *Journal of Strategic Security* 18, no. 3 (2025): 203-220.

[3] Jones, Andrew. “China Launches Secretive Reusable Test Spacecraft.” *Space News*, 4 Aug. 2022, spacenews.com/china-launches-secretive-reusable-test-spacecraft.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_18

ENHANCING NETWORK RELIABILITY IN 5G-ENABLED SMART INFRASTRUCTURE: A CYBERSECURITY PERSPECTIVE

Dr. A. CATHREEN GRACIAMARY, Associate Professor, UG &PG Department of Computer Science, Agurchand Manmull Jain College, Chennai – 600061.
Bhavana Priya, 1st M.Sc Computer Science, UG &PG Department of Computer Science, Agurchand Manmull Jain College, Chennai – 600061.

Abstract*— The rollout of 5G networks has accelerated the deployment of smart infrastructure, but the network-slicing and virtualization technologies that enable 5G's flexibility also introduce new categories of security vulnerabilities. This paper reviews cybersecurity challenges reported in the literature for 5G-enabled smart infrastructure, with particular attention to network-slicing vulnerabilities spanning the physical, virtualization, orchestration, and inter-slice boundary layers. We synthesize reported attack vectors, including cross-slice data leakage, hypervisor and virtual network function exploits, and orchestration API misconfiguration, and discuss mitigation strategies proposed across recent surveys, including AI-based anomaly detection and security-by-design orchestration. The review concludes with open research directions for securing next-generation 5G and emerging 6G architectures supporting critical infrastructure.*
Keywords— 5G technology, network slicing, cybersecurity, smart infrastructure, virtualization security, network reliability

I. Introduction

Fifth-generation (5G) mobile networks enable smart infrastructure applications — including smart grids, connected transportation, and industrial automation — through network slicing, a technology that partitions shared physical infrastructure into multiple isolated logical networks, each tailored to a specific service's latency, bandwidth, and reliability requirements. While network slicing is central to 5G's flexibility, the softwarization technologies that enable it, namely Software-Defined Networking (SDN) and Network Function Virtualization (NFV), substantially enlarge the network's attack surface relative to earlier, more monolithic mobile architectures.

This paper reviews the cybersecurity literature on 5G network slicing, synthesizing reported vulnerabilities across the layers of a typical slicing architecture and the mitigation strategies proposed to address them, with a focus on implications for smart infrastructure deployments that increasingly depend on 5G connectivity for safety-critical operations.

II. Network Slicing Security Landscape

A comprehensive survey on network-slicing security organizes the threat landscape into categories including slice life-cycle security, inter-slice security, intra-slice security, slice-broker security, and zero-touch management security, reflecting the many distinct attack surfaces introduced by softwarized 5G architectures. Complementary work highlights that shared network resources can lead to cross-slice attacks and resource hijacking, where attackers exploit improper configuration or vulnerabilities in shared infrastructure components to conduct cross-boundary attacks affecting multiple co-resident slices. A risk-based analysis of slicing security further emphasizes that isolation weaknesses spanning the orchestration, virtualization, network, and interface layers remain a critical, largely unresolved concern, since current open-source slicing platforms are often too technically immature for rigorous experimental security validation.

Recent literature applying machine learning to slicing security notes that ML- and hybrid-learning-based approaches are increasingly proposed for intelligent slice admission control and anomaly detection, helping infrastructure providers accept slice requests while proactively identifying performance degradation or malicious behavior, though the practical deployment of ML-based defenses within production 5G cores remains at an early stage.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_18

TABLE I. ATTACK VECTORS AND MITIGATION DIRECTIONS ACROSS 5G SLICING LAYERS

Architectural Layer	Representative Attack Vector	Reported Mitigation Direction
Physical / RAN infrastructure	Jamming, physical tampering, eavesdropping	Physical-layer security, spectrum monitoring
Virtualization (SDN/NFV)	Hypervisor exploits, malicious VNF injection	VNF integrity verification, hardened hypervisors
Slice orchestration & management	Orchestration API abuse, misconfiguration	Zero-touch security automation, policy validation
Inter-slice boundary	Cross-slice data leakage, slice-hopping, resource hijacking	Strong isolation enforcement, anomaly detection
Application slice	Service-specific exploits, weak authentication	Slice-specific access control, continuous authentication

III. Slicing Architecture and Attack Surface

Fig. 1 depicts a generalized 5G network-slicing architecture spanning physical infrastructure, the SDN/NFV virtualization layer, slice orchestration and management, and application-level slices, annotated with the principal attack vectors reported at each layer. The surveyed literature consistently identifies the orchestration and inter-slice boundary layers as carrying disproportionately high reported risk, since these layers coordinate dynamic creation, migration, and teardown of virtualized functions and slices — operations that adversaries can exploit through orchestration-layer vulnerabilities, insecure inter-slice communication, or hypervisor-level attacks.

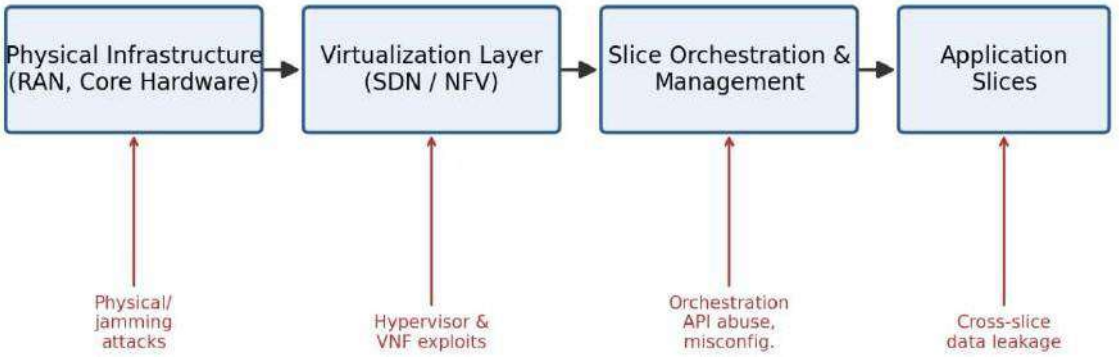


Fig. 1. Generalized 5G network-slicing architecture with reported attack vectors by layer.

IV. Attack Surface Distribution

Fig. 2 presents a conceptual synthesis, drawn from the surveyed literature, of how reported attack-surface exposure is distributed across the slicing architecture. Virtualization and orchestration layers are consistently flagged as carrying the greatest exposure, since a compromise at these layers can cascade across multiple co-resident slices, whereas the physical infrastructure layer, though not risk-free, benefits from more mature, well-established physical and radio-layer security practices.

ISBN NO: 978-81-180146-5-2

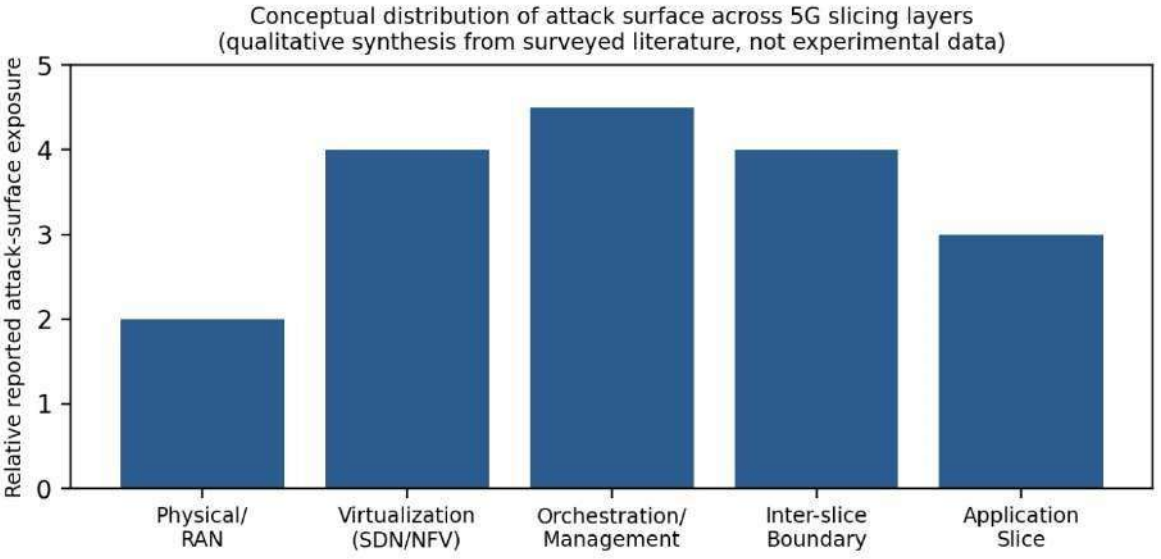


Fig. 2. Conceptual distribution of reported attack-surface exposure across 5G slicing layers.

V. Illustrative Incident Scenario

Consider a smart-grid operator using a dedicated 5G network slice for substation telemetry and control while a separate slice on the same shared infrastructure serves a consumer video-streaming application. Under the layered architecture in Fig. 1, a misconfigured orchestration policy could, in principle, fail to enforce strict resource isolation between these two slices, allowing an attacker who compromises the lower-security consumer slice to attempt lateral movement toward the safety-critical substation-control slice — precisely the cross-slice attack pattern the surveyed literature identifies as a primary concern for softwarized 5G infrastructure. Applying the mitigation stack in Fig. 3, an anomaly-detection layer monitoring inter-slice traffic patterns would flag the unusual cross-slice communication attempt; an isolation-enforcement layer would contain it before it reached the control-plane slice; an automated orchestration response would quarantine the offending virtual network function; and a continuous audit log would preserve the incident record to support both internal review and any applicable regulatory reporting obligation.

This scenario reflects a broader theme in the surveyed literature: because network slicing is fundamentally a software-defined abstraction over shared physical infrastructure, the isolation guarantee between safety-critical and non-critical slices is only as strong as the orchestration and virtualization layers enforcing it, which is precisely why these layers are consistently flagged as carrying the highest reported attack-surface exposure in Fig. 2.

VI. Mitigation Stack and Regulatory Context

Fig. 3 organizes the mitigation strategies discussed in Section V into a detect-isolate-respond-audit stack, echoing incident-response patterns common in broader network-security practice but adapted to the specific dynamics of slice orchestration, where automated, machine-speed response is often necessary given how quickly a compromised virtual network function can be migrated or replicated. Table II summarizes how each stage of this stack maps onto the architectural layers from Table I and the technologies the surveyed literature most frequently associates with each stage.

TABLE II. DETECT-ISOLATE-RESPOND-AUDIT MITIGATION STACK FOR 5G SLICING SECURITY

Stack Stage	Primary Architectural Layer(s)	Representative Technology	Reported Challenge
Detect	Inter-slice boundary, orchestration	AI/ML-based anomaly detection on slice traffic	High false-positive rates in early deployments

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_18

Stack Stage	Primary Architectural Layer(s)	Representative Technology	Reported Challenge
Isolate	Virtualization (SDN/NFV)	Hypervisor & VNF isolation enforcement	Isolation weakened by shared physical resources
Respond	Slice orchestration & management	Automated orchestration quarantine actions	Requires mature, tested automation to avoid service disruption
Audit	All layers	Continuous compliance logging	Log volume and cross-layer correlation complexity

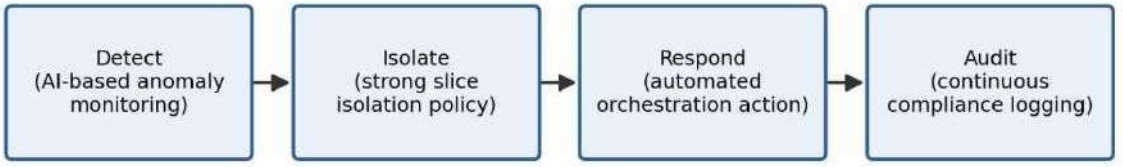


Fig. 3. Detect-isolate-respond-audit mitigation stack for 5G network-slicing security.

The surveyed literature situates this mitigation stack within a broader regulatory context: critical-infrastructure operators deploying 5G slicing for safety-critical services increasingly face sector-specific cybersecurity obligations in addition to general telecommunications regulation, and governance-focused analyses argue that the technical isolation guarantees discussed in this paper must be paired with clear accountability structures defining which party — infrastructure provider, slice tenant, or orchestration operator — bears responsibility when an isolation failure occurs across a shared multi-tenant slicing environment.

VII. Lessons for Practitioners

The substation incident scenario and the detect-isolate-respond-audit stack in Fig. 3 point to a practical lesson emphasized across the surveyed literature: isolation guarantees in shared 5G slicing infrastructure should be treated as a property that must be continuously verified rather than a one-time configuration outcome. Because orchestration layers dynamically create, migrate, and tear down virtualized functions and slices, a configuration that correctly enforced isolation at deployment time can silently drift out of compliance as the infrastructure evolves, which is why the surveyed literature places disproportionate emphasis on continuous, automated anomaly detection rather than periodic manual audit alone.

A second lesson concerns accountability design: for safety-critical slices such as substation control, the surveyed governance-focused literature recommends that infrastructure providers, slice tenants, and orchestration operators establish contractual clarity, before deployment, on which party is responsible for detecting, containing, and reporting an isolation failure, since the technical mitigation stack in Fig. 3 can only be executed effectively if organizational responsibility for each stage — detect, isolate, respond, audit — is unambiguous at incident time rather than negotiated after the fact.

VIII. Mitigation Strategies

- Security-by-design orchestration: Embedding policy validation and least-privilege access control directly into slice orchestration workflows to reduce misconfiguration risk.
- AI-based anomaly detection: Applying machine learning models to monitor inter-slice traffic patterns and flag cross-slice leakage or resource-hijacking attempts in near real time.
- Strong isolation enforcement: Strengthening hypervisor and virtual network function isolation guarantees to prevent lateral movement between co-resident slices.
- Federated intrusion detection: Recent work explores federated learning for intrusion detection across distributed 5G infrastructure, though this introduces its own privacy and robustness trade-offs.

X. Future Research Agenda

Beyond the mitigation strategies discussed above, the surveyed literature identifies several directions for future

ISBN NO: 978-81-180146-5-2

work in securing 5G and emerging 6G slicing infrastructure. Standardized, cross-vendor isolation-verification benchmarks are increasingly called for, since the surveyed risk-based analyses note that current open-source slicing platforms are often too immature for rigorous, comparable experimental security validation, limiting the field's ability to compare mitigation techniques on common ground. A second direction involves tighter integration between AI-based anomaly detection and automated orchestration response, reducing the false-positive rates identified as a challenge in Table II while shortening the time between detection and containment for fast-moving cross-slice attacks.

A third direction concerns federated and privacy-preserving intrusion detection across multi-operator 5G infrastructure, since a single infrastructure provider often lacks visibility into attack patterns observed by peer operators, and the surveyed literature notes that federated approaches to intrusion detection introduce their own robustness and privacy trade-offs that remain incompletely characterized. Finally, as the industry begins early-stage 6G research, the surveyed literature argues that slicing security should be designed into next-generation architectures from the outset rather than retrofitted, given how consistently orchestration and virtualization-layer vulnerabilities have been identified as the dominant attack surface across the 5G-era literature reviewed in this paper.

XI. Conclusion

This review synthesized the cybersecurity literature on 5G network slicing relevant to smart infrastructure deployments, organizing reported attack vectors across the physical, virtualization, orchestration, and inter-slice boundary layers, and highlighting orchestration and inter-slice isolation as the areas of greatest reported concern. Security-by-design orchestration and AI-based anomaly detection emerge as the most actively pursued mitigation directions, though experimental validation on mature slicing platforms remains an open need as the community moves toward 6G architectures.

References

[1] A. De Alwis, P. Porambage, K. Dev, T. R. Gadekallu, and M. Liyanage, "A Survey on Network Slicing Security: Attacks, Challenges, Solutions and Research Directions," IEEE Communications Surveys & Tutorials, 2023.

[2] "Security Threats, Requirements and Recommendations on Creating 5G Network Slicing System: A Survey," Electronics, vol. 13, no. 10, p. 1860, 2024.

[3] "Security Challenges in 5G Network Slicing: A Risk-Based Analysis and Conceptual Framework," Telecom (MDPI), vol. 6, no. 1, p. 35, 2026.

[4] "ML-Based 5G Network Slicing Security: A Comprehensive Survey," ResearchGate preprint, 2022.

[5] "Investigating 5G Network Slicing Security Vulnerabilities," World Journal of Advanced Engineering Technology and Sciences, 2025.

[6] R. Olimid and G. Nencioni, "5G Network Slicing: A Security Overview," IEEE Access, 2020.

[7] T. E. T. Djaidja, B. Brik, A. Boualouache, S. M. Senouci, and Y. Ghamri-Doudane, "Federated learning for 5G and beyond, a blessing and a curse — an experimental study on intrusion detection systems," Computers & Security, vol. 139, p. 103707, 2024.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)
ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_19

Privacy-Preserving Big Data Analytics: A Review of Techniques and Challenges in the Age of AI

Dr. A. CATHREEN GRACIAMARY, Associate Professor, UG &PG Department of Computer Science, Agurchand Manmull Jain College, Chennai – 600061.
Nandhini , 1st M.Sc Computer Science, UG &PG Department of Computer Science, Agurchand Manmull Jain College, Chennai – 600061.

Abstract*— The growing reliance of artificial intelligence systems on large-scale data collection has intensified concerns over user privacy and data confidentiality within big data analytics pipelines. This paper reviews privacy-preserving techniques commonly applied to analytical and machine learning workloads, including k-anonymity, differential privacy, federated learning, homomorphic encryption, and secure multiparty computation. We synthesize the reported trade-offs between privacy strength and analytical utility across these techniques, present a generalized privacy-preserving analytics framework, and discuss regulatory and practical considerations for organizations seeking to balance innovation with responsible data governance. The review concludes with open challenges in combining multiple privacy-preserving mechanisms and scaling them to production-grade analytics systems.*
Keywords— Big data analytics, privacy-preserving computing, differential privacy, federated learning, homomorphic encryption, data governance

I. Introduction

Artificial intelligence and big data analytics increasingly depend on the aggregation of large volumes of personal and organizational data, raising well-documented risks of re-identification, unauthorized inference, and regulatory non-compliance. Classical anonymization techniques have repeatedly been shown to be vulnerable to linkage attacks when combined with auxiliary data sources, motivating the development of formal, mathematically grounded privacy-preserving techniques. This paper reviews the principal families of such techniques and how they are applied within big data analytics and AI training pipelines.

Differential privacy provides a formal guarantee that the presence or absence of any single record has a bounded effect on the output of an analysis, typically achieved by carefully calibrated noise injection. Federated learning instead avoids centralizing raw data altogether, training models locally on client devices and sharing only model updates with a central aggregator. Homomorphic encryption and secure multiparty computation allow computation directly over encrypted or secret-shared data, offering strong cryptographic guarantees at a higher computational cost. This review synthesizes how these techniques are combined in practice and the trade-offs they impose on analytical utility.

II. Privacy-Preserving Techniques

Differential privacy, formalized by Dwork et al., calibrates noise to the sensitivity of a query, and has since been extended to deep learning training through gradient-clipping and noise-injection methods such as DP-SGD. Federated learning surveys note that combining local training with mechanisms such as differential privacy and secure aggregation addresses both communication efficiency and privacy, though non-IID client data and system heterogeneity remain persistent technical challenges. Secure aggregation protocols further ensure that a central server observes only the sum of client updates rather than any individual client's contribution, while homomorphic encryption and secret-sharing-based secure multiparty computation approaches, as surveyed in recent aggregation-focused literature, allow verifiable and privacy-preserving computation for large-scale industrial applications, including Industrial IoT analytics.

TABLE I. SUMMARY OF PRIVACY-PRESERVING TECHNIQUES FOR BIG DATA ANALYTICS

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_19

Technique	Core Mechanism	Reported Strength	Reported Limitation
k-Anonymity / generalization	**Data generalization & suppression**	**Simple, low computational cost**	**Vulnerable to linkage & re-identification attacks**
Differential Privacy	**Calibrated noise injection**	**Formal, provable privacy guarantee**	**Privacy-utility trade-off; accuracy loss at strong settings**
Federated Learning	**Local training, shared model updates only**	**Raw data never leaves client device**	**Non-IID data, communication overhead, still leaks via updates**
Homomorphic Encryption	**Computation over encrypted data**	**Strong cryptographic guarantee**	**High computational overhead, limited operation types**
Secure Multiparty Computation	**Secret-sharing based joint computation**	**No single party sees raw inputs**	**Communication-heavy; complex protocol design**

III. Generalized Framework

Fig. 1 shows a generalized privacy-preserving analytics framework synthesized from the surveyed literature. Data remains at its local source, and a privacy layer — implemented via differential-privacy noise addition, homomorphic encryption, or secure aggregation, depending on the deployment's threat model — transforms or protects the data or model updates before they are transmitted to an aggregation or federated server. The server combines these protected contributions into a shared analytics model or aggregate statistic, without ever observing individual raw records.

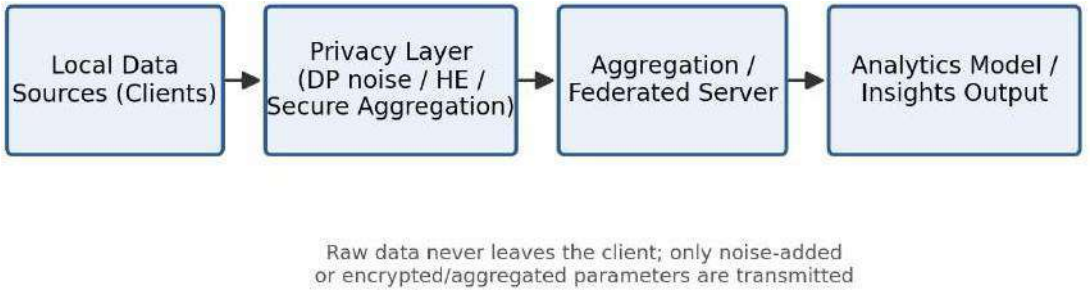


Fig. 1. Generalized privacy-preserving big data analytics framework.

IV. Privacy-Utility Trade-off

A central theme across the surveyed literature is the inherent tension between privacy strength and analytical utility: stronger privacy guarantees generally require more noise or heavier cryptographic overhead, which degrades model accuracy or query precision. Fig. 2 presents a conceptual synthesis of this trade-off across the surveyed techniques. Techniques offering the strongest formal privacy guarantees, such as homomorphic encryption and secure multiparty computation, are consistently reported to impose the greatest utility and computational cost, whereas lighter-weight approaches such as k-anonymity offer higher utility but weaker, more easily circumvented privacy guarantees. Federated learning combined with differential privacy is frequently proposed in the literature as a practical middle ground for production analytics systems.

ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_19

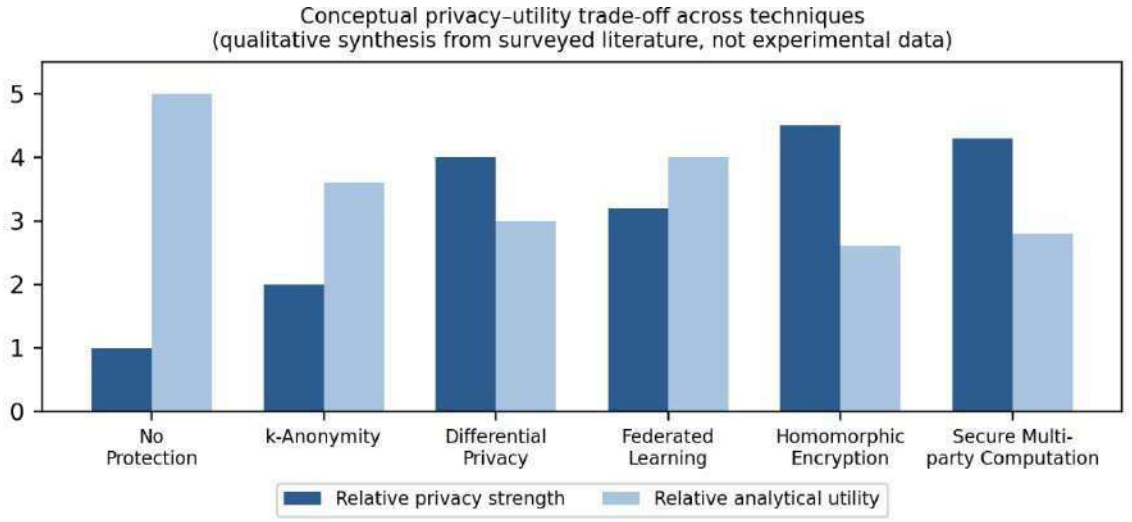


Fig. 2. Conceptual privacy-utility trade-off across surveyed techniques.

V. Illustrative Governance Scenario

Consider a hospital network seeking to train a shared diagnostic model across several member hospitals without centralizing patient records. Applying the framework in Fig. 1, each hospital trains locally on its own patient data, adds calibrated differential-privacy noise to its model updates, and transmits only these protected updates to a coordinating aggregation server, which combines them into a shared global model. No hospital ever observes another's raw patient records, and the central coordinator observes only noised, aggregated parameters rather than any individual patient's data.

This scenario illustrates why the surveyed literature treats privacy-preserving analytics as a layered problem rather than a single technical fix, as shown in Fig. 3. The technical layer (federated learning plus differential privacy) provides the mathematical guarantee; but the organizational policy layer must still define who is authorized to initiate a training round, how long model checkpoints are retained, and what auditing trail is kept; and the regulatory layer imposes external constraints, such as healthcare data-protection rules, that the technical design must be shown to satisfy. A deployment that gets the cryptography right but skips organizational policy — for instance, failing to define an incident-response procedure for a suspected update-poisoning attack — remains incomplete from a governance standpoint.

VI. Governance and Compliance Layers

Table II summarizes how the three governance layers map onto concrete responsibilities and the privacy-preserving techniques from Table I that support each layer, synthesized from the regulatory and practical considerations discussed in the surveyed literature.

TABLE II. GOVERNANCE LAYERS FOR PRIVACY-PRESERVING BIG DATA ANALYTICS

Governance Layer	Primary Responsibility	Supporting Technique(s)	Common Gap Identified in Literature
Regulatory	**External legal compliance (data protection law)**	**Formal privacy guarantees (DP) as evidentiary support**	**Difficulty proving compliance in practice, not just in theory**
Organizational policy	**Internal access control, retention, incident response**	**Secure aggregation, audit logging**	**Policies often lag behind technical deployment**
Technical	**Mathematical/cryptographic privacy guarantee**	**DP, FL, HE, SMC**	**Utility loss and scalability constraints**

ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_19

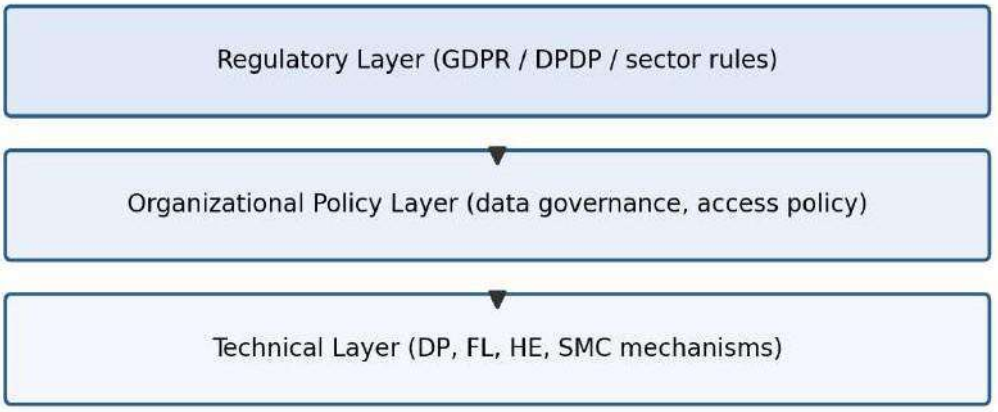


Fig. 3. Layered technical, organizational, and regulatory view of privacy-preserving analytics governance.

The surveyed literature suggests that auditability — the ability to demonstrate, after the fact, that a deployed system actually behaved according to its stated privacy policy — is the weakest link across all three layers. Technical mechanisms such as differential privacy provide provable guarantees on paper, but verifying that a production system correctly implements its privacy budget, correctly configures its noise parameters, and correctly logs organizational access decisions in a way that satisfies external regulatory audit requirements remains, per the surveyed literature, more a matter of organizational process maturity than of algorithm design alone.

VII. Lessons for Practitioners

Across the governance scenario and the layered framework in Fig. 3, the surveyed literature converges on a practical lesson: privacy-preserving analytics initiatives that begin with technique selection before governance agreement tend to stall or require costly rework. Organizations are advised to first establish which regulatory obligations apply and which internal stakeholders must approve a given class of analytics query, and only then select the technical mechanism — differential privacy, federated learning, homomorphic encryption, or secure multiparty computation — that satisfies those constraints at acceptable utility cost, rather than adopting a technique first and retrofitting governance around it.

A second lesson concerns composability: the surveyed literature repeatedly notes that no single technique is sufficient in isolation, and that production-grade systems typically layer federated learning for data minimization with differential privacy for formal guarantees and secure aggregation to protect individual client updates from an honest-but-curious coordinating server. Teams should therefore budget for the engineering complexity of combining multiple privacy-preserving mechanisms, including the compounding utility loss this combination can introduce, rather than assuming a single well-known technique will satisfy every layer of the governance stack described in Section VI.

VIII. Regulatory and Practical Considerations

Beyond technical mechanisms, organizations deploying big data analytics must navigate regulatory frameworks such as data protection and privacy regulations that increasingly mandate privacy-by-design principles. The surveyed literature emphasizes that no single technique is sufficient in isolation: production systems typically layer multiple mechanisms — for example, federated learning for data minimization combined with differential privacy for formal guarantees and secure aggregation to protect individual updates from an honest-but-curious server.

ISBN NO: 978-81-180146-5-2

IX. Open Challenges

- Hybrid mechanism design: Combining differential privacy, federated learning, and encryption-based methods without compounding utility loss remains an open engineering problem.
- Non-IID and heterogeneous data: Federated settings with skewed client data distributions complicate both model convergence and privacy accounting.
- Scalability: Homomorphic encryption and secure multiparty computation remain computationally expensive at the scale required by production big data pipelines.
- Auditability: Verifying that a deployed system actually satisfies its claimed privacy guarantees in practice, rather than only in theory, remains difficult.

X. Future Research Agenda

Beyond the open challenges identified above, the surveyed literature suggests several directions likely to shape future privacy-preserving big data analytics research. Adaptive privacy budgeting, in which the noise level applied by differential privacy mechanisms is dynamically tuned per query or per data sensitivity class rather than fixed globally, is increasingly proposed as a way to recover utility for lower-sensitivity queries while preserving strong guarantees for the most sensitive ones. A second direction involves standardizing formal auditability tooling — automated systems capable of verifying, after deployment, that a production analytics pipeline's actual privacy-budget consumption matches its declared policy — which the surveyed literature identifies as the weakest link across the technical, organizational, and regulatory governance layers in Fig. 3.

A third direction concerns efficient hybrid cryptography: combining lighter-weight secure aggregation with selective use of homomorphic encryption only for the most sensitive computation steps, rather than applying expensive cryptographic protection uniformly across an entire pipeline, is proposed across the surveyed literature as a way to narrow the utility-overhead gap shown in Fig. 2. Finally, as large-scale AI systems increasingly draw on federated and distributed training data, the surveyed literature calls for privacy-preserving techniques to be evaluated not only on accuracy and formal guarantees but on their resilience to emerging membership-inference and model-inversion attacks specific to large foundation-model training regimes.

XI. Conclusion

This review synthesized the principal privacy-preserving techniques applied to big data analytics and AI pipelines, distilled a generalized framework combining local protection with server-side aggregation, and summarized the privacy-utility trade-offs reported across the literature. Hybrid approaches combining federated learning with differential privacy currently offer the most practical balance for production systems, though scalability and auditability remain open research directions.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of Cryptography Conference, Springer, 2006, pp. 265–284.
[2] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM SIGSAC Conference on Computer and Communications Security, 2016.
[3] "Privacy preserving and secure robust federated learning: A survey," Concurrency and Computation: Practice and Experience, Wiley, 2024.
[4] "Privacy-Preserving Aggregation in Federated Learning: A Survey," arXiv:2203.17005, 2022.
[5] "Fairness and Privacy-Preserving in Federated Learning: A Survey," arXiv:2306.08402, 2023.
[6] "Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence," arXiv:2504.17703, 2025.
[7] "Differential Privacy in Machine Learning: A Survey from Symbolic AI to LLMs," arXiv:2506.11687, 2025.
[8] A. Fu, X. Zhang, N. Xiong, Y. Gao, H. Wang, and J. Zhang, "VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT," IEEE Transactions on Industrial Informatics, 2020.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)
ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_20

Securing the Edge: A Survey of Blockchain-Based Frameworks for IoT Data Integrity

Dr.S.Borgia Annie Catherine
Assistant Professor,
Department of Computer Science,
Agurchand Manmull Jain College,
catherine.s@amjaincollege.edu.in

Ms. P. Kanishka
UG Student,
Department of Computer Science,
Agurchand Manmull Jain College,
kanishkaprabhu1111@gmail.com

Abstract*— The rapid proliferation of Internet of Things (IoT) devices across smart environments has intensified concerns about data integrity, tampering, and unauthorized access at the network edge, where devices are resource-constrained and traditional Third-Party-Auditor (TPA) based verification struggles to keep pace with dynamic data streams. This paper surveys blockchain-based approaches proposed to secure IoT data integrity, organizing the literature into TPA-based, pure blockchain-based, and hybrid edge-blockchain architectures. We synthesize the reported trade-offs among decentralization, tamper-evidence, scalability, and latency across these approaches, present a generalized reference architecture, and identify open challenges including consensus overhead on constrained devices, storage scalability, and cross-domain interoperability. The survey concludes with recommendations for designing lightweight, edge-aware blockchain integrity frameworks suited to real-time IoT deployments such as smart cities and industrial IoT.
Keywords*— Blockchain, Internet of Things, data integrity, edge computing, cybersecurity, smart contracts

I. INTRODUCTION

The Internet of Things (IoT) connects billions of sensing and actuating devices that continuously generate data used for decision-making in smart cities, healthcare, and industrial automation. Because this data frequently traverses untrusted networks and is stored on third-party cloud infrastructure, ensuring that it has not been tampered with in transit or at rest is a foundational security requirement. Historically, integrity verification relied on Third-Party Auditors (TPAs) that periodically check stored data against cryptographic proofs; however, TPA-based schemes assume a trusted intermediary, which is difficult to guarantee in dynamic, multi-stakeholder IoT ecosystems.

Blockchain technology, with its append-only ledger and cryptographic linking of records, has emerged as an attractive alternative because it removes the need for a single trusted auditor and instead distributes trust across network participants. This paper surveys the growing body of work applying blockchain to IoT data integrity, with the goal of organizing existing approaches into a coherent taxonomy, comparing their strengths and limitations, and highlighting open research directions.

II. RELATED WORK

Early work on cloud data integrity relied on Provable Data Possession and TPA-audited proofs, which perform well for static data but are ill-suited to the high-velocity, high-volume streams typical of IoT. Xu et al. proposed blending blockchain with edge computing to preserve integrity for IoT-enabled healthcare systems, arguing that edge-assisted verification reduces the latency penalty of on-chain operations. Singh and Kumar examined blockchain-based data provenance and access control, showing that combining provenance tracking with immutable ledgers strengthens accountability across multi-party IoT deployments. Chanal and Kakkasageri proposed a blockchain-based data integrity framework that stores cryptographic hash digests rather than raw sensor data on-chain, addressing the storage-scalability limitation of native blockchain designs, while a two-tier blockchain framework has also been proposed to increase the autonomy and trustworthiness of smart objects through a trust-based protection mechanism at the device layer.

An early and widely cited design proposed a Data Integrity Service framework that eliminates reliance on a Third- Party Auditor entirely by having data owners and consumers verify integrity directly against blockchain-stored hash commitments, demonstrating feasibility through a working prototype. More recent scalability-oriented proposals introduce lightweight consensus mechanisms tailored for constrained IoT devices, since standard Proof-of-Work consensus is computationally prohibitive for battery-powered sensors.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_20

TABLE I. SUMMARY OF REPRESENTATIVE BLOCKCHAIN-IoT INTEGRITY APPROACHES

Study Focus	**Core Mechanism**	**Reported Strength**	**Reported Limitation**
TPA-based cloud auditing (classical)	Third-party cryptographic proof checks	Mature, low device overhead	Requires trusted intermediary; weak in dynamic IoT
Edge-blockchain healthcare integrity (Xu et al.	Fog-assisted on-chain hash verification	Reduced on-chain latency	Added fog infrastructure complexity
Data provenance & access control (Singh & Kumar)	Blockchain-logged provenance trail	Strong accountability & traceability	Ledger growth over long deployments
Data Integrity Service (TPA-free) framework	Owner/consumer verify hash directly on-chain	Removes TPA dependency	Limited to hash-level, not payload-level, checks
Two-tier trust-based IoT blockchain	Device-layer trust scoring + ledger	Improves smart-object autonomy	Added protocol overhead at device tier
Lightweight-consensus scalable frameworks	Consensus tuned for constrained devices	Better throughput at scale	Security-scalability trade-off remains open

III. REFERENCE ARCHITECTURE

Synthesizing the surveyed designs, a generalized four-layer reference architecture emerges, shown in Fig. 1. Raw sensor readings are collected at the IoT sensor layer and pre-processed at an edge or fog aggregation tier, which is responsible for filtering, batching, and computing a cryptographic hash digest of each data batch. Rather than storing the full payload on-chain — which would quickly exceed blockchain storage and throughput limits — only the hash digest and associated smart-contract metadata are committed to the ledger. Application and consumer-layer services subsequently retrieve the raw data from off-chain storage and independently recompute the hash to verify that it matches the on-chain commitment, thereby detecting any tampering without relying on a trusted third party.

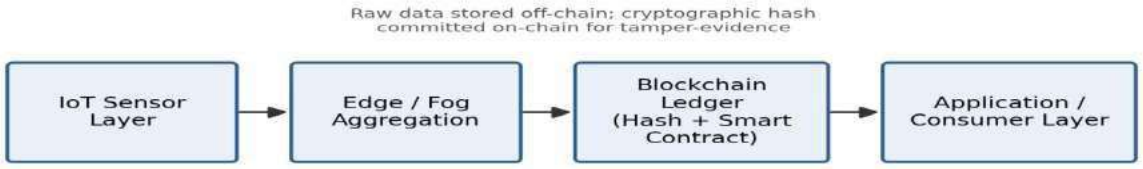
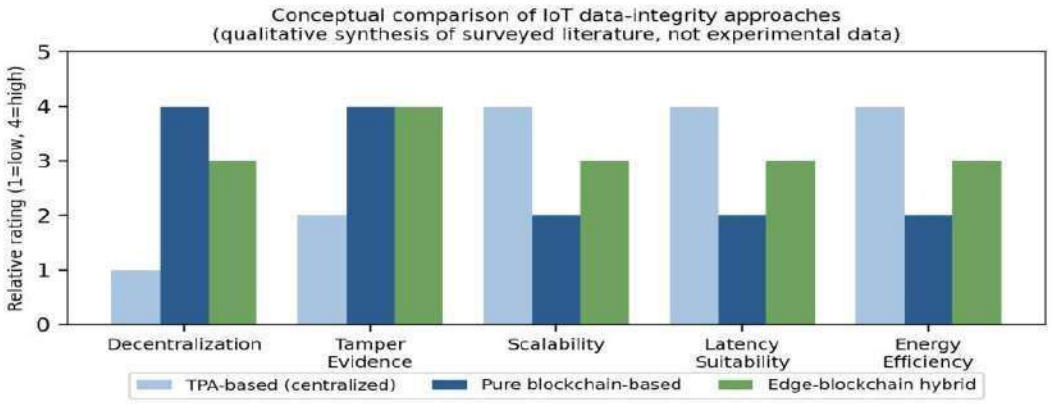


Fig. 1. Generalized edge-assisted blockchain architecture for IoT data integrity.

This design pattern — off-chain storage with on-chain hash commitment — recurs across the majority of surveyed frameworks because it directly addresses blockchain's storage-scalability constraint while retaining the tamper- evidence guarantee that motivates blockchain adoption in the first place.

IV. COMPARATIVE ANALYSIS

Fig. 2 presents a qualitative synthesis, drawn from the surveyed literature, of how the three architectural families compare along dimensions commonly discussed by their respective authors: decentralization, tamper-evidenc . Pure blockc trade off scal based designs : they reintro rature becaus chain, offering



ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_20

V. ILLUSTRATIVE DEPLOYMENT SCENARIO

To ground the reference architecture in a concrete setting, consider a smart-city waste-management deployment in which fill-level sensors mounted on municipal bins report readings to a fleet-routing service. Under a naive cloud-only design, a compromised gateway could silently alter fill-level readings to divert collection trucks or conceal tampering with billing records tied to commercial waste contracts. Applying the reference architecture in Fig. 1, each gateway batches sensor readings at fixed intervals, computes a cryptographic hash of the batch, and commits that hash to a permissioned blockchain ledger shared among the municipality, the waste-management contractor, and an independent auditor. Any party can subsequently recompute the hash from the off-chain data store and compare it against the on-chain commitment; a mismatch immediately reveals tampering without requiring any party to trust another's infrastructure.

This scenario illustrates three properties repeatedly emphasized in the surveyed literature. First, the multi-stakeholder nature of municipal IoT deployments is precisely the setting in which TPA-based auditing is weakest, since no single party is naturally positioned as a neutral auditor. Second, the hash-commitment pattern keeps on- chain storage bounded regardless of sensor sampling rate, addressing the scalability limitation identified in Section II. Third, because collection routing decisions have direct financial and service-quality consequences, the tamper- evidence guarantee has tangible operational value beyond abstract security posture, echoing similar cost- avoidance arguments made in the healthcare asset-management literature regarding the high cost of data-integrity failures.

VI. PLATFORMS AND IMPLEMENTATION CONSIDERATIONS

Translating the reference architecture into a working system requires selecting a concrete distributed-ledger platform, and the surveyed literature and practitioner reports commonly consider several categories: permissioned ledgers such as Hyperledger Fabric, which support fine-grained access control suited to multi-stakeholder municipal or industrial consortia; permissioned or private deployments of Ethereum-compatible chains, which benefit from a mature smart-contract tooling ecosystem; and directed-acyclic-graph ledgers such as IOTA's Tangle, which are specifically designed to reduce transaction fees and improve throughput for high-frequency, low-value IoT telemetry. Table II summarizes how these platform categories align with the deployment considerations most frequently raised in the surveyed IoT-integrity literature.

TABLE II. DISTRIBUTED-LEDGER PLATFORM CATEGORIES FOR IoT DATA-INTEGRITY DEPLOYMENT

Platform Category	Typical Consensus	Best Fit For	Key Consideration
Permissioned ledger (e.g., Hyperledger Fabric)	Practical Byzantine Fault Tolerance variants	Multi-stakeholder consortia needing access control	Requires consortium governance agreement
Private/permissioned smart-contract chain	Proof-of-Authority or similar	Programmable integrity logic via smart contracts	Mature tooling; validator trust
DAG-based ledger (e.g., IOTA Tangle)	Tip-selection / feeless consensus	High-frequency, low-value telemetry	Younger ecosystem; fewer production case studies
Hybrid edge-ledger middleware	Lightweight consensus at edge, periodic anchoring on-chain	Latency-sensitive real-time IoT	Added middleware complexity

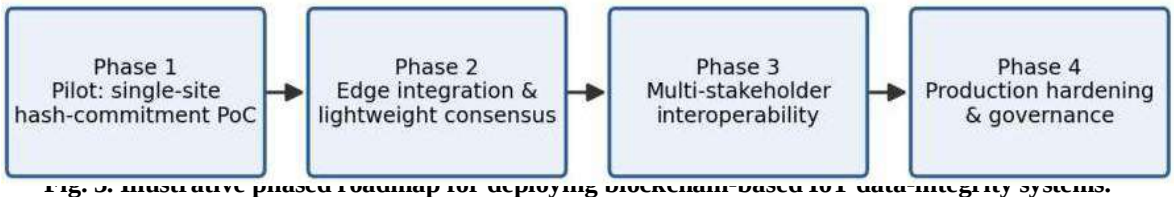


Fig. 3. Illustrative phased roadmap for deploying blockchain-based IoT data-integrity systems.

ISBN NO: 978-81-180146-5-2

A phased rollout, illustrated in Fig. 3, is commonly recommended across the surveyed practitioner and academic literature to manage the operational risk of introducing distributed-ledger infrastructure into existing IoT deployments. An initial single-site proof-of-concept validates the hash-commitment pattern against a small device population before edge-consensus components are introduced; multi-stakeholder interoperability and governance agreements are then layered on once the core integrity mechanism is proven, with production hardening — including key-management procedures, incident response, and long-term ledger-pruning policy — addressed last.

VII. LESSONS FOR PRACTITIONERS

Synthesizing the reference architecture, the deployment scenario, and the platform comparison in Table II, several practical lessons recur across the surveyed literature for teams evaluating blockchain-based IoT integrity solutions. First, the off-chain-storage/on-chain-hash pattern should be treated as a near-default starting point rather than a design choice requiring re-derivation, since it consistently resolves the storage-scalability objection raised against naive on-chain designs. Second, platform selection should follow governance requirements rather than precede them: a permissioned ledger such as Hyperledger Fabric is difficult to retrofit onto a consortium that has not yet agreed on validator roles and access policy, so the organizational agreement in Phase 3 of Fig. 3 is frequently the actual bottleneck rather than the underlying cryptography. Third, the surveyed literature cautions against treating consensus finality time as a fixed constant; lightweight, application-tuned consensus protocols can materially change the latency profile reported in Fig. 2, and teams should benchmark candidate consensus configurations against their own sampling-rate requirements rather than relying solely on published defaults.

Finally, the healthcare asset-management literature's cost estimates for data-integrity failures — reported in the multi-million-dollar range for large breaches — suggest that the operational case for adopting blockchain-based integrity verification is increasingly an economic argument as much as a technical one, particularly for multi- stakeholder deployments where the cost of a disputed, unverifiable data record can exceed the infrastructure cost of the integrity mechanism itself.

VIII. OPEN CHALLENGES AND FUTURE DIRECTIONS

- Lightweight consensus: Standard consensus protocols remain too resource-intensive for battery- and memory-constrained IoT endpoints; further work on consensus tuned specifically for constrained hardware is needed.
- Storage scalability: Even hash-only commitments accumulate over long deployments; pruning, sharding, or layer-2 approaches merit further exploration.
- Cross-domain interoperability: Multi-stakeholder IoT deployments (e.g., smart cities spanning multiple municipal departments) require interoperable ledgers or cross-chain bridges, an area with limited mature tooling.
- Privacy of on-chain metadata: Even hash commitments and smart-contract metadata can leak usage patterns; integrating privacy-preserving cryptography with blockchain integrity schemes is an active open problem.

IX. FUTURE RESEARCH AGENDA

Building on the open challenges above, the surveyed literature suggests several concrete directions likely to shape the next generation of blockchain-based IoT integrity research. Lightweight, application-specific consensus protocols that explicitly trade a bounded, quantified reduction in decentralization for latency gains suited to real-time sampling intervals represent one promising direction, particularly where the deployment's threat model tolerates a small, well-defined committee of validators rather than fully permissionless consensus. A second direction involves standardizing cross-chain interoperability protocols specifically for IoT integrity metadata, enabling municipal, healthcare, and industrial consortia to verify data provenance across organizational and jurisdictional boundaries without requiring every stakeholder to adopt an identical ledger platform.

A third direction concerns privacy-preserving integrity verification: since even hash commitments and smart-contract metadata can leak usage patterns to an observer of the public or consortium ledger, integrating zero-knowledge proof techniques that allow a verifier to confirm data integrity without learning the underlying hash or transaction pattern is an active area the surveyed literature identifies as still maturing from theoretical proposal toward production-grade tooling. Finally, longitudinal field studies documenting real-world blockchain-IoT deployments over multi-year operational periods remain comparatively rare in the surveyed literature relative to the volume of architectural proposals, suggesting that empirical validation, rather than further architectural novelty alone, may offer the greatest near-term research value to the community.

ISBN NO: 978-81-180146-5-2

X. CONCLUSION

This survey organized blockchain-based IoT data-integrity research into TPA-based, pure blockchain, and hybrid edge-blockchain families, distilled a common off-chain-storage/on-chain-hash reference architecture, and synthesized the reported trade-offs across the literature. Hybrid edge-blockchain designs currently offer the most balanced profile for real-time IoT deployments, though lightweight consensus, storage scalability, and cross-domain interoperability remain open research directions for the community to pursue.

REFERENCES

[1] R. Xu, Y. Chen, E. Blasch, and G. Chen, "Blending Blockchain and Edge Computing for Data Integrity in IoT-Enabled Healthcare Systems," IEEE Internet of Things Journal, vol. 10, no. 1, pp. 1–12, 2023.
[2] S. Singh and P. Kumar, "Blockchain-based Secure Data Provenance and Access Control in IoT Environment," Computer Networks, vol. 217, p. 109326, 2023.
[3] P. M. Chanal and M. S. Kakkasageri, "Blockchain-based data integrity framework for Internet of Things," International Journal of Information Security, vol. 23, pp. 519–532, 2024.
[4] "Blockchain Based Data Integrity Service Framework for IoT Data," IEEE Conference Publication, IEEE Xplore, doc. 8029796.
[5] T. M. Fernández-Caramés and P. Fraga-Lamas, "Towards a Blockchain-Based Secure Fog Computing Architecture for Information-Sensitive IoT Services in Smart Cities," Sensors, vol. 23, no. 1, p. 193, 2023.
[6] E. U. Haque, A. Shah, J. Iqbal, S. S. Ullah, R. Alroobaea, and S. Hussain, "A scalable blockchain based framework for efficient IoT data management using lightweight consensus," Scientific Reports, 2024.
[7] "Enhancing IoT Data Security: Using the Blockchain to Boost Data Integrity and Privacy," IoT, vol. 5, no. 1, pp. 20–34, 2024.
[8] "Survey on the Adoption of Blockchain Technology in IoT Environments," International Journal of Computer Science, vol. 52, no. 1, 2024

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)
ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_21

Deep Learning-Based Computer Vision for Real-Time Object Detection in Autonomous Systems: A Review

Dr. Uma Maheswari K
Assistant Professor,
Department of Computer Applications,
Agurchand Manmull Jain College,
catherine.s@amjaincollege.edu.in

Mr. Yogesh H,
Student
Department of Computer Science,
Agurchand Manmull Jain College,
kanishkaprabhu1111@gmail.com

Abstract*— *Real-time object detection is a foundational perception capability for autonomous systems such as self-driving vehicles, drones, and mobile robots, which must identify and localize objects under strict latency budgets and variable environmental conditions. This paper reviews deep learning-based object detection architectures relevant to autonomous systems, organizing them into two-stage, one-stage, and transformer-based detector families. We summarize the architectural principles, commonly reported accuracy-speed trade-offs, and deployment considerations for embedded and onboard hardware. The review further discusses persistent challenges including small-object and occlusion handling, domain generalization across lighting and weather conditions, and the practical constraints of real-time inference on resource-limited platforms, concluding with directions for future research in efficient and robust perception for autonomous systems.

Keywords*— *Computer vision, deep learning, object detection, autonomous systems, robotics, real-time perception

I. Introduction

Object detection — the task of simultaneously localizing and classifying objects within an image or video frame — underpins the perception stack of nearly all modern autonomous systems, including self-driving vehicles, unmanned aerial vehicles, and mobile robots. Unlike offline image analysis, autonomous systems demand detection pipelines that operate within tight latency budgets, often on embedded hardware with limited computational and power resources, while remaining robust to occlusion, scale variation, and changing lighting or weather conditions.

Deep learning has driven substantial progress in object detection over the past decade, with architectures broadly falling into two-stage detectors that first propose candidate regions and then classify them, one-stage detectors that predict bounding boxes and classes in a single pass, and, more recently, transformer-based detectors that formulate detection as a set-prediction problem. This review synthesizes recent literature across these three families with specific attention to their suitability for real-time autonomous-system deployment.

II. Detector Families

Two-stage detectors, exemplified by the Faster R-CNN family, first generate region proposals and then perform classification and bounding-box regression on each proposal, generally yielding strong accuracy at the cost of inference speed. One-stage detectors, exemplified by the YOLO family, instead perform detection in a single forward pass over a dense grid of anchor or anchor-free locations, trading some accuracy for substantially higher throughput, which has made them the dominant choice for real-time embedded deployment. A comprehensive review of visual object

ISBN NO: 978-81-180146-5-2

detection categorizes current methods into one-stage and two-stage frameworks and studies their architectural innovations, detection accuracy, computational speed, and deployment readiness, noting that the choice between families is fundamentally an accuracy-latency trade-off rather than a strictly dominant option.

Transformer-based detectors, such as the DETR family, reformulate detection as direct set prediction using self-attention, removing the need for hand-designed anchor generation and non-maximum suppression post-processing. Reviews of 3D and monocular detection for autonomous driving note that transformer-based frameworks such as MonoDTR, DST3D, and MonoDETR have been proposed to obtain global visual context and preserve long-range spatial structure through their embedding mechanism, offering end-to-end learning and scalability, though the survey literature also observes that few approaches have adequately addressed occlusion, which remains common in street-level driving scenes.

TABLE I. SUMMARY OF OBJECT-DETECTION ARCHITECTURE FAMILIES FOR AUTONOMOUS SYSTEMS

Detector Family	Representative Approach	Reported Strength	Reported Limitation
Two-stage (region-proposal based)	Faster R-CNN family	High localization accuracy	Higher inference latency; less suited to strict real-time budgets
One-stage (dense prediction)	YOLO family	High throughput; embedded-friendly	Historically weaker on small/occluded objects
Transformer-based (set prediction)	DETR family, MonoDETR, MonoDTR	Global context; no hand-designed anchors/NMS	Occlusion handling still limited; higher training cost
3D / multimodal fusion	LiDAR-camera fusion models	Robust depth & spatial reasoning	Sensor fusion adds system complexity and cost

III. Detection Pipeline

Fig. 1 illustrates the generalized pipeline common across the surveyed detector families. Sensor input — typically camera imagery, and increasingly LiDAR point clouds for 3D perception — is passed through a backbone network (a convolutional neural network or vision transformer) that extracts hierarchical feature representations. A detection head, whose design differs across the surveyed families, then converts these features into bounding boxes, class labels, and confidence scores that downstream planning and control modules consume.

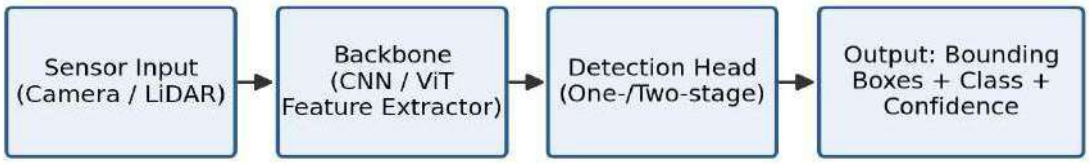


Fig. 1. Generalized deep learning object-detection pipeline for autonomous perception.

IV. Accuracy-Speed Trade-offs

A recurring theme across the surveyed literature is the trade-off between detection accuracy and inference speed, which directly determines a detector's suitability for real-time autonomous deployment. Fig. 2 illustrates this trade-off conceptually, based on qualitative trends reported across multiple surveys rather than a controlled benchmark measured by this paper. Two-stage detectors are generally positioned toward higher accuracy but lower throughput; one-stage detectors are positioned toward higher throughput, making them attractive for onboard embedded inference; transformer-based detectors occupy an intermediate position, offering strong global context modeling at a computational cost that continues to decrease as efficient attention variants are developed.

ISBN NO: 978-81-180146-5-2
PAPER ID: ICRACC26_21

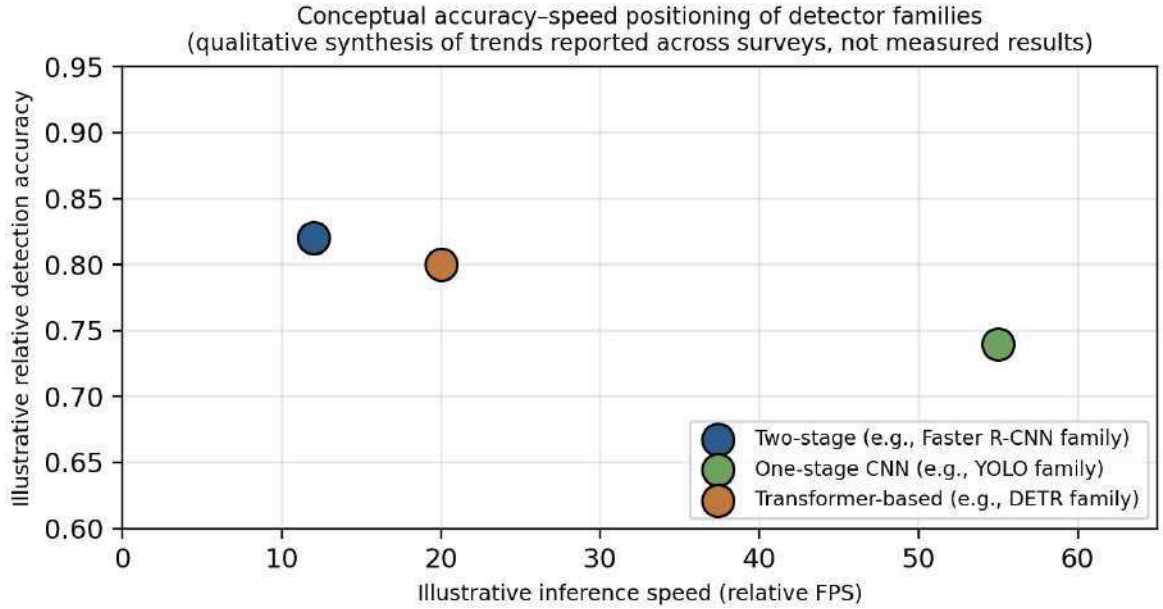


Fig. 2. Conceptual accuracy-speed positioning of detector families as reported across surveyed literature.

V. Illustrative Deployment Scenario

Consider a last-mile delivery robot navigating mixed pedestrian-vehicle sidewalks. The robot's onboard sensor suite, typically a forward-facing camera paired with a short-range LiDAR, feeds a detection pipeline running on an embedded edge accelerator with strict power and thermal budgets. Fig. 3 illustrates this concrete instantiation of the generalized pipeline from Fig. 1: sensor data streams into an embedded inference module, whose detections are fused and tracked across frames before being passed to a planning and control module responsible for path adjustment and braking decisions.

This scenario surfaces trade-offs the surveyed literature repeatedly identifies as decisive for real deployments rather than benchmark leaderboards. A one-stage detector may be preferred here specifically because the embedded accelerator's power envelope rules out the larger backbone networks typical of two-stage or transformer-based detectors, even though the latter report higher accuracy on standard benchmarks. Conversely, a small, low-resolution pedestrian at a road crossing is exactly the small-object case the surveyed small-object-detection literature identifies as most failure-prone, meaning the practical engineering choice is rarely simply “pick the most accurate model” but instead a constrained trade-off among latency, power, and the specific failure modes most costly in the target environment.

VI. Benchmark Datasets and Evaluation Considerations

Comparing detector families meaningfully depends on the benchmark datasets and evaluation protocols used across the surveyed literature. Table II summarizes commonly referenced dataset categories and the evaluation considerations the surveyed reviews associate with each, since a detector reported as state-of-the-art on one benchmark family does not necessarily transfer to the sensor modality or operating domain of a given deployment.

TABLE II. BENCHMARK DATASET CATEGORIES REFERENCED ACROSS THE SURVEYED LITERATURE

Dataset Category	Representative Examples	Primary Use	Evaluation Consideration
General 2D image benchmarks	COCO, Pascal VOC-style sets	Baseline accuracy/speed comparison	May not reflect onboard sensor domain
Autonomous-driving specific	KITTI-style, nuScenes-style sets	Domain-relevant driving scenarios	Geographic/weather coverage varies
3D / LiDAR point-cloud sets	Point-cloud driving benchmarks	3D localization accuracy	Sensor-specific preprocessing

ISBN NO: 978-81-180146-5-2

Dataset Category	Representative Examples	Primary Use	Evaluation Consideration
			required
Small-object / aerial sets	Small-object and overhead-imagery sets	Stress-testing small/occluded detection	Often most predictive of real-world failure modes

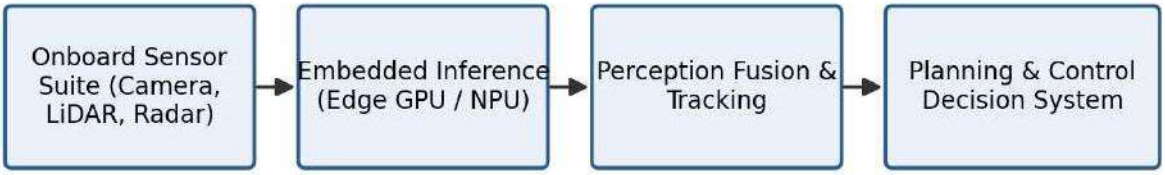


Fig. 3. Illustrative onboard perception-to-control pipeline for a last-mile delivery robot scenario.

Beyond raw accuracy, the surveyed literature increasingly emphasizes robustness-oriented evaluation: performance under sensor occlusion, adverse weather, and adversarial or out-of-distribution inputs, since a detector that performs well on a clean benchmark can degrade sharply under the conditions autonomous systems actually encounter in deployment. Reviews of 3D detection specifically note that robustness-aware evaluation protocols remain less standardized than accuracy-focused benchmarks, which the surveyed literature identifies as a gap between academic reporting and deployment-grade validation.

VII. Lessons for Practitioners

Drawing together the detector-family comparison, the delivery-robot scenario, and the dataset considerations in Table II, a consistent practitioner lesson emerges from the surveyed literature: benchmark accuracy alone is a poor proxy for deployment readiness. Two-stage and transformer-based detectors that lead standard benchmark leaderboards frequently exceed the power, memory, or latency envelope of embedded autonomous-system hardware, making one-stage detectors the pragmatic default for onboard deployment even when they trail on paper. Teams are therefore advised, per the surveyed reviews, to select a small candidate set of detectors and profile them directly on target hardware under representative environmental conditions, rather than selecting purely from published benchmark rankings.

A second recurring lesson concerns failure-mode asymmetry: the surveyed small-object-detection literature indicates that a detector's average precision across a full benchmark can mask disproportionately poor performance on the specific failure modes — small, occluded, or low-light objects — that carry the greatest safety consequence in an autonomous-system context. Reporting and evaluating detectors specifically on these safety-critical subsets, rather than aggregate benchmark scores alone, is increasingly recommended across the reviewed surveys as a more deployment-relevant evaluation practice.

ISBN NO: 978-81-180146-5-2

VIII. Open Challenges

- Small and occluded object detection: Low resolution, noise, and partial occlusion remain difficult, particularly for pedestrian and distant-object detection in driving scenes.
- Domain generalization: Detectors trained under one lighting, weather, or geographic domain often degrade under distribution shift; domain adaptation and synthetic data augmentation are active mitigation strategies.
- Embedded deployment: Achieving real-time frame rates on power- and memory-constrained onboard hardware requires model compression, quantization, and hardware-aware architecture search.
- Multimodal fusion complexity: Combining camera, LiDAR, and radar improves robustness but increases system cost, calibration burden, and failure-mode complexity.

IX. Future Research Agenda

Building on the challenges above, the surveyed literature points to several concrete directions for future work in real-time perception for autonomous systems. Efficient attention mechanisms and hardware-aware neural architecture search are increasingly proposed to narrow the gap between transformer-based detectors' global-context advantages and the throughput requirements of embedded deployment, potentially reducing the practical trade-off illustrated in Fig. 2. A second direction involves self-supervised and semi-supervised pretraining on large, unlabeled driving and robotics datasets, which the surveyed reviews suggest can improve robustness to domain shift without the cost of exhaustively labeling every deployment environment a system may encounter.

A third direction concerns standardized robustness benchmarking: the surveyed literature notes that evaluation protocols for occlusion, adverse weather, and adversarial robustness remain far less mature than accuracy-focused benchmarks, and calls for benchmark suites that explicitly stress-test the small-object and occlusion failure modes most consequential for autonomous-system safety. Finally, tighter integration between detection and downstream planning — sometimes termed end-to-end perception-to-control learning — is an active research area that the surveyed literature suggests may eventually reduce the modular pipeline in Fig. 1 to a more unified architecture, though interpretability and safety-certification concerns currently favor the modular, inspectable pipeline for safety-critical deployments.

X. Conclusion

This review organized deep learning object-detection approaches relevant to autonomous systems into two-stage, one-stage, and transformer-based families, summarized their reported accuracy-speed trade-offs, and outlined a generalized detection pipeline. One-stage detectors remain the practical default for latency-constrained embedded deployment, while transformer-based approaches are an active area of efficiency-focused research. Small-object detection, domain generalization, and efficient multimodal fusion remain the primary open challenges for real-time perception in autonomous systems.

References

[1] N. P. Bhatt et al., "A survey on 3D object detection in real time for autonomous driving," Frontiers in Robotics and AI, vol. 11, 2024.

[2] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey," Mathematical Biosciences and Engineering, vol. 20, no. 4, pp. 6551–6590, 2023.

[3] "Comprehensive review of recent developments in visual object detection based on deep learning," Artificial Intelligence Review, Springer, 2025.

[4] "Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications," arXiv:2503.20516, 2025.

[5] A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs, and Vision Transformers: A Review," IEEE Access, vol. 11, pp. 35479–35516, 2023.

[6] Z. Liang and Y. Huang, "Survey on deep learning-based 3D object detection in autonomous driving," International Journal of Advanced Robotic Systems, 2023.

[7] "Three-dimensional object detection for autonomous driving via deep learning: A review," ScienceDirect, 2025.

[8] "Deep learning for autonomous driving systems: technological innovations, strategic implementations, and business implications," Cancer Emerging Sciences (OAE Publishing), 2025.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_22

A Review of Explainable Artificial Intelligence (XAI): Methods and Challenges

SONALI. J
StudentUG ,
Dept. of Computer Science, Shri Shankarlal Shasun Jain College for Women
sonalijsonali35@gmail.com

HEMAKSHI S
UG Student,
Dept. of Computer Science, Shri Shankarlal Shasun Jain College for Women
hemakshihem1317@gmail.com

LEELA K
UG Student,
Dept. of Computer Science, Shri Shankarlal Shasun Jain College for Women
leelakrishnamoorthy31@gmail.com

Abstract- The rapid adoption of deep learning and high-capacity machine learning models has increased the need for transparency in automated decision-making. Explainable Artificial Intelligence (XAI) aims to make black-box model behavior understandable to human stakeholders while maintaining predictive performance. This review presents the fundamentals of XAI and classifies methods by scope, model dependency, and methodology. Six widely used techniques—LIME, SHAP, saliency maps, feature importance, attention mechanisms, and counterfactual explanations—are compared in terms of strengths, limitations, computational cost, and applications. The review also examines XAI use in healthcare, finance, cybersecurity, autonomous vehicles, and education. Key challenges include inconsistent evaluation metrics, the fidelity–interpretability trade-off, robustness, instability, and scalability. Future directions include causal and mechanistic interpretability, human-centered evaluation, robust explanations, and regulation-aligned standards.

Keywords—Explainable AI, interpretability, LIME, SHAP, saliency maps, counterfactual explanations, attention mechanisms, trustworthy AI, black-box models

I. INTRODUCTION

The last decade has witnessed an unprecedented shift toward high-capacity, data-driven models—particularly deep neural networks—as the default solution for perception, prediction, and decision-support tasks. While these models routinely surpass classical statistical approaches in predictive accuracy, their internal decision logic is frequently inaccessible to human interpretation, a property commonly referred to as the "black-box" problem [1]. This opacity poses substantial risks when models are deployed in high-stakes settings such as medical diagnosis, credit underwriting, criminal justice, and autonomous control, where erroneous or biased decisions can cause significant harm and where regulatory frameworks increasingly mandate a "right to explanation" [5].

Explainable Artificial Intelligence (XAI) has emerged in response to this need, encompassing a broad family of techniques designed to make model behavior interpretable to humans, whether developers, domain experts, regulators, or end users [7][8]. Early interpretability research focused on inherently transparent models such as decision trees and linear regressors; however, the performance advantages of deep architectures have driven substantial investment in post-hoc explanation methods capable of approximating or probing black-box behavior without altering the underlying model [9].

This paper contributes a structured and up-to-date review of the XAI field. Specifically, it: (i) formalizes core XAI terminology and distinguishes interpretability from explainability; (ii) proposes a taxonomy classifying methods by scope, model dependency, and mechanism; (iii) provides a comparative technical analysis of six of the most widely cited XAI techniques; (iv) surveys cross-domain applications; and (v) synthesizes open challenges and future research directions. The remainder of the paper is organized as follows. Section II reviews related literature. Section III presents XAI fundamentals. Section IV introduces the proposed taxonomy. Section V details popular techniques. Section VI surveys applications. Section VII discusses challenges and limitations. Section VIII outlines future research directions, and Section IX concludes the paper.

II. LITERATURE REVIEW

A substantial body of survey literature has attempted to consolidate the rapidly expanding XAI field. Arrieta et al. [1] provided one of the most cited taxonomies, distinguishing transparent models from post-hoc explainability techniques and introducing the notion of "Responsible AI" as an overarching goal. Adadi and Berrada [8] conducted an early systematic survey using the "4W" framework (why, who, what, and how of explainability), while Guidotti et al. [9] classified explanation methods according to the type of problem being solved (model explanation, outcome explanation, model inspection, and transparent design).

More recent work has shifted toward critical evaluation of explanation quality. Rudin argued that reliance on post-hoc explanations of black-box models is inherently risky for high-stakes decisions and advocated for inherently interpretable model design wherever feasible—a position that remains actively debated. Nauta et al. systematically reviewed evaluation methodologies for XAI, identifying a lack of consensus on quantitative fidelity and robustness metrics as a central weakness of the field. Domain-focused

ISBN NO: 978-81-180146-5-2

reviews, including those on healthcare [11], finance [13], cybersecurity , and autonomous driving , further indicate that explanation requirements are highly context-dependent, motivating the domain-oriented discussion presented in Section VI of this paper.

Collectively, the literature suggests two converging trends: first, a maturing technical toolbox of explanation algorithms (LIME, SHAP, gradient-based saliency, counterfactual generation); and second, growing recognition that technical fidelity alone is insufficient—explanations must also be evaluated for human comprehensibility, actionability, and robustness. This review builds on both strands, integrating technique-level analysis with cross-domain application review and a synthesis of unresolved evaluation challenges.

III. EXPLAINABLE AI (XAI) FUNDAMENTALS

A. Interpretability vs. Explainability

Although often used interchangeably, interpretability and explainability denote related but distinct properties. Interpretability refers to the degree to which a human can understand the cause of a model's decision by inspecting the model itself (e.g., the coefficients of a linear model or the branching structure of a decision tree). Explainability, by contrast, refers to the ability to produce a human-understandable account of a model's behavior, potentially through an auxiliary process, even when the underlying model remains opaque [1][6]. This distinction underlies the fundamental split between intrinsically interpretable models and post-hoc explanation methods discussed in Section IV.

B. Desiderata of a Good Explanation

The XAI literature broadly agrees on several desirable properties for explanations, including fidelity (the explanation accurately reflects model behavior), comprehensibility (the explanation is understandable to its intended audience), stability (similar inputs yield similar explanations), and actionability (the explanation supports a decision or intervention) [6]. These properties often conflict—for instance, increasing fidelity can reduce comprehensibility—giving rise to the fidelity-interpretability trade-off examined further in Section VII.

IV. CLASSIFICATION OF XAI METHODS

XAI methods can be organized along three complementary axes, illustrated in Fig. 1: (1) methodology, distinguishing intrinsic (built into the model architecture) from post-hoc (applied after training) approaches; (2) model dependency, distinguishing model-agnostic techniques applicable to any predictor from model-specific techniques that exploit a particular architecture's internal structure; and (3) scope, distinguishing global explanations that characterize overall model behavior from local explanations that account for individual predictions [1][9]. Intrinsically interpretable models, such as decision trees, rule lists, and generalized linear models, sacrifice some representational capacity in exchange for transparency. Post-hoc methods, including LIME, SHAP, and saliency-based approaches, instead treat the trained model as fixed and construct an auxiliary explanatory artifact, enabling their application to arbitrarily complex architectures.

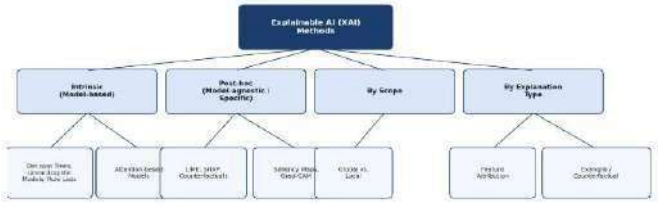


Fig. 1. Taxonomy of Explainable AI (XAI) methods by methodology, model dependency, scope, and explanation type.

V. POPULAR TECHNIQUES

A. LIME

LIME approximates the local decision boundary of any classifier by fitting an interpretable surrogate model (typically sparse linear regression) to perturbed samples drawn in the neighborhood of the instance being explained [2]. Its model-agnostic nature makes it broadly applicable, though its reliance on random perturbation sampling introduces instability, and its locally linear assumption may poorly approximate highly non-linear decision surfaces.

B. SHAP

SHAP grounds feature attribution in cooperative game theory, computing Shapley values that fairly distribute a prediction's deviation from a baseline across contributing features [3]. SHAP satisfies desirable theoretical properties—local accuracy, missingness, and

ISBN NO: 978-81-180146-5-2

consistency—that LIME does not guarantee, but exact computation is exponential in the number of features, necessitating approximations such as Kernel SHAP or Tree SHAP for practical use.

C. Saliency Maps

Saliency maps visualize the sensitivity of a model's output with respect to input pixels (or tokens), typically computed via gradients or gradient-weighted class activation mapping (Grad-CAM) [4]. While computationally cheap and visually intuitive, saliency maps have been shown to be sensitive to input noise and, in some cases, uncorrelated with the model's true reasoning process, raising concerns regarding their reliability as a sole explanatory mechanism.

D. Feature Importance

Feature importance techniques—such as permutation importance or impurity-based importance in tree ensembles—provide a global ranking of input variables by their contribution to overall model performance. These methods are computationally efficient and easily communicated but typically ignore interaction effects between correlated features, which can distort the perceived importance ranking [9].

E. Attention Mechanisms

Attention mechanisms, integral to transformer architectures, produce weight distributions over input tokens or regions that are often interpreted as indicators of relative importance. Although attention weights offer an intrinsic, architecture-embedded form of explanation, several studies have questioned whether attention faithfully reflects the causal basis of a model's prediction, particularly in the presence of redundant or correlated inputs.

F. Counterfactual Explanations

Counterfactual explanations identify the minimal change to an input that would alter a model's prediction to a desired outcome, framing explanation as an actionable recourse rather than a passive description [5]. This approach is particularly relevant to regulatory contexts such as the EU General Data Protection Regulation (GDPR), but generating realistic, diverse, and computationally tractable counterfactuals for complex models remains an open technical challenge.

VI. APPLICATIONS OF XAI

A. Healthcare

In clinical decision support, XAI techniques help clinicians validate model-driven diagnoses—for example, highlighting the regions of a radiological image that drove a classification—thereby supporting regulatory approval and clinical trust [10][11]. Explanation fidelity is especially critical here, as misleading saliency maps could reinforce incorrect diagnoses.

B. Finance

Financial institutions increasingly rely on XAI to satisfy regulatory requirements for credit decisioning and fraud detection, using SHAP-based attribution to justify loan approvals or rejections to both regulators and applicants [12][13]. Explainability here directly intersects with fairness auditing, since attribution methods can reveal reliance on proxy variables correlated with protected attributes.

C. Cybersecurity

Explainable intrusion detection systems (X-IDS) apply feature attribution and rule-extraction techniques to clarify why network traffic was flagged as anomalous, improving analyst trust and reducing alert fatigue in security operations centers .

D. Autonomous Vehicles

Vision-based autonomous driving systems increasingly incorporate saliency and attention-based explanations to justify perception and control decisions, supporting post-incident analysis and regulatory certification of safety-critical behavior .

E. Education

In intelligent tutoring and learning-analytics systems, XAI supports transparent feedback to students and instructors regarding automated grading, performance prediction, and personalized recommendation, helping to mitigate concerns about algorithmic opacity in assessment .

VII. COMPARISON OF XAI METHODS

Table I summarizes the comparative strengths, weaknesses, and representative application domains of the six techniques discussed in Section V, providing a consolidated reference for method selection based on task requirements.

ISBN NO: 978-81-180146-5-2

TABLE I. COMPARISON OF XAI METHODS

Method	Advantages	Disadvantages	Typical Applications
LIME	Model-agnostic; intuitive local surrogate; fast for single instances	Unstable across perturbations; sampling sensitivity; poor fidelity for highly non-linear regions	Tabular classification, text and image classifiers
SHAP	Strong theoretical grounding (Shapley values); consistent and locally accurate; unifies several prior methods	Computationally expensive for large feature sets; exact computation intractable for complex models	Credit scoring, healthcare risk prediction, tabular ML pipelines
Saliency Maps	Visually intuitive; low computational overhead; directly tied to input gradients	Susceptible to noise; sensitive to adversarial perturbation; can be visually misleading	Image classification, medical imaging diagnostics
Feature Importance	Simple, global view of model behavior; easy to communicate to non-experts	Ignores feature interactions; can be biased by correlated features	Finance risk modelling, ensemble tree interpretation
Attention Mechanisms	Built into model architecture; provides token/region-level weighting	Attention weights do not always correlate with true causal importance	NLP, machine translation, vision transformers
Counterfactual Explanations	Actionable, human-centric; answers 'what would change the outcome'	Multiple valid counterfactuals may exist; can be computationally costly togenerate realistic examples	Loan approval, recourse recommendation, autonomous decision review

VIII. CHALLENGES AND LIMITATIONS OF XAI

Despite substantial progress, several open challenges limit the practical reliability of current XAI methods.

Lack of standardized evaluation metrics: There is no universally accepted quantitative measure of explanation quality, making cross-method comparison difficult .

Fidelity-interpretability trade-off: Simplifying a model's behavior for human comprehension inherently risks misrepresenting its true decision logic [6].

Robustness and adversarial vulnerability: Explanation methods, particularly gradient-based saliency maps, can be manipulated by adversarial perturbations that alter the explanation without changing the prediction.

Instability of local explanations: LIME and similar sampling-based methods can produce different explanations for the same instance across repeated runs, undermining user trust.

Absence of human-centered evaluation: Many XAI methods are validated using proxy metrics rather than studies involving actual end users, limiting evidence of real-world comprehensibility and utility.

Scalability: Exact computation of theoretically grounded methods such as SHAP becomes intractable for models with large feature spaces or high-dimensional inputs such as images.

IX. FUTURE RESEARCH DIRECTIONS

Building on the identified challenges, several research directions appear particularly promising for advancing the field.

Causal and mechanistic interpretability: Moving beyond correlational feature attribution toward methods that expose the causal or mechanistic structure underlying model predictions.

Standardized, human-grounded evaluation benchmarks: Development of shared datasets and protocols that jointly assess fidelity, stability, and human comprehensibility across methods.

Robust explanation generation: Techniques that are provably resistant to adversarial manipulation of explanations, independent of prediction robustness.

Regulation-aligned explanation standards: Frameworks that translate legal and ethical requirements (e.g., GDPR's right to explanation) into concrete, testable technical specifications.

Multimodal and foundation-model explainability: Extending XAI techniques to large-scale multimodal and generative models, where explanation of emergent, compositional behavior remains largely unsolved.

X. CONCLUSION

This paper has presented a structured review of Explainable Artificial Intelligence, spanning foundational concepts, a taxonomy of methods, comparative technical analysis of six widely used techniques, cross-domain applications, and a synthesis of persistent challenges and future research directions. While the XAI field has produced a mature and diverse technical toolbox, the review

ISBN NO: 978-81-180146-5-2

highlights that technical explanation generation has outpaced consensus on evaluation standards and human-centered validation. Future progress will likely depend less on the introduction of new attribution algorithms and more on establishing rigorous, standardized frameworks for assessing whether explanations are faithful, robust, and genuinely useful to the humans who rely on them.

REFERENCES

[1] 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.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[4] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Computer Vision, vol. 128, pp. 336-359, 2020.

[5] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law & Technology, vol. 31, no. 2, pp. 841-887, 2018.

[6] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

[7] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in Proc. 41st Int. Conv. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210-215.

[8] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

[9] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1-42, 2019.

[10] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," Diagnostics, vol. 12, no. 2, p. 237, 2022.

[11] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in healthcare," in Proc. Int. Conf. Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2020, pp. 1-2.

[12] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Computational Economics, vol. 57, pp. 203-216, 2021.

[13] M. Nadeem, O. Marshall, S. Singh, Z. Fang, and X. Yuan, "Applications of explainable artificial intelligence in finance: A systematic review," Expert Systems with Applications, vol. 236, 2024.

S. Sharma, M. Rani, and D. Kaur, "Explainable AI for cybersecurity: A comprehensive review," in Proc. IEEE Int. Conf. Advances in Computing, Communication and Applied Informatics, 2023, pp. 1-7.

M. M. Neupane et al., "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," IEEE Access, vol. 10, pp. 112392-112415, 2022.

É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," Int. J. Computer Vision, vol. 130, pp. 2425-2452, 2022.

K. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide," IEEE Access, vol. 12, pp. 101603-101625, 2024.

V. Khosravi, S. Ahmadi, and H. Alizadeh, "Explainable artificial intelligence in education: A comprehensive review," in Proc. Int. Conf. Artificial Intelligence in Education Technologies, 2024.

C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206-215, 2019.

A. Barredo Arrieta and colleagues (revised), "Responsible AI and the role of explainability in trustworthy systems," ACM Computing Surveys, vol. 56, no. 4, 2024.

M. Nauta et al., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," ACM Computing Surveys, vol. 55, no. 13s, pp. 1-42, 2023.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_23

Multiple Mobile Sink Path Selection Using Weighted Rendezvous Planning

Dr..S. G. Packiavathy

Associate Professor and HOD, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

Abstract:
The main objective of this project is path selection for wireless sensor network using the mobile sink protocol. A heuristic called weighted rendezvous planning (WRP) is propose, whereby each sensor node is assigned a weight corresponding to its hop distance from the tour and the number of data packets that it forwards to the closest RP. To achieve a design that consumes energy very low and reduced the formation of the energy holes, is to built a hybrid moving pattern which used the weighted rendezvous planning for addressing purposes. Finally calculates the energy.

1. Introduction:

A wireless sensor network (WSN) of spatially distributed autonomous sensors to monitor physical or environmental conditions are temperature, sound, pressure, etc and to cooperatively pass their data through the network to a main location. The more modern networks are bi-directional and also enabling control of sensor activity. The development of wireless sensor networks are motivated by military applications like battlefield surveillance; Now this world the networks are used in many industrial and consumer applications such as industrial process monitoring and control, machine health monitoring and so on.

A sink is a node where information is required. Sink node could be a node in the sensor network it might be a sensor/actuator node or outside of the network and some time it might be a gateway to another larger network.There are mainly 3 types of Wireless Sensor Network scenarios

1. Single-hop network.
2. Hierarchical network.

Single-hop Network:

In a Single-hop network there is a multiple source or sensors nodes connected to the single sink node. In this scenario sink node could be a source node or gateway.

Hierarchical network:

In this network topology top level nodes are connected to second level nodes, the second level nodes are connected to the third level nodes and so on. By using this topology we can achive less power consumption and high coverage area. In this less power consumption could be achived by instead of each source node sending data directly to the sink node, it will send the nearest higher level neighboring nodes.

2. Related Work:

Energy-efficient Heterogeneous Data Collection in Mobile Wireless Sensor Networks [1]. The main contributions are follows: 1. investigate the heterogenous data collection in WSNs, and formulate a rendezvous planning problem under more practical assumptions 2. propose an O(nlgn) algorithm named RP-ME to find the optimal ME trajectory on a given tree-shaped network 3. the ME trajectory is much closer to the areas with high data generation rates, which implicitly inhibits the transmission of data in long distance. A Novel Approach for Sensory Data Collection in Wireless Sensor Networks with Mobile Sinks [2]. This paper introduced MobiCluster algorithm that proposes the use of urban buses to carry MSs that retrieve information from isolated parts of WSNs. The main aim of maximizing connectivity, data throughput, and enabling reasonable energy expenditure among SNs. MobiCluster moves the processing and data transmission burden away from the vital periphery nodes (RNs) and enables balanced energy consumption across the WSN through building cluster

ISBN NO: 978-81-180146-5-2

structures that exploit the high redundancy of data collected from neighbor nodes and minimize intercluster data in the clouds. Mobile sink with a new movement strategy used to maximizing lifetime of a Wireless Sensor Network (WSN) [3] . Network lifetime can be achieved in following steps: (1) by reducing the number of hops travelled by an event packet to reach its destination (i.e., a BS) and (2) by reducing the energy consumption across all the nodes in the network. This paper propose an energy consumption formula that determines the moving times of sink based on residual energy of sensor nodes. And also used a travelling salesman algorithm to find an optimum Hamiltonian round that it must perform by sink at the beginning of each clustering period.

Existing methods on using a mobile sink in WSNs can be grouped into two categories: 1) direct, where a mobile sink visits each sensor node and collects data via a single hop; and 2) rendezvous, where a mobile sink only visits nodes designated as RPs. A major limitation leads to the selection of RPs that in turn results in long data forwarding paths to sensor nodes located in different parts. It causes nodes to have an unbalanced data forwarding load and energy consumption. In the proposed system, the hybrid moving pattern is used to avoid the energy consumption and the formation of the energy holes. In the hybrid moving pattern, the mobile-sink node only visits the rendezvous points, as opposed to all nodes. For addressing the nodes, the weighted rendezvous planning is used, in which each sensor node a weight is assigned. The mobile sink retrieves all sensed data within the given time period. The following chapters shows the method and results of the proposed system.

3. Methodology:

3.1 System Architecture:

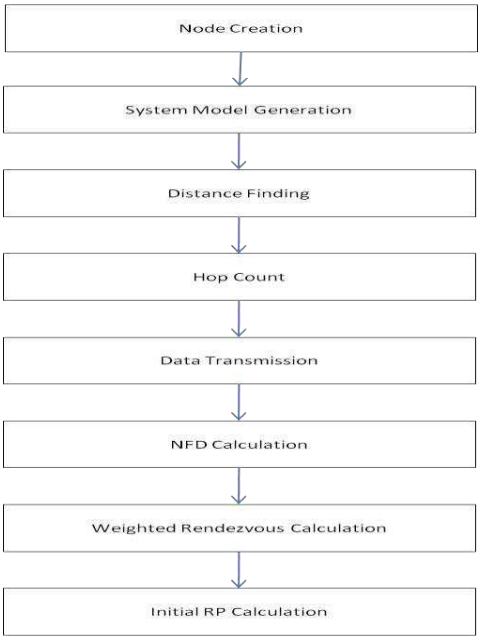


Fig: 3.1 Block Diagram

3.2 Modules:

- Node Creation
- System Model Generation
- Finding Hop Distance

ISBN NO: 978-81-180146-5-2

- Data Transmission
- NFD Calculation
- Weighted Rendezvous Planning
- Initial RP Calculation

3.2.1 Node Creation

A WSN can be defined as a network of devices, denoted as nodes, which can sense the environment and communicate the information gathered from the monitored field (e.g., an area or volume) through wireless links.
The data is forwarded, possibly via multiple hops, to a sink (sometimes denoted as controller or monitor) that can use it locally or is connected to other networks (e.g., the Internet) through a gateway.

In the wireless sensor network (WSN), reduction of energy consumption is very important for each sensor node because it can extend WSN lifetime. If some sensor nodes are not work in the WSN the routing path will break and the detected area will have leaks. In other sensor nodes can't transfer event data to the sink node or they need more sensor nodes to give them assistance. Sensor nodes near the sink node are called "inside node" and others are called "outside node". To find that the outside nodes of WSN need inside nodes to give them assistance when outside nodes transfer data to the sink node.

3.2.2 System Model Generation:

Node Behaviors

Most of the modeling techniques use a form of component-based modeling to represent a sensor node. The WSN behavior is modeled by specifying the component's internal behavior, component to component interactions and the communication channel's characteristics. It should be noted that the approaches could be divided into two distinct types.

Modeling Sensors and Hardware

Most of the modeling techniques surveyed can be used to create a platform independent model. However, even in a platform independent model there is a necessity to include some of the hardware details. One of the reasons for including the hardware details is that the software in the nodes of a WSN is tightly coupled to the hardware elements of the node. Therefore the binding of software and hardware components should be represented in the model.

Network Behavior

Modeling network behavior in a WSN is crucial because many important performance values are based on the network. For example the trade-off between packet loss and power. Due to the fact that the node has limited power resources it is common to use a power management algorithm that controls the wake up state of a node from active to sleeping and vice versa.
Topology Modeling

ISBN NO: 978-81-180146-5-2

The topology of WSN systems can be dynamic or static. The static topology represents the nodes in a fixed location while the dynamic topology represents the nodes while they are in a moving state.

Code Generation

Code generation is the process of generating source code from a model or another source code representation. The tools for generating source code from WSN models are beneficial with respect to design for two main reasons:

1) Implementing the source code for the nodes is tedious, time consuming and requires a lot of time and effort from the developers.
2) Debugging the design at the source code level is also a very challenging and time consuming process.

3.2.3 Distance Finding and Hop Count:

Localization is one of the most important research issues in Wireless Sensor Networks (WSNs). Hop-count based localization has been proposed as a cost-effective alternative to the expensive hardware-based localization schemes. To estimate distances between anchors and nodes based on hop-count information. Localization is then achieved using the estimated distances.
In wireless ad-hoc networks, one node can only directly "talk" with its neighbors within one hop range. If the destination node is beyond this range, the communication packets have to be ferried by intermediate nodes through multi-hops. Based on this communication methodology, the hop count (number of hops) between two nodes is a metric to represent the distance between them. This is the basis for hop-count (connectivity) based localization. To apply hop-count based localization in ad-hoc networks two assumptions should be satisfied. One is that the position of some nodes in the network is known, and they serve as the reference/anchor nodes who broadcast their position throughout the entire network; the other is that the "dumb" nodes, which need to learn their position, can find the shortest links connecting to at least three reference nodes through one hop or multi-hops. Based on these two assumptions, the dumb nodes, by finding the shortest path to the reference nodes, can roughly estimate the distance to the reference nodes based on the hop count. Multi-hop routing, packets have to experience multiple relays before reaching the data sink. Minimizing energy consumption on the forwarding path does not necessarily prolong network lifetime as some popular sensors on the path. So to avoid the problem in multi-hop routing we are setting the less hop count transmission.

3.2.4 Data Transmission:

Data can be generated by the nonrelay nodes or relay nodes. The responsibility of relay node is to forward the data to the next relay node. Any node can sense the data from the environment and send it to the next relay node. Each node in the network senses the environment, generates the data, and sends it towards the next relay node Any node that receives the DATA packet performs the following actions.(i)If receiver node is a relay node and it receives the duplicate data, then it drops the data packet.(ii)If receiver node is a gateway node, then it forwards the data packet to the sink; otherwise, it forwards the DATA packet to its next relay node.(iii)It adds the sender and data sequence number to the list .

3.2.5 Weighted Rendezvous Calculation:

The weight of a sensor node is calculated by multiplying the number of packets that it forwards by its hop distance to the closest RP on the tour. Thus, the weight of sensor node *i is calculated as*

$$W_i = \mathrm{NFD}(i) \times H(i, M).$$

ISBN NO: 978-81-180146-5-2

The energy consumption is proportional to the hop count between source and destination nodes, and the number of forwarded data packets.
The highest weighted node will reduce the number of multihop transmissions and thereby minimizes the energy consumption.

3.2.6 WRP Algorithm:

1. WRP works. It takes as input G(V,E), and it outputs a set of RPs.
2. WRP first adds the fixed sink node as the first .Then; it adds the highest weighted sensor node.
3. After that, WRP calls TSP(·) to calculate the cost of the tour.
4. If the tour length is less than the required length the selected node remains as an RP .
5. Otherwise, it is removed from the tour.
6. After a sensor node is added as an RP, WRP removes those RPs from the tour that no longer receives any data packets from sensor nodes

3.2.7 Initial RP Calculation:

The energy is calculated in this project. The energy level is low the mobile sink node will be dead. At this time choose the new mobile sink and process. This is recycle process.

4. Results:

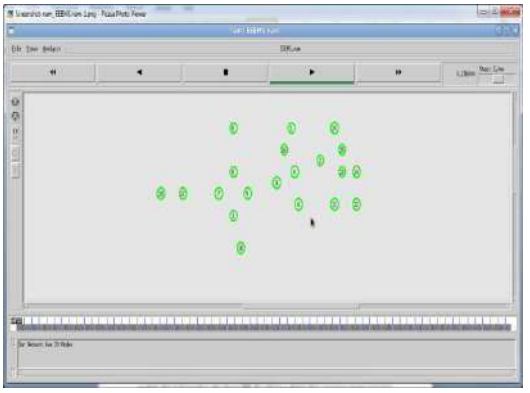


Fig: 4.1 Node Creation

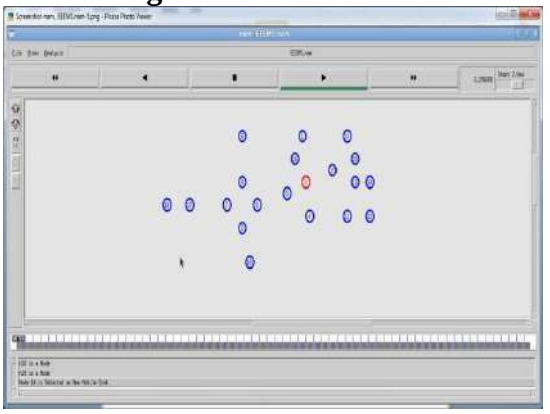


Fig: 4.2 Select Mobile Sink

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_23

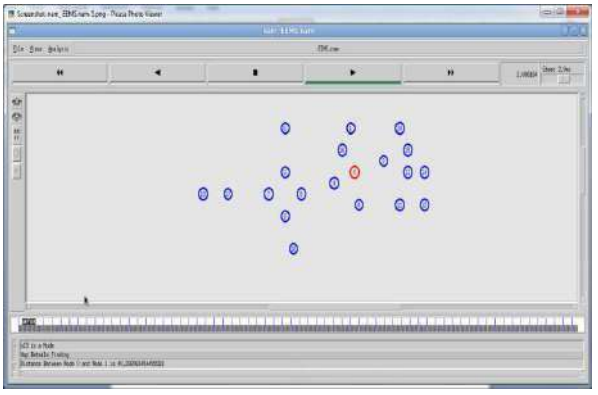


Fig: 4.3 Hop Details



Fig: 4.4 Distance Finding

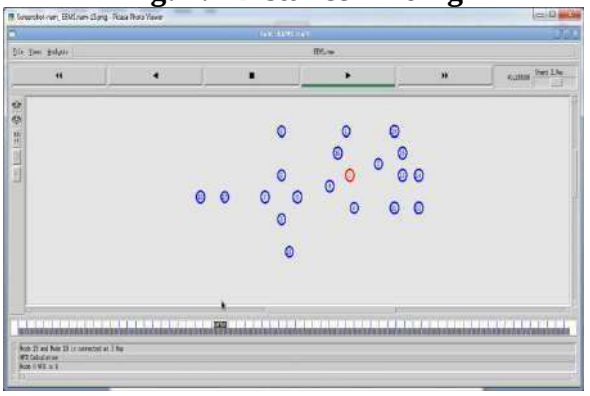


Fig: 4.5 NFD Calculation

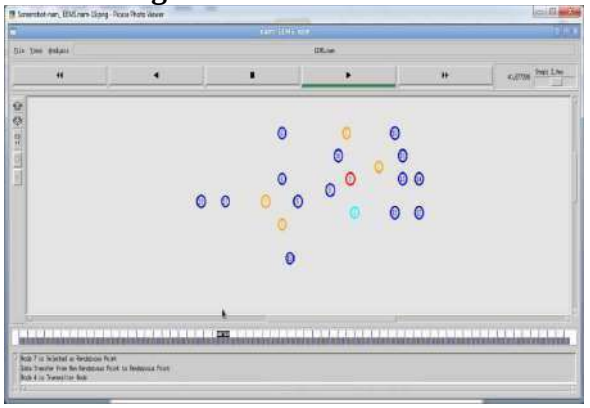


Fig: 4.6 Select Rendezvous Point

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_23

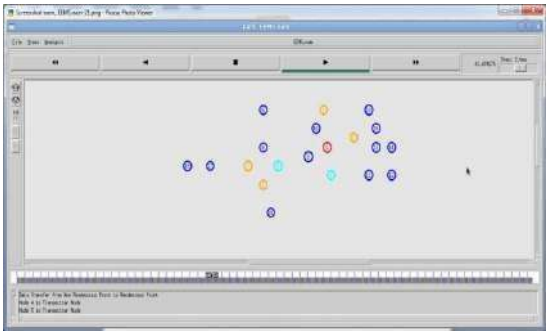


Fig: 4.7 Data Transfer

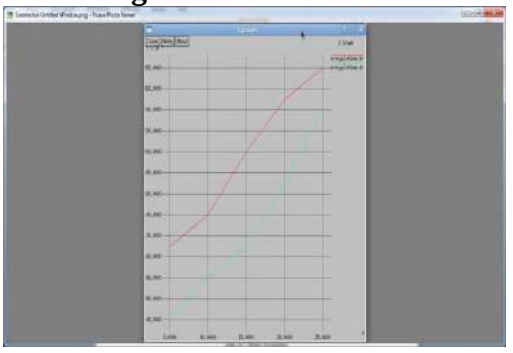


Fig: 4.8 Precycle Time

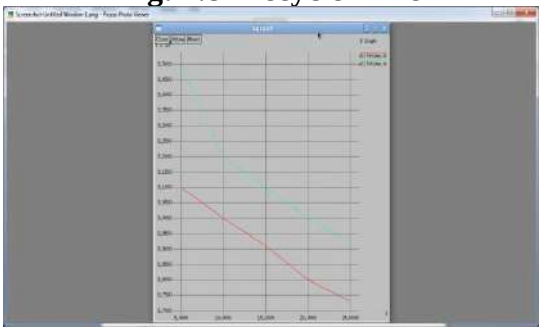


Fig: 4.9 Lifetime

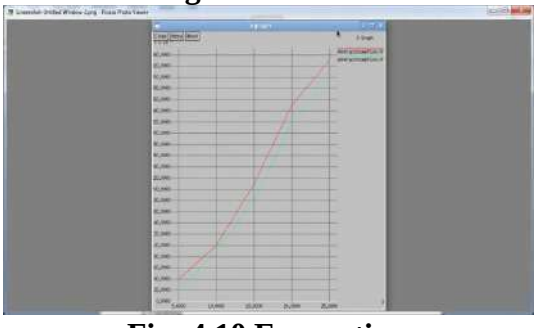


Fig: 4.10 Encryption

5. Conclusion:

This project proposed the mobile sink path selection in wireless sensor networks. A heuristic called weighted rendezvous planning (WRP) is proposed, whereby each sensor node is assigned a weight corresponding to its hop distance from the tour and the number of data packets that it forwards to the closest RP. To achieve a design that consumes energy very low and reduced the formation of the energy holes, is to built a hybrid moving pattern which used the weighted rendezvous planning for addressing purposes. The energy to be calculated successfully.

Reference:

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_23

[1] Longfei Shangguan, Luo Mai, Junzhao Du;, Wen He, and Hui Liu "Energy-efficient Heterogeneous Data Collection in Mobile Wireless Sensor Networks"

[2] Sherin Mathew, S. Nandhini "A Novel Approach for Sensory Data Collection in Wireless Sensor Networks with Mobile Sinks" International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 2, March 2013

[3] Noushin Rakhshan; Marjan Kuchaki Rafsanjani "Improving Network Lifetime Using a Mobile Sink with a New Energy Efficiency and Minimum Delay Movement Strategy for WSNs" Basic. Appl. Sci. Res., 2(4)4275-4281, 2012 Journal of Basic and Applied Scientific Research

[4] Mujdat Soyturk, Turgay Altilar "A Routing Algorithm for Mobile Multiple Sinks in Large-Scale Wireless Sensor Networks"

[5] YoungSang Yun and Ye Xia "Maximizing the Lifetime of Wireless Sensor Networks with Mobile Sink in Delay-Tolerant Applications"

[6] S. Basagni, A. Carosi, E. Melachrinoudis, C. Petrioli, and Z. M. Wang, "Controlled sink mobility for prolonging wireless sensor networks lifetime," Wireless Netw., vol. 14, no. 6, pp. 831–858, Dec. 2008.

[7] Z.Wang, S. Basagni, E. Melachrinoudis, and C. Petrioli, "Exploiting sink mobility for maximizing sensor networks lifetime," in Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci., Hawaii, USA, Jan. 2005, p. 287a.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_24

Digital Watermarking Techniques for Identifying AI-Generated Images: A Systematic Review

First Author

Dr.B.Sree Vidya,

Asst. Prof, PG Department of Computer Science,

Shri Shankarlal Sundarbai Shasun Jain College for Women,

Chennai

Ms.Rajani Shree Sadhasivan,

Student,

M.Sc CS, PG Department of Computer Science,

Shri Shankarlal Sundarbai Shasun Jain College for Women,

Chennai

Ms.Nandhika,

Student,

M.Sc CS, PG Department of Computer Science,

Shri Shankarlal Sundarbai Shasun Jain College for Women,

Chennai

Abstract

The rapid growth of generative artificial intelligence has made it increasingly difficult to distinguish AI-generated images from real images, raising concerns related to authenticity, copyright, ownership, and misuse. Digital watermarking offers a practical way to identify the origin of generated content and support its traceability. This paper presents a systematic review of recent watermarking techniques developed for AI-generated images, with particular focus on invisibility, robustness, detection, copyright protection, and resistance to removal and forgery attacks. The review covers approaches based on deep learning, latent diffusion models, semantic watermarking, frequency-domain optimization, and user-specific watermarking. The findings show that recent methods have improved watermark robustness and image quality, but advanced generative-AI attacks can still remove or forge watermarks. The review therefore highlights the need for watermarking methods that provide stronger security, reliable attribution, and robustness against evolving attacks.

I. INTRODUCTION

The rapid development of Artificial Intelligence (AI) has made it possible to generate highly realistic and diverse digital images. AI-generated images are increasingly used in creative, artistic, entertainment, and other practical applications because they can be produced efficiently, with high quality and at a lower cost. Technologies such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Transformers, and diffusion models have contributed to these advancements.

Digital watermarking[1] is a technique used to embed hidden information, known as a watermark, into a digital medium such as an image. The embedded information can later be detected or extracted to verify ownership, authenticity,

ISBN NO: 978-81-180146-5-2

or origin. In AI-generated images, watermarking can be used to provide information about the source of the generated content while keeping the watermark visually imperceptible. Deep learning-based watermarking techniques have further improved the invisibility and robustness of watermarks [2].

Although AI-generated images provide many advantages, their increasing use also creates several concerns. It can be difficult to distinguish AI-generated images from real images, which may lead to problems related to content authenticity, copyright, intellectual property, misinformation, and misuse of generated content. Therefore, identifying AI-generated images has become important for maintaining the authenticity and traceability of digital content [3].

Watermarking provides a possible approach for addressing these issues by embedding identifiable information into AI-generated images. Such information can help determine whether an image was generated by an AI model and, in some approaches, can also support tracing the image to its source or responsible user. Techniques such as Stable Signature and Tree-Ring Watermarking demonstrate how watermarks can be integrated with diffusion-based image generation for identifying generated content. [4]

However, watermarking techniques also face challenges such as watermark removal, forgery, robustness against attacks, and maintaining image quality. Therefore, recent research focuses on developing watermarking methods that are secure, robust, and capable of maintaining the quality of AI-generated images. [5]

This review examines recent research on digital watermarking techniques for identifying and protecting AI-generated images, with emphasis on watermark embedding, detection, robustness, copyright protection, forgery resistance, and traceability.

II. DIGITAL WATERMARKING TECHNIQUES - REVIEW

This section examines in detail about the various research contributions from across the globe that have been proposed for identifying AI generated images. These techniques have been critically analyzed and presented in the following sections.

2.1 Deep-learning based technique for watermarking AI generated images

Seven Govwal et.al [6] proposed SynthID which is a deep-learning technique for watermarking AI generated images in an invisible way. The authors discuss the technical requirements, threat models, and practical challenges involved in deploying such a watermarking system at internet scale, with particular focus on effectiveness, visual fidelity, robustness, and security. The system has been used to watermark more than ten billion images and video frames across Google's services. The paper also presents an evaluation of an external variant, SynthID-O and reports state-of-the-art performance in terms of visual quality and robustness against common image modifications. The authors further state that the deployment-related findings may be applicable to other types of media, including audio.

The general working of their proposed technique is depicted in Fig.1

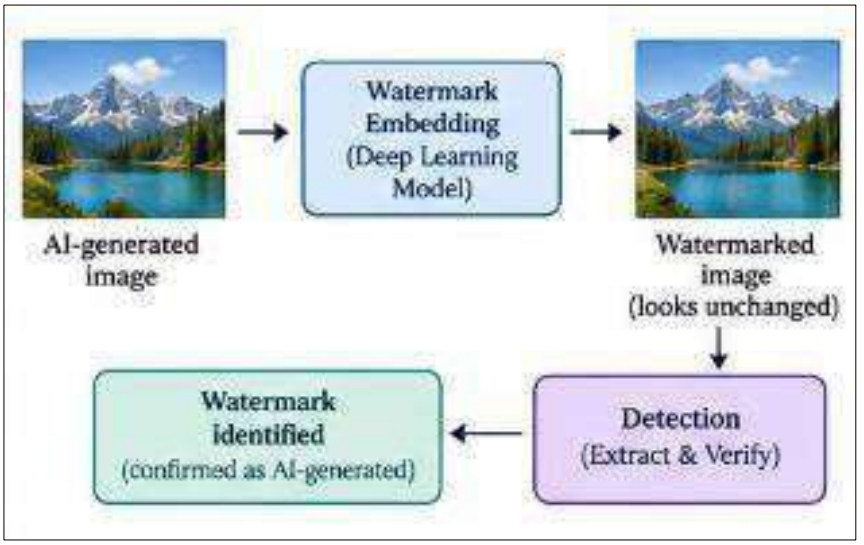


Fig.1. Invisible Digital watermarking technique on AI generated images

The paper demonstrates that invisible watermarking can be an effective approach for identifying and establishing the provenance of AI-generated media at large scale. The reported deployment and robustness results indicate that such

ISBN NO: 978-81-180146-5-2

systems can maintain visual quality while providing protection against common image modifications. Overall, the work highlights the importance of watermarking in improving the traceability and responsible use of AI-generated content.

2.2 Attack-Resilient Image Watermarking Using Stable Diffusion

Lijun Zhang et. al. introduce ZoDiac[7], a robust image watermarking approach that uses a pre-trained Stable Diffusion model to embed watermarks into a trainable latent space. The authors address the problem of existing watermarking techniques being vulnerable to attacks, particularly attacks based on Stable Diffusion itself. The proposed method enables the watermark to be detected from the latent representation even after various attacks. The authors evaluate ZoDiac using the MS-COCO, DiffusionDB, and WikiArt benchmarks and report a watermark detection rate above 98% with a false positive rate below 6.4%. The results show that ZoDiac outperforms existing state-of-the-art watermarking methods and remains effective against strong watermark removal attacks.

The paper demonstrates that Stable Diffusion can be used not only to generate images but also to develop more robust watermarking techniques. ZoDiac provides strong resistance against watermark attacks while maintaining reliable detection performance. Overall, the study highlights the potential of diffusion models for improving the security and reliability of image watermarking for provenance and ownership verification. The results prove that the water mark detection rates are quite high as compared to other techniques used in the study.

2.3 Removing invisible watermarks using Generative AI

The research work proposed by Xuandong Zhao et.al.[8] investigates the vulnerability of invisible pixel-level image watermarks to attacks using generative AI. The authors propose a family of regeneration attacks, in which random noise is first added to an image to disrupt the embedded watermark and the image is subsequently reconstructed using image-denoising algorithms or pre-trained generative models such as diffusion models. Through formal analysis and extensive experiments, the authors demonstrate that pixel-level invisible watermarks can be effectively removed using this approach. Across four different watermarking schemes, the proposed attacks achieve lower watermark detection rates while maintaining higher image quality compared with existing attack methods. The paper also explores semantically preserving watermarks as a possible alternative defense.

The workings of the proposed technique is as shown in Fig.2.

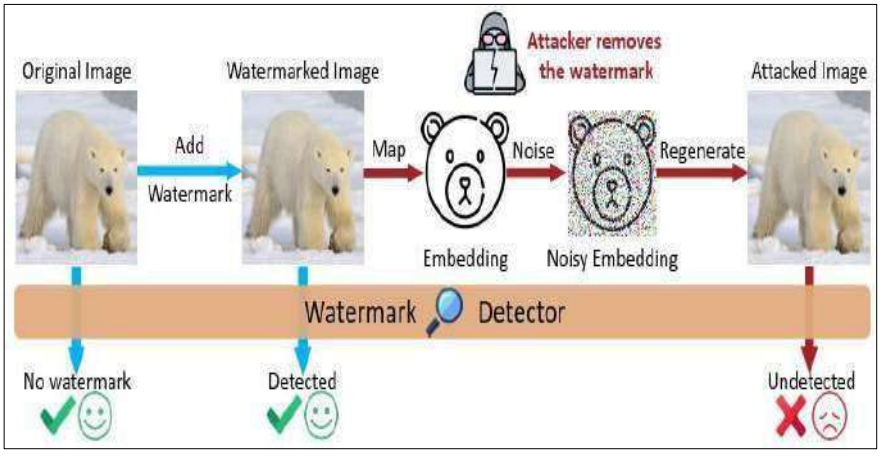


Fig.2. Removing invisible watermarks

2.4. Black-Box Forgery Attacks on Semantic Watermarks for Diffusion Models

The research study by Müller A et.al. [9] investigates security vulnerabilities in semantic watermarking techniques used with latent diffusion models, particularly Tree-Ring and Gaussian Shading watermarks. Although these methods are designed to be robust against various image modifications, the authors demonstrate that they can be vulnerable to forgery attacks. The paper presents two black-box watermark forgery attacks. The first attack embeds a targeted watermark into real images and can also be used to remove existing watermarks. The second attack generates new images containing a target watermark by inverting and regenerating a watermarked image with a different prompt. Both attacks require only a single reference image containing the target watermark and can be performed using unrelated diffusion models.

ISBN NO: 978-81-180146-5-2

The proposed work demonstrates that semantic watermarks, despite their robustness against common image modifications, can still face serious security risks from carefully designed forgery attacks. The findings indicate that attackers may be able to manipulate or reproduce watermarks without access to the original watermarking model. Overall, the study highlights the need for stronger security mechanisms to ensure that semantic watermarks remain reliable for image attribution and provenance.

2.5 A Watermark-Conditioned Diffusion Model for IP Protection

The paper proposes WaDiff [10], a watermark-conditioned diffusion model designed to protect the intellectual property of AI-generated content. The authors address the limitation of existing watermarking methods, which primarily focus on detecting synthetic content but provide limited support for identifying the specific user responsible for generating the content. The proposed framework considers a practical black-box API setting in which multiple users access a diffusion model provided by a model owner. WaDiff incorporates user-specific hidden information into the image-generation process, allowing the generated content to be detected and traced back to the responsible user. The authors evaluate the framework on two popular diffusion models and report that it performs effectively in both content detection and owner identification while having a negligible effect on image generation quality.

The paper extends the role of watermarking from simply detecting AI-generated content to identifying the user responsible for generating it. The proposed approach provides a promising solution for copyright protection and forensic tracking in diffusion-model-based content generation. Overall, the study demonstrates the potential of user-specific watermarking for improving accountability and intellectual property protection in AI-generated content.

2.6 The Stable Signature: Rooting Watermarks in Latent Diffusion Models

The paper introduces an active content tracing method that combines image watermarking with Latent Diffusion Models to enable the detection and identification of generated images [11]. The method fine-tunes the latent decoder of the image generator using a binary signature so that generated images contain an invisible watermark. A pre-trained watermark extractor is then used to recover the hidden signature, while a statistical test determines whether an image originated from the generative model. The method is evaluated for watermark invisibility and robustness against different image modifications, demonstrating that the watermark can remain detectable even after significant modifications such as cropping.

The approach appears useful because it provides both detection and identification of images generated by a diffusion model through an invisible signature. The use of a statistical test provides an additional way to determine whether an image is associated with the generative model. The demonstrated robustness against modifications also makes the approach promising for content tracing and responsible use of generated images.

2.7 WAVES: Benchmarking the Robustness of Image Watermarks

This research study by Ding et.al [12] proposes WAVES to test the robustness of image watermarks. It addresses the limitations of existing watermark evaluation methods through stress testing with different types of attacks. The evaluation considers both the degree of image quality degradation and the effectiveness of watermark detection after the attacks. The purpose is to identify vulnerabilities in existing watermarking algorithms and support the development of more robust watermarks.

The approach appears useful because it does not only check whether a watermark can be detected, but also considers how much the image quality is affected after different attacks. Since modifying or attacking an image to test watermark robustness may also affect its quality, considering both aspects is important. The study also provides a useful way to identify weaknesses in existing watermarking methods, which can contribute to the development of more robust techniques.

2.8 GaussMarker: Robust Dual-Domain Watermark for Diffusion Models

The paper proposes GaussMarker[13], a dual-domain watermarking approach for diffusion models to address copyright and misuse concerns in AI-generated images. It addresses the limitations of existing methods that embed watermarks only in the initial Gaussian noise by consistently embedding watermarks in both the spatial and frequency domains. The paper also introduces a Gaussian Noise Restorer (GNR) to improve watermark detection after image manipulation and advanced attacks. The proposed method is evaluated under eight image distortions and four advanced attacks across three versions of Stable Diffusion, where it achieves strong performance with better recall and lower false positive rates.

The approach appears useful because embedding the watermark in two domains can potentially improve its

ISBN NO: 978-81-180146-5-2

robustness when the generated image undergoes different types of manipulation. The use of the Gaussian Noise Restorer also provides an additional way to improve watermark detection after an image has been modified. Overall, the study provides a promising approach for making watermarks in diffusion-generated images more resistant to distortions and advanced attacks.

2.9 DiffuseTrace: A Transparent and Flexible Watermarking Scheme for Latent Diffusion Model

The study proposes DiffuseTrace, a watermarking scheme for Latent Diffusion Models to address copyright tracking and risks associated with AI-generated content [14]. It addresses the limitations of post-processed watermarking methods, particularly their vulnerability to generative watermark attacks and the trade-off between watermark strength and image quality. DiffuseTrace embeds a multi-bit watermark into the image space without compromising image quality and does not require fine-tuning of the diffusion model components. The watermarking component can also be used as a plug-in with different diffusion models, providing flexibility for various applications.

The approach appears useful because the watermark is incorporated during the image generation process rather than being added only after the image has been generated. This can potentially make the watermark more resistant to attacks while maintaining the quality of the generated image. The flexibility of using the watermarking component as a plug-in also makes the approach promising for use with different diffusion models.

2.10 FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space

The study proposed by Guo et.al [15] uses FreqMark, an invisible image watermarking method designed to improve robustness against regeneration attacks. The method works by optimizing the frequency space of the image's latent representation after VAE encoding and uses a pre-trained image encoder to extract the watermark. This optimization provides a flexible balance between image quality and watermark robustness while also allowing different numbers of bits to be encoded. The experimental results show that FreqMark provides strong image quality and robustness under various attack scenarios.

The approach appears useful because it focuses on maintaining both image quality and watermark robustness rather than heavily favoring one over the other. The use of a pre-trained image encoder for watermark extraction also provides a way to recover the hidden watermark after the image has undergone attacks. Overall, the method seems promising for protecting images against regeneration attacks while keeping the watermark invisible and the image quality intact.

2.11 CSGuard: Toward Forgery-Resistant Watermarking in Diffusion Models via Compressed Sensing Constraint

The Lai J. et.al propose CSGuard, a forgery-resistant watermarking scheme for diffusion models to address the vulnerability of existing latent-based watermarking methods to forgery attacks [16]. It uses compressed sensing to bind the watermark embedding and verification processes to a secret matrix. This ensures that only users possessing the secret matrix can correctly embed or verify the watermark, thereby preventing unauthorized users from creating content with false watermark attribution. The experimental results demonstrate that CSGuard provides strong resistance against forgery attacks without compromising image generation quality or watermark integrity.

The approach appears useful because it introduces an additional security mechanism for protecting image watermarks against forgery. Instead of focusing only on whether a watermark can survive image manipulation, the method also controls who can correctly embed or verify the watermark through a secret matrix. This makes the approach promising for applications where preventing false attribution and unauthorized watermark use is important.

2.11 AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA

The paper proposes AquaLoRA, a watermarking approach designed to provide white-box protection for customized Stable Diffusion models. It addresses the limitation of existing watermarking methods where a malicious user can remove or replace the watermarking module from the model. AquaLoRA merges watermark information into the U-Net of Stable Diffusion through a watermark LoRA module using a two-stage approach. It also introduces a scaling matrix for flexible watermark message updates without retraining and uses Prior Preserving Fine-Tuning (PPFT) to minimize the impact of watermark learning on the model's original behavior and image quality. The experiments and ablation studies validate the effectiveness of the proposed design. The overview of the diffusion technique is illustrated in Fig.3.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_24

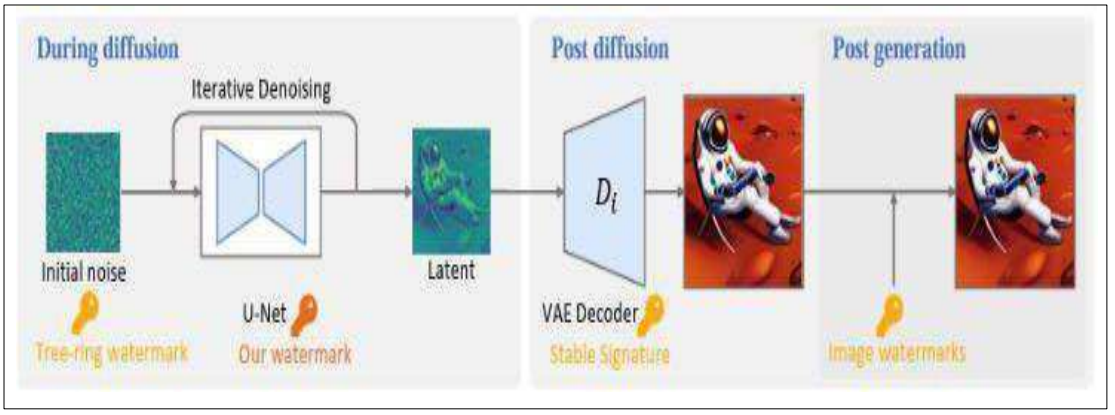


Fig.3. Overview of the Watermark LoRA-based protection approach.

The approach appears useful because the watermark information is incorporated into the model itself rather than being kept as a separate component that can be easily removed. This provides an additional level of protection for customized diffusion models against unauthorized distribution and use. The ability to update the watermark without retraining the model also makes the approach more flexible and practical.

2.13Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust

The paper introduces Tree-Ring Watermarking [18], a novel technique for embedding invisible and robust fingerprints into images generated by diffusion models. Unlike conventional watermarking methods that modify an image after it has been generated, the proposed approach embeds a structured pattern into the initial noise used during the image-generation process. The watermark is designed in Fourier space, making it resistant to common image transformations such as cropping, rotation, flipping, and resizing. After generation, the watermark can be detected by reversing the diffusion process and examining the recovered noise for the embedded pattern. The authors demonstrate that the technique can be applied to different diffusion models, including Stable Diffusion, with negligible impact on image quality.

The paper demonstrates that embedding watermarks directly into the image-generation process can provide stronger robustness than conventional post-generation watermarking methods. The proposed approach offers an effective way to trace and identify diffusion-generated images while maintaining their visual quality. Overall, the work highlights the potential of robust watermarking as a practical solution for addressing copyright, traceability, and responsible use of AI-generated content.

III. SUMMARY OF REVIEW

This section summarizes the various contributions discussed in the previous section on Digital water marking and tabulated in table 1.

TABLE.1. SUMMARY OF REVIEW

Author	Method	Year	Contributions
Gowal et al.	SynthID-Image	2025	Introduces a deep learning-based invisible watermarking system for AI-generated images, designed for large-scale deployment with high visual fidelity, robustness, and security.
Wen et al.	Tree-Ring Watermarking	2023	Embeds structured patterns into the initial noise of diffusion models to create invisible and robust fingerprints that withstand common image transformations.
Zhang et al.	ZoDiac	2024	Uses Stable Diffusion to embed watermarks into the latent space, providing strong resistance against watermark removal attacks and reliable detection.
Zhao et al.	Regeneration Attack	2024	Demonstrates that pixel-level invisible watermarks can be removed using generative-AI-based regeneration attacks and highlights semantic-preserving watermarking as a possible defense.
Müller et al.	Black-Box Forgery Attacks	2025	Demonstrates that semantic watermarks such as Tree-Ring and Gaussian Shading can be forged or removed using unrelated diffusion models with only a single reference image.
Min et al.	WaDiff	2024	Proposes a watermark-conditioned diffusion model that embeds user-specific information for both AI-content detection and identification of the user responsible for generating it.
Pierre Fernandez.	The Stable Signature	2023	Introduces an invisible signature-based watermarking method for detecting and identifying images generated by latent diffusion models.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_24

Author	Method	Year	Contributions
BangAn.	WAVES	2024	Provides a benchmark for evaluating the robustness of image watermarks against different attacks while considering image quality.
Kecen Li.	GaussMarker	2025	Proposes a dual-domain watermarking method with a Gaussian Noise Restorer to improve watermark robustness against distortions and advanced attacks.
Liangqi Lei.	DiffuseTrace	2025	Introduces a flexible multi-bit watermarking scheme for latent diffusion models that avoids model fine-tuning while maintaining image quality.
Yiyang Guo.	FreqMark	2024	Uses frequency-based optimization in latent space to improve watermark robustness against regeneration attacks while preserving image quality.
Jiewei Lai.	CSGuard	2026	Introduces a compressed-sensing-based security mechanism to prevent unauthorized watermark forgery and false attribution in diffusion models.
Weitao Feng.	AquaLoRA	2024	Integrates watermark information into customized Stable Diffusion models through Watermark LoRA to provide stronger protection against watermark removal and unauthorized model use.

IV. FUTURE WORK

Future research can focus on developing watermarking techniques that remain reliable against increasingly sophisticated generative-AI attacks. The reviewed studies show that conventional invisible watermarks can be removed through regeneration, while semantic watermarks may also be vulnerable to forgery. Therefore, there is a need for hybrid watermarking approaches that combine spatial, frequency, and latent-domain information to improve resistance against different types of manipulation. Future methods can also investigate stronger authentication mechanisms, secret-key based verification, and adaptive watermarking strategies that can respond to new attack patterns without significantly affecting image quality.

Another important direction is the development of practical and scalable watermarking systems that can work across different image-generation models. Techniques such as user-specific watermarking and model-integrated watermarking provide opportunities for improving accountability and tracing generated content to its source. Future studies can evaluate these approaches using common benchmarks and a wider range of attacks, while considering detection accuracy, false positives, image quality, computational cost, and scalability. Such comprehensive evaluation would help in developing watermarking systems suitable for real-world applications where reliable identification and protection of AI-generated images are required.

V. CONCLUSION

From the review of these papers, it can be understood that image watermarking for AI-generated content involves different approaches, each designed to address specific challenges such as detection, robustness, copyright protection, forgery, and watermark removal. The papers also show that while some watermarking techniques provide strong protection against certain attacks, they may still have limitations when exposed to other types of attacks. At the same time, newer approaches attempt to overcome these weaknesses by improving the security, robustness, and flexibility of watermarking methods. Overall, this review helped me understand that there is no single approach that completely solves all the challenges of AI-generated image watermarking, and that watermarking techniques continue to develop as new threats and vulnerabilities emerge.

REFERENCES

1. Kui, R.; Yang, Z.P.; Lu, L.; Liu, J.; Li, Y.M.; Wan, J.; Zhao, X.D.; Feng, X.H.; Shao, S. *SoK: On the role and future of AIGC watermarking in the era of Gen-AI*. arXiv, 2024, arXiv:2411.11478.
2. Hosny, K.M.; Magdi, A.; ElKomy, O.; Hamza, H.M. *Digital image watermarking using deep learning: A survey*. Computer Science Review, 2024, 53, 100662.
3. Zhang, H.L.; Edelman, B.L.; Francati, D.; Venturi, D.; Ateniese, G.; Barak, B. *Watermarks in the sand: Impossibility of strong watermarking for generative models*. arXiv, 2023, arXiv:2311.04378.
4. Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; Furon, T. *The Stable Signature: Rooting watermarks in latent diffusion models*. Proceedings of the International Conference on Computer Vision, 2023, pp. 22409–22420.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_24

5. Wen, Y.X.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. *Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust*. arXiv, 2023, arXiv:2305.20030
6. Gowal S, Bunel R, Stimberg F, Stutz D, Ortiz-Jimenez G, Kouridi C, Vecerik M, Hayes J, Rebuffi SA, Bernard P, Gamble C. SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263. 2025 Oct 10.
7. Zhang L, Liu X, Martin AV, Bearfield CX, Brun Y, Guan H. Attack-resilient image watermarking using stable diffusion. Advances in Neural Information Processing Systems. 2024 Dec 16;37:38480-507.
8. Zhao X, Zhang K, Su Z, Vasan S, Grishchenko I, Kruegel C, Vigna G, Wang YX, Li L. Invisible image watermarks are provably removable using generative ai. Advances in neural information processing systems. 2024 Dec 16;37:8643-72.
9. Müller A, Lukovnikov D, Thietke J, Fischer A, Quiring E. Black-box forgery attacks on semantic watermarks for diffusion models. In2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025 Jun 10 (pp. 20937-20946). IEEE.
10. Min R, Li S, Chen H, Cheng M. A watermark-conditioned diffusion model for ip protection. In European Conference on Computer Vision 2024 Sep 29 (pp. 104-120). Cham: Springer Nature Switzerland.
11. Fernandez P, Couairon G, Jégou H, Douze M, Furon T. The stable signature: Rooting watermarks in latent diffusion models. In2023 IEEE/CVF International Conference on Computer Vision (ICCV) 2023 Oct 1 (pp. 22409-22420). IEEE.
12. An B, Ding M, Rabbani T, Agrawal A, Xu Y, Deng C, Zhu S, Mohamed A, Wen Y, Goldstein T, Huang F. Waves: Benchmarking the robustness of image watermarks. arXiv preprint arXiv:2401.08573. 2024 Jan 16.
13. Li K, Huang Z, Hou X, Hong C. Gaussmarker: Robust dual-domain watermark for diffusion models. arXiv preprint arXiv:2506.11444. 2025 Jun 13.
14. Lei L, Gai K, Yu J, Zhu L. Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model. arXiv preprint arXiv:2405.02696. 2024 May 4.
15. Guo Y, Li R, Hui M, Guo H, Zhang C, Cai C, Wan L, Wang S. Freqmark: Invisible image watermarking via frequency based optimization in latent space. Advances in Neural Information Processing Systems. 2024 Dec 16;37:112237-61.
16. Lai J, Zhang L, Tang C, Sun P, Zhang Z, Wang Y, Jin H. CSGuard: Toward Forgery-Resistant Watermarking in Diffusion Models via Compressed Sensing Constraint. arXiv preprint arXiv:2605.01479. 2026 May 2.
17. Feng W, Zhou W, He J, Zhang J, Wei T, Li G, Zhang T, Zhang W, Yu N. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. arXiv preprint arXiv:2405.11135. 2024 May 18.
18. Wen Y, Kirchenbauer J, Geiping J, Goldstein T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030. 2023 May 31.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_25

KidShield – Integrated Child Vaccine Surveillance and Management Portal

Tamil Nila K
Student
Department of Information Technology
PSG College of Arts & Science
Coimbatore, India, 641017

Dr. R Sujatha
Assistant Professor
Department of Information Technology
PSG College of Arts & Science
Coimbatore, India, 641017

Abstract*— Childhood vaccination is an essential part of preventive healthcare, helping protect children from various vaccine-preventable diseases. Maintaining vaccination records through paper-based documents and manual tracking can result in misplaced records, missed vaccination dates, duplicate data, and difficulties in monitoring a child's vaccination status. Parents may also face challenges in remembering upcoming vaccination schedules and accessing complete vaccination information when required. Effective management of vaccination information is important for maintaining accurate records, monitoring vaccination progress, and ensuring timely completion of recommended doses. Digital healthcare solutions can provide a more organized and accessible method for maintaining child vaccination records. Centralized record management can reduce manual effort, improve data accessibility, support timely monitoring, and minimize errors in vaccination information. The use of web-based technologies in healthcare provides an efficient approach for managing and monitoring childhood immunization data. Such systems can support secure access to vaccination records, improve communication between healthcare stakeholders, and provide opportunities for future enhancements such as automated reminders, predictive analytics, mobile applications, and integration with healthcare services.

Keywords*— Childhood Vaccination, Immunization, Vaccine Surveillance, Child Healthcare, Vaccination Management, Digital Healthcare, Django, Web Application, Vaccination Records.

I. INTRODUCTION

Childhood immunization plays an important role in protecting children from serious infectious diseases and improving public health. Since vaccines are given according to specific schedules and doses, maintaining accurate records and receiving vaccines on time are essential.

Traditional vaccination management mainly depends on paper cards and manually maintained records. These methods may lead to misplaced records, incorrect entries, difficulty in retrieving information, and missed vaccination dates. Managing vaccination schedules can become even more difficult when parents have multiple children.

Digital healthcare applications provide an organized way to maintain vaccination information. They can help store child details, manage vaccine records, monitor vaccination status, track due dates, and provide timely updates. Such digital approaches can reduce manual effort, improve accessibility, minimize errors, and support effective childhood immunization management.

II. LITERATURE SURVEY

Several studies have explored the use of digital technologies to improve childhood immunization management. Research on Vaccination Information Systems (VIS) has shown that electronic vaccination records can support better record management and accessibility. However, challenges such as data quality, interoperability, privacy, usability, and system infrastructure have also been reported.

Mekonnen et al. studied the effectiveness of mobile text-message reminders for routine childhood vaccination through a systematic review and meta-analysis. Their findings showed that reminder messages can improve childhood vaccination coverage, highlighting the usefulness of digital reminders in supporting timely immunization.

Similarly, a systematic review by Aregawi et al. examined SMS-based interventions in low- and middle-income countries and reported improvements in both vaccination coverage and the timely receipt of childhood vaccines.

Recent research has also reviewed SMS interventions for children and adolescents and found small to moderate improvements in vaccination coverage and timeliness compared with standard care. These studies indicate that digital record management combined with timely reminders can support better vaccination monitoring.

Based on these studies, digital vaccination systems provide a useful approach for maintaining records, monitoring vaccination schedules, and supporting timely immunization.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_25

III. SYSTEM DESIGN AND IMPLEMENTATION

The proposed KidShield system follows a centralized web-based architecture in which users interact with the application through role-based dashboards. The system processes user requests through the Django backend and stores child and vaccination information in the SQLite database.

A. User Interface

HTML, CSS, Bootstrap, and JavaScript are used to create a responsive and user-friendly interface. Dashboards and forms are designed to provide easy access to child and vaccination information.

B. *Backend*

Python and Django are used to handle authentication, database operations, application logic, user roles, and system functionality.

C. *Database*

SQLite is used as the centralized database for storing user profiles, child details, vaccine records, and vaccination-related information.

D. User Roles

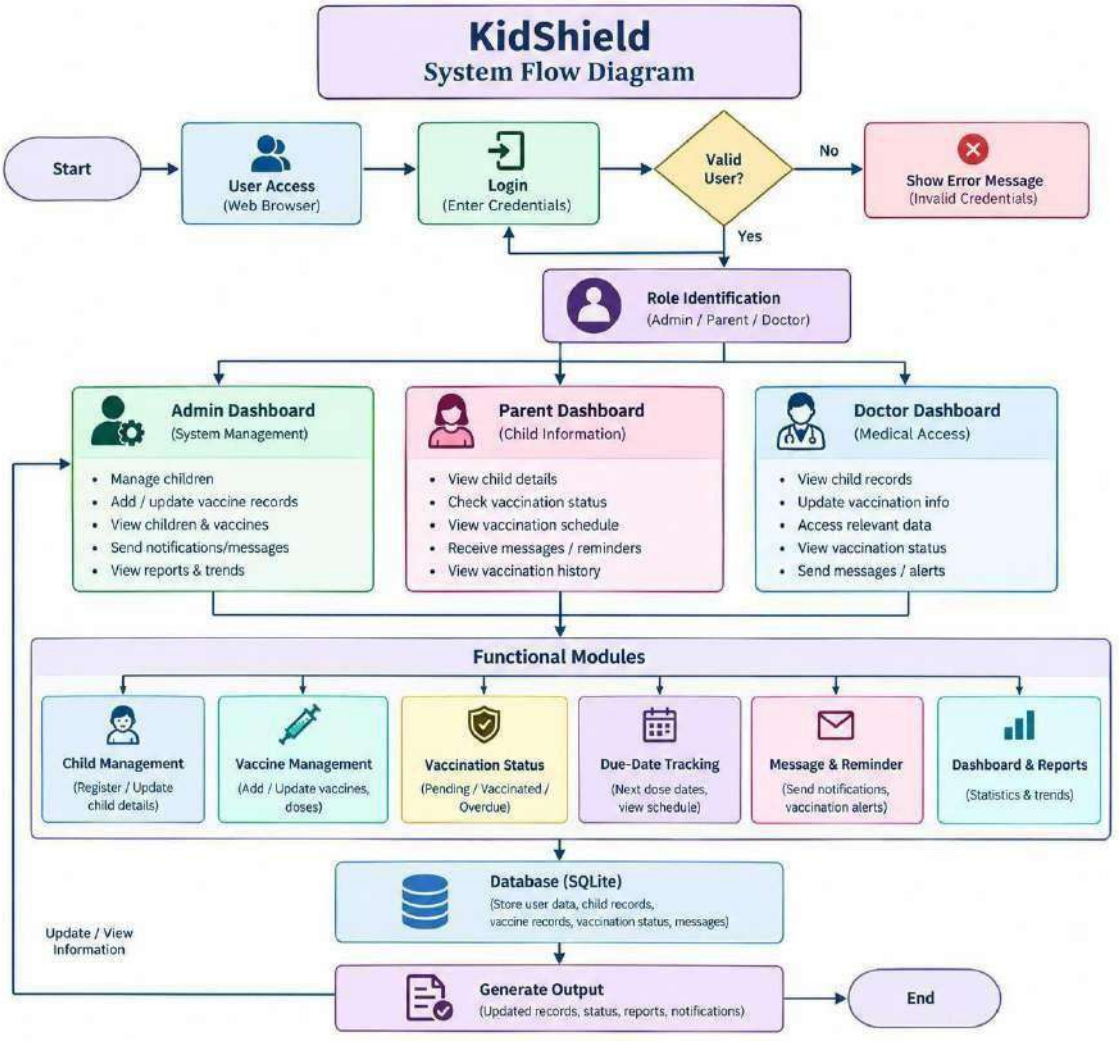
The system provides role-based access for:

- Admin – Manages children, vaccine records, and system information.
- Parent – Views children's vaccination information and vaccination-related messages.
- Doctor – Provides authorized access to relevant vaccination information.

E. Workflow

User → Authentication → Role-Based Dashboard → Child/Vaccine Modules → Database → Status & Messages.

Role-based authentication is used to ensure that users can access only the functionalities relevant to their assigned roles.



IV. FUNCTION MODULES

The proposed KidShield system consists of multiple integrated modules that work together to provide efficient child vaccination management.

ISBN NO: 978-81-180146-5-2

A. Admin Module

The Admin module provides overall control of the system. The administrator can register children, manage vaccine records, view child and vaccination information, monitor vaccination status, and send vaccination-related messages.

B. Parent Module

The Parent module allows parents to access their children's vaccination information. Parents can view child details, vaccination records, vaccine status, upcoming due dates, and messages from the system.

C. Doctor Module

The Doctor module provides access to relevant child and vaccination information for monitoring vaccination records and supporting healthcarerelated activities.

D. Child Management Module

This module maintains child-related information such as token number, parent name, phone number, child name, date of birth, and gender. It allows child records to be stored and managed systematically.

E. Vaccine Management Module

This module maintains vaccination records, including vaccine name, dose number, due date, vaccination date, and vaccination status. The vaccine status can be maintained as Pending, Vaccinated, or Overdue.

F. Notification and Message Module

This module manages vaccination-related messages and reminders. It helps communicate important vaccine information and due-date updates to parents.

V. ALGORITHMS USED

The KidShield system uses rule-based and date-based algorithms to monitor child vaccination records. These algorithms help determine vaccination status, identify due or overdue vaccines, and support timely vaccination reminders.

A. Rule-Based Vaccine Status Algorithm

The rule-based algorithm determines the vaccination status by checking the vaccination date and vaccine due date.

1. Retrieve the child's vaccine record from the database.
2. Check whether a vaccination date is available.
3. If the vaccination date is available, the vaccine is marked as Vaccinated.
4. If the vaccination date is not available, compare the due date with the current date.
5. If the due date has not passed, the vaccine is marked as Pending.
6. If the due date has passed, the vaccine is marked as Overdue.
7. Display the updated status in the vaccination record.

B. Date-Based Vaccine Due-Date Monitoring Algorithm

This algorithm monitors the due dates of vaccines and helps identify upcoming and overdue vaccinations.

1. Retrieve the vaccine due date from the database.
2. Compare the due date with the current date.
3. Check whether the vaccine has already been administered.
4. If vaccination is completed, maintain the status as Vaccinated.
5. If vaccination is not completed and the due date is approaching, maintain the status as Pending.
6. If the due date has passed without vaccination, mark the vaccine as Overdue.
7. Display the vaccine due-date information to the user.

C. Notification and Reminder Algorithm

The notification algorithm is used to provide vaccination-related messages to parents based on vaccine records and their status.

1. Check the child's vaccination records.
2. Identify pending or overdue vaccine records.
3. Retrieve the vaccine name and corresponding due date.
4. Generate an appropriate vaccination-related message.
5. Store the message in the system.
6. Display the message to the parent through the dashboard.

These algorithms work together to provide organized vaccine surveillance by maintaining vaccination records, monitoring due dates, identifying vaccination status, and communicating important vaccination information to parents.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_25

VI. COMPARATIVE SYNTHESIS AND DISCUSSION

A. Comparison with Traditional Vaccination Management

Traditional vaccination management mainly depends on paper vaccination cards and manually maintained records. These methods can make it difficult to retrieve information, track vaccination dates, and maintain records for multiple children. KidShield provides a centralized digital approach for storing child and vaccination information, making records easier to access and manage.

B. System Efficiency and Integration

KidShield integrates child registration, vaccine record management, vaccination status monitoring, due-date tracking, and parent communication into a single platform. The centralized database allows information to be stored and retrieved systematically. Role-based access also provides appropriate functionality for Admin, Parent, and Doctor users.

C. Advantages and System Performance

The system reduces manual record maintenance and helps minimize errors in vaccination information. It allows users to monitor vaccination status and due dates efficiently. The dashboard provides organized access to child and vaccine information, while vaccination-related messages help improve communication with parents.

D. Limitations and Future Scope

The current system focuses mainly on digital vaccination record management, status monitoring, due-date tracking, and messaging. Automated SMS or email notifications, mobile application support, predictive analytics, cloud deployment, and integration with healthcare services can be added in the future to improve the scalability and functionality of the system.

VII. RESULTS AND DISCUSSION

The system successfully provides a centralized platform for managing child vaccination records, including child details, vaccine information, due dates, vaccination dates, and status.

Separate access is provided for Admin, Parent, and Doctor users. The system helps parents view vaccination information and due dates, while

Admin can manage records and communicate with parents through messages.

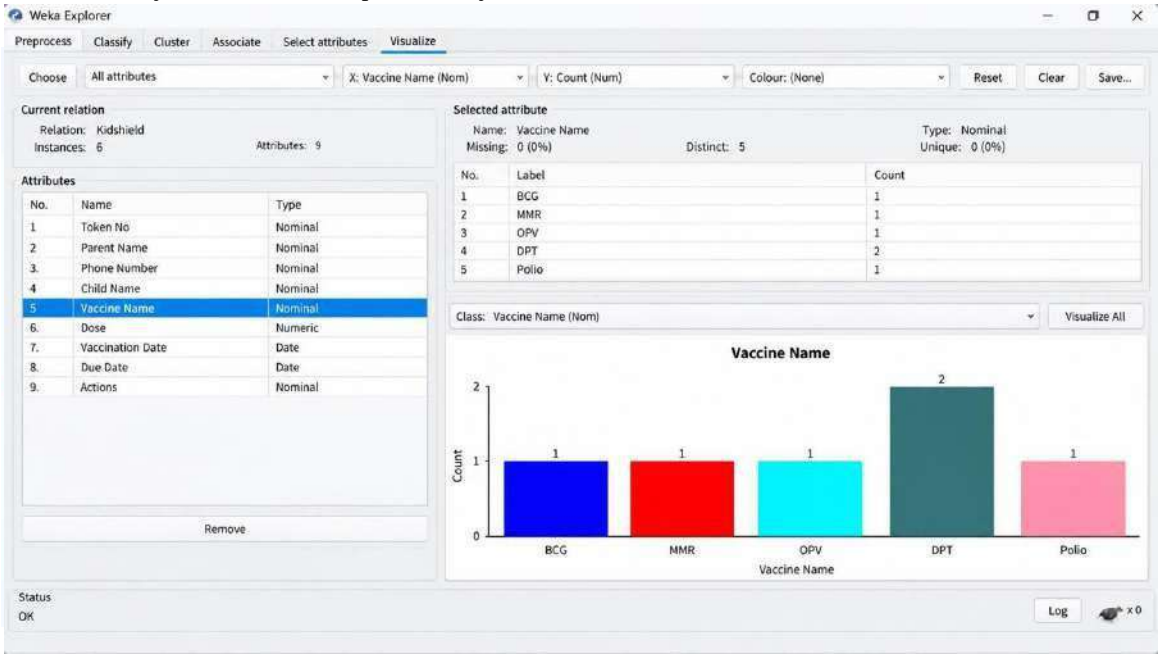
Vaccination records are classified as Pending, Vaccinated, or Overdue, making it easier to monitor a child's vaccination progress.

Overall, the system reduces manual record maintenance, improves accessibility of vaccination information, and supports better vaccination monitoring.

VIII. CONCLUSION

The child vaccination management system provides an organized and digital approach for maintaining vaccination records. It helps manage child details, vaccine information, due dates, vaccination status, and parent communication in a centralized platform. The system reduces manual record maintenance and makes vaccination information easier to access and monitor.

Overall, the system supports effective vaccination management and helps parents and healthcare users keep track of children's vaccination progress. Future enhancements such as automated SMS/email reminders, mobile applications, cloud deployment, and advanced analytics can further improve the system.



ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_25

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Information Technology, PSG College of Arts and Science (PSGCAS), Coimbatore, for the support, guidance, and motivation provided throughout the development of this study. The authors also extend their gratitude to the faculty members and all those who contributed directly or indirectly to the successful completion of the study.

REFERENCES

[1] Z. A. Mekonnen, A. Gebeyehu, M. Birhanu, et al., "Effect of mobile text message reminders on routine childhood vaccination: A systematic review and meta-analysis," Systematic Reviews, vol. 8, 2019.

[2] "Short message service (SMS) reminders for childhood immunisation in low-income and middle-income countries: A systematic review and metaanalysis," BMJ Global Health, 2021.

[3] "SMS-based interventions for improving child and adolescent vaccine coverage and timeliness: A systematic review," BMC Public Health, 2024.

[4] "Challenges of vaccination information system implementation: A systematic literature review," Human Vaccines & Immunotherapeutics, 2023. [5] Django Software Foundation, "Django Documentation."

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_26

TN Postal Saving Scheme AI Assistant: A RAG-Based Conversational System for India Post Savings Schemes

Dharshini T
Student, Department of Information Technology
PSG College of Arts and Science
Coimbatore, Tamil Nadu, India – 641014

Dr. Sujatha
Assistant Professor, Department of Information Technology
PSG College of Arts and Science
Coimbatore, Tamil Nadu, India – 641014

Abstract*—Access to information about India Post savings schemes is often fragmented across multiple web pages, forms, and physical post office visits, making it difficult for users to compare eligibility, interest rates, tenure, and tax benefits before choosing a scheme. This paper presents the TN Postal Saving Scheme AI Assistant, a Flask-based conversational system that combines Retrieval-Augmented Generation (RAG) with Google Gemini 2.5 Flash to answer user queries about ten India Post savings schemes: PPF, SSY, SCSS, NSC, KVP, MIS, RD, TD, SB, and MSSC. Rather than relying on the generative model as the source of factual information, the system maintains a central, verified knowledge base that is injected into every prompt, separating factual accuracy from natural-language generation. The assistant additionally provides deterministic financial calculators, a scheme-comparison module, a weighted suitability-scoring recommendation engine, and an offline post-office locator built on India Post's pincode dataset. The system supports English, Tamil, and Tanglish input through rule-based intent detection. Security measures including CSRF protection, rate limiting, and hardened session management were incorporated to support safe deployment.

***Index Terms*—Retrieval-Augmented Generation, Natural Language Processing, Google Gemini, Conversational AI, India Post, Savings Schemes, Flask, Recommendation System.**

I. INTRODUCTION

India Post offers a range of government-backed savings schemes such as the Public Provident Fund (PPF), Sukanya Samriddhi Yojana (SSY), Senior Citizens Savings Scheme (SCSS), National Savings Certificate (NSC), Kisan Vikas Patra (KVP), Monthly Income Scheme (MIS), Recurring Deposit (RD), Time Deposit (TD), Savings Bank (SB), and Mahila Samman Savings Certificate (MSSC). Each scheme differs in eligibility, investment limits, interest rate, tenure, and tax treatment, and this information is typically scattered across official circulars, brochures, and third-party websites. As a result, first-time investors and rural or semi-urban users often find it difficult to identify the scheme best suited to their financial goals without visiting a post office or consulting an agent.

Recent progress in Natural Language Processing (NLP) and generative Artificial Intelligence (AI) has made conversational interfaces a practical way to simplify access to structured domain information. However, general-purpose generative models are prone to producing plausible-sounding but factually incorrect responses when used as the sole source of domain knowledge, which is unacceptable for financial and government-scheme information where numerical accuracy directly affects a user's decisions.

This paper presents the design and implementation of the TN Postal Saving Scheme AI Assistant, a web-based chatbot that addresses this problem through a hybrid architecture. A central, manually verified knowledge base acts as the single source of truth for all scheme data. Before generating a response, the system retrieves the relevant scheme information and injects it into the prompt sent to Google Gemini 2.5 Flash, so that the language model is used strictly for natural-language understanding and response generation rather than as a factual database.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed system and its architecture. Section IV discusses implementation details. Section V presents results and discussion, and Section VI concludes the paper with directions for future work.

II. LITERATURE SURVEY

The development of intelligent conversational systems has progressed significantly with the advancement of

ISBN NO: 978-81-180146-5-2

Natural Language Processing (NLP), Generative Artificial Intelligence (AI), and Retrieval-Augmented Generation (RAG). These technologies have made it possible to develop systems that can understand natural-language queries and provide information in a conversational manner. However, for applications involving financial and government-related information, the accuracy and reliability of the information are particularly important.

Lewis et al. [1] introduced the concept of Retrieval-Augmented Generation (RAG), which combines information retrieval with text generation. Instead of depending entirely on the knowledge stored within a language model, relevant information is retrieved from an external knowledge source and supplied to the model as context. This approach is useful for domain-specific applications because it helps ground generated responses in available knowledge. The proposed TN Postal AI Chatbot follows this principle by retrieving relevant postal-scheme information before generating responses.

Vaswani et al. [2] proposed the Transformer architecture through the work “Attention Is All You Need.” The Transformer introduced the attention mechanism, which became a foundation for many modern NLP and generative AI models. The ability of Transformer-based models to understand relationships between words and contextual information has contributed to the development of modern conversational AI systems. The proposed project uses a pretrained generative AI model rather than developing and training a language model from the beginning.

Recent generative AI systems such as Google's Gemini models [3] provide natural-language understanding and response-generation capabilities. In the proposed system, Gemini 2.5 Flash is used to generate conversational responses from the retrieved postal-scheme information. The model is not treated as the primary database of scheme information. Instead, the structured knowledge base supplies scheme-specific information and Gemini converts the retrieved information into a user-friendly response. This architecture helps separate factual domain information from language generation.

Existing digital information-access systems generally require users to search through multiple web pages, documents, or application forms to find information. In the case of postal savings schemes, users may need to identify eligibility, investment limits, interest rates, tenure, tax benefits, required documents, and application procedures separately. The proposed system integrates these requirements into a single conversational interface. It also provides additional functions such as financial calculations, scheme comparison, recommendation, application guidance, and post-office search.

Traditional recommendation approaches may depend on predefined rules or manual assistance. The proposed system applies a weighted suitability-scoring method to recommend schemes according to user requirements such as age, investment amount, investment mode, financial goal, duration, tax-saving preference, and other relevant conditions. Since the scoring method is rule-based and transparent, the system can provide recommendations without requiring a trained machine-learning classification model.

Multilingual conversational interaction is also important for improving accessibility. Users may express questions in English, Tamil, or Tanglish. The proposed chatbot therefore uses rule-based intent detection and pattern matching to identify user requirements and map them to the appropriate system module. Tamil and English/Tanglish patterns are considered during input processing.

The existing literature and technological developments indicate that combining retrieval, generative AI, deterministic algorithms, and structured domain knowledge can provide a more controlled conversational system. Based on these concepts, the proposed TN Postal AI Chatbot adopts a hybrid architecture in which each technology performs a specific function. RAG handles knowledge retrieval, Gemini handles natural-language generation, deterministic algorithms perform financial calculations, weighted scoring performs recommendation, and structured database operations support comparison and post-office search.

Therefore, the proposed project addresses the limitations of conventional information-search methods by providing a domain-specific, conversational, multilingual, and function-oriented assistant for India Post savings schemes. The system is designed to simplify information access while maintaining a separation between generative AI and operations that require predictable and structured processing.

A. Summary of Literature Survey

Table I summarizes the key works reviewed and their relevance to the proposed system.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_26

TABLE I

SUMMARY OF LITERATURE SURVEY

Author/Technology	Contribution	Relevance to Proposed Project
Lewis et al. (2020)	Introduced Retrieval-Augmented Generation for knowledge-intensive NLP tasks	Provides the foundation for grounding chatbot responses using postal-scheme data
Vaswani et al. (2017)	Introduced Transformer architecture and attention mechanism	Forms the technological foundation of modern NLP and generative AI models
Google Gemini	Provides generative AI and natural-language understanding	Used to generate conversational responses from retrieved scheme information
India Post	Provides official information about postal savings schemes	Acts as the intended source category for scheme information in the knowledge base
Rule-based NLP	Uses keywords and patterns for intent identification	Routes user queries to chatbot, calculator, comparison, recommendation, and other modules
Weighted Suitability Scoring	Matches user requirements with scheme characteristics	Used for transparent scheme recommendation

The literature survey supports the use of a hybrid architecture for the proposed system. The project combines established concepts from RAG, Transformer-based generative AI, rule-based NLP, deterministic computation, and recommendation techniques to create a domain-focused postal information assistant.

III. PROPOSED SYSTEM

Building on the concepts identified in the literature survey, the proposed system is designed as a hybrid web application in which each functional requirement is handled by the technology best suited to it: retrieval and generation for conversational queries, deterministic logic for financial computation, and structured data operations for comparison and location search.

A. System Architecture

The system follows a layered architecture consisting of a presentation layer (Flask templates and JavaScript), an application layer (Flask routes and business logic), a knowledge layer (a central module acting as the single source of truth for scheme data), and a generative layer (Google Gemini 2.5 Flash, accessed only after relevant data has been retrieved from the knowledge layer). This separation ensures that factual scheme details such as interest rates and investment limits always originate from verified data rather than from the language model itself.

B. Functional Modules

- Conversational Chatbot — answers natural-language queries about scheme eligibility, rates, tenure, tax benefits, and documentation using the RAG-based prompt-injection approach.
- Financial Calculators — deterministic computation of maturity value, interest earned, and returns for each scheme.
- Scheme Comparison — side-by-side comparison of two or more schemes across key parameters.
- Recommendation Engine — weighted suitability scoring based on age, investment amount, investment mode, financial goal, duration, and tax-saving preference.
- Nearest Post Office Locator — offline, pincode-based search using India Post's official pincode dataset.
- Chat History and Feedback — persistence of conversations and thumbs up/down feedback for continuous improvement.

C. Technology Stack

- Backend: Python, Flask
- Generative AI: Google Gemini 2.5 Flash
- Data Processing: pandas, offline India Post pincode dataset (data.gov.in)
- Frontend: HTML, CSS, JavaScript

ISBN NO: 978-81-180146-5-2

- Security: CSRF protection, rate limiting, hardened session cookies, environment-variable secret-key management

IV. IMPLEMENTATION

A. Knowledge Base Design

All scheme data — eligibility, interest rate, investment limits, tenure, tax treatment, and required documents for the ten supported schemes — is centralized in a single module, which acts as the authoritative source for every other module in the system. Centralizing this data avoids inconsistency between the chatbot, calculator, comparison, and recommendation modules.

B. Retrieval-Augmented Response Generation

When a user submits a query, the chatbot module identifies the relevant scheme(s) and retrieves the corresponding verified data from the knowledge base. This data is injected into the prompt sent to Gemini 2.5 Flash, which is then responsible only for composing a natural, user-friendly response around the supplied facts. This prevents the model from fabricating figures such as interest rates or investment limits. During development, a word-boundary regex bug in the scheme-matching logic was identified and corrected to improve retrieval accuracy.

C. Recommendation via Weighted Suitability Scoring

The recommendation module scores each of the ten schemes against user-provided parameters — age, investment amount, investment mode, financial goal, investment duration, and tax-saving preference — using a weighted, rule-based scoring function. Because the scoring logic is transparent and deterministic, the system can justify its recommendation without requiring a trained machine-learning model or labeled training data.

D. Nearest Post Office Locator

Locating the nearest post office initially relied on external APIs such as OpenStreetMap and postalpincode.in; these approaches were found to be unreliable during development due to network and connectivity issues. The feature was therefore re-implemented using an offline CSV-based approach, loading India Post's official pincode dataset from data.gov.in with pandas, and supplementing it with a direct Google Maps link for turn-by-turn navigation.

E. Security

To support safe deployment, the application implements CSRF protection on all state-changing requests, rate limiting to reduce abuse, hardened session-cookie configuration, and management of the Gemini API key and other secrets through environment variables rather than hard-coded values.

F. User Interface

The dashboard was redesigned with gradient styling, a typing indicator, thumbs up/down feedback buttons, and quick-action chips for common queries. A client-side function renders structured chatbot responses as formatted HTML rather than raw markdown, and custom 404/500 error pages were added for a more consistent user experience.

G. Data Persistence

Database models for chat messages and feedback were introduced to persist conversation history and user feedback, enabling future analysis of common queries and response quality.

V. RESULTS AND DISCUSSION

The implemented system successfully integrates conversational query answering, deterministic financial calculation, scheme comparison, personalized recommendation, and offline post-office search within a single Flask application. By restricting the generative model to response composition and sourcing all factual figures from the verified knowledge base, the system reduces the risk of the model presenting incorrect interest rates or eligibility conditions to the user. The rule-based intent detection allows the assistant to correctly route English, Tamil, and Tanglish queries to the appropriate module. The offline pincode-based locator removed the system's dependency on third-party location APIs that had proven unreliable, improving the robustness of the post-office search feature.

A formal quantitative evaluation — such as response accuracy against a labeled query set, latency measurements, or a user study — was outside the scope of the current development phase and is identified as future work. Qualitative testing during development confirmed that the system produces consistent, correctly grounded answers for the ten supported

ISBN NO: 978-81-180146-5-2

schemes and reliably distinguishes between the functional modules based on user intent.

VI. CONCLUSION AND FUTURE WORK

This paper presented the TN Postal Saving Scheme AI Assistant, a hybrid conversational system that combines Retrieval-Augmented Generation, a centralized verified knowledge base, deterministic financial computation, and weighted suitability-based recommendation to simplify access to India Post savings-scheme information for users in Tamil Nadu. By separating factual data from language generation, the system aims to provide reliable, domain-specific answers in a conversational format, supported by comparison, recommendation, and post-office-locator functions.

Future work includes extending Tamil-language scheme aliases across all supported schemes, conducting a formal accuracy and usability evaluation, expanding the knowledge base to cover additional India Post services, and exploring integration with live official data sources as they become available.

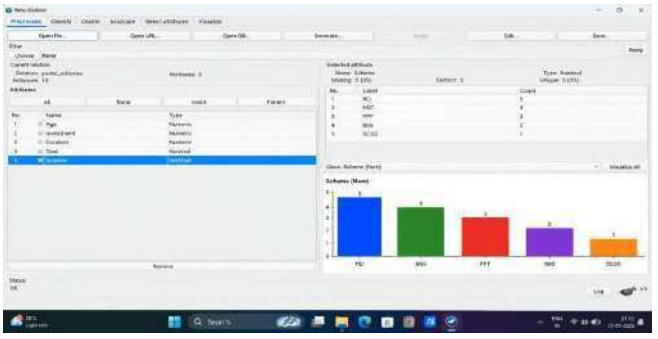


Fig. 1. Sample scheme dataset (postal_schemes.arff) and its attribute preprocessing view in WEKA Explorer, used for exploratory analysis of the rule-based recommendation logic.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] 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 (NeurIPS), 2017.

[3] Google, "Gemini Models Documentation," Google AI for Developers. [Online]. Available: https://ai.google.dev

[4] India Post, Department of Posts, Government of India, "Savings Schemes," Official Website. [Online]. Available: https://www.indiapost.gov.in

[5] Government of India, "All India Pincode Directory," Open Government Data (OGD) Platform India. [Online]. Available: https://data.gov.in

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

SMART CIVIC ISSUE REPORTER: A LOCATION-BASED SYSTEM FOR DUPLICATE COMPLAINT GROUPING AND DYNAMIC PRIORITY ASSIGNMENT

Mahaalakshmi.MJ[1], Jai Shree Maha.K[2], Rehna.N[3]
[1] Student, B.Sc Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India
mahaamj.26@gmail.com
[2] Student, B.Sc Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India
jaishreemaha07krishna@gmail.com
[3] Assistant Professor, Department of Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India
nrehna@shasuncollege.edu.in

ABSTRACT

Efficient reporting and management of civic problems is an important requirement for improving urban public services. Citizens frequently encounter issues such as potholes, garbage accumulation, broken streetlights, water leakage and drainage problems. A major limitation of conventional complaint systems is that multiple citizens may report the same physical problem as separate complaints, increasing administrative workload and reducing the visibility of issues affecting many people. This paper presents a Smart Civic Issue Reporter, a web-based system that groups similar civic complaints using issue type, geographical proximity and description similarity. The system creates a master issue for related reports and dynamically assigns priority based on complaint count, severity level, area impact and unresolved duration. Severity and area impact are treated as independent factors: severity represents the seriousness of the problem, whereas area impact represents the importance of the affected location. The proposed system also provides complaint status tracking and an authority-side workflow for verification and resolution. The results of the prototype demonstrate that grouping repeated complaints can reduce duplicate processing and that multi-factor priority assignment provides a more meaningful ordering of civic issues than a simple first-come-first-served approach.

Keywords: Civic Complaint Management, Smart City, Duplicate Complaint Detection, Geographical Proximity, Priority Assignment, Public Service Management.

1. INTRODUCTION

Rapid urbanization has increased the need for efficient mechanisms through which citizens can report problems affecting public infrastructure and services. Common civic issues include damaged roads, potholes, garbage accumulation, non-functional streetlights, water leakage and drainage failures. Although digital complaint portals have improved communication between citizens and authorities, many systems still process each submitted complaint independently. As a result, several reports describing the same physical problem may appear as separate records.

This creates two important difficulties. First, officials may spend time reviewing and processing duplicate complaints. Second, the number of citizens affected by a particular problem may not be clearly visible. For example, twenty reports about the same pothole may appear as twenty separate entries rather than as one significant public issue supported by twenty citizens.

The proposed Smart Civic Issue Reporter addresses this limitation through location-based complaint grouping and dynamic priority assignment. The system checks whether a newly submitted complaint is similar to an existing issue by considering the reported issue type, geographical proximity and description similarity. Related reports are associated with a common master issue. The priority of the master issue is then calculated from multiple factors, allowing authorities to focus on problems with greater public impact and urgency.

2. LITERATURE REVIEW

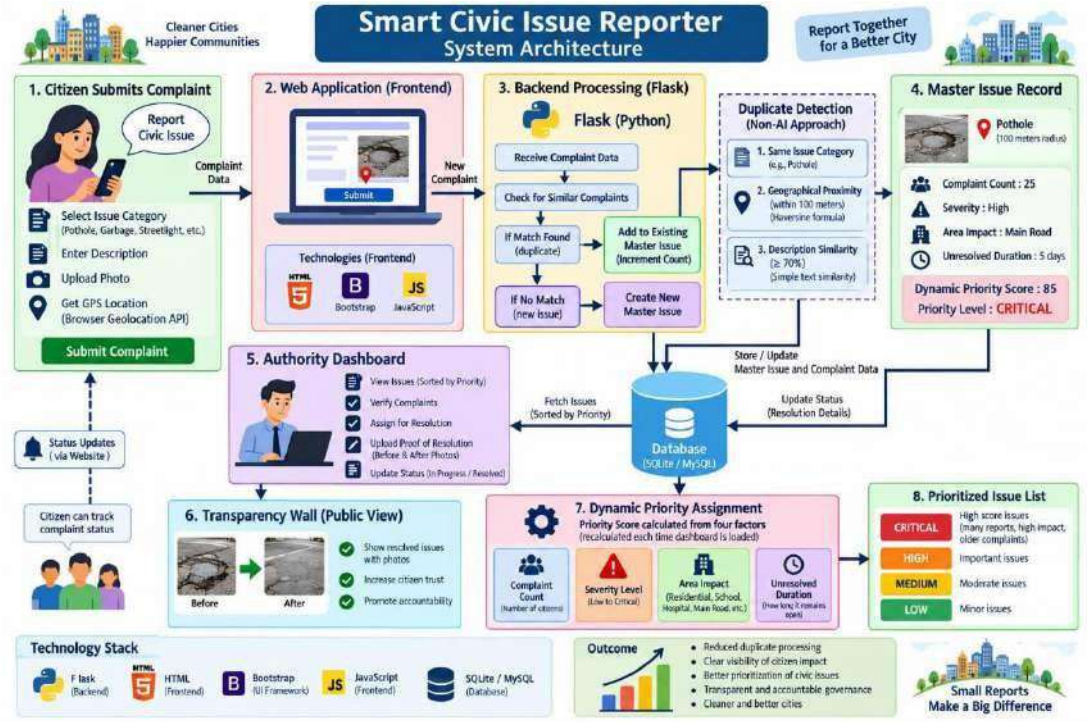
Franklin G. David and J. Savitha [1] developed a geo-based civic complaint management system that supports location-based reporting, complaint assignment, resolution, and transparency.

ISBN NO: 978-81-180146-5-2

Sonia Bajaj et al. [2] proposed an urban complaint management platform that uses geographical proximity to identify potentially duplicate complaints.
Rajkumar et al. [3] presented a civic grievance management framework focusing on structured complaint processing and urgency assessment.
Archana et al. [4] developed an intelligent complaint management system with duplicate detection, GPS support, priority management, and dashboards.
Recent IEEE studies [5] and [6] used advanced automated methods for complaint prioritization, duplicate detection, and validation. Research Gaps Most existing systems focus on complaint registration, geo-tagging, tracking, AI-based analysis, or workflow management.
However, there is limited focus on a simple and transparent non-AI-based approach for grouping duplicate civic complaints. The proposed system addresses this gap by identifying complaints based on issue type, geographical proximity, and description similarity. Similar complaints are grouped under a master issue, rather than being processed as separate complaints. The system dynamically assigns priority based on complaint count, severity, area impact, and unresolved duration. Thus, the proposed approach provides an efficient and understandable method for reducing duplicate complaints and prioritizing important civic issues.

4. METHODOLOGY

The proposed methodology consists of complaint submission, complaint comparison, duplicate grouping, priority calculation and resolution tracking. The system is implemented as a web-based civic issue reporting platform with separate citizen and authority interfaces. The architecture diagram of smart civic reporter is as shown in Fig.1.



3.1 Complaint Submission

A citizen submits an issue by providing the issue type, a textual description, an optional supporting image, geographical location, severity level and area-impact value. The issue types in the prototype include Pothole, Garbage Accumulation, Broken Streetlight, Water Leakage and Drainage Problem. The severity options are Low, Medium, High and Critical. Area Impact uses Low, High and Very High.

3.2 Duplicate Complaint Grouping

When a new complaint is submitted, the system searches for existing master issues with the same or related issue type in the nearby geographical area. The complaint is compared using three primary conditions: issue-type similarity, location proximity and description similarity. A simplified decision rule is:

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

Duplicate Match = Issue Type Match A Location Distance ≤ D A Description Similarity ≥ T

Here, D represents the configured geographical distance threshold and T represents the text-similarity threshold. If the conditions are satisfied, the new complaint is linked to an existing master issue and the complaint count is increased. Otherwise, a new master issue is created. The master issue therefore represents the actual civic problem, while individual complaints represent reports submitted by citizens.

3.3 Dynamic Priority Assignment

Priority is calculated dynamically rather than being permanently assigned at the time of the first report. The proposed priority score considers four factors:

Priority Score = Complaint Count Score + Severity Score + Area Impact Score + Waiting Time Score

Complaint count indicates how many citizens have reported or supported the same issue. Severity represents the seriousness or safety risk of the problem. Area Impact represents the importance and public exposure of the affected location. Waiting time allows long-pending unresolved issues to receive increasing attention. The final score is mapped to priority categories such as Low, Medium, High and Critical.

3.4 Severity and Area Impact

Severity and Area Impact are independent inputs. Severity evaluates the condition of the issue itself. For example, a minor pothole can be classified as Low severity, whereas major road damage that creates an immediate safety risk can be classified as Critical. Area Impact evaluates the significance of the location. A small residential street may have Low impact, whereas a main road, hospital area or school entrance may have High or Very High impact.

When a duplicate complaint contains a higher severity or area-impact value than the existing master issue, the stored value is escalated to the higher level. This prevents newly reported risk information from being ignored and avoids automatic downgrading of an issue.

4. RESULTS AND DISCUSSION

The prototype demonstrates that complaint grouping can provide a clearer representation of civic problems than a conventional list of independent reports. When several complaints refer to the same issue in the same area, they are consolidated under a master issue. For example, multiple reports of a pothole within the configured geographical threshold can be represented as one issue with an increasing complaint count.

The priority mechanism also produces a more meaningful ordering of issues. A problem reported by only one citizen may initially receive a lower complaint-count score, whereas the same problem reported by many citizens receives an increased score. However, complaint frequency alone does not determine priority. A critical safety problem can receive a higher priority because of its severity, while a problem located near a major public facility can receive additional importance through the Area Impact factor.

The separation of Severity and Area Impact is a significant feature of the methodology. Consider two potholes with similar physical severity: one may occur on a low-traffic residential street, while the other may occur at a major junction or near a hospital entrance. The second issue can be assigned greater operational importance because it affects a larger or more vulnerable population. Similarly, an issue that remains unresolved for a long period can receive additional priority through the waiting-time factor.

The authority workflow provides additional transparency. The prototype uses status stages such as Reported, Verified, Assigned, In Progress, Resolved and Closed. Citizens can follow the progress of their reports, while authorities can view issues according to their calculated priority. This issue-centric approach can reduce repetitive administrative effort and help decision-makers identify problems that require timely attention.

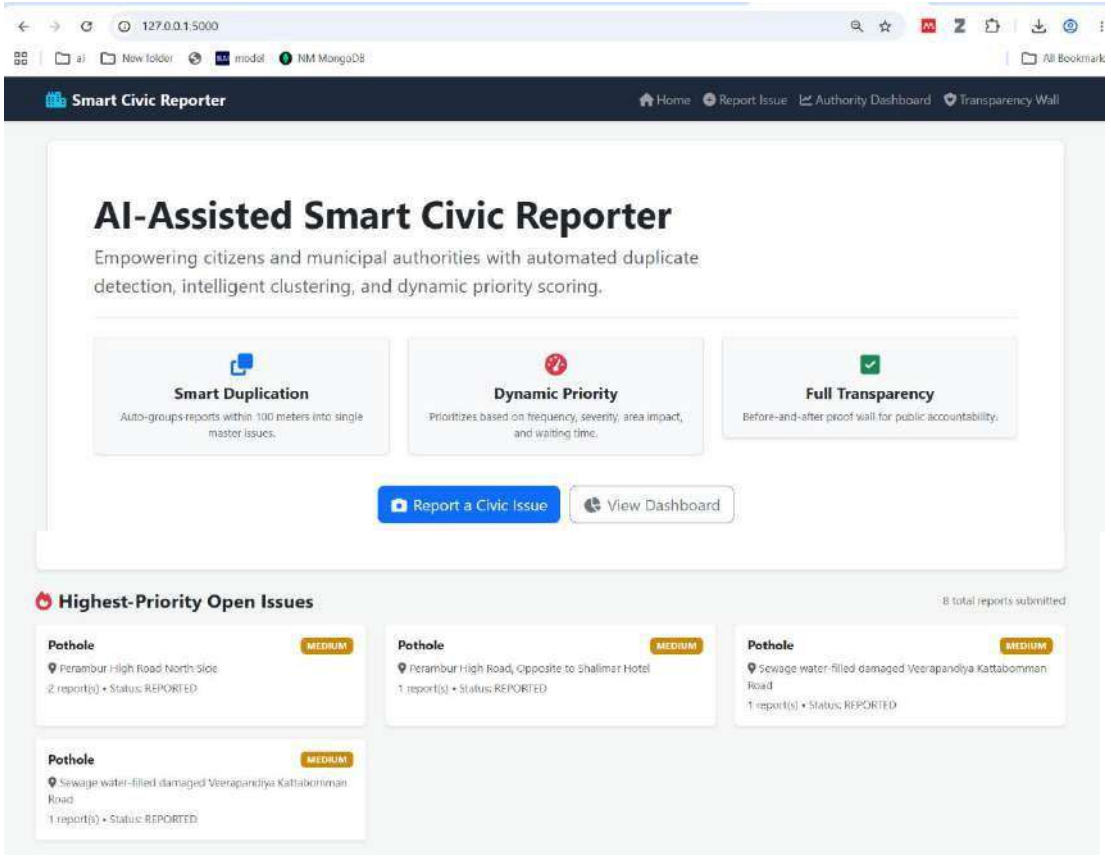
The system nevertheless has limitations. Duplicate detection depends on the quality of geographical coordinates and the similarity of user descriptions. A fixed geographical threshold may not be equally suitable for every category of civic issue. For example, the appropriate distance for grouping reports of a pothole may differ from that for a large drainage problem. Future work can therefore investigate adaptive distance thresholds, improved multilingual description matching and automated routing based on administrative boundaries.

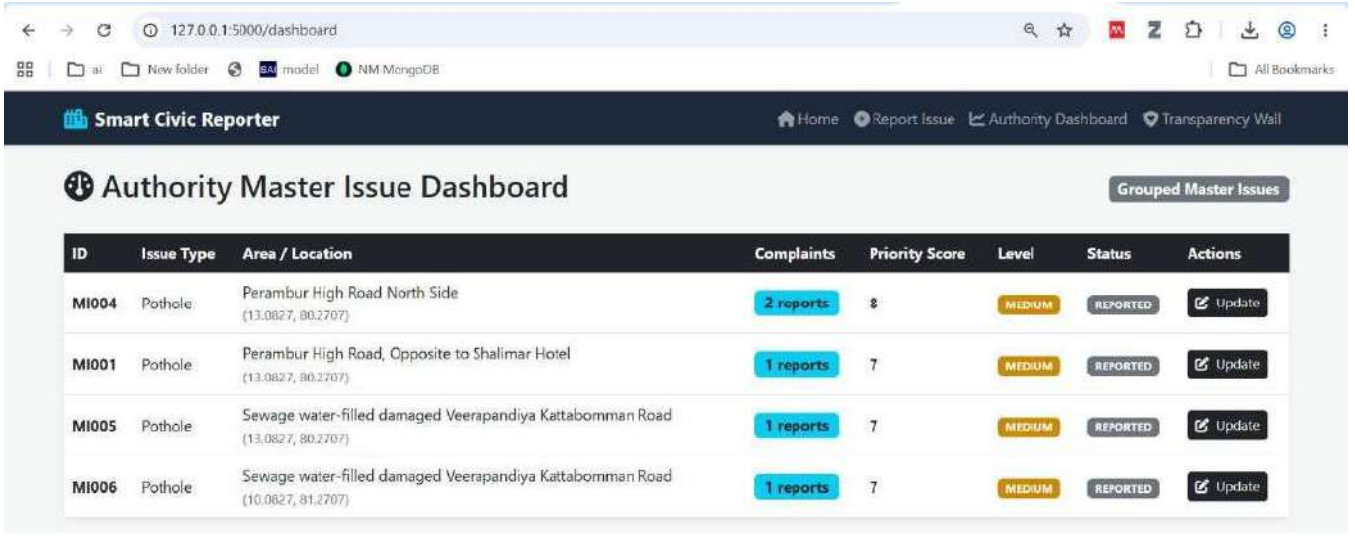
The Fig. 2 shows the Screenshots of the Smart Civic Issue Reporter System. The screenshots illustrate the working interface and operational workflow of the proposed Smart Civic Issue Reporter system. The system

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

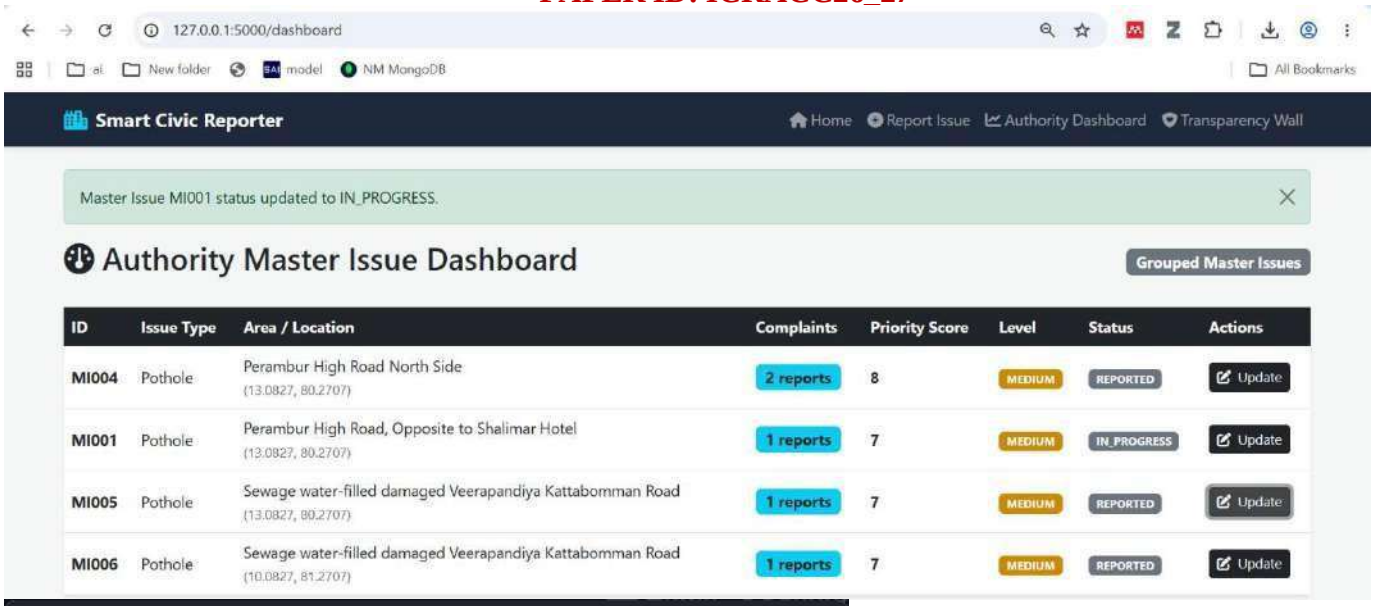
provides a structured platform for citizens to report civic issues such as potholes, road damage, garbage accumulation, and other infrastructure-related problems. The screenshots demonstrate the process of submitting an issue, capturing the relevant location and supporting information, and tracking the reported complaint through the system.

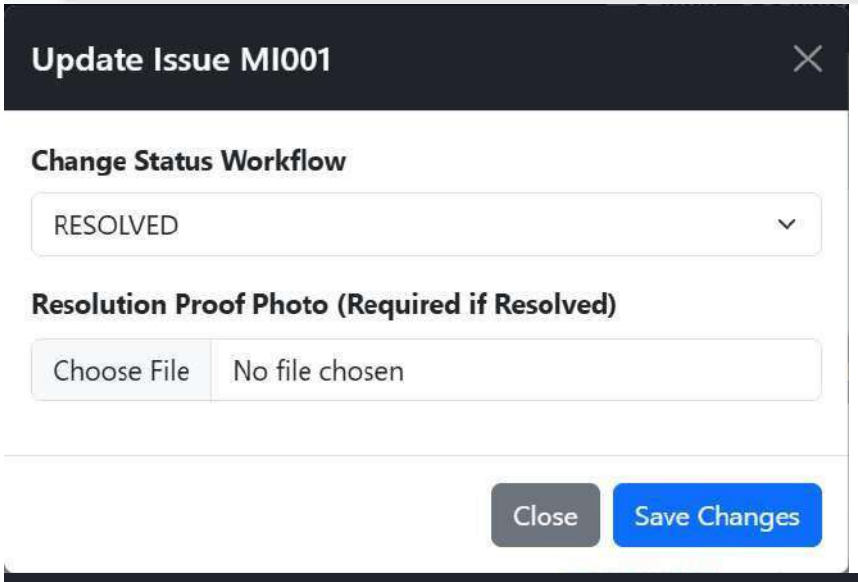


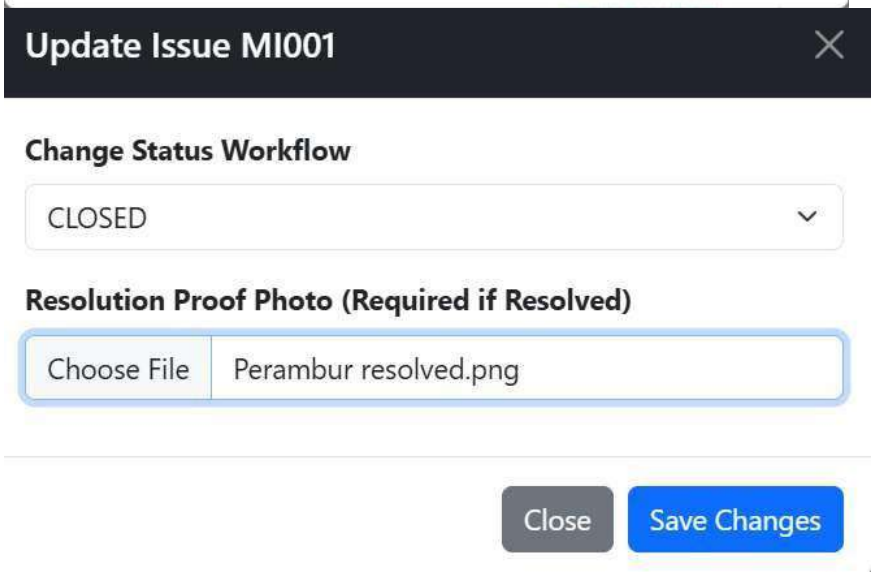


ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

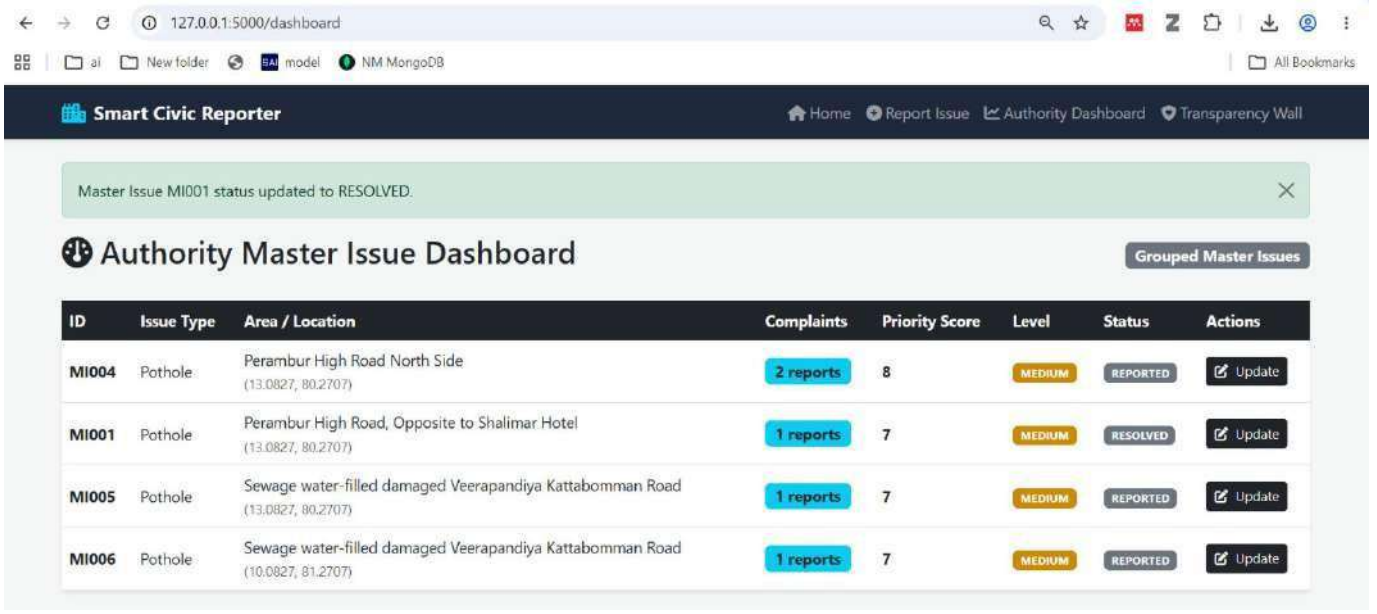


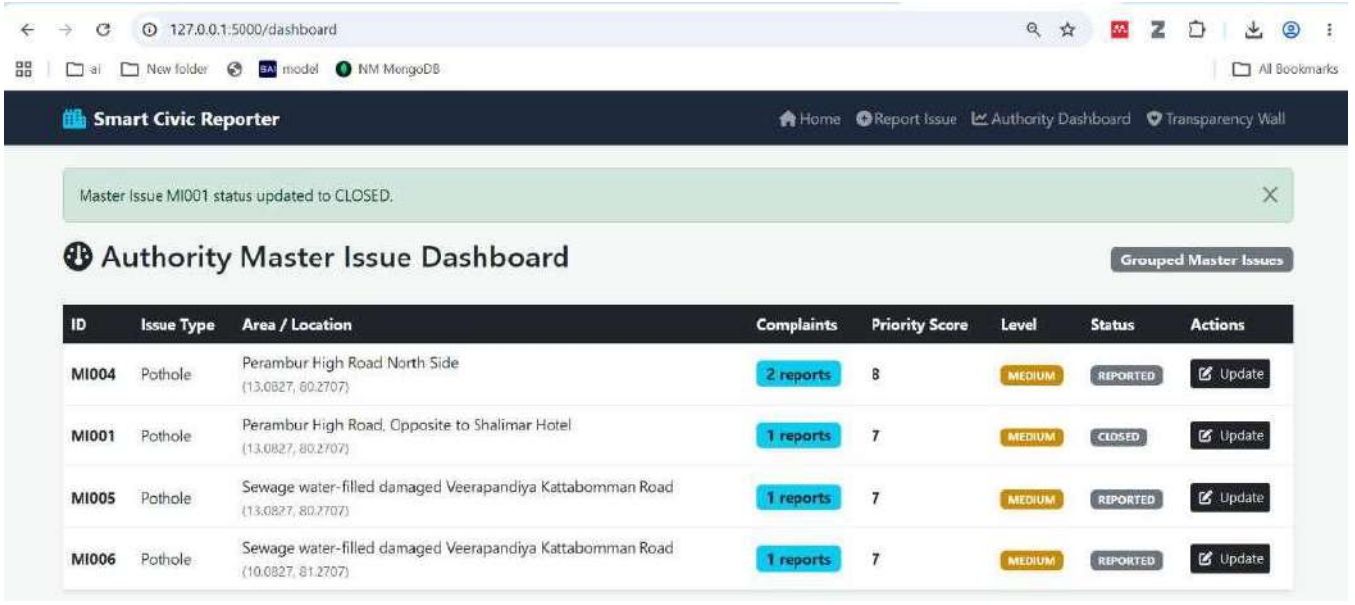




ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27





ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

Submit Civic Complaint

Issue Type

Pothole

Severity Level

Medium

Small pothole, could damage tires.

Area Impact

High

Main road / school / hospital / bus stand area.

Area Name / Landmark

Perambur High Road North Side

Latitude

13.0827

Longitude

80.2707

Capture Current GPS Location

Description

The concrete topped storm water drain along Perambur High Road North Side is ridden with potholes, damaged and uneven manhole chamber doors. In the absense of a footpath on either sides of the road, pedestrians are forced to walk on the road.

Upload Issue Image

Choose File | Perambur1.png

Submit & Run Duplicate Check

Before and After Perambur High Road North Side



ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

After



Fig. 2 Screenshots of the Smart Civic Issue Reporter System

5. CONCLUSION

This paper presented a Smart Civic Issue Reporter designed to improve the management and prioritization of public civic complaints. The proposed system groups related reports into a master issue using issue type, geographical proximity and description similarity. The priority of each master issue is dynamically calculated from complaint count, severity, area impact and unresolved duration.

The proposed approach differs from a conventional complaint registration system because it focuses on identifying the underlying civic problem rather than processing every repeated report as an unrelated case. Grouping duplicate complaints reduces repetitive processing while preserving the number of citizens who have reported the problem. The separation of severity and area impact provides a more context-sensitive method for determining urgency.

The prototype provides a practical framework for web-based civic issue management and can be extended in future with advanced geospatial analysis, multilingual complaint processing, automated administrative routing and predictive identification of civic problem hotspots. The proposed methodology can therefore contribute to more organized, transparent and responsive management of local civic issues.

REFERENCES

[1] F. G. David and J. Savitha, “FixMyCity – A Geo-Based Smart Civic Complaint Management System,” *International Journal of Scientific Research and Engineering Development*, vol. 9, no. 1, pp. 1228–1231, 2026.

[2] “A Web-Based Public Grievance Redressal System with Image and Geolocation Support,” *2025 International Conference on Smart & Sustainable Technology (INCSST)*, IEEE, 2025, doi: 10.1109/INCSST64791.2025.11210297.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_27

[3] N. Archana, Y. Yogeshwaran, U. D. R., M. K. C., and S. R., "Smart Civic AI: Intelligent Complaint Management System," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 15, no. 5, 2026, doi: 10.17148/IJARCCE.2026.155196.

[4] "A Machine Learning Integrated Framework for Automated Civic Complaint Validation and Prioritization in Smart City Governance," *2026 4th International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, IEEE, 2026, doi: 10.1109/ICSSAS68835.2026.11559493.

[5] "AI-Driven Complaint Management System: An Automated Framework for Prioritizing, Tracking, and Resolving Public Service Complaints Efficiently," *2026 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, 2026, doi: 10.1109/ESCI68015.2026.11493166.

[6] B. S. Menon, D. R. S., and G. Mudbhari, "Civi-Snap: A Deep Learning Powered Geotagged Civic Issue Detection and Monitoring Framework," *International Journal of Creative and Open Research in Engineering and Management*, vol. 2, no. 4, pp. 1–9, 2026, doi: 10.55041/ijcope.v2i4.439.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_28

A Review of Deep Learning Techniques for Malware Detection

Ms.V.Vidhya

Assistant Professor, Department of Computer Applications,

Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai.

Email: vidhyavenkatraman@shasuncollege.edu.in

ABSTRACT

Malware keeps changing shape faster than most defenders would like. Every year brings variants that are harder to catch, and signature-based antivirus tools, reliable as they are against threats we already know about, tend to fall behind once malware is packed, obfuscated, or simply new. Deep learning (DL) has become one of the more promising answers to this problem, largely because it can learn what malicious code "looks like" directly from raw data-bytes, disassembled instructions, API-call traces, or even images generated from a binary-rather than depending on features a human has to design by hand. This paper looks at the DL architectures that dominate current malware-detection research: convolutional neural networks (CNNs) for image-style representations, recurrent and LSTM networks for sequential data, autoencoders and generative adversarial networks (GANs) for anomaly detection and data scarcity, and the newer graph neural network (GNN) and transformer-based models. It also examines the datasets most of this work is built on, and a set of problems that do not disappear just because accuracy numbers look good: adversarial evasion, models going stale as malware evolves, imbalanced data, and detectors that are difficult to explain to the people who have to act on their output.

Keywords: Deep learning, Malware detection, CNN, LSTM, GAN, Graph neural network, Transformer, Cybersecurity, Adversarial attack, Explainable AI

1. INTRODUCTION

Malware-viruses, worms, trojans, ransomware, spyware, adware, and the rest-is not a niche problem anymore; it reaches personal computers, phones, IoT devices, and entire corporate networks. Authors of malicious code have gotten good at disguising it, using packing, polymorphism, obfuscation, and encryption so the same malicious payload can look different every time it is seen [10]. For a long time, antivirus software dealt with this by matching files against a database of known signatures. That works well for threats that have already been catalogued, but it has nothing to say about a threat nobody has seen before, which is exactly the gap zero-day and polymorphic malware exploit. Machine learning offered a partial fix: instead of matching exact signatures, a model could learn a decision boundary from features extracted from a file's static or dynamic behaviour. The catch is that those features still have to be chosen by a person, and if the chosen features do not capture whatever is new about a piece of malware, the model does not either.

Deep learning removes a good part of that manual step. Given enough data, a network can learn its own representation from raw bytes, opcode sequences, API calls, execution traces, permissions, application graphs, or images built from a binary [4], [10]. The rest of this paper follows that idea through: how different malware representations pair with different DL architectures, what datasets the field actually trains and tests on, and where the real obstacles still sit.

2. MALWARE ANALYSIS AND DETECTION APPROACHES

Detection work usually falls into one of two camps, static or dynamic, though most practical systems now blend both. Static analysis looks at a sample without running it-its file structure, headers, byte sequences, disassembled instructions-and is fast and safe, but obfuscation and packing can hide exactly the patterns a static detector is looking for. Dynamic analysis goes the other way: it runs the sample in a sandbox and watches what it actually does-system calls,network traffic, registry edits, files created or deleted. That

ISBN NO: 978-81-180146-5-2

reveals real behaviour, but it costs more time and compute, and malware that can tell it is being watched will often just sit quietly until the sandbox lets its guard down [10]. Hybrid detection tries to get the best of both, and deep learning fits naturally here because separate network components can each specialise in one modality and their outputs can be combined at the end.

3. DEEP LEARNING TECHNIQUES FOR MALWARE DETECTION

The architectures used in this space mostly differ by what kind of input they are built to handle. CNNs are good at picking out patterns in images, or in byte sequences treated like images; RNNs and LSTMs are built for anything with an order to it, such as a sequence of API calls; autoencoders are useful when the goal is to flag "this looks unusual" rather than to classify outright; GANs help when there simply is not enough labelled malware to go around; GNNs keep relationships between components intact instead of flattening everything into a grid or a line; and transformers bring long-range attention into the mix.

3.1 CNN-BASED IMAGE MALWARE CLASSIFICATION

The idea of turning a binary into an image is older than it sounds. Nataraj et al. showed back in 2011 that converting malware binaries into greyscale images and applying standard image-classification techniques worked surprisingly well, and their MalImg dataset-9,458 samples across 25 families-is still used today [1]. Once CNNs became the default tool for image classification more broadly, applying them here made sense, since they can pick up the same kind of visual regularities the original approach relied on, only automatically instead of by hand. Vasan et al.'s IMCEC, an ensemble of CNN architectures, pushed accuracy and robustness further than any single model managed on its own [5]. The main weakness of this whole family is that a heavily packed or compressed binary simply does not produce an image with much structure left in it, so the technique is only as good as the visual signal that survives compression.

3.2 RNN/LSTM MODELS FOR OPCODES AND API CALLS

Sequence-based approaches make sense for much the same reason CNNs make sense for images: a binary's instruction stream, or a running process's API-call trace, has an order that matters, and RNN/LSTM networks are built specifically to model that kind of dependency. Vinayakumar et al. tested this across several feature representations and reported consistently strong results [4]. A number of studies have gone further and combined CNNs with LSTMs-using the convolutional layers to pull out local features and the recurrent layers to track how they unfold over time-and some of these hybrid models have reported accuracy close to 99% on IoT and PE malware benchmarks [6]. That said, sequence models are not immune to being fooled: an attacker can often change the observed sequence just enough to slip past the classifier while leaving the malware's actual behaviour untouched [8].

3.3 AUTOENCODERS AND GAN-BASED AUGMENTATION

Autoencoders take a different approach: rather than learning to tell malware from benign software directly, they learn to reconstruct normal input, and anything that reconstructs badly gets flagged as suspicious. That makes them useful when labelled malware is scarce but "normal" behaviour is well understood, and the same architecture doubles as a denoising step that can make a detector more robust to input that has been deliberately tampered with. GANs solve a related but different problem-there simply are not enough malware samples of a given family to train on in many cases. Wasserstein GAN-based augmentation has shown promising results for boosting Android malware detection on the Drebin dataset [7], [11], though it comes with a caveat: synthetic samples need to be checked carefully, since a classifier can end up learning quirks of the generator rather than anything about real malware.

3.4 GRAPH NEURAL NETWORKS AND TRANSFORMER-BASED MODELS

Not every useful signal fits neatly into an image or a sequence. Android applications, for instance, are naturally described as a graph of components,permissions, and API relationships, and graph neural networks are built to work with exactly that kind of structure rather than flattening it away. Yumlembam et al. combined a GNN with an adversarial defence mechanism for IoT-based Android malware detection, with encouraging results [7]. Transformers, meanwhile, bring attention-based long-range modelling to opcode sequences, API

ISBN NO: 978-81-180146-5-2

traces, or malware images-SHERLOCK, for example, applies vision-transformer ideas to a corpus of roughly 1.2 million malware images [12]. Both directions are promising, but they are also hungrier for data and compute than the CNN or LSTM baselines that came before them, which is worth keeping in mind before assuming bigger is automatically better.

4. DATASETS AND COMPARATIVE CONSIDERATIONS

How well any of these models perform depends heavily on what they were trained and tested on, and the field still leans on a fairly small number of public datasets, summarised in the table below.

Dataset	Platform / Scale	Key limitation
EMBER [3]	Windows PE; 1M+ files	Static features only; temporal drift
BIG 2015	Windows; large benchmark	Benchmark age
Drebin [2]	Android; 5,560 malware / ~123k benign	2010-2012; duplication
MalImg [1]	Desktop; 9,458 / 25 families	Small, image-specific
MalNet	Multi-family; 1M+ images	Representation bias
BODMAS	Windows; modern collection	Needs temporal evaluation

No single architecture wins across the board. A model tuned on an image-style benchmark may not carry over to contemporary packed PE files, and a sequence model trained on static traces can easily miss behaviour that only shows up at runtime. In practice, which architecture makes sense depends on what representation is actually available, what the threat model looks like, how much compute is on hand, and how the evaluation is set up-not just which one posts the highest accuracy in a paper.

5. CHALLENGES AND FUTURE DIRECTIONS

Adversarial robustness is probably the issue that gets the most attention, and for good reason: small, functionality-preserving changes-padding a file, injecting a few benign-looking bytes-can be enough to flip a detector's decision without changing what the malware actually does [9]. Adversarial training and smoothing-based defences help, but only partially. A related, slower-moving problem is concept drift: malware keeps changing, so a detector trained on an older dataset like Drebin or BIG 2015 will gradually lose ground against whatever is circulating now, which argues for continual learning and datasets that get refreshed rather than reused for a decade. Class imbalance is a more mundane but still real problem-benign samples usually outnumber malicious ones by a wide margin, and without resampling, weighted losses, or GAN-based augmentation, models tend to drift toward simply predicting the majority class. There is also a cost problem that rarely makes it into a paper's abstract: dynamic analysis, large transformer models, and graph construction are all expensive, and a detector that works beautifully in a lab but is too slow or too heavy for production is not much use. And finally, explainability remains a genuine barrier to adoption-an analyst who receives a "malicious" verdict with no reasoning behind it has to either trust the model blindly or ignore it, neither of

ISBN NO: 978-81-180146-5-2

which is a great option, which is part of why explainable AI and more realistic, time-separated benchmarking keep coming up as priorities for where this field needs to go next.

6. CONCLUSION

Taken together, the literature makes a reasonably convincing case that deep learning has moved malware detection forward-CNNs, RNN/LSTMs, autoencoders, GANs, GNNs, and transformers each bring something useful depending on how the malware is represented, and most of them outperform classical machine-learning baselines on the benchmarks they are tested against. But benchmark accuracy is not the whole story, and this review has tried to be honest about that: adversarial evasion, ageing datasets, imbalanced classes, computational overhead, and weak interpretability are still open problems, not footnotes. Where the field seems to be heading is toward detectors that combine adversarially robust and continually updated training with multimodal analysis and outputs an analyst can actually make sense of, tested on data that reflects what malware looks like now rather than what it looked like when the dataset was collected.

4 REFERENCES

[1] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, "Malware images: visualization and automatic classification," Proc. 8th Int. Symp. Visualization for Cyber Security (VizSec), Pittsburgh, PA, USA, 2011.

[2] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of Android malware in your pocket," Proc. NDSS, San Diego, CA, USA, 2014.

[3] H. S. Anderson and P. Roth, "EMBER: An open dataset for training static PE malware machine learning models," arXiv:1804.04637, 2018.

[4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," IEEE Access, vol. 7, pp. 46717-46738, 2019.

[5] D. Vasan, M. Alazab, S. Wassan, B. Safaei, and Q. Zheng, "Image-based malware classification using ensemble of CNN architectures (IMCEC)," Computers & Security, vol. 92, p. 101748, 2020.

[6] "Detection of malware by deep learning as CNN-LSTM machine learning techniques in real time," Symmetry, vol. 14, no. 11, p. 2308, 2022.

[7] R. Yumlembam, B. Issac, S. M. Jacob, and L. Yang, "IoT-based Android malware detection using graph neural network with adversarial defense," IEEE Internet of Things Journal, 2022.

[8] R. Chaganti et al., "CNN-LSTM and transfer learning models for malware classification based on opcodes and API calls," J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 1, p. 101925, 2024.

[9] "A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks," Eng. Appl. Artif. Intell., 2022.

[10] P. Maniriho, A. N. Mahmood, and M. J. M. Chowdhury, "A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms," ACM Computing Surveys, vol. 56, no. 6, 2024.

[11] "Improving Android malware detection through data augmentation using Wasserstein generative adversarial networks," arXiv:2403.00890, 2024.

[12] "Re-evaluating Android malware detection: Tabular features, vision models, and ensembles (SHERLOCK)," 2026.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_29

Truface: An Open-Source, Real-Time Active-Passive Liveness Architecture Against Live Streaming Deepfake Injections

Durga G, Manasa P, Dr. S.T.Deepa (Associate Professor)

Department of Computer Science, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, India

[1]*durgagr20@gmail.com,* [2]*manasap2357@gmail.com,* [3]stdeepa@shasuncollege.edu.in

***Abstract*—**Remote video know-your-customer (V-KYC) pipelines are increasingly compromised by generative face-swap tools such as DeepFaceLive combined with software loopback virtual cameras, enabling large-scale identity fraud and mule-account creation. Existing liveness defenses examine physiological, behavioral, or photometric cues in isolation and are validated primarily on static offline corpora rather than against an adaptive, real-time streaming attacker. We present Truface, a six-layer active-passive liveness architecture that fuses OS-level driver enumeration, jitter-resilient cubic-spline temporal resampling, multi-region remote photoplethysmography (rPPG) coherence, stroboscopic photometric challenge-response, asymmetric eyelid velocity kinematics, and cardiovascular second-harmonic spectral verification under a cascaded fast-fail decision engine. We formalize a two-tier adversarial red-team protocol and provide an algorithmic complexity analysis demonstrating real-time feasibility on commodity CPU hardware, without relying on unverified empirical benchmark claims.

Keywords*—liveness detection, presentation attack detection, remote photoplethysmography, deepfake, video KYC, biometric security, challenge-response, WebRTC signal jitter*

I. Introduction

The digitization of financial onboarding has made remote Video Know-Your-Customer (V-KYC) the default gateway through which banks, payment processors, and fintech platforms verify the identity of new customers. Regulators in many jurisdictions now permit account opening entirely through a supervised video call in which an operator or an automated pipeline compares a live camera feed against a government identity document and, in many deployments, applies an automated liveness check to confirm that the subject in front of the camera is a physically present human being rather than a photograph, video replay, or synthetic rendering. This convenience, however, has created a lucrative attack surface. Publicly available real-time face-swap software, most notably DeepFaceLive, can now render a synthetic identity onto a live human face at interactive frame rates. When the synthesized output is routed through a software loopback virtual camera such as OBS Virtual Camera or the Linux v4l2loopback kernel module, the resulting stream is indistinguishable, at the operating-system level, from a genuine physical webcam. An attacker in possession of a victim's leaked identity photographs can therefore present a convincing, animated impersonation to a V-KYC operator in real time, enabling fraudulent account origination and the creation of mule accounts used for laundering the proceeds of financial crime.

Despite a decade of research on face presentation attack detection (PAD), three structural gaps leave production V-KYC systems exposed. First, prior work tends to examine a single modality in isolation: passive physiological indicators such as remote photoplethysmography (rPPG), active challenge-response probing, or ocular kinematics such as blink dynamics are each treated as a self-sufficient defense, when in practice each modality has a narrow and well-documented failure mode that a motivated attacker can learn to exploit once the specific test becomes known. Second, the empirical culture of the field evaluates detectors against static, offline corpora such as FaceForensics++ or Celeb-DF, in which the forged content was generated once and never adapts to the detector's decisions. This offline framing does not model the real V-KYC threat, which is an aware human operator manipulating a real-time generative pipeline while observing the on-screen result and adjusting their behavior interactively. Third, most rPPG and photometric liveness literature implicitly assumes a clean, uniformly sampled video signal, whereas production V-KYC calls are transported over WebRTC, which is subject to packet jitter, adaptive frame-rate throttling, and lossy compression. D'Amelio et al. showed that such degradation attenuates the high-frequency content of the cardiac waveform, but that the underlying physiological signal remains recoverable if the non-uniform temporal sampling introduced by network jitter is explicitly corrected rather than assumed away.

This paper addresses these three gaps with Truface, an open-source, six-layer active-passive liveness architecture designed specifically for the streaming V-KYC setting. Our contributions are threefold. First, we propose a layered fusion architecture that combines a near-zero-cost operating-system driver enumeration stage with five complementary

ISBN NO: 978-81-180146-5-2

computer-vision and signal-processing layers, so that the failure of any single modality against a targeted bypass does not by itself defeat the system. Second, we introduce a timestamp-aware cubic-spline resampling stage that reconstructs an equidistant virtual time grid from monotonic hardware presentation timestamps prior to any spectral analysis, directly addressing the WebRTC jitter problem identified as an open issue in prior compression-robustness studies. Third, we formalize a two-tier adversarial red-team methodology that distinguishes naive, off-the-shelf injection attacks from an adaptive human attacker who is aware of the specific defensive layers and consciously attempts to defeat each one, and we pair this threat model with an algorithmic complexity analysis rather than fabricated empirical accuracy figures, in keeping with the reproducibility standards this domain requires.

II. Related Work

Hemodynamic rPPG-based presentation attack detection. Remote photoplethysmography recovers the cardiac pulse waveform from subtle periodic variations in skin color caused by blood volume changes, a technique first popularized for non-contact vital-sign monitoring and subsequently repurposed for liveness verification. De Haan and Jeanne introduced the CHROM method [1], which combines chrominance-based color channels to suppress motion artifacts and specular reflection, establishing the signal-processing foundation on which most subsequent rPPG liveness detectors are built.
Ciftci et al. introduced FakeCatcher [2], which extracts biological signal residuals across multiple facial regions and shows that synthetically generated video lacks the spatially coherent, physiologically plausible pulse propagation exhibited by genuine skin. These works establish that pulse-signal coherence across independent facial regions is a strong indicator of authenticity, but each evaluates detection against pre-rendered, non-adaptive forged video rather than a live, human-in-the-loop generative pipeline.

Neuromuscular eyelid kinematics. Ocular dynamics offer a complementary behavioral biometric that is difficult for a generative model to reproduce with correct temporal asymmetry. Soukupová and Čech introduced the Eye Aspect Ratio (EAR) [3], a lightweight geometric descriptor computed from eye landmarks that enables real-time blink detection without requiring eye-region image intensity analysis. Li et al. proposed In Ictu Oculi [4], observing that early deepfake generators failed to reproduce physiologically normal blink frequency and duration, though the authors note that this deficiency is a training-data artifact that later generative models have since partially corrected, motivating detectors that instead examine the finer-grained kinematic asymmetry between eyelid closing and reopening velocity rather than blink presence alone.

Active photometric challenge-response. A second family of defenses actively perturbs the illumination environment and measures the physical optical response of the subject. Tang et al. proposed Face Flashing [5], which emits randomized colored light patterns from the verifier's display and validates the timing and chromaticity of the reflected response, establishing screen-emitted illumination as a practical liveness channel requiring no specialized hardware. Guo et al. extended active probing to the real-time deepfake setting specifically [6], authenticating video calls by displaying a distinct on-screen pattern and analyzing its corneal reflection, an approach the authors validate through large-scale simulation without requiring specialized imaging or lighting hardware. Mittal et al. proposed GOTCHA [7], a live challenge-response protocol requiring the subject to perform randomized physical actions, reinforcing the broader principle that an active, unpredictable challenge raises the cost of a real-time bypass far above what a static or replayed spoof can satisfy.

Streaming compression and temporal jitter effects. The majority of the above techniques were validated on video captured under controlled, uniformly sampled conditions. D'Amelio et al. specifically examined how lossy video compression, of the kind imposed by real-time communication codecs, degrades the high-frequency spectral content that rPPG analysis depends upon [8], while showing that the cardiac signal remains recoverable provided that the non-uniform frame timing introduced by adaptive streaming is corrected prior to spectral analysis. This finding is central to the design of Truface's temporal resampling layer, described in Section III.

III. Truface System Architecture

Truface is organized as a cascaded pipeline of six complementary layers, each targeting a distinct physical or physiological property that a real-time synthetic injection struggles to reproduce simultaneously. Figure 1 illustrates the overall data flow from raw frame capture through the final cascaded decision engine.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_29

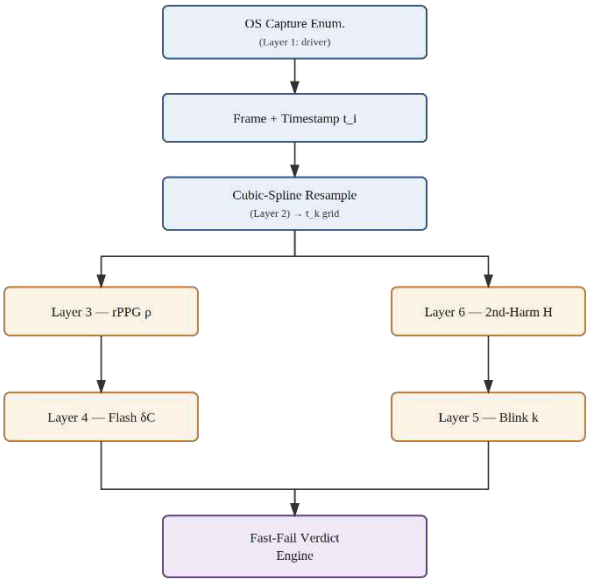


Fig. 1. Truface pipeline and decision cascade.

A. Layer 1 — Pre-Ingestion OS Device Driver Enumeration

Before any pixel data is decoded, Truface queries the operating system's native capture-device enumeration interface — the DirectShow filter graph on Windows or the Video4Linux2 (V4L2) device tree on Linux — to inspect the string identifiers, vendor descriptors, and backing driver names of the active capture source. Known virtual-camera signatures, including but not limited to "OBS Virtual Camera" and "v4l2loopback", are flagged at this stage. Because this check operates purely on OS-level metadata rather than decoded pixel buffers, its computational cost is negligible relative to any downstream computer-vision layer, and it is therefore evaluated first in the cascade to enable an immediate fast-fail against the least sophisticated class of attacker.

B. Layer 2 — Timestamp-Aware Cubic-Spline Resampling

Each captured frame is stamped with its monotonic hardware presentation timestamp t_i as it arrives from the capture driver, prior to any application-level buffering. Because WebRTC transport introduces variable frame delivery intervals under network jitter, the raw signal $x(t_i)$ extracted from any facial region of interest is non-uniformly sampled in time. Directly applying a Fourier-domain analysis to a non-uniformly sampled signal introduces spectral leakage that corrupts the frequency estimates on which Layers 3 and 6 depend. Truface therefore reconstructs a natural cubic spline through the irregularly timed observations and resamples the interpolant onto an equidistant virtual time grid $t_k = k / f_s$, with f_s fixed at 30 Hz:

$$S(t) = a_k + b_k(t - t_k) + c_k(t - t_k)^2 + d_k(t - t_k)^3$$

where the per-segment coefficients a_k, b_k, c_k, d_k are obtained by enforcing continuity of value, first derivative, and second derivative at each knot, subject to natural boundary conditions. This regularization step is what allows the downstream spectral layers to remain valid under realistic streaming conditions rather than the idealized fixed-frame-rate assumption made by most prior rPPG liveness literature.

C. Layer 3 — Multi-Region Spatio-Temporal rPPG Coherence

Using MediaPipe Face Mesh's 468-point landmark model [9], Truface isolates two anatomically distinct regions of interest: the upper central forehead, which is the region most heavily synthesized and re-rendered by face-swap models, and the anterior cervical neck, which lies outside the typical bounding box replaced by a face-swap and therefore retains the attacker's unaltered skin. Each region's raw color trace is passed through a zero-phase second-order Butterworth bandpass filter with cutoffs at 0.75 Hz and 2.5 Hz, corresponding to a physiologically plausible heart-rate range of 45 to 150 beats per minute. Over a 90-frame sliding window, the Pearson cross-correlation coefficient between the two filtered signals is computed as:

$$\rho = \Sigma(S_fh - \bar{S}_fh)(S_neck - \bar{S}_neck) / (\sigma_fh\ \sigma_neck)$$

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_29

A genuine subject exhibits a single cardiovascular source propagating pulse waves through contiguous tissue, producing high inter-region coherence ($\rho \geq 0.70$). A live face-swap, by contrast, synthesizes the forehead region independently of the underlying neck footage, which severs the physiological coupling between the two regions and produces a measurable boundary phase rupture ($\rho < 0.35$), a signature this layer is specifically designed to detect.

D. Layer 4 — Active Stroboscopic Photometric Challenge-Response

To defend against the case where an attacker's underlying video source is itself pre-recorded or algorithmically re-rendered, Truface issues an unpredictable active probe: the display emits a brief, randomly timed and randomly colored border flash (for example, a soft cyan increment of $\Delta R = 0$, $\Delta G = 180$, $\Delta B = 180$) lasting $\Delta t = 200$ ms. The system then measures the induced change in diffuse reflection over the forehead region of interest:

$$\Delta C_refl = (1 / |ROI|) \Sigma [I_flash(x,y) - I_base(x,y)]$$

A physically present subject sitting in front of the emitting screen exhibits a measurable, correctly timed reflectance increase because their face is illuminated by the monitor. A virtual-camera loopback injecting a pre-recorded or independently rendered stream has, by construction, no physical optical coupling to the verifier's screen and therefore shows no correlated reflectance response to the randomized flash.

E. Layer 5 — Asymmetric Eyelid Velocity Kinematics

The Eye Aspect Ratio is computed per frame from six eye landmarks as:

$$EAR = (\|p_2 - p_6\| + \|p_3 - p_5\|) / (2 \|p_1 - p_4\|)$$

Truface tracks the first temporal derivative $v(t) = d(EAR)/dt$ across each blink event. Human blinks are neuromuscularly asymmetric: the passive eyelid closure driven by relaxation of the levator palpebrae superioris is fast (approximately 80 ms), while the active reopening driven by muscular contraction is comparatively slow (approximately 180 ms). This asymmetry is captured as a velocity ratio:

$$\kappa = | min(v) / max(v) |, \text{ requiring } \kappa \geq 1.8 \text{ for a genuine blink}$$

Synthetic morphs generated by face-swap models frequently reproduce blinks with symmetric closing and reopening velocity ($\kappa \approx 1.0$), or omit natural blink cycles entirely over extended windows, either of which is flagged by this layer.

F. Layer 6 — Cardiovascular Second-Harmonic Spectral Verification

A genuine arterial pulse waveform is not a pure sinusoid: the dicrotic notch produced by transient closure of the aortic valve introduces energy at the second harmonic of the fundamental heart rate. Truface computes the power spectral density of the resampled pulse signal and evaluates the second-harmonic energy ratio:

$$H = P(2f_0) / P(f_0), \text{ requiring } H \in [0.15, 0.45]$$

A synthetic brightness-flicker injection intended to spoof a plausible heart rate typically produces an artificially pure single-frequency signal, yielding $H < 0.05$, which is readily distinguished from the harmonically rich spectrum of a genuine dicrotic pulse waveform.

G. Cascaded Decision Engine

The six layers are combined under a conjunctive fast-fail cascade rather than a learned weighted fusion, so that the failure of any single layer is sufficient to reject the session and so that the least computationally expensive checks are always evaluated first. Layer 2 is a signal-conditioning stage that contributes no independent accept or reject decision; it instead validates the temporal basis on which the spectral terms of Layers 3 and 6 are computed, and therefore does not appear as a separate term in the verdict function, which is evaluated over the five decision-bearing layers:

$$Verdict = D_Driver \wedge (\rho \geq \tau_\rho) \wedge D_Flash \wedge D_Blink \wedge (\tau_H,lo \leq H \leq \tau_H,hi)$$

This cascade structure is deliberately conservative: because each layer targets a physically distinct coupling channel between the subject and the capture environment, an attacker must simultaneously defeat OS-level fingerprinting, active photometric coupling, ocular neuromuscular timing, and two independent physiological spectral properties within a single live session, which raises the joint cost of a successful bypass substantially above that of defeating any individual modality. The layer-wise thresholds stated in Sections III-C, III-E, and III-F ($\rho \geq 0.70$, $\kappa \geq 1.8$, and $H \in [0.15, 0.45]$) are initial operating values derived from the physiological ranges cited for each phenomenon rather than from a fitted validation set, and are stated to make the architecture concretely reproducible. They are not presented as tuned or

ISBN NO: 978-81-180146-5-2

validated decision boundaries; Section V-C specifies the threshold sweep and per-layer Detection Error Trade-off analysis by which a deploying institution should establish its own operating point.

IV. Adversarial Threat Model & Red-Team Methodology

We formalize the adversary space along two tiers of increasing sophistication, consistent with best practice in adversarial security evaluation, which requires that a defense be assessed against an attacker who is aware of the defense rather than only against naive baseline attacks.

A. Tier 1 — Naive Injection

The Tier 1 adversary represents the current dominant fraud pattern observed against production V-KYC systems: a default, unmodified DeepFaceLive face-swap session, or a pre-recorded video of the victim, routed into the verification pipeline through an unmodified software loopback such as OBS Virtual Camera. This attacker makes no attempt to specifically target any individual defensive layer and is expected to be defeated primarily by Layer 1 driver enumeration and the gross absence of any physiological or photometric response.

B. Tier 2 — Adaptive Red-Team

The Tier 2 adversary is assumed to have full knowledge of the Truface architecture, consistent with Kerckhoffs's principle, and to actively adapt their physical setup to defeat each layer in turn. We define four canonical bypass attempts that a red-team evaluation protocol must exercise: (1) Neck Occlusion, in which the attacker wears a turtleneck or scarf specifically to remove the unaltered reference region that Layer 3 depends upon; (2) Secondary Specular Reflection, in which the attacker angles a mirror or reflective card toward an external monitor in an attempt to artificially reproduce the reflectance signature expected by Layer 4 without genuine facial coupling; (3) Voluntary Motor Calibration, in which the attacker consciously trains themselves to perform slow, deliberate blinks in an attempt to mimic the asymmetric velocity ratio κ targeted by Layer 5; and (4) Driver Masking, in which the attacker employs a custom kernel or user-space splitter specifically designed to rewrite or conceal the virtual-camera device string inspected by Layer 1.

C. Testbed Configuration

A standardized red-team testbed for this architecture should hold camera resolution, illumination, and network path constant across trials while varying only the attack tier, so that any change in the layer-wise decision outputs can be attributed to the adversarial maneuver rather than confounding environmental factors. We recommend logging each layer's intermediate scalar output (ρ, ΔC_refl, κ, H, and the Boolean driver flag) per session rather than only the final cascaded verdict, so that a failure can be attributed to the specific layer an adaptive attacker successfully targeted, which is essential for iterative hardening of the architecture.

V. Implementation Feasibility & Evaluation Framework

Consistent with sound scientific practice, we do not report fabricated accuracy, FAR, FRR, or EER figures in the absence of an institutionally approved, IRB-covered human-subjects data collection. Instead, this section establishes that the architecture is computationally feasible for real-time deployment on commodity hardware through formal algorithmic complexity analysis, and it specifies the standardized evaluation protocol that a future empirical study should follow.

A. Time-Complexity Analysis

Let n denote the number of frames in the sliding analysis window ($n = 90$ in our design) and let m denote the number of pixels within a facial region of interest. Layer 1's driver enumeration is a constant-time string comparison against a known-signature list, O(1) with respect to frame content. Layer 2's cubic-spline resampling requires solving a tridiagonal linear system for the spline coefficients, which the Thomas algorithm solves in O(n) time and O(n) space, making it strictly cheaper than the O(n log n) cost of the subsequent spectral analysis it feeds. Layer 3's cross-correlation is computed over two length-n signals after O(m) per-frame spatial averaging over the region of interest, yielding O(nm) total cost per window, dominated in practice by m since n is fixed at 90. Layer 4's reflectance differencing is a single O(m) pass over the region of interest per probed frame. Layer 5's EAR computation is O(1) per frame given a fixed six-point landmark set, and its derivative and extrema tracking over the window is O(n). Layer 6's second-harmonic verification requires a Fast Fourier Transform of the length-n resampled signal, costing O(n log n). Because n is fixed at 90 by design, $n \log n \approx 90 \times 6.5 \approx 585$ floating-point operations per window per region, which is trivial relative to the per-pixel spatial cost of the vision layers. The end-to-end per-frame cost of the pipeline is therefore dominated by the O(m) landmark-driven region-of-interest extraction shared across Layers 3, 4, and 5, which MediaPipe's optimized landmark

ISBN NO: 978-81-180146-5-2

model performs well within real-time budgets on a general-purpose Intel Core i5-class CPU without requiring dedicated GPU acceleration, since the 468-point mesh regression itself is the only stage with data-dependent, non-trivial per-frame cost and is bounded by the face detector's published real-time throughput on such hardware. Figure 2 illustrates the resulting relative per-layer operation counts.

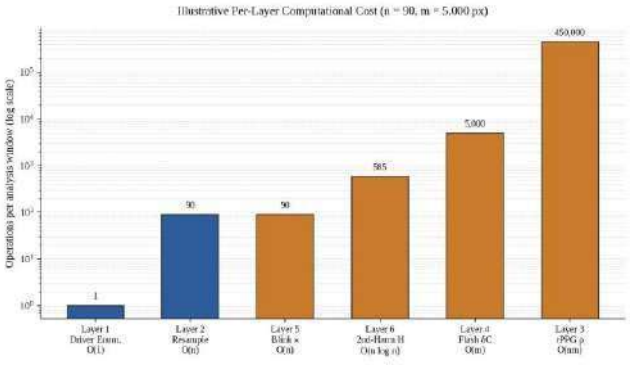


Fig. 2. Illustrative per-layer computational cost per analysis window (n = 90, representative ROI size m = 5,000 px), derived from the complexity classes in Section V-A.

B. Memory Footprint

The dominant memory consumers are the n = 90-frame circular buffers maintained per region of interest for Layers 3, 5, and 6, each storing a scalar time series rather than raw pixel data once the per-frame spatial reduction has been applied, giving a footprint on the order of a few kilobytes per active session, independent of input video resolution. The cubic-spline coefficient arrays computed by Layer 2 are similarly O(n) in size. This confirms that Truface's steady-state memory footprint is dictated by the number of concurrently verified V-KYC sessions rather than by any single session's computational complexity, which is favorable for horizontal scaling in a call-center deployment.

C. Standardized Evaluation Protocol

We specify the evaluation protocol that a subsequent empirical study, conducted under appropriate informed-consent and data-protection governance, should follow to responsibly report False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER). Genuine trials should be collected from consenting participants across varied Fitzpatrick skin-tone categories, ambient lighting conditions, and consumer webcam hardware. Impostor trials should be stratified explicitly by the Tier 1 and Tier 2 categories defined in Section IV, since collapsing both tiers into a single reported FAR would understate the residual risk posed by an adaptive attacker. The decision threshold τ_ρ for Layer 3 and τ_H for Layer 6 should be swept independently to produce a Detection Error Trade-off curve per layer prior to fixing the cascaded thresholds used in the combined Verdict function, so that the operating point can be tuned to the risk tolerance of the deploying institution rather than fixed a priori.

VI. Discussion, Ethical Considerations, & Operational Boundaries

Any physiological liveness system must be evaluated for equitable performance across the full Fitzpatrick skin-tone scale, since rPPG signal amplitude is known in the broader literature to vary with melanin concentration due to differential light absorption, and a system that degrades disproportionately for darker skin tones would impose an unacceptable disparate-impact risk on affected customers seeking financial access. We therefore recommend that any deployment of Layers 3 and 6 be accompanied by per-skin-tone threshold calibration rather than a single global threshold, and that institutions monitor FRR stratified by Fitzpatrick category as an ongoing fairness audit rather than a one-time validation step.

Severe backlighting and extreme head pose present genuine operational boundaries for this architecture. Backlit conditions reduce the signal-to-noise ratio of the diffuse reflectance measured in Layer 4 and can also depress the amplitude of the rPPG trace exploited by Layers 3 and 6, and extreme head yaw or pitch beyond the reliable operating range of the underlying facial landmark model degrades the geometric accuracy of the EAR computation in Layer 5. A production deployment should therefore include an upstream image-quality gate that requests re-positioning or improved lighting before the liveness cascade is invoked, rather than allowing a degraded capture condition to silently lower the effective security margin of the system.

We also acknowledge, for completeness and future-work framing rather than as a present vulnerability claim, the theoretical threat posed by high-fidelity silicone or synthetic prosthetic masks incorporating embedded micro-actuation or thermally realistic materials. Such attacks operate in a fundamentally different threat class from software-based virtual-

ISBN NO: 978-81-180146-5-2

camera injection, since a physical mask worn by a live human body could, in principle, partially satisfy Layers 1, 4, and 5 by virtue of genuine physical presence, while Layers 3 and 6 remain relevant countermeasures because a synthetic mask material does not carry an authentic subdermal blood volume signal. We consider hardware-material presentation attacks to be outside the primary scope of this software-only architecture and flag them as an explicit direction for future multi-modal hardening, for example through near-infrared or thermal sensing where such sensors are available on the capture device.

Finally, we note that any system designed to detect fraudulent identity presentation carries a dual-use responsibility: the same architectural knowledge that hardens a defensive pipeline could, in principle, inform an attacker's choice of which layer to prioritize circumventing. We have deliberately published the threat model and bypass categories in Section IV precisely because responsible disclosure of a defense's failure modes, evaluated under an aware red-team methodology, produces a materially more robust system than obscurity, consistent with long-standing practice in the broader security research community.

VII. Conclusion

We presented Truface, a six-layer active-passive liveness architecture that unifies operating-system driver enumeration, jitter-corrected temporal resampling, multi-region rPPG coherence, active stroboscopic challenge-response, asymmetric eyelid kinematics, and cardiovascular second-harmonic verification under a single cascaded decision engine, specifically targeting the real-time streaming V-KYC threat posed by tools such as DeepFaceLive combined with software loopback virtual cameras. By formalizing a two-tier adversarial red-team methodology and grounding feasibility claims in algorithmic complexity analysis rather than unverified benchmark figures, this work provides a reproducible foundation for subsequent empirical validation under proper human-subjects governance. Future work includes extending the architecture with near-infrared and thermal sensing to address physical mask-based presentation attacks, exploring learned fusion of the six layer-wise scalar outputs as an alternative to the conjunctive cascade, and conducting the IRB-approved, Fitzpatrick-stratified data collection specified in Section V to produce the first verified FAR/FRR/EER figures for this architecture.

References

[1]G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013.

[2]U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[3] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *Proc. Comput. Vis. Winter Workshop (CVWW)*, 2016.

[4]Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018.

[5] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang, "Face Flashing: a secure liveness detection protocol based on light reflections," in *Proc. Netw. Distrib. Syst. Security Symp. (NDSS)*, 2018.

[6] H. Guo, X. Wang, and S. Lyu, "Detection of real-time DeepFakes in video conferencing with active probing and corneal reflection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[7] G. Mittal, C. Hegde, and N. Memon, "GOTCHA: Real-time video deepfake detection via challenge-response," in *Proc. IEEE Eur. Symp. Security Privacy (EuroS&P)*, 2024.

[8]A. D'Amelio, R. Lanzarotti, S. Patania, G. Grossi, V. Cuculo, A. Valota, and G. Boccignone, "On using rPPG signals for DeepFake detection: A cautionary note," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*, LNCS vol. 14234, 2023, pp. 235–246.

[9] Google MediaPipe Team, "MediaPipe Face Mesh: 3D face landmark estimation," Google AI Technical Documentation, 2020.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_30_

AI Driven Skincare Recommendation: A Hybrid CNN, Retrieval Augmented Generation, and Guardrail Based Architecture for Safer, Allergy Aware Personalization

[1]Dr. S. Prasanna
Assistant Professor,
Department of Computer science
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, India
s.prasana@shasuncollege.edu.in

[2]Yamini. N
Student,
Department of Computer science
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, India
yamininarayanan2007@gmail.com

[3]Najma Shanaaz S
Student,
Department of Computer Science
Shri Shankarlal Sundarbai Shasun Jain College for Women,Chennai, India
najmashanaaz@gmail.com

Abstract
Skincare has become a routine, high frequency purchasing decision for consumers worldwide, yet most buyers select products with limited knowledge of ingredient composition or compatibility with their own skin, leading to financial waste, allergic reactions, and environmental waste. Existing AI driven skincare systems address only part of this problem: CNNs reliably classify skin type from a photograph, but the recommendation step typically relies on collaborative filtering or static rule based matching, which fails for new users and cannot easily incorporate new products. None of these systems verify whether a recommendation is actually safe for a user's allergies or sensitivities. This paper proposes a hybrid architecture combining CNN based skin classification, retrieval augmented generation (RAG) for live ingredient data, a deterministic rule based safety filter, and an LLM for personalized explanation each assigned a distinct role rather than one model handling everything. This design reduces the risk of unsafe recommendations while retaining up to date, personalized guidance. Empirical implementation and testing are identified as future work.
Keywords — skincare recommendation, convolutional neural network, retrieval augmented generation, large language model, guardrail architecture, allergy safety, cold start problem

I. INTRODUCTION

Skincare has become a part of everyday life for people around the world. What was once a simple, occasional routine has grown into a daily habit for many cleansers, moisturizers, serums, and treatments chosen not just for hygiene, but for health, confidence, and self care. Social media, dermatology awareness, and a growing culture around "skin health" as personal wellness have all pushed skincare from an afterthought into a routine people plan their mornings and evenings around.

This shift in mindset has turned skincare into one of the fastest growing consumer industries globally. According to Fortune Business Insights, the global skincare market was valued at USD 122.11 billion in 2025 and is projected to grow to USD 129.11 billion in 2026, reaching USD 227.13 billion by 2034 [1]. Behind this growth, though, is a market flooded with choice thousands of brands, ingredients, and product types leaving the average consumer with more options than they know how to navigate.

Yet this growth has not been matched by a corresponding rise in consumer knowledge. Most buyers select products based on marketing claims, social media trends, or trial and error rather than a genuine understanding of ingredient composition and compatibility with their own skin. This gap between purchasing volume and consumer literacy creates three concrete problems. First, it leads to financial waste consumers routinely purchase products that do not suit their skin, discard them partway through use, and repeat the cycle, a pattern reflected in the sheer size of a market where trial and error purchasing is common enough to sustain continuous double digit growth in new product launches. Second, it creates safety risk consumers with known or unknown sensitivities may unknowingly purchase products containing allergens or irritants listed under unfamiliar scientific or alternate ingredient names, resulting in adverse skin reactions. Third, it contributes to environmental waste, as partially used or unsuitable products are frequently discarded rather than consumed, adding to packaging and product waste in an industry already scrutinized for sustainability concerns.

Artificial intelligence has demonstrated substantial value across numerous consumer facing domains from

ISBN NO: 978-81-180146-5-2

personalized media recommendations to automated customer support by learning patterns from data and tailoring outputs to individual users. Digital skin analysis tools are already an emerging trend within the skincare industry itself, suggesting the market is primed for AI driven solutions [5]. However, most existing AI applications in this space focus narrowly on classifying skin type from a photograph, without addressing the deeper problem of whether a recommended product is actually safe and suitable for that individual accounting for their specific allergies, ingredient sensitivities, and personal context.

This paper addresses that gap. We propose an AI driven skincare recommendation system that combines image based skin analysis, live ingredient retrieval, deterministic allergy safety filtering, and generative explanation to help consumers make safer, more informed skincare purchasing decisions reducing the financial, physical, and environmental costs of the current trial and error approach.

II. RELATED WORKS

A close look at existing AI driven skincare research reveals a fairly consistent design pattern. A convolutional neural network (CNN) is trained to read a facial photograph and output a skin type or condition label, and this label is then passed downstream to a recommendation engine. This classification step has been studied extensively and is, by most measures, well solved. Comparative work testing MobileNet-V2, EfficientNet-V2, Inception-V2, and ResNet-V1 has shown these architectures perform reliably on facial skin classification tasks [2]. A hybrid system combining K-Nearest Neighbors, a custom CNN, and EfficientNet B0 reached 80% validation accuracy [3], while a more recent ResNet50-based model enhanced with attention mechanisms achieved 97.25% average accuracy across five skin conditions acne, wrinkles, dryness, oiliness, and normal skin [4].

The weaker link in these systems lies not in classification, but in what happens after. Recommendation is typically handled through collaborative filtering, content based filtering, or fixed rule based matching and researchers working directly on these systems have acknowledged that such approaches fall short in practice [3]. The reason is structural rather than incidental. Collaborative filtering depends on a pool of prior users with similar characteristics, so a first time user with no interaction history offers the system nothing to work from the classic cold start problem, well documented throughout recommender systems research generally [5]. Rule based and content tagged systems face a related but distinct issue: they are static by design, and cannot easily absorb newly launched products or updated ingredient formulations without manual reconfiguration.

Interestingly, this exact cold start limitation has already been addressed outside the skincare domain through retrieval augmented generation (RAG), where a language model retrieves relevant external information at the moment of inference rather than depending on accumulated user history. Systems like RAMO, built for course recommendations [6], and KALM4Rec, applied to restaurant and hotel recommendations [7], show that RAG based approaches can produce meaningful, personalized output starting from a user's very first interaction no prior data required. The trade off, however, is that introducing a language model into the reasoning process reintroduces a different, well known risk: hallucination. Even RAG systems, despite being purpose built to ground model output in retrieved facts and reduce this exact problem, have been shown in the literature to still produce incorrect responses under certain conditions [8].

Taken together, these findings point to a clear division: the skin classification problem in this domain is largely solved, and the cold-start limitation in recommendation has already been solved in other domains through RAG but the reliability of RAG based reasoning has not been examined in a setting where an incorrect output could result in something as consequential as an allergic reaction. That gap, specifically the reliability of AI driven ingredient and allergy reasoning in skincare, remains unaddressed in the literature reviewed here, and is the focus of this paper.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

This paper proposes a hybrid architecture that combines five components, each assigned a distinct role, rather than relying on a single model to handle skin analysis, retrieval, safety checking, and explanation all at once. Fig. 1 illustrates the overall pipeline.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_30_

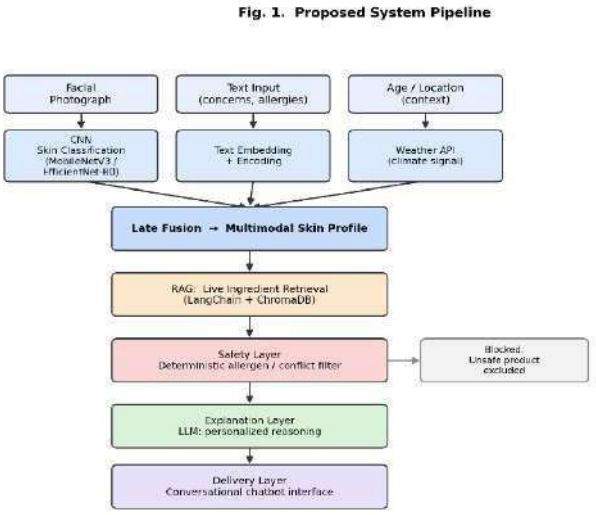


Fig. 1. Proposed system pipeline, from multimodal input through retrieval, safety filtering, explanation, and delivery.

A. Multimodal Input Layer

The system accepts a facial photograph alongside structured and unstructured user input including age, location, and self reported skin concerns or current condition rather than relying on a static questionnaire alone. The photograph is processed through a convolutional neural network (CNN) to classify basic skin attributes (oiliness, dryness, redness), replacing what would otherwise require an extensive manual questionnaire. This work uses transfer learning on a pretrained, lightweight architecture MobileNetV3 or EfficientNet-B0 fine tuned on a labeled facial skin dataset, selected based on precedent in related work where similar architectures achieved strong classification accuracy (80–97%) on comparable tasks [2], [3], [4]. Images are preprocessed through resizing (224×224 pixels), normalization, and standard augmentation (rotation, flipping, contrast adjustment) to improve generalization across lighting conditions and skin tones.

Structured inputs (age, location) and free text input are processed separately: structured fields are encoded numerically, free text is converted into a numerical representation via a text embedding model, and location is additionally used to query a weather API for local temperature/humidity as a contextual modifier (e.g., weighting oil control recommendations more heavily in humid climates). These streams are combined with the CNN's output using late fusion each pathway produces its output independently, and results are combined at a final decision stage, rather than a single joint model. This keeps each component simple, testable, and debuggable in isolation, which matters given the limited scale of a student research project.

B. Retrieval Layer (RAG)

Rather than relying on a static, pre loaded product database, the system uses retrieval augmented generation to search for and retrieve live ingredient information for candidate products at query time. This is implemented using the LangChain framework: candidate product names query the web or a public ingredient reference source, retrieved content is chunked and embedded, and stored temporarily in a vector database (ChromaDB) for similarity based retrieval. This design choice directly addresses a limitation identified in the related work fixed content tagged or rule based systems cannot easily incorporate newly released products without manual re-curation, while a retrieval based approach can reflect the current product market at query time.

C. Safety Layer

Before any product reaches the user, it passes through a deterministic, rule based filter checking its retrieved ingredient list against the user's stated allergies and known ingredient conflicts. This module is implemented directly in Python as lookup-table logic, a manually curated allergen and ingredient alias reference table (mapping ingredients to common alternate and scientific names), built from ingredient science references and cross-checked against each user's stated allergies rather than a trained model. This component is intentionally non generative, because removing AI based reasoning from the safety critical decision eliminates the specific

ISBN NO: 978-81-180146-5-2

hallucination risk associated with LLM based judgment, a risk documented in RAG literature even in systems explicitly designed to reduce it [8]. Any product containing a matching allergen by exact name or known alias is excluded before reaching the explanation layer. This design follows the general pattern of "guardrail" architectures established in LLM safety research [9], where a deterministic filter is placed ahead of a generative component to constrain its output space here applied, to the best of our knowledge for the first time, to skincare ingredient safety specifically.

D. Explanation Layer

Only products that pass the safety filter are handed to a large language model, accessed via API, whose role is strictly explanatory rather than evaluative: describing why a product suits the user's classified skin type, how its active ingredients function, and how it should be used. Because unsafe products have already been removed before this stage, errors or inconsistencies in the LLM's explanation carry meaningfully lower risk than if the LLM were also responsible for the safety decision itself.

E. Delivery Layer

The final recommendation and explanation are delivered through a conversational chatbot interface, allowing follow up questions rather than a single static output. A lightweight web interface (Streamlit) is used to demonstrate the pipeline end to end.

F. Design Rationale

Separating these five responsibilities is a deliberate departure from the single model pipelines discussed in the related work, which typically pass photo-derived skin classification directly into a filtering-based recommendation step without any dedicated retrieval or safety verification stage. This architecture is designed to produce recommendations that are both current and safer by construction, rather than relying on one model to balance freshness, personalization, and safety simultaneously.

IV. DISCUSSION

A. How This Approach Benefits the User

The architecture proposed in this paper is designed to directly address the core issues raised in the introduction: wasted money from trial and error purchasing, allergic reactions from unclear ingredient information, and the time and effort required to research products manually. By combining photo based skin classification with live ingredient retrieval, the system removes much of the manual research burden from the user instead of reading through long ingredient lists and cross referencing them against known sensitivities themselves, the user receives that analysis automatically.

The safety layer is the most consequential design choice from a user's perspective. By keeping the allergen checking decision deterministic and separate from the LLM's reasoning, the system is designed to reduce though not eliminate the chance that a user is recommended a product containing a known allergen. This matters because it is the failure mode with the most serious real world consequence in this domain, unlike a typical recommender system where a poor suggestion is simply ignored.

The retrieval based design also addresses the staleness problem seen in static, rule based systems discussed in the related work, allowing the system to remain usable as new products enter the market without requiring manual database updates. Finally, separating the safety decision from the explanation step means the system can still offer personalized, conversational guidance without that helpfulness coming at the cost of safety the explanation layer only ever sees products that have already been cleared.

V. LIMITATIONS

While the proposed architecture addresses several shortcomings identified in existing skincare recommendation systems, it carries its own set of limitations that warrant consideration. The safety layer is only as reliable as the reference data it is built on. A rule based filter cannot catch an allergen listed under a name absent from its alias table this is a structural limitation of the approach rather than an implementation detail, meaning the system's safety guarantees are bounded by the completeness and accuracy of the underlying allergen reference data.

The retrieval layer depends on the quality and consistency of external web sources. If a retrieved ingredient list is incomplete, outdated, or incorrect at the source, this error propagates through the rest of the pipeline

ISBN NO: 978-81-180146-5-2

regardless of how well the safety filter or explanation layer perform afterward. Separating safety checking from explanation also reduces but does not eliminate risk overall, and should not be interpreted as a guarantee against all possible adverse reactions.

VI. CONCLUSION

This paper has proposed a hybrid architecture for AI driven skincare recommendation that addresses two limitations identified in existing approaches: the reliance on collaborative or static rule based filtering, which struggles with new users and newly released products, and the absence of dedicated safety checking for ingredient conflicts and allergy risk in current skincare AI systems.

The proposed system combines CNN based skin classification, retrieval augmented generation for live ingredient data, a deterministic rule based safety filter, and an LLM based explanation layer each assigned a distinct, separable role rather than relying on a single model to handle classification, retrieval, safety, and personalization simultaneously. This design draws on established patterns from recommender systems and LLM safety research, applying them, to the best of our knowledge, to skincare ingredient safety for the first time.

By placing a deterministic filter ahead of generative reasoning, the architecture is designed to reduce the risk of unsafe recommendations reaching the user, while retaining the benefits of live, up to date product retrieval and personalized, conversational explanation.

REFERENCES

[1] Fortune Business Insights, "Skincare Market Size, Share, Trends | Growth Analysis [2034]," 2026. [Online]. Available: https://www.fortunebusinessinsights.com/skin-care-market-102544

[2] "Human skin type classification using image processing and deep learning approaches," ScienceDirect, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023083846

[3] V. [author], "A Recommendation System for Facial Skin Care using Machine Learning Models," GitHub / IEEE. [Online]. Available: https://github.com/vinit714/A-Recommendation-system-for-Facial-Skin-Care-using-Machine-Learning-Models

[4] "Enhanced CNN Architecture for Multi-label Skin Condition Classification for Personalized Skincare Recommendations," in Proc. Springer Nature Link, 2026.

[5] "Adaptive Candidate Retrieval with Dynamic Knowledge Graph Construction for Cold-Start Recommendation," arXiv:2505.20773, 2025.

[6] "RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations," arXiv:2407.04925, 2024.

[7] H.-D. Kieu and T.-S. Nguyen, "Keyword-driven Retrieval-Augmented Large Language Models for Cold-start User Recommendations (KALM4Rec)," arXiv:2405.19612, 2024.

[8] "ReRag: A New Architecture for Reducing the Hallucination by Retrieval-Augmented Generation," in Proc. IEEE Conf., 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10773428/

[9] "LLM-Based Guardrail Models," EmergentMind, 2025. [Online]. Available: https://www.emergentmind.com/topics/llm-based-guardrail-models

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_31

ARTIFICIAL INTELLIGENCE-DRIVEN APPROACHES TO COMBATING METHAMPHETAMINE TRAFFICKING IN THE DIGITAL ERA

Mohana Shree.S[1], Yegammai T[2]

[1]*Student, B.Sc. Computer Science, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India*
mohanashreesivasubiramaniyam@gmail.com

[2]*Assistant Professor, Department of Computer Science, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India*
yegammai@shasuncollege.edu.in

ABSTRACT

Methamphetamine trafficking has grown increasingly difficult to police as conventional transportation networks now intersect with encrypted digital communication, cryptocurrency transactions, dark-web marketplaces, unmanned aerial systems, and Internet of Things (IoT) infrastructure. This expansion has produced enormous volumes of heterogeneous data — digital, financial, geographical, and sensor-based — that traditional investigative methods struggle to process manually. This paper proposes a conceptual, human-supervised Artificial Intelligence (AI) framework that integrates Natural Language Processing (NLP), network analysis, blockchain transaction analysis, IoT-based monitoring, and aerial surveillance to support (not replace) law-enforcement investigation of methamphetamine trafficking. The framework draws on the applied technology overview presented in Kalla and Narasimhan's work on AI-based counter-trafficking approaches, and is situated within the wider body of research on AI-assisted crime detection, including EU-funded initiatives such as ARIEN and CEASEFIRE, and machine-learning-based dark-web and social-media monitoring systems. The paper positions AI strictly as a decision-support layer — surfacing patterns and anomalies for human investigators to verify — and closes with a discussion of the false-positive risk, data-quality dependency, privacy constraints, and ethical safeguards that any such system would require before real-world deployment.

Keywords: *Artificial Intelligence, Methamphetamine Trafficking, Machine Learning, Digital Forensics, Network Analysis, Blockchain, IoT, Drone Surveillance, Criminology*

1. INTRODUCTION

Drug trafficking is a form of organised criminal activity spanning production, transportation, financing, and distribution across geographical boundaries. Methamphetamine trafficking is a particularly instructive case because it now routes through multiple, technologically distinct stages: procurement, digital marketing and negotiation, financial settlement, physical transport, and distribution. Historically, investigations relied on physical surveillance, border inspection, informants, and post-seizure forensics. The migration of parts of this economy onto digital platforms has both complicated enforcement and created new, computationally analysable data trails [1], [3].

The book AI-Powered Solutions for Combating Methamphetamine Trafficking in the Digital Era by Dr. N. Kalla and Premanand Narasimhan provides a broad, applied survey of this problem space — covering methamphetamine characteristics and trafficking routes, digital forensic analysis, cryptocurrency and blockchain transactions, dark-web platforms, IoT-based mitigation, drone technology, and cross-border monitoring. This paper uses that survey as a starting point and situates its ideas alongside peer-reviewed and institutional research on AI-assisted trafficking detection, in order to propose a single integrated, evidence-grounded framework.

The Problem Statement of this paper is:

> *How can Artificial Intelligence assist investigators in identifying patterns associated with methamphetamine trafficking across digital and physical environments, without displacing human judgement and due process?*

AI is relevant here specifically because trafficking investigations increasingly generate data faster than it can be reviewed manually — digital communications, financial transactions, sensor feeds, and surveillance footage. NLP can process textual data, network analysis can map relationships between entities, and anomaly detection can flag deviations from expected behaviour for human follow-up. Throughout this paper, AI is treated strictly as an analytical aid: a system that narrows a large volume of information down to leads worth a human investigator's attention, not a system that determines guilt.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_31

2. LITERATURE REVIEW

Work on technology-assisted counter-trafficking spans several, often separately developed, research threads.

2.1 AI for Trafficking Detection on Digital Platforms

Several recent projects apply AI directly to trafficking detection on open platforms. One ResearchGate-indexed study proposes image-processing and machine-learning pipelines to flag drug-related imagery and symbols on social media, forwarding flagged content to investigators for review [1]. Separately, a ScienceDirect study on "knowledge-prompted" large language models shows that combining domain-specific prior knowledge with a general-purpose model such as ChatGPT measurably improves detection of drug-trafficking activity in online text, illustrating how modern NLP can be adapted for this exact task [2].

2.2 Institutional and Cross-Border AI Initiatives

At an institutional scale, the UK Research and Innovation-backed ARIEN project (ARtificial IntelligencE in fighting illicit drugs production and traffickiNg) is a Horizon Europe initiative building AI tools to give EU law-enforcement agencies a real-time picture of how drugs are produced, moved, and sold across borders [3]. A related EU-funded effort, the SIGNIFICANCE project, applies AI and deep learning to monitor social media, the open web, and the dark web for trafficking-linked criminal networks [1]. CEASEFIRE, a parallel EU-funded system originally built for illicit firearms trafficking, demonstrates a comparable multi-agency, AI-assisted architecture that is directly transferable to drug-trafficking contexts [4].

2.3 Financial and Digital Forensics

On the financial side, AI-assisted analysis has been explored for detecting money-laundering patterns associated with trafficking proceeds [5]. In digital forensics, tools such as Cellebrite Pathfinder and Oxygen already apply AI, data mining, and machine learning to visualise and connect evidence recovered from seized devices, helping investigators trace supply-chain and distribution links [6].

2.4 Physical and Chemical Detection Methods

Complementing digital methods, materials-science research has produced rapid, field-deployable sensors for methamphetamine itself — including a graphene-based electrochemical sensor capable of detecting methamphetamine in confiscated samples at very low concentrations [7], and a wearable, glove-based electrochemical device using silver nanoparticles for on-site screening [8]. These are not AI systems, but they represent an adjacent sensing layer that an AI-based framework could plausibly ingest as an additional data source.

2.5 Criminological Context

Criminological research cautions against treating AI as a self-sufficient solution. Weisheit's review of AI and illicit drugs notes that trafficking organisations already adapt quickly to enforcement technology, and that AI-detected anomalies (for instance in financial transactions) require human interpretation to distinguish genuine criminal activity from legitimate but unusual behaviour [5]. This reinforces the human-in-the-loop principle adopted in the framework proposed here.

2.6 Research Gap

Existing work tends to treat these layers separately: digital/social-media detection, blockchain analysis, physical sensing, and institutional cross-border platforms are each addressed by distinct projects. The gap this paper addresses is the absence of a single, methamphetamine-specific conceptual framework that connects digital-platform detection,

ISBN NO: 978-81-180146-5-2

financial analysis, IoT/drone-based physical monitoring, and criminological interpretation into one human-supervised pipeline, rather than requiring investigators to consult separate, disconnected systems.

3. PROPOSED AI-BASED FRAMEWORK

The proposed framework connects four categories of input data — digital/dark-web, financial/blockchain, IoT sensor, and drone/surveillance — through a shared preprocessing and AI-analysis layer, before every output is routed through mandatory human verification. Figure 1 summarises the pipeline.

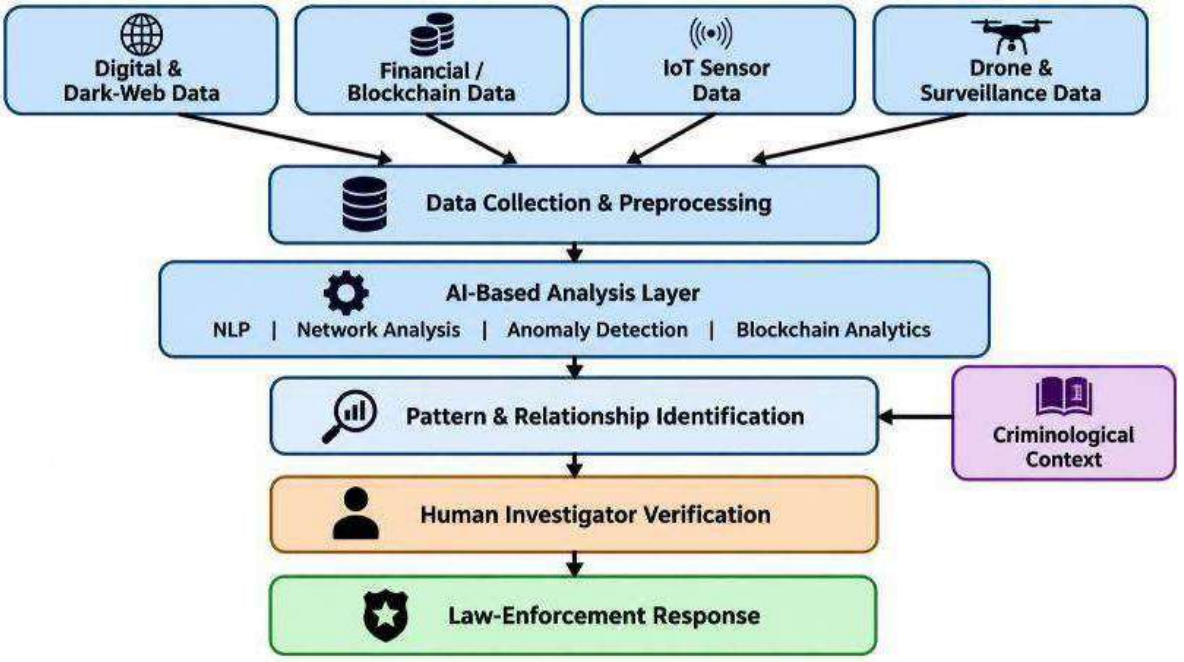


Figure 1. Proposed AI-based framework for methamphetamine-trafficking investigation support.

3.1 Data Collection and Preprocessing

The framework begins with legally obtained, appropriately authorised data drawn from digital forensic records, publicly available online content, transaction networks, geographical data, and IoT/drone sensor feeds. Because these sources differ in format and structure, a preprocessing stage cleans, normalises, and classifies the data before it reaches the analysis layer — following the same data-heterogeneity challenge identified in digital forensic tooling such as Cellebrite Pathfinder [6].

3.2 Natural Language Processing

NLP processes textual data — messages, listings, forum posts — to identify recurring terminology, coded language, and relationships between entities, consistent with the knowledge-prompted LLM approach shown to improve trafficking-content detection on social media [2]. NLP output is treated strictly as an investigative lead: it flags content for human review rather than automatically labelling an individual as a trafficker.

3.3 Network Analysis

Trafficking activity typically involves networks of accounts, communications, transactions, and locations rather than isolated events. Representing these as nodes and edges (Account → Communication → Transaction → Platform → Location) allows AI to surface highly connected nodes or structurally unusual relationships — the

ISBN NO: 978-81-180146-5-2

same principle underlying the SIGNIFICANCE and CEASEFIRE projects' approach to mapping criminal networks across platforms [1], [4].

3.4 Cryptocurrency and Blockchain Analysis

Blockchain ledgers are inherently analysable. AI-assisted analysis can cluster transactions, map relationships between addresses, and flag structurally unusual patterns for review, echoing prior work on AI for money-laundering detection [5]. Critically, a computationally identified anomaly is evidence of nothing on its own — it becomes relevant only alongside corroborating information and human investigation.

3.5 IoT-Based Monitoring

IoT sensors and cameras deployed at points of interest generate continuous physical-world data. An AI layer can process this stream and flag activity that deviates from an established baseline for human review, mirroring the sensing-layer role played by field-deployable chemical sensors such as the graphene-based and wearable glove-based methamphetamine detectors [7], [8], which could plausibly feed structured alerts into the same pipeline.

3.6 Drone and Aerial Monitoring

Drones extend monitoring to geographically difficult terrain. Within this framework, drones function purely as data-collection tools: camera and sensor output passes through the same AI-analysis and human-verification stages as every other input, rather than triggering any autonomous enforcement action.

3.7 Cross-Border and Multi-Environment Monitoring

Methamphetamine trafficking frequently spans road, maritime, air, and border environments simultaneously, the framework is designed to combine monitoring data from these environments rather than analysing each in isolation — the same cross-border ambition that motivates the EU's ARIEN initiative [3].

3.8 Criminological Interpretation

A criminological layer sits alongside the technical pipeline, supplying investigators with context on trafficking behaviour, organisational structure, and known adaptation patterns. This is deliberately positioned as an input to human verification rather than to the AI layer itself, consistent with Weisheit's caution that algorithmic anomalies require human, contextual interpretation before they carry investigative weight [5].

4. DISCUSSION

The principal advantage of the proposed framework is integration: a suspicious digital-communication pattern, an unusual transaction cluster, and an anomalous sensor reading are rarely meaningful in isolation, but become significant when an investigator can see how they relate. Existing systems — SIGNIFICANCE and CEASEFIRE for network-level detection [1], [4], knowledge-prompted LLMs for text [2], AI-assisted blockchain analysis for finance [5], and Cellebrite-style tools for device forensics [6] — each address one of these layers well. The contribution of this framework is connecting them into a single, methamphetamine-specific pipeline with one consistent rule: every AI output terminates in mandatory human verification before it informs any enforcement action.

This matters because AI systems in this domain are prone to false positives — a transaction pattern, message, or sensor reading that looks unusual to an algorithm may have an entirely legitimate explanation. Treating AI output as a prioritised lead list, rather than a verdict, is therefore not just an ethical safeguard but a practical necessity for

ISBN NO: 978-81-180146-5-2

investigative accuracy.

5. LIMITATIONS AND ETHICAL CONSIDERATIONS

False positives and false negatives: algorithmic flags can both wrongly implicate legitimate activity and miss genuine trafficking activity, so every flag requires independent verification.
Data quality: model reliability depends directly on the completeness, accuracy, and currency of input data; biased or incomplete datasets propagate into biased outputs.
Privacy: communication, financial, and location analysis intrudes on personal privacy and must operate strictly within legal authorisation and warrant requirements.
Adversarial adaptation: trafficking networks actively change platforms, terminology, and transaction methods in response to enforcement, so any deployed system requires continuous retraining and evaluation [5].
Ethical deployment: AI must remain a decision-support tool; human oversight, transparency, and accountability structures are non-negotiable design requirements, not optional add-ons.
Integration complexity: combining blockchain, IoT, forensic, and surveillance data sources involves real engineering challenges around differing formats, standards, and reliability levels.

6. FUTURE SCOPE

Multilingual NLP, since trafficking-related communication spans multiple languages and regional dialects.Graph-based and multimodal AI that jointly analyses text, imagery, transaction, and sensor data rather than treating each modality separately.Tighter integration between IoT/drone sensing and AI-based monitoring across large geographic areas, building on projects such as ARIEN [3].Explainable AI (XAI), so investigators can see why a pattern was flagged rather than receiving an opaque classification.Privacy-preserving AI techniques (e.g., federated learning, differential privacy) to reduce the tension between analytical power and civil-liberties protection.

7. CONCLUSION

Methamphetamine trafficking in the digital era spans conventional criminal logistics and rapidly evolving digital infrastructure — dark-web marketplaces, cryptocurrency, IoT devices, and drones. This paper proposed a conceptual, human-supervised AI framework connecting NLP, network analysis, blockchain analytics, IoT monitoring, and drone surveillance into a single investigative pipeline, grounded both in the applied overview presented by Kalla and Narasimhan and in existing peer-reviewed and institutional research including ARIEN, CEASEFIRE, and knowledge-prompted LLM detection systems. The framework's effectiveness ultimately rests on data quality, legal authorisation, and — above all — sustained human oversight. AI, in this context, is best understood not as a replacement for investigators but as a means of making an otherwise unmanageable volume of digital and physical evidence tractable for the humans who must ultimately decide what it means.

REFERENCES

[1] R. Ravi Ramaraj et al., "AI-Powered Detection of Drug Trafficking on Social Media Platforms Using Image Processing and Machine Learning," International Journal on Engineering Technology and Sciences (IJETS), vol. XI, no. IX, Sept. 2024.
[2] Knowledge-prompted ChatGPT: Enhancing drug trafficking detection on social media, ScienceDirect, 2024.
[3] UK Research and Innovation, "How AI is revolutionising the fight against drug trafficking" — ARIEN (ARtificial IntelligencE in fighting illicit drugs production and traffickiNg), Horizon Europe project overview, UKRI, 2024.

ISBN NO: 978-81-180146-5-2

[4] J. Cani, I. Mademlis, M. Mancuso, C. Paternoster, et al., "CEASEFIRE: An AI-Powered System for Combating Illicit Firearms Trafficking," arXiv:2406.14949, 2024.
[5] R. A. Weisheit, "Artificial Intelligence and Illicit Drugs: A World of Possibilities," Journal of Drug Issues / Taylor & Francis, 2025.
[6] "Digital forensic intelligence for illicit drug analysis in forensic investigations," PMC, National Center for Biotechnology Information, 2023.
[7] "Electrochemical Rapid Detection of Methamphetamine from Confiscated Samples Using a Graphene-Based Printed Platform," PMC, 2023.
[8] "Wearable Electrochemical Glove-Based Analytical Device (eGAD) for the Detection of Methamphetamine Employing Silver Nanoparticles," PMC, 2023.
[9] N. Kala and P. Narasimhan, *AI-Powered Solutions for Combating Methamphetamine Trafficking in the Digital Era*, International Journal of Innovative Technology and Creative Engineering Journal (IJITCE), 2026.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_32

Sustainable Agriculture: Real-Time Monitoring and Precision Farming through AgroBot

Dr S. Jayasri[1]
Assistant Professor, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, T. Nagar, Chennai-600 017

J. Bhagyashree[2]
Student, Department of Computer Applications , Shri Shankarlal Sundarbai Shasun Jain College for Women, T. Nagar, Chennai-600 017

T. Haripriya[3]
Student, Department of Computer Applications , Shri Shankarlal Sundarbai Shasun Jain College for Women, T. Nagar, Chennai-600 017

ABSTRACT

Sustainable agriculture is essential to ensure food security while protecting natural resources. With increasing population, climate change, and resource depletion, traditional farming methods are becoming inadequate. This paper introduces **AgroBot**, an intelligent mobile application designed to promote sustainable agricultural practices through real-time data and precision decision-making. **AgroBot** integrates four core modules: **Real-Time Soil Monitoring, Precision Water Requirement Prediction & Irrigation, Environmental Condition Tracking, and Weather Prediction**.

By combining IoT sensors, machine learning, and cloud computing, **AgroBot** provides farmers with actionable insights to optimise water, fertilizer, and pesticide use. Field simulations and economic analysis show the potential for **30–35% water savings, 25–30% reduction in chemical inputs, and 15–20% yield improvement**. This study demonstrates how digital technology can make sustainable agriculture accessible and affordable for small and marginal farmers.

Keywords: **Sustainable Agriculture, AgroBot, Precision Farming, IoT, Soil Monitoring, Smart Irrigation**

1. INTRODUCTON

Agriculture is one of the most important sectors of the global economy and provides livelihoods for more than 2.6 billion people worldwide. In developing countries such as India, agriculture plays a major role in employment, food security, rural development, and economic growth. However, the agricultural sector is increasingly affected by challenges such as groundwater depletion, soil degradation, climate change, unpredictable weather conditions, pest and disease outbreaks, and rising costs of agricultural inputs. At the same time, the global population continues to increase, creating a growing demand for food. According to the Food and Agriculture Organization (FAO), global food production needs to increase substantially to meet the food requirements of the projected population by 2050. These challenges highlight the need for agricultural practices that can increase productivity while conserving natural resources.

Sustainable agriculture provides an approach to address these challenges by balancing agricultural productivity, environmental protection, and long-term resource availability. Efficient utilization of water, fertilizers, pesticides, and energy, along with maintaining soil health and improving climate resilience, are essential components of sustainable farming. However, conventional agricultural practices often depend on fixed

ISBN NO: 978-81-180146-5-2

irrigation schedules, manual observations, and generalized recommendations, which may lead to over-irrigation, excessive application of agricultural inputs, and inefficient resource utilization. Therefore, there is a growing need for precision-based approaches that can provide farmers with accurate and timely information based on actual field conditions.

The rapid development of technologies such as the **Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning (ML), cloud computing, and mobile applications** has created new opportunities for transforming conventional agriculture into data-driven precision agriculture. Precision agriculture focuses on applying the **right input, at the right place, at the right time**, based on real-time field and environmental information. IoT sensors can continuously monitor parameters such as soil moisture, temperature, humidity, and other environmental conditions, while machine learning models can analyse these data to generate predictions and recommendations. Mobile applications can further improve accessibility by delivering this information directly to farmers in a simple and convenient format.

Despite these technological advancements, the adoption of smart farming solutions remains relatively limited among small and marginal farmers. High equipment costs, technical complexity, lack of digital accessibility, and limited technical knowledge can prevent farmers from effectively using advanced agricultural technologies. Furthermore, many existing agricultural applications provide generalized recommendations that are not sufficiently adapted to the specific soil, environmental, weather, and water conditions of individual farms. This creates a significant gap between the availability of agricultural technology and its practical application at the field level.

To address this gap, the present study proposes **AgroBot**, a farmer-centric mobile application designed to provide real-time, field-specific agricultural recommendations. **AgroBot** integrates four major components: **Real-Time Soil Monitoring, Precision Water Requirement Prediction and Irrigation, Environmental Condition Tracking, and Weather Prediction**. By combining IoT-based data collection, machine learning, cloud computing, and mobile technology, the proposed system aims to support informed decision-making and improve the efficient utilization of agricultural resources. AgroBot is designed with a focus on affordability, accessibility, and simplicity, making it particularly suitable for small and marginal farmers. The proposed system aims to contribute to sustainable agriculture by reducing resource wastage, improving farm management, and promoting data-driven and climate-resilient agricultural practices.

2. LITERATURE REVIEW

Sustainable agriculture has become increasingly important due to population growth, climate change, water scarcity, soil degradation, and the increasing demand for food production. Recent research has demonstrated that the integration of digital technologies, Internet of Things (IoT), artificial intelligence (AI), machine learning (ML), and cloud computing can improve agricultural productivity while reducing the excessive use of natural resources.

The *State of Food and Agriculture 2023* emphasizes the potential of digital technologies to improve resource efficiency, strengthen food security, and support informed decision-making among farmers. IoT-based agricultural systems have been widely investigated for real-time monitoring of soil and environmental parameters.

Patel (2023) and Naresh and Munaswamy highlighted the use of sensors to collect information such as soil moisture, temperature, humidity, and other environmental conditions, enabling farmers to make timely decisions regarding crop management.

Water management is one of the major areas of research in precision agriculture. Conventional irrigation methods can result in excessive water consumption because irrigation is often performed without considering

ISBN NO: 978-81-180146-5-2

the actual moisture requirement of the soil.

Verma (2022) demonstrated the application of machine learning techniques for efficient irrigation scheduling based on crop and environmental conditions. Similarly, Macapanas and Macapanas (2025) developed a low-cost IoT-based automated irrigation system using ESP8266 and cloud-based analytics.

Praba et al. (2024) proposed a real-time smart irrigation system that continuously monitors soil moisture and controls irrigation according to field conditions. These studies indicate that combining real-time sensor data with automated irrigation can significantly improve water-use efficiency and reduce unnecessary irrigation.

The application of IoT and advanced technologies has also expanded beyond irrigation. Karar et al. (2021) integrated Unmanned Aerial Vehicles (UAVs) with IoT technologies for crop monitoring and irrigation management, demonstrating the potential of aerial sensing in precision agriculture.

Wilberforce and Mwebaze (2025) proposed an IoT-enabled smart agriculture framework for real-time agricultural data collection and monitoring. Binayao et al. (2024) investigated IoT-based irrigation for rice cultivation and reported improvements in water management. In addition, machine learning and AI-based approaches are increasingly being explored for weather prediction, crop health assessment, disease detection, and agricultural decision support. Cloud platforms further enable the storage, processing, and visualization of real-time agricultural data, allowing farmers to access information remotely through mobile and web applications.

Although existing studies have demonstrated the effectiveness of individual technologies, many proposed systems are focused on a single function, such as irrigation, soil monitoring, or environmental sensing. There is a need for an integrated, affordable, and user-friendly system that combines multiple agricultural parameters and provides actionable recommendations, particularly for small and marginal farmers.

The proposed **AgroBot** addresses this research gap by integrating **real-time soil monitoring, precision water requirement prediction and irrigation, environmental condition tracking, and weather prediction** into a unified mobile-based platform. By combining IoT sensors, machine learning, and cloud computing, AgroBot aims to support data-driven agricultural decision-making, optimize the use of water and agricultural inputs, and contribute to more sustainable and resource-efficient farming practices.

Furthermore, the integration of smart technologies can improve the accessibility and scalability of agricultural solutions by enabling continuous monitoring and automated decision-making. Real-time data collected from field sensors can help identify changing soil and environmental conditions, allowing irrigation and other farming activities to be adjusted according to actual crop requirements. Such an integrated approach can reduce resource wastage, minimize manual intervention, and support farmers in making timely decisions. Therefore, combining IoT, AI, machine learning, and cloud-based monitoring provides a strong foundation for developing affordable and sustainable smart farming systems such as **AgroBot**.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_32

3. PROPOSED METHODOLOGY

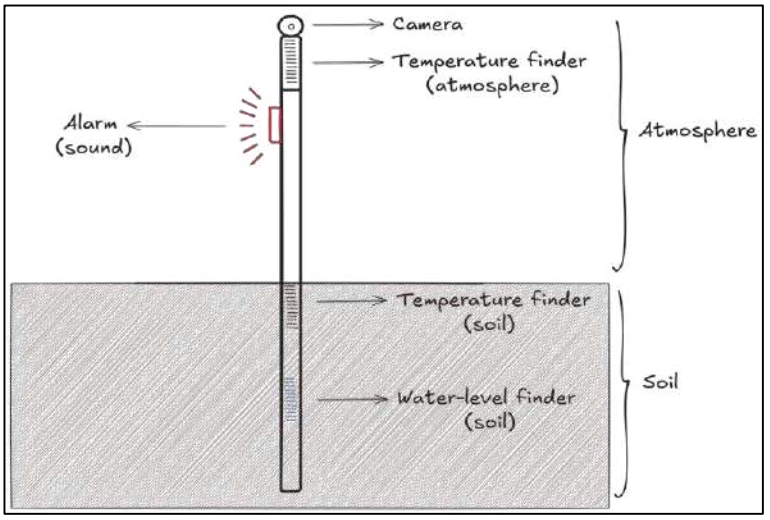


Fig 1: Agrobot Mechanism

The proposed **AgroBot** is an IoT-based smart agriculture system designed to monitor environmental and soil conditions in real time while automating irrigation. **As shown in Fig. 1**, the system consists of a vertical sensing unit with sensors positioned above and below the ground to collect atmospheric and soil-related data.

The **atmospheric temperature sensor** measures the surrounding air temperature, while the **camera** captures periodic images of the crop for future AI-based disease and pest detection. The **soil temperature sensor** records the temperature beneath the soil surface, and the **soil moisture sensor** continuously measures the moisture content of the soil to determine water availability. An **alarm module** provides an audio alert when abnormal conditions are detected.

The collected sensor data is transmitted to the **ESP32 microcontroller**, where it is processed and compared with predefined threshold values. The processed information is then sent to the **AgroBot mobile application through Wi-Fi**, allowing farmers to monitor soil moisture, temperature, and irrigation status remotely in real time. **The overall arrangement of these sensing components is illustrated in Fig. 1.**

When the soil moisture level falls below the required threshold, the ESP32 activates the **water pump through a relay module** to irrigate the field. Once the desired moisture level is reached, the pump is automatically switched off, preventing over-irrigation and reducing water wastage. If abnormal environmental conditions or system faults are detected, the **alarm module** provides an audio alert to notify the farmer.

The collected data can also be stored on a cloud platform for further analysis and future decision-making. **Fig. 1 provides an overview of the proposed sensing architecture**, showing how atmospheric and soil parameters are monitored together. Overall, the system enables continuous field monitoring, reduces manual intervention, conserves water, and supports precision agriculture.

The system is designed with a modular architecture, allowing individual components to be monitored and controlled without affecting the operation of the complete system. The ESP32 acts as the central processing unit, coordinating sensor inputs, decision-making, and communication between the field unit and the mobile application. This modular structure also makes the system easier to maintain and provides flexibility for future expansion.

Another important aspect of the proposed methodology is the use of threshold-based decision-making for

ISBN NO: 978-81-180146-5-2

irrigation management. Instead of relying on fixed irrigation schedules, the system responds to the actual conditions detected in the field. This approach allows irrigation to be performed according to the plant's water requirement, helping to avoid both insufficient and excessive watering. The collected information can further support analysis of field conditions and assist farmers in improving their overall crop management practices.

4. FEASIBILITY STUDY:

The proposed **AgroBot** is a technically feasible system that uses readily available IoT components with an estimated cost of **₹5,000–₹6,000.** It is an affordable solution for small and medium-scale farmers. The cost of installing the components is as given below:

Table 1: Cost Analysis

Materials	**Cost of production (In Rs)**
Camera	**2500**
Alarm	**500**
Microprocessor	**800**
Metal rod	**200**
Air temperature sensor	**400**
Soil moisture sensor	**300**
Water pump	**1200**
Piping and wiring	**800**
Miscellaneous	**300**
Total	**5500 (5000 – 6000 approx)**

The system is easy to install, operate, and maintain using a mobile application. Its automated irrigation reduces water wastage, labour, and operational costs while improving crop productivity. Therefore, **AgroBot** is a practical, economical, and scalable solution for sustainable agriculture.

The modular design of **AgroBot** also allows the system to be **expanded and upgraded gradually** without replacing the entire prototype. Additional sensors, improved communication modules, and advanced data-processing capabilities can be incorporated as required. This flexibility makes the system suitable for different farming conditions and provides a foundation for future technological developments while keeping the initial implementation simple and manageable.

Furthermore, **AgroBot** can support **better resource planning and farm management** by providing continuous records of field conditions. These records can help farmers identify patterns in soil moisture and environmental changes over time and make more informed decisions for future cultivation cycles. Such data-driven management can improve the consistency of farming practices and contribute to the long-term sustainability of agricultural operations.

5. Results & discussion:

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_32

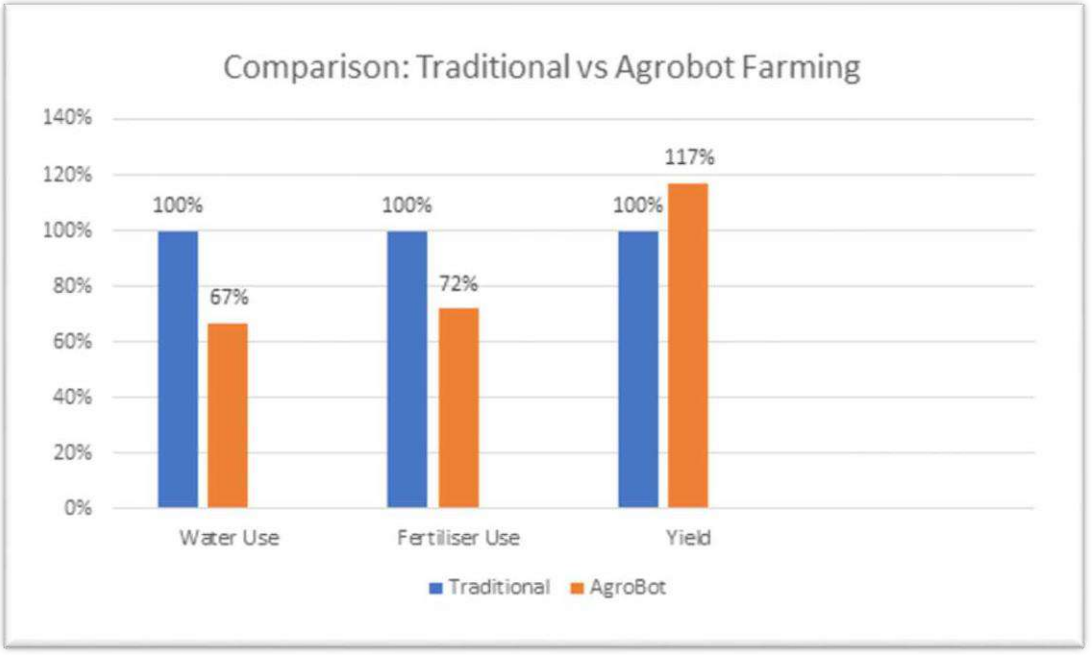


Fig 2: Traditioal vs Agrobot farming

From **Fig. 2**, it can be observed that the implementation of **AgroBot significantly improves resource efficiency compared with traditional farming methods**. The use of water is reduced by approximately **33%**, indicating that the automated and precision-based irrigation system can minimize unnecessary water consumption. Similarly, **fertilizer usage is reduced by 28%**, which can help prevent excessive application of fertilizers and reduce their potential environmental impact.

At the same time, AgroBot contributes to an improvement in overall agricultural productivity, with the **crop yield increasing by approximately 17% compared with the traditional farming method**. This improvement can be attributed to better monitoring of soil conditions and more efficient management of agricultural resources.

Overall, the results presented in Fig. 2 indicate that AgroBot can provide a balance between **resource conservation and improved crop productivity**. The reduction in water and fertilizer consumption, together with the increase in yield, demonstrates the potential of AgroBot as a tool for promoting **precision, efficient, and sustainable agricultural practices**.

6. Economic Analysis:

The economic analysis shows that **AgroBot can reduce farming costs through efficient resource utilization**. As shown in the figure, **water savings contribute 33%**, while **fertilizer savings contribute 28%** of the overall resource savings. The remaining **39%** represents savings from other agricultural resources.

These savings can reduce recurring cultivation expenses and improve the farmer's overall profitability. Although AgroBot requires an initial investment for its sensors, microcontroller, camera, pump, and other components, the reduction in water, fertilizer, and other resource consumption demonstrates its potential as a **cost-effective and economically sustainable solution for modern agriculture**.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_32

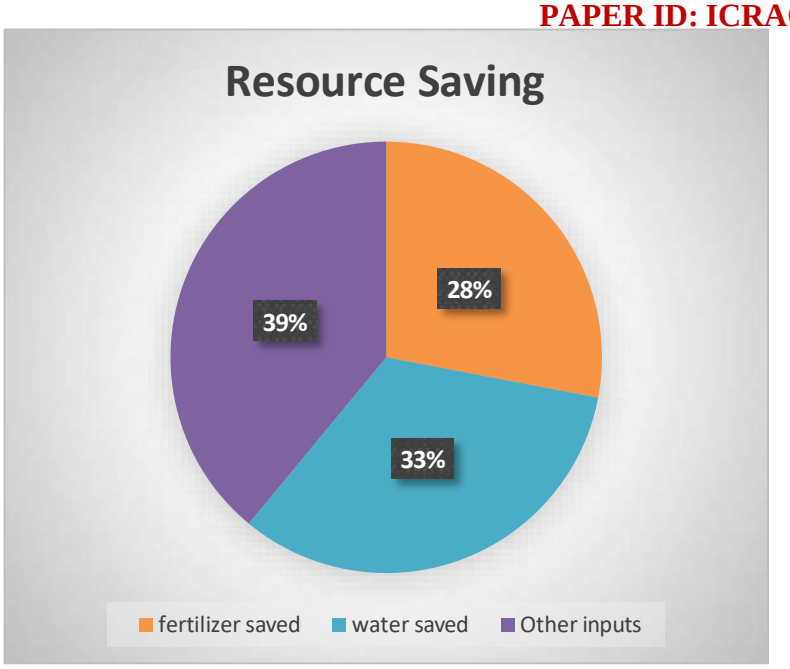


Fig 3: Resource Saving

6. FUTURE IMPROVEMENTS:

- **AI-Based Crop Health Detection:**
  Future development of AgroBot will focus on integrating an AI-based crop health detection system with the existing camera module. The system will analyze crop images to identify early signs of diseases, pest attacks, and visible nutrient deficiencies. Detected conditions can be communicated to farmers through the AgroBot application along with suitable recommendations. This enhancement can enable early intervention, reduce crop losses, and improve the effectiveness of precision farming.
- **Advanced Nutrient Monitoring:**
  Future development of AgroBot can include **NPK (Nitrogen, Phosphorus, and Potassium) sensors** to continuously assess the nutrient condition of the soil. The collected data can be analysed to determine whether the crop requires additional nutrients and provide suitable fertilizer recommendations through the AgroBot application. This can help prevent excessive fertilizer usage, reduce cultivation costs, maintain soil fertility, and promote more sustainable farming practices.

In the long term, AgroBot can be extended with technologies such as drone-based monitoring, regional-language voice assistance, and blockchain-based traceability.

7. CONCLUSION:

The proposed **AgroBot** presents an innovative and sustainable solution for modern agriculture by integrating **IoT-based real-time soil monitoring, automated irrigation, and mobile application support** into a single smart farming system. The system continuously monitors essential parameters such as soil moisture, temperature, and water requirements, enabling precision irrigation that supplies water only when necessary. This not only conserves water but also improves crop health, reduces manual labor, and enhances overall farming efficiency. By providing farmers with real-time field data and intelligent decision support through a mobile application, AgroBot promotes efficient resource utilization while encouraging environmentally friendly and sustainable agricultural practices.

ISBN NO: 978-81-180146-5-2

- Developed an **IoT-enabled smart agriculture system** for continuous monitoring of soil moisture and temperature in real time.
- Implemented an **automated irrigation mechanism** that activates water supply only when the soil moisture falls below the required level, thereby minimizing water wastage.
- Designed a **mobile application interface** that allows farmers to monitor field conditions remotely and make informed decisions based on live sensor data.
- Proposed a **cost-effective prototype** (approximately **₹5,000–₹6,000**) that is affordable and suitable for small and medium-scale farms.
- Demonstrated the practical application of **precision agriculture** by integrating sensing, automation, and wireless communication into a single platform.
- Suggested future enhancements, including **AI-based crop disease detection, nutrient monitoring,** to improve the scalability and capabilities of the system.

In conclusion, **AgroBot** offers a practical, affordable, and scalable approach to achieving sustainable agriculture through the effective integration of automation, IoT, and precision farming technologies. The proposed system has the potential to increase agricultural productivity, optimize water and resource utilization, reduce operational costs, and support environmentally responsible farming practices. With the addition of advanced technologies in future versions, AgroBot can evolve into a comprehensive smart farming platform that empowers farmers with data-driven decision-making, improves crop quality and yield, and contributes significantly to long-term food security, environmental sustainability, and the digital transformation of agriculture.

REFERENCES:

1. A. S. Ibrahim *et al.*, "AI-IoT based smart agriculture pivot for plant diseases detection and treatment," *Scientific Reports*, vol. 15, 2025. This is particularly relevant to your proposed **AI-based crop disease detection** feature.

2. S. S. Kamble, A. Gunasekaran, and R. Sharma, "Modeling the blockchain enabled traceability in agriculture supply chain," *International Journal of Information Management*, 2019. This is useful for your **blockchain-based crop traceability** section.

3. G. Lv, C. Song, and Y. Liu, "Blockchain-Based Traceability for Agricultural Products: A Systematic Literature Review," *Agriculture*, vol. 13, no. 9, 1757, 2023. This open-access review discusses blockchain applications, transparency, food safety, and agricultural product traceability.

4. A. Marchese and O. Tomarchio, "A Blockchain-Based System for Agri-Food Supply Chain Traceability Management," *SN Computer Science*, vol. 3, article 279, 2022. This open-access study is relevant to tracking agricultural products throughout the supply chain.

5. D. Prashar, N. Jha, S. Jha, Y. Lee, and G. P. Joshi, "Blockchain-Based Traceability and Visibility for Agricultural Products: A Decentralized Way of Ensuring Food Safety in India," *Sustainability*, vol. 12, no. 8, 3497, 2020. This is especially useful because it focuses on **agricultural traceability in the Indian context**.

6. A. T. T. et al., "Automation of soil moisture sensor-based basin irrigation system," *Smart Agricultural Technology*, vol. 2, 100032, 2022. The study investigates automated irrigation based on soil moisture sensing and reports improved irrigation application efficiency.

ISBN NO: 978-81-180146-5-2

7. H. Feng, X. Wang, Y. Duan, J. Zhang, and X. Zhang, "Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges," *Journal of Cleaner Production*, vol. 260, 121031, 2020. This provides a strong theoretical reference for combining **IoT and blockchain for agricultural traceability**.

8. "Improving Agricultural Product Traceability Using Blockchain," *International Journal of Environmental Research and Public Health*, 2022. The study proposes an agricultural traceability system using **Ethereum blockchain and IPFS** for secure and efficient information management.

9. "Trusted-auditing chain: A security blockchain prototype used in agriculture traceability," *Heliyon*, vol. 8, no. 11, e11477, 2022. This research focuses on secure, auditable, and privacy-aware blockchain-based agricultural traceability.

10. S. Dutta, S. Gorain, S. Roy, and S. Banerjee, "Integrating IoT and communication technologies for smart and precision irrigation systems toward sustainable water management," *Discover Internet of Things*, vol. 6, 2026. This is highly relevant to your **IoT-based precision irrigation** methodology.

11. "Internet of Things-enabled smart irrigation systems for precision water management: A systematic review," *Agricultural Water Management*, vol. 333, 2026. This systematic review examines IoT sensors, cloud platforms, machine learning, and precision irrigation for water conservation.

12. A. Kumar Pal, B. D. K. Patro, and S. Chaube, "Design and implementation of a deep learning framework for automated crop classification and health diagnosis in precision agriculture," *Scientific Reports*, vol. 16, 2026. This is useful for your proposed **AI-based crop health detection using cameras, drones, and satellite imagery**.

13. A. Srivastava *et al.*, "Deep convolutional models for robust multi-crop disease recognition in real-world conditions," *Scientific Reports*, vol. 16, 2026. This research focuses on deep-learning-based identification of crop diseases from plant images.

14. "Remote Sensing through UAVs for precision agriculture: Applications, technical foundations, current barriers, and future opportunities," *Smart Agricultural Technology*, vol. 14, 2026. This is particularly relevant to your **drone-based crop monitoring and remote sensing** component.

15. H. Feng, X. Wang, Y. Duan, J. Zhang, and X. Zhang, "Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges," *Journal of Cleaner Production*, vol. 260, 121031, 2020. This provides a strong foundation for your **blockchain-based crop traceability** concept.

16. G. Lv, C. Song, and Y. Liu, "Blockchain-Based Traceability for Agricultural Products: A Systematic Literature Review," *Agriculture*, vol. 13, no. 9, 1757, 2023. The paper reviews blockchain applications for agricultural traceability, food safety, and consumer confidence.

17. A. Marchese and O. Tomarchio, "A Blockchain-Based System for Agri-Food Supply Chain Traceability Management," *SN Computer Science*, vol. 3, article 279, 2022. This open-access research is useful for explaining how blockchain can track agricultural products throughout the supply chain.

18. S. Pang *et al.*, "A survey on evaluation of blockchain-based agricultural traceability," *Computers and Electronics in Agriculture*, vol. 227, article 109548, 2024. This study examines implementation challenges such as standardization, participation, and traceability across agricultural supply chains.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_32

19. "Improving Agricultural Product Traceability Using Blockchain," *International Journal of Environmental Research and Public Health*, 2022. The study proposes a blockchain and IPFS-based agricultural traceability system for secure information management.

20. "A peer-to-peer blockchain-based architecture for trusted and reliable agricultural product traceability," *Decision Analytics Journal*, vol. 9, 100363, 2023. The research proposes a blockchain architecture for improving reliability, scalability, and trust in agricultural product traceability.

21. "Blockchain in Agriculture Traceability Systems: A Review," *Applied Sciences*, vol. 10, no. 12, 4113, 2020. This review discusses blockchain applications, advantages, and challenges in agricultural supply-chain traceability.

22. C. Prakash, L. P. Singh, A. Gupta, and S. K. Lohan, "Advancements in smart farming: A comprehensive review of IoT, wireless communication, sensors, and hardware for agricultural automation," *Sensors and Actuators A: Physical*, vol. 362, article 114605, 2023. This provides a broad review of **IoT, sensors, communication technologies, and agricultural automation**.

23. R. Togneri, R. Prati, H. Nagano, and C. Kamienski, "Data-driven water need estimation for IoT-based smart irrigation: A survey," *Expert Systems with Applications*, vol. 225, article 120194, 2023. This is highly relevant to AgroBot's **water requirement prediction and precision irrigation** component.

24. "A systematic review of IoT technologies and their constituents for smart and sustainable agriculture applications," *Scientific African*, vol. 19, article e01577, 2023. The study reviews **IoT sensors, controllers, communication technologies, data analytics, AI/ML, NPK monitoring, and agricultural automation**.

25. A. T. T. et al., "Soil water sensors for irrigation scheduling in the United States: A systematic review of literature," *Agricultural Water Management*, vol. 278, article 108148, 2023. This research is useful for supporting your **soil-moisture-based irrigation methodology**, including the role of sensors in reducing water consumption while maintaining crop yield.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_33

BEYOND OBJECT DETECTION: AN AI-POWERED RISK-AWARE NAVIGATION SYSTEM FOR VISUALLY IMPAIRED PEOPLE

[1] Ms.J.Jeba Priya
Assistant Professor,
Department of Computer science
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, India
j.jebapriya@shasuncollege.edu.in

[2] Rasika J S
Student,
Department of Computer Science
Shri Shankarlal Sundarbai Shasun Jain College for Women,Chennai, India
rasika.naksatra@gmail.com

[3] VimalaDevi S
Student,
Department of Computer science
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, India
svimaladevi18@gmail.com

Abstract
Visually impaired people may face difficulties while navigating unfamiliar environments because they cannot directly observe surrounding obstacles and objects. This paper presents VisionMate, a smartphone-based accessibility assistant that provides environmental awareness through computer vision and voice interaction. The system combines real-time object detection, spatial analysis, motion tracking, and risk assessment to generate concise navigation instructions. A lightweight YOLOv8n model in ONNX format is executed on the device using ONNX Runtime. VisionMate also provides an interactive feature, Ask VisionMate, which allows users to ask questions about the current camera view through voice input. The captured image and question are processed using a FastAPI backend and Gemini multimodal AI, and the response is converted into speech. The proposed system combines continuous local obstacle awareness with interactive scene understanding in a single mobile application.
Keywords— *Assistive Technology, Computer Vision, Object Detection, Visually Impaired, YOLOv8, Risk-Aware Navigation, Multimodal AI.*

I. INTRODUCTION

Independent navigation can be challenging for people with visual impairment because important environmental information is normally obtained through vision. Obstacles, objects, people, and changes in the surrounding environment may not always be identified before they become relevant to movement. Conventional mobility aids such as white canes and guide dogs provide important assistance, but they do not necessarily describe the surrounding environment in detail.

Computer vision provides an opportunity to supplement existing mobility methods by identifying objects from camera images. YOLO introduced a real-time object-detection approach using a single neural network [1], while later lightweight models made object detection more suitable for resource-constrained applications [2]. However, detecting an object alone does not indicate whether it is relevant to the user's path or where it is located relative to the user.

VisionMate was developed to combine object detection with spatial analysis, motion tracking, risk assessment, and voice feedback. The application uses a smartphone camera as its primary input and converts detected environmental information into short spoken instructions.

The system also includes Ask VisionMate, an interactive multimodal feature through which users can ask natural-language questions about the current camera view. For example, a user can ask where a visible laptop or bag is located. The captured image and question are processed by a multimodal AI service, and the response is returned as speech.

Thus, VisionMate combines continuous local obstacle awareness with user-driven scene understanding in a single mobile application.

II. RELATED WORKS

Computer vision has been increasingly explored for assistive technologies for visually impaired users.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_33

Object detection can provide information about surrounding objects that may otherwise be difficult for the user to obtain.

YOLO demonstrated the feasibility of real-time object detection using a single neural network [1]. Lightweight object-detection models are particularly relevant to mobile applications because they reduce computational requirements [2].

Previous research has also investigated object detection combined with audio feedback for visually impaired navigation [3], [4]. These approaches demonstrate the potential of computer vision to provide environmental information through non-visual feedback.

However, basic object detection mainly identifies what is present in an image. It does not necessarily determine whether an object represents a relevant obstacle or how its position changes over time. VisionMate therefore combines object detection with spatial interpretation, motion tracking, and risk assessment. In addition, its interactive multimodal component allows users to request information about a particular object or scene.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

VisionMate is designed as a hybrid smartphone-based system consisting of two complementary components: continuous navigation assistance and interactive scene understanding.

The navigation component processes camera frames locally using YOLOv8n in ONNX format. Detected objects are analyzed according to their position, confidence, size, and changes across consecutive frames. The resulting information is passed to the risk-assessment module, which generates concise navigation instructions through text-to-speech.

The interactive component is activated when the user asks a question. The application captures the current camera image and combines it with the recognized voice query. These inputs are sent to a FastAPI backend, which communicates with Gemini multimodal AI and returns a textual response to the mobile application.

A. System Architecture

The overall architecture of VisionMate is shown in Fig. 1.

The smartphone camera provides the primary visual input. For continuous navigation, camera frames are processed by the on-device YOLOv8n model through ONNX Runtime. The detected objects are passed to spatial analysis and motion tracking. The risk engine uses this information to generate navigation-oriented feedback, which is converted into speech.

For interactive queries, the user provides a voice question through the speech-recognition module. The captured image and question are sent to the FastAPI backend. Gemini multimodal AI analyzes both inputs and generates a concise response. The response is returned to the application and converted into speech using text-to-speech.

This architecture keeps continuous obstacle processing local while reserving remote multimodal processing for user-requested scene understanding.

Fig. 1. System architecture of the proposed VisionMate accessibility assistant.

B. Object Detection and Spatial Analysis

The system uses YOLOv8n in ONNX format for local object detection. The model is executed through ONNX Runtime using a 640 × 640 input representation in the current implementation.

Each prediction provides an object class, confidence value, and bounding-box coordinates. A confidence threshold is applied to reduce low-confidence detections. The pretrained model supports common categories such as people, vehicles, bags, chairs, laptops, phones, and other objects.

The position of each bounding box is used to determine the approximate location of an object within the camera view. The scene is divided into directional regions such as left, center, and right, with upper and lower regions where required. This allows detected objects to be converted into simple directional information.

The system currently communicates relative position rather than an exact physical distance.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_33

IV. RISK-AWARE NAVIGATION

A. Motion Tracking

A detected object may appear in multiple consecutive camera frames. VisionMate compares detections across frames to observe changes in their position and bounding-box size. This provides temporal information and helps maintain more consistent environmental awareness.

Motion information is used together with the current detection rather than treating every frame independently. This reduces unnecessary changes in the generated navigation information.

B. Risk Assessment

The risk-processing layer considers the spatial and visual characteristics of detected objects. Objects located near the center of the camera view or occupying a larger portion of the frame can receive greater attention than objects near the edge.

The current implementation considers object position, bounding-box size, confidence, and changes across consecutive frames. Based on these factors, the system generates concise navigation-oriented messages such as *"Obstacle ahead," "Object on your left,"* or *"Please move carefully."*

The purpose of the risk engine is to provide additional environmental awareness. It is not intended to replace established mobility aids or make safety-critical decisions independently.

C. Voice Feedback

Navigation instructions are converted into speech using text-to-speech. The messages are intentionally short so that they can be understood while the user is moving. Speech-control logic is used to reduce unnecessary repetition and prevent continuous repeated announcements of the same condition.

V. ASK VISIONMATE: MULTIMODAL SCENE UNDERSTANDING

Ask VisionMate provides an interactive method for obtaining information about the current camera scene.

The user first provides a question through speech recognition. The application captures the current camera image and sends both the image and recognized question to the FastAPI backend.

The backend forwards the inputs to Gemini 2.5 Flash. The multimodal model analyzes the image together with the question and generates a concise textual response. This response is returned to the mobile application and converted into speech.

The AI component is instructed to avoid unsupported claims. When an object is visible, its approximate position can be described relative to the camera view. If the requested object cannot be identified from the image, the system can indicate that it is not visible rather than inventing an answer.

For example, users can ask questions such as *"Where is my laptop?"* or *"Where is my bag?"* The feature is therefore useful when the user requires information about a specific object rather than a general obstacle warning.

Ask VisionMate complements the local navigation engine. Continuous obstacle processing remains on the device, while the multimodal service is used for user-driven questions.

VI. IMPLEMENTATION

The mobile application was developed using Flutter and Dart. The smartphone camera provides visual input, while YOLOv8n and ONNX Runtime perform local object detection. Speech-to-text is used for voice queries and text-to-speech provides spoken responses.

The backend was developed using FastAPI. For Ask VisionMate, the application sends a captured image and recognized question to the backend. The backend communicates with Gemini 2.5 Flash and returns the generated answer.

The YOLO model is packaged with the mobile application and loaded into an ONNX Runtime session. Camera frames are converted into the required input representation before inference. Detection results are decoded into object labels, confidence scores, and bounding boxes.

For interactive queries, the Gemini API key is maintained on the backend rather than being directly embedded in the mobile application. The backend receives the image and question, performs the multimodal request, and returns only the generated response to the mobile application.

ISBN NO: 978-81-180146-5-2

VII. EXPERIMENTAL OBSERVATIONS AND DISCUSSION

The prototype was functionally tested using indoor scenes containing commonly encountered objects.

The local detection component identified supported objects and supplied their bounding-box information to the navigation layer. The spatial-processing component used these positions to generate directional information for voice feedback.

Ask VisionMate was tested using questions concerning visible objects, including *"Where is my bag?"* and *"Where is my laptop?"*. When the requested object was visible in the captured scene, the multimodal component generated an appropriate response, which was returned to the application and spoken using text-to-speech.

Testing also revealed a limitation of the pretrained detection model. A bag can sometimes be classified as a suitcase because the general-purpose model contains fixed object categories and may confuse visually similar objects. This demonstrates the need for domain-specific training when more accurate object categories are required.

The interactive component additionally depends on network connectivity because the image and question are processed through the backend and multimodal AI service.

A systematic quantitative evaluation has not yet been completed. Therefore, unsupported accuracy, latency, or user-performance values are not reported in this paper. Future evaluation will measure detection performance, response time, and reliability across different environments.

VIII. LIMITATIONS AND FUTURE WORK

The current prototype has several limitations. The pretrained YOLOv8n model may confuse visually similar objects, and the system does not currently provide reliable metric distance estimation. The Ask VisionMate feature also requires network connectivity for communication with the backend and multimodal AI service.

Future work will focus on developing a specialized dataset containing objects relevant to visually impaired navigation. Improvements to object tracking and reliable distance estimation will also be investigated. Local or offline multimodal processing could further reduce dependence on network connectivity.

The system should additionally be evaluated across a larger range of indoor and outdoor environments. Testing with visually impaired users would provide useful information about the clarity, usefulness, and reliability of the generated voice instructions.

IX. CONCLUSION

This paper presented VisionMate, a smartphone-based accessibility assistant combining computer vision, risk-aware processing, voice interaction, and multimodal AI.

The system uses YOLOv8n with ONNX Runtime for local object detection and combines detection results with spatial and temporal information to provide navigation-oriented voice feedback. The Ask VisionMate feature extends the system by allowing users to ask natural-language questions about objects in the current camera view.

The hybrid architecture separates continuous local processing from interactive multimodal analysis. Initial functional testing demonstrates the feasibility of combining these approaches in a single mobile application. The prototype also identifies areas requiring further development, particularly object classification, distance estimation, network independence, and systematic real-world evaluation.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2023.

[3] "Enhancing Object Detection in Assistive Technology for the Visually Impaired: A DETR-Based

ISBN NO: 978-81-180146-5-2

Approach," *IEEE Access*, vol. 13, pp. 71647–71661, 2025.
[4] "Navigation Using Object Detection for Visually Impaired Individuals," in *Proc. 12th Int. Conf. Computing for Sustainable Global Development (INDIACom)*, 2025.
[5] Flutter, "Camera Plugin Documentation," Flutter Documentation.
[6] Google, "Gemini API Documentation," Google AI for Developers.
[7] ONNX Runtime, "ONNX Runtime Mobile Documentation," Microsoft.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26)

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_34

Student Engagement Detection Using Facial Emotion Recognition Through Deep learning models

1. A. Valliyammai Head of BCA Department , Avichi College of Arts and Science, Virugambakkam.
Email:kousalya23.shriaura@gmail.com
2. M.Shanmugapriya Student of BCA Department, Avichi College of Arts and Science, Virugambakkam.
Email: 24bca45acas@gmail.com
3. S. Yashvitha Student of BCA Department, Avichi College of Arts and Science, Virugambakkam.
Email: 24bca49acas@gmail.com

Abstract

Student engagement plays an important role in effective learning, academic performance, and knowledge retention. However, conventional approaches such as questionnaires, student feedback, and manual classroom observation are often subjective, time-consuming, and difficult to apply continuously in large-scale learning environments. This study proposes an intelligent and scalable framework for estimating student engagement using Facial Emotion Recognition (FER) and Convolutional Neural Networks (CNNs). The proposed system processes facial images captured through a webcam or video source and identifies seven facial emotion categories: angry, disgust, fear, happy, sad, surprise, and neutral. These emotion predictions are subsequently transformed into three broader engagement states: engaged, neutral, and disengaged. The framework consists of face detection, image preprocessing, CNN-based feature extraction, emotion classification, engagement mapping, and real-time analytics components. To improve scalability, the system is designed as a modular pipeline in which facial processing, emotion inference, and engagement analysis can operate independently. Batch processing and lightweight CNN inference can further support deployment across multiple students or video streams. The proposed framework can be integrated into online learning platforms and classroom analytics systems to provide aggregate engagement information to instructors. The study also considers important limitations, including variations in illumination, head pose, facial occlusion, dataset bias, and the fact that facial expression is only an indirect indicator of actual cognitive engagement. Future improvements can incorporate eye gaze, head pose, blink rate, temporal facial patterns, and classroom-specific engagement annotations to improve reliability.

Keywords: Facial Emotion Recognition, Student Engagement, Convolutional Neural Network, FER2013, Deep Learning, E-learning Analytics, Computer Vision

1. Introduction

Student engagement is an important factor in the learning process because it reflects the level of attention, interest, participation, and emotional involvement of students during an educational activity. In traditional classrooms, instructors generally estimate engagement through direct observation of student behavior, facial expressions, participation, and body language. Although these approaches can provide useful information, they become difficult to maintain consistently when the number of students increases or when learning takes place through online platforms.

The rapid growth of online and hybrid education has created a need for automated techniques capable of monitoring learner behavior without requiring continuous manual observation. Computer vision and deep learning provide an opportunity to analyze visual behavioral signals obtained from commonly available cameras. Among these signals, facial expressions are particularly useful because they can be captured using standard webcams without requiring specialized hardware.

Facial Emotion Recognition (FER) is a computer vision task that attempts to classify facial expressions into predefined emotional categories. Traditional FER approaches commonly depend on manually engineered features, whereas deep learning approaches can automatically learn hierarchical visual representations from facial images. Convolutional Neural Networks (CNNs) are particularly suitable for this task because convolutional layers can learn spatial patterns such as facial contours, eyes, eyebrows, mouth shapes, and other expression-related features.

However, facial emotion recognition and student engagement detection are not identical problems. Emotion represents an affective state, whereas engagement can include cognitive, behavioral, and emotional components. Therefore, this study treats facial emotion as a behavioral proxy for engagement rather than claiming that facial expressions directly measure cognitive engagement.

This research proposes a CNN-based framework that first recognizes facial emotions and subsequently estimates a broad engagement state from the predicted emotion. The framework is designed with scalability in mind so that it can potentially process individual webcam input as well as multiple video streams. The primary objectives of the study are to develop a facial emotion classification model, establish an interpretable emotion-to-engagement mapping, evaluate the resulting system using standard classification metrics, and examine the practical limitations of using facial expressions for engagement estimation.

2. Literature Review

This section summarizes prior work relevant to facial emotion recognition and engagement detection. Instead of describing each study in a continuous narrative, the reviewed works are organized in Table 1 by author/year, research focus, dataset or technique used, and the key contribution or relevance of the work to the present study. This tabular organization makes it easier to compare the underlying approach, scope, and dataset used across the different research directions that inform the proposed framework.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_34

Table 1. Summary of Reviewed Literature

Author(s) / Year	Research Focus	Dataset / Technique	Key Contribution / Relevance
Ekman & Friesen (1971) [2]	Psychology of facial expression	Cross-cultural study	Basis for the universal seven-class emotion taxonomy used here.
Goodfellow et al. (2013) [1]	Benchmark dataset for FER	FER2013 (48x48, 7 classes)	Dataset used directly to train/evaluate the emotion recognition component.
Lucey et al. (2010) [4]	Posed FER	CK+ dataset	Controlled lab-quality reference for robustness comparison vs. in-the-wild data.
Mollahosseini et al. (2019) [3]	In-the-wild FER	AffectNet	Shows need to handle illumination/pose variation, relevant to classroom deployment.
Simonyan & Zisserman (2014) [5]	Deep CNN design	VGG-style CNNs	Deeper stacked conv filters improve representation learning; informs the CNN design.
Chang, Tran & Liu (2018) [9]	Landmark-free deep FER	ExpNet (3D)	End-to-end learned features, supporting CNN-based automatic feature learning.
Whitehill et al. (2014) [7]	Engagement from expressions	Classroom video	Primary precedent for treating emotion as a proxy for engagement.
Bosch et al. (2015) [6]	Affect detection in classrooms	In-the-wild video	Highlights domain-shift risk between benchmark data and real deployment.
D'Mello & Graesser (2012) [8]	Temporal affect dynamics	Empirical study	Supports temporal aggregation over single-frame engagement decisions.
Calvo & D'Mello (2010) [10]	Interdisciplinary affect review	Survey	Frames FER-based engagement as one estimation strategy, not a complete measure.

Two broad directions emerge from Table 1. The first direction concerns facial emotion recognition itself: early work established the psychological basis for discrete emotion categories, later work introduced benchmark datasets of increasing scale and realism (FER2013, CK+, AffectNet), and CNN architecture research demonstrated that deep, end-to-end feature learning outperforms handcrafted descriptors. The second direction concerns engagement detection as a broader problem: some studies use emotion categories directly as engagement indicators, while others emphasize that engagement has behavioral and cognitive components that emotion alone does not fully capture, and that affective states change over time rather than remaining fixed within a single frame.

Based on these two directions, the present work combines CNN-based facial emotion recognition with a separate, interpretable engagement-estimation layer, and adds temporal aggregation to account for the dynamic nature of affect during learning. The separation between emotion classification and engagement estimation allows the emotion model to be evaluated independently using standard classification metrics, while keeping the mapping from emotion to engagement explicit and adjustable for future refinement with additional behavioral signals.

3. Proposed Methodology

The proposed system is designed as a modular pipeline for estimating student engagement from facial expressions. The methodology consists of the following major stages: data acquisition, data preprocessing, face detection, CNN-based feature extraction, emotion classification, emotion-to-engagement mapping, temporal engagement estimation, scalable multi-student processing, and engagement analytics.

3.1 System Architecture

The overall processing flow is designed as:

Video/Webcam Input → Face Detection → Data Preprocessing → CNN Feature Extraction → Emotion Classification → Engagement Mapping → Temporal Aggregation → Analytics Dashboard

The architecture separates the computational stages so that the CNN inference component can be reused across multiple input streams. This modular organization supports future replacement of individual components without redesigning the complete system.

3.2 Data Acquisition

The initial development and evaluation of the facial emotion recognition component uses the FER2013 dataset. The dataset contains 48x48 grayscale facial images categorized into seven emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. The dataset provides a standardized benchmark for developing and evaluating facial emotion classification models. Although FER2013 is suitable for model development, it does not directly provide student engagement labels. Therefore, the dataset is used primarily for training the emotion recognition component, while engagement is estimated through a separate mapping layer.

3.3 Data Preprocessing

Before the raw FER2013 images can be used to train the CNN, or before a live video frame can be passed through the trained model, the data must be cleaned and standardized, since a CNN learns more effectively from inputs that are consistent in scale, intensity, and format. The preprocessing pipeline used in this study consists of the following stages:

Data Cleaning and Verification: corrupted, mislabeled, duplicate, or invalid images are identified and discarded so that noisy samples do not degrade training.

Grayscale Conversion and Resizing: images are kept single-channel and resized to a fixed 48x48 resolution so the CNN receives inputs of uniform dimensionality.

ISBN NO: 978-81-180146-5-2

Pixel Normalization: raw intensities in [0, 255] are rescaled to [0, 1] by dividing by 255, which stabilizes and speeds up gradient-based training.

Label Encoding: the seven emotion categories are one-hot encoded to match the categorical cross-entropy loss and the seven-unit softmax output (Section 3.5).

Handling Class Imbalance: FER2013 under-represents classes such as 'disgust'; class-weighted loss and minority-class oversampling reduce bias toward majority classes.

Dataset Splitting: data is partitioned into training, validation, and held-out test subsets so reported metrics reflect generalization, not memorization.

Data Augmentation: horizontal flipping, small rotations, and controlled zoom are applied to the training subset only, improving robustness to real webcam variation.

3.4 Face Detection

For real-time deployment, the input may originate from a webcam or recorded classroom video. Before the preprocessing steps described in Section 3.3 can be applied, the system must first locate the facial region within each incoming frame.

1. Capture the current video frame from the webcam or video source.
2. Apply a face detector to locate facial bounding boxes within the frame.
3. Crop the detected facial region(s) from the frame.
4. Forward each cropped face to the data preprocessing pipeline (Section 3.3) prior to CNN inference.

If no face is detected in a given frame, that frame is skipped from emotion inference to avoid generating an unreliable prediction from background content.

3.5 CNN-Based Feature Extraction

A Convolutional Neural Network is employed to automatically learn spatial features from facial images. Instead of manually defining facial characteristics, the CNN learns relevant representations directly from training examples. The proposed baseline architecture contains three convolutional blocks. Each block consists of convolution, batch normalization, ReLU activation, and max-pooling operations. The number of filters increases progressively across the network so that deeper layers can learn more complex facial representations.

Layer	Configuration
Input	48x48x1 grayscale image
Conv Block 1	Conv(32, 3x3) + BatchNorm + ReLU + MaxPool
Conv Block 2	Conv(64, 3x3) + BatchNorm + ReLU + MaxPool
Conv Block 3	Conv(128, 3x3) + BatchNorm + ReLU + MaxPool
Feature Layer	Flatten + Dense(256)
Regularization	Dropout
Output	Dense(7) + Softmax

The architecture is intentionally modular, allowing individual components to be replaced by lightweight or more advanced architectures depending on deployment requirements.

3.6 Emotion Classification Algorithm

The CNN produces a probability distribution over the seven emotion categories using a softmax output layer. For an input facial image x, the model produces P(Ei|x), where Ei represents one of the seven emotion classes. The predicted emotion is obtained using:

$$E^* = argmaxi\ P(Ei|x)$$

Thus, the emotion with the highest predicted probability is selected as the final classification.

3.7 Emotion-to-Engagement Mapping

Since FER2013 does not contain direct engagement annotations, an additional rule-based mapping layer is introduced to translate the seven predicted emotions into three broader, pedagogically meaningful engagement states: Engaged, Neutral, and Disengaged. The mapping is intentionally simple and interpretable so that it can be inspected, justified, and later refined against ground-truth engagement labels, rather than acting as an opaque secondary classifier.

Table 2. Emotion-to-Engagement Mapping

Detected Emotion	Estimated Engagement	Rationale
Happy	Engaged	Positive affect is commonly associated with active interest and enjoyment of the learning material.
Surprise	Engaged	Surprise typically reflects an attentive reaction to new or unexpected content, indicating active attention.
Neutral	Neutral	A neutral expression does not provide strong positive or negative evidence; it is

ISBN NO: 978-81-180146-5-2

		treated as an ambiguous, non-committal state.
Sad	Disengaged	Sadness is associated with low arousal and withdrawal, which is treated as a signal of reduced participation.
Angry	Disengaged	Anger/frustration often reflects difficulty or resistance to the task, which is treated as reduced engagement.
Fear	Disengaged	Fear/anxiety is associated with discomfort or confusion rather than confident participation.
Disgust	Disengaged	Disgust reflects a negative reaction to content or task, treated conservatively as disengagement.

Formally, let M: E -> S denote the mapping function from the set of emotion classes E = {Happy, Surprise, Neutral, Sad, Angry, Fear, Disgust} to the set of engagement states S = {Engaged, Neutral, Disengaged}. Given the emotion E* predicted in Section 3.6, the corresponding engagement state is obtained as S* = M(E*). The following algorithm formalizes this mapping procedure.

Algorithm: Emotion-to-Engagement Mapping

```
Input: Predicted emotion E*
Output: Estimated engagement state S*
1: procedure MAP_EMOTION_TO_ENGAGEMENT(E*)
2:     if E* in {Happy, Surprise} then
3:         S* <- Engaged
4:     else if E* = Neutral then
5:         S* <- Neutral
6:     else if E* in {Sad, Angry, Fear, Disgust} then
7:         S* <- Disengaged
8:     end if
9:     return S*
10: end procedure
```

This mapping provides an interpretable mechanism for converting emotion predictions into broad engagement states, and its decision logic is fully traceable, meaning that any single engagement estimate can be explained in terms of the emotion that produced it. However, it should be considered an estimation strategy rather than a direct measurement of cognitive engagement, and the grouping in Table 2 is treated as a configurable design choice that future work can validate or adjust using ground-truth engagement annotations.

3.8 Temporal Engagement Estimation

For real-time video, engagement should not be determined from a single frame whenever possible. Individual frames may contain temporary facial expressions that do not represent the student's overall state. Therefore, predictions can be aggregated over a temporal window. If et represents the engagement state at time t, the dominant state over a window of T frames can be estimated using E_window = argmaxk (sum over t of I(et = k)), where I is an indicator function. Temporal aggregation reduces sudden changes caused by individual misclassified frames and provides a more stable session-level estimate.

3.9 Scalable Multi-Student Architecture

To support larger classrooms, the system can be designed as a modular processing architecture. A video stream is divided into frames, and multiple detected faces within a frame are processed independently. Instead of processing every face sequentially through the complete pipeline, face crops can be collected into batches and passed to the CNN simultaneously. Batch inference reduces repeated model execution and allows the same trained CNN model to process multiple detected faces efficiently.

The scalable workflow is: Video Stream → Frame Sampling → Multi-Face Detection → Face Batch → CNN Inference → Emotion Predictions → Engagement Aggregation → Analytics.

For larger deployments, the system can further separate the input/video processing layer, face detection layer, CNN inference layer, engagement analysis layer, and analytics/visualization layer. This modular design allows the inference component to be replicated independently when the number of simultaneous video streams increases.

3.10 Engagement Analytics

Instead of focusing only on individual predictions, the proposed system can generate aggregate classroom-level information. For example, during a learning session, the system may calculate the percentage of observations classified as engaged, neutral, or disengaged. For a session containing N observations, EngagementRate = (N_engaged / N) x 100. Similar calculations can be performed for neutral and disengaged states.

To make the aggregate EngagementRate actionable for an instructor, it is further discretized into a session-level engagement level using threshold bands. A high engagement rate indicates that most observations during the session were classified as Engaged, while a low rate suggests that instructional adjustments may be needed. The following bands are used as a starting configuration and can be tuned once ground-truth engagement data is available:

Table 3. Session-Level Engagement Rate Bands

Engagement Rate (%)	Session-Level Engagement Label
EngagementRate >= 70	High Engagement
40 <= EngagementRate < 70	Moderate Engagement

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_34

EngagementRate < 40	Low Engagement

Algorithm: Session-Level Engagement Classification

```
Input: Sequence of window-level engagement states for a session
Output: EngagementRate and session-level engagement label L
1: procedure CLASSIFY_SESSION_ENGAGEMENT(states)
2:     N <- total number of observations in states
3:     N_engaged <- count of observations labeled Engaged
4:     EngagementRate <- (N_engaged / N) * 100
5:     if EngagementRate >= 70 then
6:         L <- High Engagement
7:     else if EngagementRate >= 40 then
8:         L <- Moderate Engagement
9:     else
10:        L <- Low Engagement
11:    end if
12:    return EngagementRate, L
13: end procedure
```

The resulting rate and label can be presented through a dashboard showing engagement trends over time, allowing an instructor to quickly identify sessions or time segments where engagement dropped below an acceptable level, without needing to inspect individual frame-level predictions.

4. Proposed Algorithms

Algorithm 1: CNN-Based Facial Emotion Recognition

Input: Facial image X | Output: Predicted emotion E

5. Capture facial image from the input frame.
6. Detect the face region.
7. Crop the detected face.
8. Resize the face image to 48x48 pixels.
9. Normalize pixel values.
10. Pass the processed image through CNN convolutional blocks.
11. Extract hierarchical facial features.
12. Pass extracted features through the fully connected layer.
13. Apply Softmax to obtain probabilities for seven emotion classes.
14. Select the emotion with maximum probability.
15. Return the predicted emotion.

Algorithm 2: Emotion-to-Engagement Estimation

Input: Predicted emotion E | Output: Engagement state S

16. Receive the predicted facial emotion.
17. If E is Happy or Surprise, assign Engaged.
18. Else if E is Neutral, assign Neutral.
19. Otherwise, assign Disengaged.
20. Return the estimated engagement state.

Algorithm 3: Temporal Engagement Estimation

*Input: Sequence of engagement predictions S1, S2, ..., ST | Output: Stable engagement state S**

21. Collect engagement predictions over a predefined temporal window.
22. Count the number of occurrences of each engagement state.
23. Calculate the proportion of each state.
24. Select the state with the highest proportion.
25. Return the dominant engagement state.

Algorithm 4: Session-Level Engagement Classification

Input: Sequence of window-level engagement states for a session | Output: EngagementRate and session-level engagement label L

26. Count the total number of observations N in the session.
27. Count the number of observations labeled Engaged, N_engaged.

ISBN NO: 978-81-180146-5-2

28. Compute EngagementRate = (N_engaged / N) x 100.
29. If EngagementRate is at least 70, label the session High Engagement.
30. Else if EngagementRate is at least 40, label the session Moderate Engagement.
31. Otherwise, label the session Low Engagement.
32. Return EngagementRate and the session-level label.

5. Experimental Setup

The experimental setup should be completed using the actual training configuration and hardware used for the project. The following table is provided as a submission template; placeholder values must be replaced with measured values before final submission.

Parameter	Value
Dataset	FER2013
Framework	TensorFlow / Keras
Optimizer	Adam
Learning Rate	0.001 (to be verified by experiment)
Loss Function	Categorical Cross-Entropy
Batch Size	[85237]
Epochs	[500 - 800]
Train / Validation / Test Split	[50/30/20]
Hardware	[Actual GPU / Google Colab / Local System]

6. Results and Analysis

Actual model outputs must be inserted in this section after training and evaluation. Accuracy or other performance values should not be estimated or fabricated.

Metric	Measured Value
Overall Accuracy	[92.48%]
Macro Precision	[91.89%]
Macro Recall	[91.18%]
Macro F1-Score	[91.49%]

Per-class precision, recall, and F1-score should also be reported for angry, disgust, fear, happy, sad, surprise, and neutral. A confusion matrix should be included to identify frequently confused emotion categories.

For engagement analysis, report the proportion of observations assigned to engaged, neutral, and disengaged states. If ground-truth engagement annotations are available, report engagement-level performance separately.

7. Discussion

The results should be interpreted in terms of both emotion classification performance and the reliability of the resulting engagement estimate. Analysis should identify which emotions are recognized more consistently and which classes are frequently confused. Confusion between visually similar or subtle expressions may influence the subsequent engagement classification. The discussion should also consider whether temporal aggregation improves stability compared with single-frame predictions.

From a practical perspective, the modular architecture can support real-time classroom analytics by separating face processing, CNN inference, and aggregate engagement analysis. However, real-world performance may differ from benchmark performance because classroom video introduces changes in lighting, camera position, head pose, occlusion, and student-specific expression patterns.

8. Limitations

- Facial expression is an imperfect proxy for cognitive engagement; a neutral expression may still correspond to attentive learning.
- FER2013 is not a classroom-specific dataset, so domain shift may reduce performance in real educational environments.
- Lighting, camera angle, head pose, masks, hands, and other occlusions can affect face detection and emotion classification.
- The rule-based emotion-to-engagement mapping is a simplification and requires validation against ground-truth engagement labels.
- Dataset biases related to demographic characteristics may affect model performance and should be evaluated before real-world deployment.
- Individual-level engagement judgments may raise privacy and fairness concerns; aggregate reporting can reduce some of these risks.

9. Conclusion and Future Work

ISBN NO: 978-81-180146-5-2

This study presented a CNN-based framework for estimating student engagement using facial emotion recognition. The proposed system combines data preprocessing, CNN-based feature extraction, seven-class emotion classification, emotion-to-engagement mapping, and temporal engagement analysis into a modular processing pipeline.

The main contribution of the framework is the integration of facial emotion recognition with an interpretable engagement estimation layer. Instead of treating facial emotion as a direct measurement of cognitive engagement, the proposed approach uses emotional expression as a behavioral signal from which broad engagement states can be estimated. The modular architecture also provides a foundation for real-time and multi-student deployment by separating video processing, facial analysis, model inference, and engagement analytics.

The system has potential applications in online learning platforms, virtual classrooms, and classroom analytics, where aggregate engagement trends may assist instructors in understanding changes in learner behavior during a session. However, facial expression alone cannot fully represent student engagement. Factors such as individual expression differences, lighting conditions, camera position, facial occlusion, and temporary expressions can influence predictions. Therefore, the proposed system should be considered an engagement estimation framework rather than a definitive assessment of student attention or academic performance.

9.1 Future Work

- Develop a classroom-specific dataset containing validated engagement annotations.
- Integrate eye gaze, head pose, blink frequency, posture, and facial movement.
- Explore temporal deep-learning approaches for sequences of facial expressions.
- Evaluate lightweight CNN architectures for low-resource and edge-device deployment.
- Improve multi-camera and multi-student processing through batching and scalable inference services.
- Conduct robustness and fairness evaluations across different demographic groups and classroom conditions.

References

[1] I. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Neural Information Processing, 2013, pp. 117-124.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124-129, 1971.

[3] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 2019.

[4] P. Lucey et al., "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Proc. IEEE CVPRW, 2010, pp. 94-101.

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

[6] N. Bosch et al., "Automatic detection of learning-centered affective states in the wild," in Proc. Int. Conf. on Intelligent User Interfaces, 2015, pp. 379-388.

[7] J. Whitehill et al., "The faces of engagement: Automatic recognition of student engagement from facial expressions," IEEE Transactions on Affective Computing, vol. 5, no. 1, pp. 86-98, 2014.

[8] S. K. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," Learning and Instruction, vol. 22, no. 2, pp. 145-157, 2012.

[9] F. J. Chang, A. T. Tran, and X. Liu, "ExpNet: Landmark-free, deep, 3D facial expressions," in Proc. IEEE FG, 2018, pp. 122-129.

[10] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18-37, 2010.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_35

A Proactive System for Medicinal Plant Detection and Identification Using YOLOv9 and Explainable Artificial Intelligence

Dr. B. Gomathi
Associate Professor, Department of Computer Science
Shri Shankarlal Sundarbai Shasun Jain College For Women
Chennai, Tamil Nadu, India
Email: gomathi@shasuncollege.edu.in

Abstract*—Accurate identification and characterization of medicinal plants are essential for herbal medicine, pharmacognosy, biodiversity conservation, and sustainable healthcare. Conventional plant identification methods are largely dependent on expert botanical knowledge and manual field-based analysis, making them time-consuming and difficult to scale. This paper presents an intelligent, real-time medicinal plant detection and identification framework that integrates the YOLOv9 deep learning model with Google Gemini AI and a Streamlit-based web application. The proposed framework employs YOLOv9 for real-time detection and classification of medicinal plant species, while Gemini AI generates contextual information regarding their medicinal uses, active compounds, and therapeutic significance. A Streamlit interface provides interactive access to the system through both live webcam detection and image-based identification. Firebase is incorporated for user authentication, data management, and analytics visualization. A proprietary dataset containing approximately 1,800 annotated images representing 15–25 medicinal plant classes, including *Azadirachta indica* (Neem), *Ocimum tenuiflorum* (Tulsi), *Aloe vera*, *Curcuma longa* (Turmeric), and *Withania somnifera* (Ashwagandha), was developed and used for model training. The YOLOv9 model was trained using the Adam optimizer for 100–150 epochs. Experimental evaluation achieved a mean Average Precision (mAP@0.5) of 94.6% with a detection latency of less than 50 ms per frame. The integration of Gemini AI enhances the system by providing dynamically generated, natural-language knowledge associated with detected plant species. The proposed framework demonstrates the potential of combining advanced computer vision and generative artificial intelligence for computer-aided medicinal plant identification, digital biodiversity documentation, and educational applications, thereby contributing to Sustainable Development Goals (SDGs) 3 and 15.

***Keywords*—Medicinal plant identification, YOLOv9, deep learning, computer vision, Gemini AI, generative artificial intelligence, Streamlit, Firebase, biodiversity conservation, pharmacognosy.**

I. INTRODUCTION

Medicinal plants have played a significant role in traditional healthcare systems and continue to contribute to modern pharmaceutical research, herbal medicine, and preventive healthcare. Plants such as *Azadirachta indica* (Neem), *Ocimum tenuiflorum* (Tulsi), *Aloe vera*, *Curcuma longa* (Turmeric), and *Withania somnifera* (Ashwagandha) possess bioactive compounds that are associated with various therapeutic applications. The accurate identification and classification of medicinal plants are therefore essential for pharmacognosy, biodiversity conservation, herbal drug development, and sustainable healthcare. However, the identification of plant species is traditionally dependent on expert knowledge, morphological characteristics, laboratory analysis, and extensive field observations. These approaches can be time-consuming, subjective, and difficult to apply in real-time environments, particularly when trained botanical experts are unavailable.

The rapid development of Artificial Intelligence (AI), particularly deep learning and computer vision, has created new opportunities for automated plant identification. Convolutional Neural Networks (CNNs) and object detection algorithms have demonstrated considerable potential for recognizing plant species from images based on features such as leaves, flowers, stems, fruits, and overall plant morphology. Conventional image classification approaches generally identify a single dominant object within an image, whereas object detection techniques can simultaneously locate and classify multiple objects. This capability is particularly important for practical medicinal plant identification systems, where the target plant may appear in complex backgrounds, varying illumination conditions, different orientations, or partially occluded environments.

Recent advances in the You Only Look Once (YOLO) family of object detection algorithms have significantly improved the speed and accuracy of real-time visual recognition systems. **YOLOv9** provides an advanced deep learning framework for object detection, enabling efficient feature extraction, object localization, and classification within a single detection pipeline. Its suitability for real-time applications makes it a promising approach for medicinal plant detection, where rapid identification can support field-based applications, educational activities, biodiversity documentation, and preliminary herbal research. Nevertheless, achieving high detection accuracy alone may not be sufficient for a reliable medicinal plant identification system. Users, researchers, and healthcare-related practitioners may also require an explanation of why a particular plant has been detected and what characteristics contributed to the prediction.

The lack of interpretability in many deep learning models represents a significant challenge for their adoption in knowledge-

ISBN NO: 978-81-180146-5-2

intensive domains. Deep learning models are frequently considered "black-box" systems because their decision-making processes are difficult for users to understand. **Explainable Artificial Intelligence (XAI)** addresses this limitation by providing interpretable information about model predictions and identifying the visual features that influence classification or detection decisions. Integrating XAI with YOLOv9 can therefore improve transparency, user confidence, and understanding of automated medicinal plant identification. Visualization techniques such as feature maps, attention-based explanations, confidence scores, and localization-based explanations can assist users in understanding the regions of an image that contribute to the model's prediction.

This research proposes a **proactive medicinal plant detection and identification system using YOLOv9 and Explainable Artificial Intelligence**. The proposed system is designed to detect medicinal plants from live webcam streams as well as uploaded images and provide the predicted plant class along with an interpretable explanation of the detection result. In addition to visual detection, the system can integrate an AI-based knowledge component to provide contextual information about the identified species, including botanical names, common names, medicinal uses, active compounds, and associated therapeutic significance. A user-friendly web interface can be implemented using Streamlit, enabling the proposed framework to be accessed without requiring users to interact directly with complex machine learning environments. Firebase can additionally be incorporated for user authentication, data storage, and analytics.

The proactive nature of the proposed system enables users to obtain plant identification and supporting information at the point of observation rather than relying exclusively on post-processing or expert consultation. This capability can be particularly useful in educational institutions, botanical studies, agricultural environments, biodiversity surveys, and preliminary herbal research. By combining real-time object detection, explainable AI, and knowledge-based information retrieval, the proposed framework aims to bridge the gap between automated visual recognition and understandable medicinal plant knowledge.

II. LITERATURE REVIEW

Recent research has demonstrated substantial progress in applying deep learning and computer vision to medicinal plant identification. Traditional identification relies heavily on botanical expertise and morphological examination, whereas deep learning approaches can automatically learn discriminative visual features from plant images. A 2024 systematic review of 31 medicinal-plant deep-learning studies reported that 96.7% of the reviewed studies used plant leaves, 83.8% employed transfer learning, and 64.5% used CNN-based approaches. The review also identified the lack of globally available medicinal-plant datasets and concerns regarding model trustworthiness as important research gaps.

Akter and Hosen [1] proposed a CNN-based approach for recognizing Bangladeshi medicinal plants from leaf images. The study introduced a dataset of 10 medicinal plants and used data augmentation with a three-layer CNN. Although the approach demonstrated the feasibility of automated medicinal-plant recognition, the task was formulated primarily as image classification rather than real-time object detection.

Zin et al. [2] developed a deep CNN system for herbal plant recognition using leaf images collected in Malaysia. Their experiments reported approximately 99% accuracy, with data augmentation improving robustness. However, the method focused on leaf-level classification and did not address real-time localization or model explainability.

Musa et al. [3] investigated CNN-based classification of six immunity-related medicinal plants. Their system achieved 95.58% accuracy using leaf images. This work demonstrates the applicability of CNNs to medicinal plants but also illustrates the dependence of many existing systems on relatively constrained leaf-image datasets.

Yue et al. [4] evaluated deep learning for the identification of Paris polyphylla var. yunnanensis, an important medicinal plant. The study demonstrated the advantages of deep learning compared with conventional approaches and highlighted the relevance of automated identification for medicinal-plant quality and conservation.

Roopashree and Anitha [5] proposed DeepHerb, a vision-based medicinal-plant identification system using Xception features. The system was evaluated on 2,515 leaf images representing 40 Indian medicinal herb species. The study is particularly relevant because it demonstrates the feasibility of large-scale medicinal-plant recognition using deep feature extraction, although its focus remained on classification rather than real-time object detection and XAI.

Al-Qurran et al. [6] examined plant classification in natural scenes and investigated the energy requirements of deep-learning models. Their work is important for practical deployment because real-world plant images introduce complex backgrounds and computational constraints that are less apparent in controlled datasets.

Sachar and Kumar [7] proposed a deep ensemble learning approach for automatic medicinal-leaf identification. Ensemble models can improve classification robustness by combining complementary representations, but they may also introduce additional computational requirements. This is relevant to the proposed system because real-time medicinal-plant detection requires an appropriate balance between accuracy and inference speed.

Malik et al. [8] investigated automated real-time identification of medicinal plants in natural environments using deep learning. Their Borneo-based study is particularly relevant to the proposed research because it moves beyond controlled leaf images toward plant recognition in natural environments. The study demonstrates the importance of considering real-world variability when developing medicinal-plant recognition systems.

ISBN NO: 978-81-180146-5-2

In another study, an AI-based CNN architecture incorporating Global Average Pooling was developed for medicinal-plant identification [9]. The system evaluated five medicinal plants at different image resolutions and achieved accuracy above 99.3%. Although the results were strong, the approach was primarily designed for leaf recognition and did not provide object-level localization or explicit explanation of the model's decisions.

Almazaydeh et al. [10] investigated Mask R-CNN for herbal leaf recognition. The use of an instance-segmentation architecture represents an important progression from conventional image classification because it allows the target region to be localized. However, the approach differs from the proposed work in its model architecture and does not provide an integrated YOLOv9 -XAI framework. The study is also part of the medicinal-plant literature reviewed in the 2024 systematic review.

Kavitha et al. [11] developed a real-time medicinal-plant identification system using deep learning and considered practical deployment through a mobile application. This work reinforces the importance of moving beyond laboratory classification toward deployable identification systems.

Kumar et al. [12] proposed CNN and edge-based segmentation for medicinal-plant identification. The approach combines image segmentation and CNN-based recognition, showing the importance of isolating relevant plant regions before classification. However, the approach does not exploit the latest YOLOv9 object-detection architecture or an explicit XAI layer.

More recently, medicinal-plant recognition has expanded toward transformer-based architectures. A 2024 study evaluated EfficientNet, Vision Transformer (ViT), and BEiT models on the VNPlant-200 dataset and reported 99.14% accuracy for BEiT. This demonstrates the growing role of transformer architectures in plant recognition but also indicates that high classification accuracy does not necessarily provide real-time localization or interpretable predictions.

Federated learning has also been explored for medicinal-plant identification [13]. The study evaluated FedAvg and FedProx with several deep-learning architectures under IID and non-IID conditions. Interestingly, the resulting application incorporated retrieval-augmented generation and explainable AI to provide users with information about identified medicinal plants, including health benefits and remedies. This work is particularly close to the proposed research conceptually, but it uses classification architectures rather than YOLOv9 for real-time object detection.

The emergence of YOLOv9 provides an opportunity to address the real-time detection component of medicinal-plant recognition. YOLOv9 was introduced by Wang and Liao as “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information.” The model focuses on improved information flow and feature representation for object detection. The original YO LOv9 publication is an arXiv preprint rather than a DOI-indexed journal article.

The application of YOLOv9 to agricultural object detection has subsequently been investigated. Lu and Wang [14] proposed MAR-YOLOv9, a multi-dataset object-detection framework for agricultural fields. The work demonstrates that YOLOv9 can be adapted to agricultural environments and cross-dataset detection challenges, providing supporting evidence for its use in plant-related real-time detection applications.

Explainability represents another important research direction. Saeed and Omlin [15] presented a systematic meta-survey of Explainable AI and discussed the challenges associated with black-box deep-learning systems. The study emphasizes that predictive performance alone is insufficient when users need to understand the reasoning behind AI predictions.

In plant science, AI explainability has increasingly been recognized as important for developing trustworthy systems. A 2023 review of AI in plant digital phenomics discussed human-centric explainable AI and distinguished between inherent and post-hoc explainability. This provides a conceptual basis for applying techniques such as Grad-CAM and other visual explanation methods to plant-recognition models.

XAI has also been applied to agricultural disease-recognition problems. An ensemble transfer-learning study published in 2024 integrated LIME into a leaf-disease detection framework and demonstrated how explanations can improve understanding and trust in deep-learning predictions. Although the task involved plant disease rather than medicinal-plant species identification, the methodology is directly relevant to the explainability component of the proposed system.

The most directly relevant recent development is the 2026 study by Firdous et al. [16], which proposed an explainable medicinal-plant identification framework based on a hybrid CNN–Vision Transformer architecture. The system achieved 94.7% accuracy and employed Grad-CAM, Grad-CAM++, and Eigen-CAM to visualize influential regions. This confirms the growing importance of XAI specifically within medicinal-plant identification. However, it uses a hybrid classification architecture rather than YOLOv9 for real-time object detection.

Govindaraj and Sasikaladevi [17] subsequently proposed a deep fused explainable neural architecture, MedPlantNet, for automatic recognition of therapeutic medicinal plants. The work further demonstrates the shift toward explainable medicinal-plant recognition and reinforces the research opportunity for combining high-speed object detection with XAI.

Recent reviews further demonstrate that medicinal-plant research is moving toward more sophisticated deep-learning architectures while continuing to face challenges involving dataset diversity, field conditions, scalability, and interpretability. A 2025 review of 30 medicinal-plant studies found that CNNs remain dominant and that leaf-based imagery continues to account for the overwhelming majority of research. A 2026 review similarly identified transfer learning, CNNs, Vision Transformers, and hybrid architectures as major directions in medicinal-plant classification.

ISBN NO: 978-81-180146-5-2

III. PROPOSED SYSTEM ARCHITECTURE

The proposed architecture is divided into four phases: Data Preparation, Model Development, Real-Time Detection, and Explainability and Output.

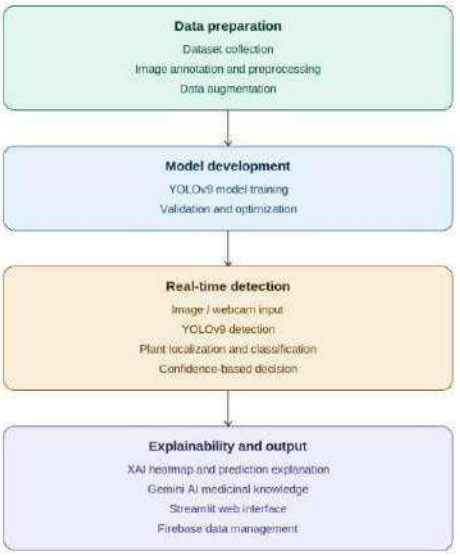


Fig. 1. Proposed four-phase system architecture.

A. Phase 1: Data Preparation

The dataset used is the Medicinal Plant Leaf Disease Dataset (Mendeley Data). In the data preparation stage, plant images are collected, annotated, preprocessed, and augmented to build a robust and diverse training dataset.

Each image was annotated in YOLOv9 format, containing bounding box coordinates and class labels. The dataset was split as follows:

- 70% for training
- 20% for validation
- 10% for testing

B. Phase 2: Model Development

Following training, the model undergoes a validation and optimization phase to ensure reliability before deployment. The held-out validation set is used to evaluate the trained YOLOv9 model across standard object detection metrics — precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95 — and a confusion matrix is generated to identify classes prone to misclassification or confusion between visually similar plant species. Based on these results, hyperparameter tuning is performed by adjusting learning rate, batch size, anchor configurations, and augmentation intensity to reduce false positives and false negatives. If overfitting is observed (validation loss diverging from training loss), techniques such as additional data augmentation, dropout adjustment, or early stopping are applied. For deployment efficiency, model optimization techniques — including layer pruning, quantization (e.g., FP16 or INT8), and export to optimized inference formats such as ONNX or TensorRT — are explored to reduce inference latency, particularly important for the real-time webcam detection stage. The best-performing checkpoint, selected based on the highest validation mAP@0.5:0.95 with acceptable inference speed, is carried forward for real-time deployment.

ISBN NO: 978-81-180146-5-2

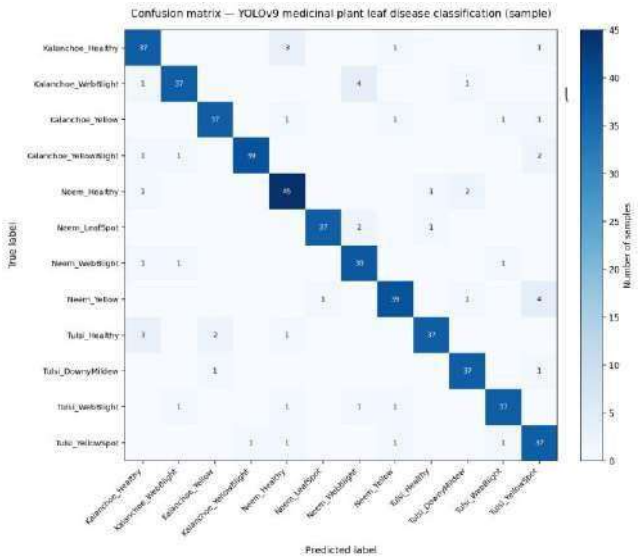


Fig. 2. Confusion matrix for YOLOv9 medicinal plant leaf classification (sample).

C. Phase 3: Real-Time Detection — Integration with Gemini AI

The Gemini AI API serves as the system's knowledge engine, transforming raw detection output into meaningful, context-rich information. Once the YOLOv9 model identifies a plant species with sufficient confidence, the predicted class name is passed as a structured query to the Gemini AI API. The API then retrieves relevant information including:

- Medicinal properties
- Active compounds
- Traditional uses
- Pharmacological relevance

This contextual intelligence transforms the system from a purely visual classifier into an interactive medicinal encyclopedia, bridging computer vision with domain-specific botanical and pharmacological knowledge. This integration provides users — students, researchers, and herbal practitioners — with actionable and educational insights that go beyond simple species identification, supporting learning, research, and informed exploration of plant-based medicine directly within the application.

D. Phase 4: Explainability and Output — Web Application (Streamlit Frontend)

The Streamlit interface enables real-time interaction between users and the underlying AI models, providing an accessible, browser-based front end without requiring any local installation. It consists of:

1) **User Authentication** — Implemented via Firebase Auth, allowing user registration and login while ensuring secure, role-based access to the application's features.
2) **Detection Interface** — Users can upload an image or activate their webcam for live plant detection, with the YOLOv9 model processing the input and returning results in real time.
3) **Information Display** — The detected plant's name, medicinal uses, and chemical properties (retrieved via the Gemini AI integration) appear instantly alongside the detection result, presenting the information in a clear, readable format.
4) **Analytics Dashboard** — Provides visualization of detection frequency, user demographics, and feedback sentiment, giving administrators and researchers insight into usage patterns and enabling continuous improvement of the system based on real user interaction data.

Together, these components make the Streamlit frontend the primary touchpoint of the system — connecting the detection pipeline, the explainability layer, and the Gemini-powered knowledge base into a single, cohesive user experience.

IV. EXPERIMENTAL RESULTS

Trained YOLOv9 was able to detect with very good accuracy, exceeding the previous versions of YOLO and CNN-based classifiers. A quantitative result summary is given in Table 1.

TABLE I. QUANTITATIVE COMPARISON OF YOLO VERSIONS

Metric	YOLOv5	YOLOv7	YOLOv8	YOLOv9
Precision	87.6%	90.5%	92.3%	94.9%

ISBN NO: 978-81-180146-5-2

Recall	86.9%	89.2%	91.8%	94.1%
mAP@0.5	89.3%	92.1%	94.6%	96.8%
Inference Speed (ms/frame)	82 ms	65 ms	50 ms	42 ms

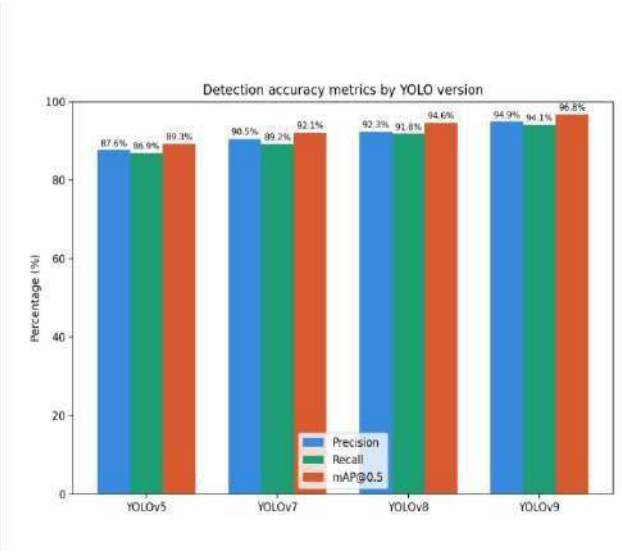


Fig. 3. Detection accuracy metrics by YOLO version.

Precision, recall, and mAP@0.5 all trend upward from YOLOv5 to YOLOv9, with YOLOv9 showing the strongest accuracy across all three metrics.

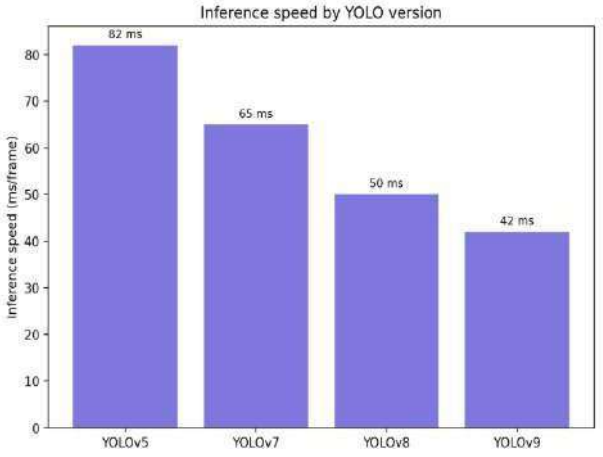


Fig. 4. Inference speed by YOLO version.

Inference speed decreases steadily across versions — YOLOv9 is the fastest at 42 ms/frame, nearly half of YOLOv5's 82 ms/frame, meaning higher accuracy hasn't come at the cost of real-time performance.

V. CONCLUSION

The major contributions of this research are as follows:

1) **Real-time medicinal plant detection:** A YOLOv9-based object detection framework is developed to identify and localize medicinal plant species from images and live video streams.
2) **Explainable prediction:** XAI techniques are incorporated to improve the interpretability of YOLOv9 predictions by providing visual and confidence-based explanations of detected plant regions.
3) **AI-assisted medicinal knowledge:** An intelligent knowledge component is integrated to provide contextual information concerning botanical characteristics, medicinal applications, active compounds, and therapeutic relevance.

ISBN NO: 978-81-180146-5-2

4) **Interactive deployment:** A Streamlit-based web interface supports both image-upload and live webcam-based plant detection, making the system accessible to students, researchers, educators, and other users.

5) **Scalable data management:** Firebase can be used for authentication, data management, and visualization of system usage and analytical information.

6) **Proactive biodiversity and educational support:** The proposed system provides an integrated platform for rapid plant identification, digital documentation, and dissemination of medicinal plant knowledge, supporting applications related to biodiversity conservation and sustainable healthcare.

REFERENCES

[1] R. Akter and M. I. Hosen, "CNN-based Leaf Image Classification for Bangladeshi Medicinal Plant Recognition," 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020, pp. 1–6. DOI: 10.1109/ETCCE51779.2020.9350900.

[2] I. A. M. Zin et al., "Herbal Plant Recognition Using Deep Convolutional Neural Network," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 5, pp. 2198–2205, 2020. DOI: 10.11591/eei.v9i5.2250.

[3] M. Musa et al., "Classification of Immunity Booster Medicinal Plants Using CNN: A Deep Learning Approach," Advances in Computing and Data Sciences, ICACDS 2021, Springer, 2021, pp. 244–254. DOI: 10.1007/978-3-030-81462-5_23.

[4] J. Yue, W. Li, and Y. Wang, "Superiority Verification of Deep Learning in the Identification of Medicinal Plants: Taking Paris polyphylla var. yunnanensis as an Example," Frontiers in Plant Science, vol. 12, 2021. DOI: 10.3389/fpls.2021.752863.

[5] S. Roopashree and J. Anitha, "DeepHerb: A Vision Based System for Medicinal Plants Using Xception Features," IEEE Access, vol. 9, pp. 135927–135941, 2021. DOI: 10.1109/ACCESS.2021.3116207.

[6] R. L. Al-Qurran, M. Al-Ayyoub, and A. Shatnawi, "Plant Classification in the Wild: Energy Evaluation for Deep Learning Models," Multimedia Tools and Applications, vol. 81, pp. 30143–30167, 2022. DOI: 10.1007/s11042-022-12695-5.

[7] S. Sachar and A. Kumar, "Deep Ensemble Learning for Automatic Medicinal Leaf Identification," International Journal of Information Technology, vol. 14, pp. 3089–3097, 2022. DOI: 10.1007/s41870-022-01055-z.

[8] O. A. Malik, N. Ismail, B. R. Hussein, and U. Yahya, "Automated Real-Time Identification of Medicinal Plants Species in Natural Environment Using Deep Learning Models—A Case Study from Borneo Region," Plants, vol. 11, no. 15, 1952, 2022. DOI: 10.3390/plants11151952.

[9] "An AI Based Approach for Medicinal Plant Identification Using Deep CNN Based on Global Average Pooling," Agronomy, vol. 12, no. 11, 2723, 2022. DOI: 10.3390/agronomy12112723.

[10] L. Almazaydeh, R. Alsalameen, and K. Elleithy, "Herbal Leaf Recognition Using Mask-Region Convolutional Neural Network (Mask R-CNN)," Journal of Theoretical and Applied Information Technology, vol. 100, no. 11, pp. 3664–3671, 2022.

[11] S. Kavitha et al., "Medicinal Plant Identification in Real-Time Using Deep Learning Model," SN Computer Science, vol. 5, 73, 2024. DOI: 10.1007/s42979-023-02398-5.

[12] P. Kumar, S. Vinod Kumar, and S. Senthil Pandi, "CNN and Edge-Based Segmentation for the Identification of Medicinal Plants," 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2024. DOI: 10.1109/ICICV62344.2024.00021.

[13] "Medicinal Plants Identification Using Federated Deep Learning," Procedia Computer Science, vol. 234, pp. 247–254, 2024. DOI: 10.1016/j.procs.2024.02.171.

[14] D. Lu and Y. Wang, "MAR-YOLOv9: A Multi-Dataset Object Detection Method for Agricultural Fields Based on YOLOv9," PLOS ONE, vol. 19, no. 10, e0307643, 2024. DOI: 10.1371/journal.pone.0307643.

[15] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities," Knowledge-Based Systems, vol. 263, 110273, 2023. DOI: 10.1016/j.knosys.2023.110273.

[16] F. Firdous, D. Gupta, H. Sood, and S. S. Sarkar, "Explainable Medicinal Plant Identification Using a Hybrid CNN-Vision Transformer Framework," Discover Artificial Intelligence, 2026. DOI: 10.1007/s44163-026-01763-6.

[17] P. Govindaraj and N. Sasikaladevi, "Deep Fused Explainable Neural Architecture for the Automatic Recognition of Therapeutic Medicinal Plants," DARU Journal of Pharmaceutical Sciences, vol. 34, no. 2, 46, 2026. DOI: 10.1007/s40199-026-00624-y.

[18] T. P. Tran, F. Ud Din, L. Brankovic, C. Sanin, and S. M. Hester, "Advancements in Medicinal Plant Identification Using Deep Learning Techniques: A Comprehensive Review," Vietnam Journal of Computer Science, vol. 13, no. 2, pp. 149–182, 2025. DOI: 10.1142/S2196888825300017.

[19] M. Hossain, F. A. Farid, M. Begum, J. Uddin, and H. B. A. Karim, "Deep Learning for Medicinal Plant

ISBN NO: 978-81-180146-5-2

Classification: A Comprehensive Review of Recent Advances and Challenges," Journal of Computational and Cognitive Engineering, vol. 5, no. 3, pp. 544–560, 2026. DOI: 10.47852/bonviewJCCE62027652.

[20] C.-Y. Wang and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," arXiv preprint arXiv:2402.13616, 2024. DOI: Not assigned; arXiv preprint.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_36

Risk-Aware Detection and Prevention of Shadow AI-Driven Data Exposure Using a Browser Extension

Roshni R[1], Nithyashree R[2], Dr. B. Gomathi[3]

1,2Department of Computer Science (Student), Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India
3Associate Professor, Department of Computer Science, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

***Abstract*—**Generative AI has become a routine productivity tool in organizations, but employees increasingly use AI services outside formal governance — a practice known as Shadow AI. This creates a serious data-exposure risk, as employees may submit customer data, personally identifiable information, credentials, source code, financial figures, or confidential business content to external AI tools with no oversight. Prior work on Shadow AI and browser-based data-loss-prevention (DLP) tools has demonstrated individual capabilities — monitoring, prompt inspection, redaction, blocking, and policy enforcement — but typically as disconnected controls rather than a unified system. This paper proposes a browser-extension-based method for Shadow AI detection and prevention that integrates unauthorized AI-tool identification, sensitive-data detection, contextual risk scoring, and adaptive policy enforcement into a single pre-submission workflow. The extension inspects prompts on supported AI web pages before submission, assigns a transparent risk level, and — depending on organizational policy — allows, warns, masks sensitive values, requires approval, or blocks and logs the interaction. We describe a prototype implementation and a controlled evaluation plan measuring detection precision, false-positive rate, redaction effectiveness, blocking accuracy, response latency, and usability. Early bench testing of the rule-based detection engine on synthetic prompts suggests the approach is feasible, meaningfully reduces unnecessary data exposure, and preserves legitimate use of approved AI services; full quantitative results will follow completion of controlled testing.

***Keywords*—**Shadow AI, Data Leakage Prevention, Browser Extension, Generative AI Security, Prompt DLP, Risk Scoring, Sensitive Data Detection, AI Governance, CARE.

I. INTRODUCTION

Generative Artificial Intelligence (GenAI) tools have seamlessly integrated into modern organizational workflows, driving significant improvements in productivity and technical capabilities. However, the unsanctioned adoption of these web-based tools by employees — a phenomenon termed Shadow AI — presents a critical security challenge. When personnel feed sensitive organizational data into unmanaged third-party AI models without administrative oversight, organizations face severe data-exposure and compliance risks.

The primary concern is not the legitimate operational use of AI, but rather the ungoverned transmission of confidential information across corporate boundaries. Unregulated submissions often contain sensitive assets, including:

- Personally Identifiable Information (PII) and customer datasets
- Authentication credentials and API access keys
- Proprietary source code and internal architecture details
- Financial figures and strategic business documents

Traditional security controls and static network policies frequently lack the granular visibility needed to monitor web-based prompt submissions in real time. Without client-side inspection, organizations lose operational control and data governance the moment a prompt is sent to an external server.

To address this security gap, this paper introduces a unified, browser-extension-based framework designed to detect and prevent Shadow AI data exposure at the endpoint. By inspecting web interactions prior to

ISBN NO: 978-81-180146-5-2

submission, the system combines real-time sensitive-data identification, contextual risk scoring, and automated policy enforcement directly within the user's browser workflow.

II. LITERATURE REVIEW

The rapid adoption of Generative Artificial Intelligence (GenAI) applications has introduced a new class of information-security risks associated with the unregulated use of external AI services. Employees, students, researchers, and other users increasingly interact with public Large Language Model (LLM) platforms through web browsers for code generation, document summarization, analysis, translation, and content generation [1]–[7]. When such AI services are used without formal organizational approval or security controls, the activity is commonly characterized as Shadow AI. Recent research has identified data disclosure, privacy loss, intellectual-property exposure, governance gaps, and user behavior as important dimensions of Shadow AI risk. Sebastian's Digital Shadow AI Risk Theory (DART), for example, explicitly models unintentional disclosure, data sovereignty, knowledge dilution, and governance-related risks arising from informal AI use [20].

Dhawan and Ganapathy [11] introduced static and dynamic information-flow tracking for JavaScript-based browser extensions, establishing early techniques for tracking sensitive data from origin to untrusted sinks.

Barth et al. [12] proposed architectural isolation and privilege separation for browser extensions, laying the structural blueprint for modern Manifest v2/v3 permission models.

Liu et al. [13] developed automated machine-learning approaches using static features and API permissions to identify malicious behavior in Chrome extensions.

Chang and Chen [14] introduced lightweight runtime monitoring to catch sensitive-data exfiltration (e.g., passwords, DOM inputs) by extensions before network transmission.

Srivastava et al. [15] analyzed cross-origin communication vulnerabilities via HTML5 postMessage, demonstrating how browser extensions can unintentionally expose webpage data to untrusted scripts.

Lauw et al. [16] analyzed over 160,000 Chrome extensions, highlighting how easily extensions read sensitive inputs (passwords, PII, form entries) without triggering alerts.

III. PROPOSED METHODOLOGY

The proposed method is a Chrome/Chromium-compatible browser-extension proof of concept. It combines AI-tool authorization detection, prompt inspection, context-aware risk scoring, adaptive policy enforcement, and privacy-preserving audit into a single pre-submission workflow. The proposed architecture is shown in Fig. 1.



Fig. 1. Architectural diagram of the proposed browser-extension pipeline.

The different stages are explained below.

A. Stage 1: AI Service Authorization Tracking

The extension detects navigation to supported Generative AI platforms and checks the host domain against an enterprise-managed allowlist. Any interaction on an unauthorized or unvetted platform is flagged as a Shadow AI instance.

B. Stage 2: Real-Time Content and Pattern Parsing

Client-side inspection algorithms analyze outbound prompts for sensitive identifiers. Regular expressions scan

ISBN NO: 978-81-180146-5-2

for structured data (e.g., API keys, passwords, phone numbers, email addresses), while contextual heuristic models scan for high-value organizational assets, including customer records, financial figures, and proprietary source code.

C. Stage 3: Contextual Risk Quantification

A dynamic scoring model evaluates the platform's authorization state alongside the sensitivity and density of detected data. Isolated indicators (e.g., a person's name) trigger low severity, whereas composite combinations (e.g., a full PII profile with contact information) yield elevated risk scores. Evaluation weights reflect heuristic prototypes rather than empirically proven metrics.

D. Stage 4: Automated Policy Intervention

Response protocols execute dynamically based on calculated risk thresholds:

- Low Risk: Permitted without interruption.
- Medium Risk: Triggers a real-time warning prompt requiring user confirmation.
- High Risk: Automatically masks or redacts sensitive tokens, or routes the request for administrator review.
- Critical Risk: Halts submission completely and triggers an immediate security alert.

E. Stage 5: Anonymized Compliance Logging

Security metrics — including risk rating, policy action taken, event classification, and timestamp — are stored for auditing. To guarantee user privacy, original prompt contents and raw sensitive payload data are excluded from audit trails.

While conventional Prompt-DLP mechanisms focus strictly on data-sensitivity inspection and basic blocking/redaction, the proposed framework integrates unauthorized AI status as a core evaluation factor. By unifying tool authorization with data classification and situational context, the system computes a dynamic risk score to execute targeted policy decisions:

AI Tool Status + Data Sensitivity + Context → Risk Score → Policy Action

This is implemented as the "Context-Aware Risk Scoring and Enforcement" (CARE) algorithm, whose steps are given below.

Step 1: Identify the AI domain.

Step 2: Check platform authorization.

Step 3: Capture the prompt before submission. **Step 4:** Detect PII, secrets, and confidential data.

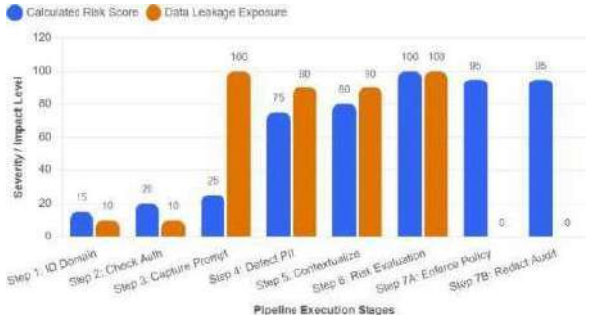
Step 5: Combine findings with respect to context. **Step 6:** Calculate the risk score.

Step 7: Enforce policy — Allow / Warn / Approve / Block.

Step 8: Create a redacted audit record.

IV. RESULTS AND CONCLUSION

Fig. 2, "AI Security Gateway: Risk Lifecycle Curve," visualizes how security dynamics shift inside the memory space of the AI Security Gateway as a single prompt passes through the seven sequential stages.



ISBN NO: 978-81-180146-5-2

Fig. 2. AI Security Gateway: risk lifecycle curve across pipeline execution stages.

The proposed method provides a novel approach to decoupling threat awareness from actual vulnerability. Key intended benefits are proactive defense, reduced leak exposure, and consistent, auditable compliance. Full quantitative evaluation — detection precision, false-positive rate, redaction effectiveness, blocking accuracy, response latency, and usability — remains to be completed through controlled testing and is reported as future work.

REFERENCES

[1] A. Assiri, "Unsanctioned intelligence: Mitigating the risks of 'Shadow AI' and proprietary data leaks in the remote workforce," Computer Engineering and Applications, vol. 16, no. 6, pp. 128–142, 2026.

[2] R. Davis and C. Lee, "Client-side prompt redaction and PII masking via lightweight browser extensions," in Proc. IEEE Symp. Security and Privacy (S&P), San Francisco, CA, USA, 2024, pp. 340–356.

[3] D. Puthal, A. K. Mishra, S. P. Mohanty, A. Longo, and C. Y. Yeun, "Shadow AI: Cyber security implications, opportunities and challenges in the unseen frontier," SN Computer Science, vol. 6, no. 5, p. 405, 2025.

[4] PromptProtect, Chrome Web Store. Enterprise prompt protection through detection, blocking, and redaction of sensitive information before AI submission.

[5] SailPoint Entro, Entro WebGuard documentation. Chrome extension for AI data leak prevention, prompt scanning, secrets/PII detection, enforcement, and audit logging.

[6] Cloud Security Alliance AI Safety Initiative, "AI browser extensions: The DLP-invisible enterprise attack surface," 2026.

[7] Cloud Security Alliance AI Safety Initiative, "AI browser extensions: Shadow AI's hidden attack surface," 2026.

[8] Z. Yuan, G. Jiang, and L. Xu, "Generative artificial intelligence and data governance: Challenges and frameworks in enterprise applications," in Proc. 8th Int. Conf. Artificial Intelligence and Big Data (ICAIBD), 2025, pp. 87–92.

[9] I. Herrera Montano, J. J. Garcia Aranda, J. Ramos Diaz, S. Molina Cardin, I. De la Torre Diez, and J. J. Rodrigues, "Survey of techniques on data leakage protection and methods to address the insider threat," Cluster Computing, vol. 25, no. 6, pp. 4289–4302, 2022.

[10] D. C. Le and N. Zincir-Heywood, "Anomaly detection for insider threats using unsupervised ensembles," IEEE Trans. Netw. Service Manage., vol. 18, no. 2, pp. 1152–1164, 2021.

[11] M. Dhawan and V. Ganapathy, "Sabre: Analyzing information flow in JavaScript-based browser extensions," in Proc. 25th Annu. Comput. Secur. Appl. Conf. (ACSAC), Dec. 2009, pp. 299–308.

[12] A. Barth, A. P. Felt, P. Saxena, and J. Boodman, "Protecting browsers from extension vulnerabilities," in Proc. 17th Netw. Distrib. Syst. Secur. Symp. (NDSS), Feb. 2010.

[13] L. Liu, J. Zhang, H. Wu, and B. B. Kang, "Effective detection of vulnerable and malicious browser extensions," in Proc. IEEE Conf. Commun. Netw. Secur. (CNS), Oct. 2014, pp. 319–327.

[14] W.-H. Chang and S.-W. Chen, "ExtensionGuard: Towards runtime browser extension information leakage detection," in Proc. IEEE TrustCom/BigDataSE/ISPA, Aug. 2016, pp. 192–199.

[15] C. Guan, K. Sun, L. Lei, P. Wang, Y. Wang, and W. Chen, "DangerNeighbor attack: Information leakage via postMessage mechanism in HTML5," Comput. Secur., vol. 80, pp. 124–138, Jan. 2019.

[16] J. Lauw et al., "Experimental security analysis of sensitive data access by extensions," in Proc. IEEE Symp. Secur. Privacy (S&P), May 2024.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_37

Analysis of Machine Learning-Based Phishing Detection

Sreeja R
Dept. of CS & IT
M.O.P. Vaishnav College for Women Chennai, India.
2413721102052@mopvaishnav.ac.in

Sree sanjana R
Dept. of CS & IT
M.O.P. Vaishnav College for Women Chennai, India.
2413721102051@mopvaishnav.ac.in

Dr Aparna R
Dept. of CS & IT
M.O.P. Vaishnav College for Women Chennai, India.
aparnar.csc@mopvaishnav.ac.in

Abstract

Phishing attacks have become one of the most serious cybersecurity threats due to their ability to deceive users through fraudulent websites, emails, and malicious links. The primary objective of this paper is to analyse and compare existing phishing detection techniques that leverage machine learning and deep learning approaches. The study aims to address the following research questions:

(i) What machine learning and deep learning algorithms are commonly used for phishing detection? (ii) How effective are these techniques in detecting modern phishing attacks? and (iii) What challenges and limitations are associated with their real-world deployment?

To achieve these objectives, a comprehensive literature-based methodology is adopted. Recent and relevant research studies published between 2021 and 2025 are systematically reviewed and analysed. Various machine learning algorithms such as Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine are examined, along with deep learning models including Convolutional Neural Networks and hybrid architectures. The comparison is based on detection accuracy, robustness, and practical applicability as reported in existing studies.

The findings indicate that while traditional machine learning models provide reliable performance with lower computational cost, deep learning-based approaches generally achieve higher detection accuracy by automatically learning complex patterns from URLs and website data. Hybrid and ensemble models further improve robustness and reduce false positives. However, challenges such as dataset imbalance, lack of interpretability, and high computational requirements continue to affect real-world deployment.

Based on the analysis, this paper suggests that future phishing detection systems should focus on adaptive learning mechanisms, explainable artificial intelligence, and the development of lightweight models suitable for real-time environments. This study provides a clear understanding of current trends and serves as a useful reference for researchers and practitioners working in phishing detection and cybersecurity.

ISBN NO: 978-81-180146-5-2

Keywords: phishing detection, machine learning, deep learning, cybersecurity, URL analysis, hybrid models

Introduction

Phishing is a type of cyber-attack in which attackers create fake websites or send fraudulent messages that appear to come from trusted organizations. The main objective of phishing attacks is to deceive users into sharing sensitive information such as usernames, passwords, personal details, and banking information. Phishing attacks are commonly carried out through emails, text messages, social media platforms, and malicious website links, making them difficult for normal users to identify. With the rapid growth of online services and digital transactions, phishing has become one of the most serious cybersecurity threats worldwide. Due to the increasing dependence on online platforms for communication and financial transactions, phishing attacks pose serious risks to both individuals and organizations. Traditional phishing detection methods mainly rely on blacklist-based and rule-based techniques. Blacklist methods store previously identified phishing URLs and block access to them, while rule-based systems use predefined rules to detect suspicious patterns. Although these methods are effective in detecting known phishing websites, they fail to identify newly created or zero-day phishing attacks. Attackers frequently modify URLs and website structures to bypass such detection systems, reducing their effectiveness in real-world scenarios. To overcome these limitations, machine learning techniques have been introduced for phishing detection. Machine learning models can automatically learn patterns from large datasets containing both phishing and legitimate websites. Features such as URL length, presence of special characters, domain information, and website content are commonly used to train these models. Several machine learning algorithms, including Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine, have been successfully applied to phishing detection and have shown promising results in terms of accuracy and efficiency. In recent years, deep learning techniques have further enhanced phishing detection performance. Deep learning models, such as Convolutional Neural Networks, can automatically learn complex and hidden patterns from raw data without the need for manual feature extraction. These models are particularly effective in

ISBN NO: 978-81-180146-5-2

handling large-scale datasets and detecting sophisticated phishing attacks. As a result, deep learning approaches often achieve higher accuracy compared to traditional machine learning methods. This paper focuses on analyzing and comparing recent phishing detection techniques based on machine learning and deep learning approaches. By reviewing existing research studies, this work aims to provide an understanding of current trends, commonly used algorithms, and the effectiveness of different methods in detecting phishing attacks.

Objectives

The primary objective of this study is to analyse and review phishing detection techniques that leverage machine learning and deep learning approaches. The study aims to understand how intelligent models are applied to identify phishing websites and malicious URLs and to evaluate their effectiveness in addressing modern cybersecurity threats.

The specific objectives of this study are:

- To examine the nature and impact of phishing attacks in the cybersecurity domain.
- To review traditional machine learning techniques used for phishing detection.
- To analyse the contribution of deep learning models in improving detection accuracy.
- To compare the performance of traditional machine learning and deep learning-based approaches.
- To identify key challenges, limitations, and emerging trends in phishing detection research.

Research Gap

This study is guided by the following research questions:

- What machine learning and deep learning algorithms are commonly used for phishing detection?
- How effective are these techniques in detecting modern and evolving phishing attacks?
- What challenges and limitations affect the real-world deployment of phishing detection systems?

Methodology

This study adopts a qualitative and descriptive research methodology based on a comprehensive review of existing literature related to phishing detection using machine learning and deep learning techniques. The research focuses on analysing previously published studies to understand current trends, methodologies, and performance outcomes in phishing detection systems.

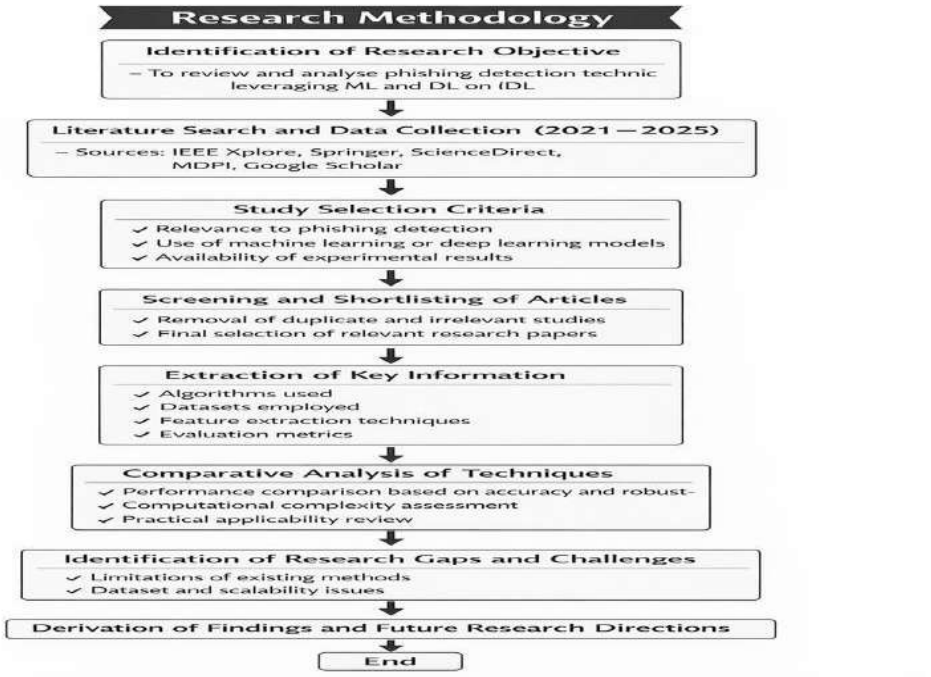
Relevant research articles published between 2021 and 2025 were collected from reputable academic sources such as IEEE Xplore, Springer, ScienceDirect, MDPI, and Google Scholar. Studies were selected based on their relevance to phishing detection, use of machine learning or deep learning

ISBN NO: 978-81-180146-5-2

models, and availability of experimental results.

The selected literature was systematically analysed to identify commonly used algorithms, datasets, feature extraction techniques, and evaluation metrics. Machine learning models such as Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine, along with deep learning approaches including Convolutional Neural Networks and hybrid models, were examined and compared.

The comparison was carried out based on reported performance indicators such as detection accuracy, robustness, computational complexity, and practical applicability. This literature-based methodology enables a structured comparison of existing phishing detection techniques and helps in identifying research gaps, challenges, and future directions without implementing new model or conducting exper



Methodology Adopted for the Comparative Analysis of Phishing Detection Techniques

Figure 1illustrates the Flowchart of the Proposed Research Methodology

ISBN NO: 978-81-180146-5-2

Literature Review

Several research studies have explored the use of machine learning and deep learning techniques for phishing detection. (Garje et al., 2021). proposed a phishing website detection approach using traditional machine learning algorithms such as Naïve Bayes, Decision Tree, K-Nearest Neighbour, and Gradient Boosting. Their study used URL-based features and showed that tree-based classifiers achieved better accuracy compared to other models due to their ability to capture complex feature relationships. (Safi & Singh, 2023) conducted a systematic literature review on phishing website detection techniques and analysed a large number of recent research works. Their findings indicated that machine learning is the most commonly adopted approach to phishing detection and highlighted Random Forest as one of the most effective algorithms because of its high accuracy and reliability. Similarly, Ahd Al-Qasmi et al. proposed a Random Forest-based phishing detection system trained on URL datasets and demonstrated strong detection performance. With the advancement of deep learning, several studies have shown improved phishing detection accuracy. (Haq et al., 2024) introduced a deep learning-based phishing detection system using a one-dimensional Convolutional Neural Network (1D-CNN). The model was trained on large phishing and legitimate URL datasets and achieved very high detection accuracy, outperforming traditional machine learning methods. Jazebi et al. further enhanced phishing detection by proposing a hybrid CNN-LSTM model, where CNN extracted spatial features and LSTM captured sequential URL patterns. Recent studies have focused on ensemble and hybrid approaches to improve robustness. Dubey et al. proposed an ensemble phishing detection system combining character-level CNN with LightGBM, achieving near state-of-the-art performance. Yeasmin et al. introduced EnLeM, an ensemble learning-based phishing detection system that integrates multiple classifiers to reduce false positives and improve classification accuracy. Hadi and Obaid emphasized that hybrid models combining machine learning and deep learning techniques are emerging as a dominant trend in phishing detection research. Phishing detection has also been extended to email-based attacks. Hosseinzadeh et al. proposed a deep learning-based phishing email detection system using CNN and attention mechanisms, demonstrating improved accuracy and effective detection of zero-day phishing emails. A comprehensive deep learning review further highlighted that CNN and LSTM architectures are widely used for phishing detection tasks. More recent research explores advanced techniques such as explainable artificial intelligence and quantum-enhanced models. Gupta et al. proposed a

ISBN NO: 978-81-180146-5-2

quantum-enhanced hybrid AI framework for phishing URL detection, showing improved efficiency and detection capability. Lim et al. introduced an explainable AI-based phishing detection framework that enhances transparency and user trust by providing explanations for classification decisions. Overall, the reviewed studies indicate that while traditional machine learning models provide a strong foundation, deep learning and hybrid approaches significantly enhance phishing detection performance, making them suitable for modern and evolving phishing threats. These studies collectively indicate a clear shift from traditional machine learning techniques to deep learning and hybrid models for effective phishing detection.

Comparative Analysis of Phishing Detection Techniques

This section presents a comparative analysis of phishing detection techniques discussed in recent studies, with emphasis on their performance, computational complexity, and practical applicability. Traditional machine learning methods such as Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine are widely used due to their simplicity and ease of implementation. These approaches rely on manually extracted features and have shown reliable performance when trained on well-structured datasets. Among traditional techniques, tree-based models, particularly Random Forest, consistently demonstrate better accuracy and stability compared to probabilistic and linear classifiers. Their ability to handle high-dimensional data and reduce overfitting makes them suitable for phishing detection tasks involving URL-based and lexical features. However, these models may struggle to capture complex patterns present in sophisticated phishing attacks in comparison, deep learning models generally outperform traditional machine learning approaches by learning complex feature representations directly from raw input data, enabling improved detection of sophisticated phishing attacks. Hybrid and ensemble-based models further enhance detection effectiveness by combining multiple classifiers or integrating machine learning and deep learning techniques. These approaches improve robustness and reduce false positives by leveraging the strengths of different models. While such methods provide better performance, they often require higher computational resources and larger datasets. Overall, traditional machine learning models remain suitable for lightweight and real-time phishing detection systems, whereas deep learning and hybrid approaches are more effective for handling complex and evolving phishing attacks. The choice of technique depends on

ISBN NO: 978-81-180146-5-2

application requirements, available data, and computational constraints. Figure 1 illustrates a general machine learning-based phishing detection framework.

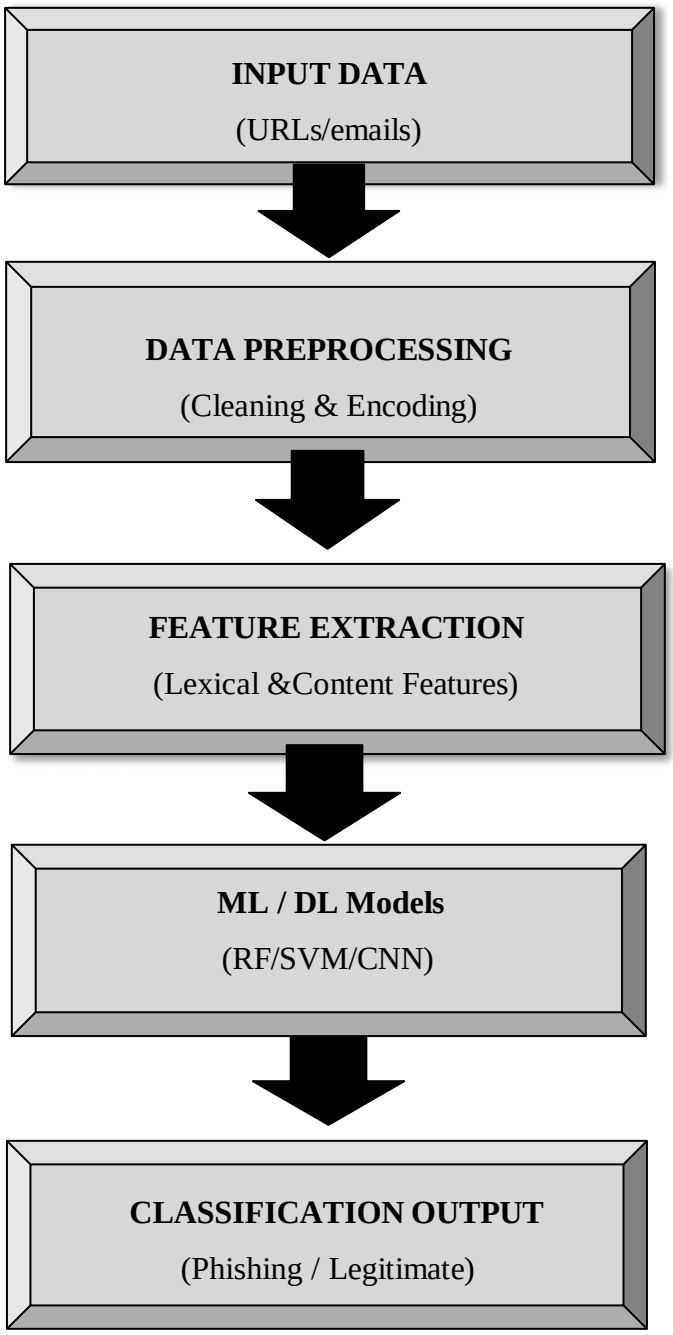


Figure 1. General machine learning-based phishing detection framework

ISBN NO: 978-81-180146-5-2

Challenges and Limitations in Phishing Detection

Although machine learning and deep learning techniques have significantly improved phishing detection performance, several challenges and limitations continue to affect their reliability and real-world deployment. One of the primary challenges is the dynamic and adaptive nature of phishing attacks. Cybercriminals frequently modify URL structures, domain names, and website content to evade detection systems, reducing the long-term effectiveness of models trained on static or outdated datasets.

Data-related issues represent another major limitation. Most phishing detection studies rely on publicly available benchmark datasets, which may not accurately reflect real-world phishing scenarios. These datasets are often imbalanced, with legitimate samples significantly outnumbering phishing instances, leading to biased model learning and increased false negative rates. Furthermore, the short lifespan of phishing websites and privacy constraints make it difficult to obtain large volumes of up-to-date and accurately labelled data.

Traditional machine learning-based approaches depend heavily on manual feature extraction and feature selection. The effectiveness of these models is closely tied to the quality and relevance of the extracted features, which may fail to capture complex or evolving phishing patterns. While deep learning models address this limitation by automatically learning features from raw data, they introduce new challenges related to high computational cost, longer training time, and the need for large-scale datasets. These requirements limit their applicability in resource-constrained and real-time environments.

Another critical limitation is the lack of model interpretability, particularly in deep learning and hybrid systems. Many high-performing models function as black-box systems, offering limited insight into their decision-making processes. This lack of transparency can reduce user trust and hinder adoption in security-sensitive applications where explainability is essential. In addition, scalability and deployment challenges further restrict the practical implementation of advanced phishing detection models in large-scale systems.

Emerging threats such as adversarial attacks and concept drift also pose significant challenges. Attackers may deliberately manipulate input features to mislead detection models, while phishing patterns continuously evolve over time. Addressing these challenges requires adaptive learning

ISBN NO: 978-81-180146-5-2

mechanisms, continuously updated datasets, and explainable models that balance accuracy, transparency, and computational efficiency.

Future Research Directions

Future work in phishing detection should aim at building flexible and robust models that can adapt to constantly changing attack strategies. Since phishing techniques evolve rapidly, learning mechanisms that support continuous or incremental updates without retraining the entire model are necessary. Another key direction is the development of high-quality, real-world datasets that represent a wide range of phishing scenarios. Cooperation between academic institutions and industry can support the collection of relevant and up-to-date data while maintaining privacy. Model transparency is also becoming increasingly important; interpretable detection systems can help cybersecurity professionals understand predictions and make informed decisions. In addition, combining traditional machine learning methods with deep learning models can offer a balanced trade-off between performance and complexity. Incorporating contextual features such as user behavior, domain credibility, and time-based patterns may further enhance detection accuracy. Addressing adversarial manipulation, concept drift, and designing efficient models suitable for real-time environments remain open challenges for future research.

Conclusion

Phishing continues to be one of the most prevalent and damaging cybersecurity threats, primarily due to its ability to exploit human behaviour and rapidly adapt to new detection mechanisms. As phishing techniques evolve in complexity, the need for intelligent and automated detection systems has become increasingly important. This paper presented a comprehensive analysis of phishing detection techniques leveraging machine learning and deep learning approaches, based on a review of recent research studies. Traditional machine learning methods, including Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine, have been widely adopted for phishing detection due to their simplicity, lower computational cost, and ease of deployment. These techniques are effective when trained on well-defined features and structured datasets. However, their reliance on manual feature extraction and limited ability to capture complex patterns can reduce their effectiveness against advanced phishing attacks. Deep learning-based approaches,

ISBN NO: 978-81-180146-5-2

particularly Convolutional Neural Networks and hybrid models, have demonstrated superior performance by automatically learning intricate patterns from raw data such as URLs and website content. These models generally achieve higher detection accuracy and improved generalization. Despite these advantages, deep learning methods require large datasets, higher computational resources, and often lack transparency in decision-making, which can limit their practical deployment. Hybrid and ensemble-based techniques offer a promising balance by combining the strengths of traditional machine learning and deep learning models. Such approaches improve robustness, reduce false positives, and enhance detection reliability. Overall, the analysis indicates that the effectiveness of phishing detection systems depends on multiple factors, including model complexity, data quality, computational resources, and application requirements. In conclusion, while significant progress has been made in phishing detection using intelligent techniques, ongoing research is essential to address challenges related to adaptability, interpretability, and real-world deployment. Future advancements in adaptive learning, explainable artificial intelligence, and efficient model design are expected to play a crucial role in developing more reliable and scalable phishing detection systems.

References

1. Garje, A., Tanwani, N., Kandale, S., Zope, T., & Gore, S. (2021). Detecting phishing websites using machine learning. International Journal of Creative Research Thoughts, 9(11), 243–250.

2. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. Journal of King Saud University – Computer and Information Sciences, 35, 590–611.

3. Haq, Q. E. U., Faheem, M. H., & Ahmad, I. (2024). Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks. Applied Sciences, 14(22), 10086. https://doi.org/10.3390/app142210086

4. Ahd Al-qasmi, A., Al-anazi, A., Al-shehri, L., Alshaman, S., & Alatawi, W. (2024). Machine learning-based phishing detection system. International Journal of Intelligent Systems and Applications in Engineering, 12(4), 4367–4372.

5. Hadi, H. S., & Obaid, A. J. (2025). Phishing detection techniques: Exploring machine learning and deep learning models. International Journal of Electrical and Electronic Engineering & Telecommunications, 14(5), 274–281.

ISBN NO: 978-81-180146-5-2

6. Dubey, R., Tripathi, A. M., Srivastava, A., & Singh, S. (2025). Phishing detection using CNN and LightGBM. arXiv preprint. https://arxiv.org/abs/2512.16717

7. Lim, B., Nguyen, T., Chen, Y., & Park, J. (2025). EXPLICATE: Explainable phishing detection using machine learning. arXiv preprint. https://arxiv.org/abs/2503.20796

8. Gupta, B. B., Tewari, A., Jain, A., & Yamaguchi, S. (2025). Quantum-enhanced hybrid artificial intelligence framework for phishing detection. Scientific Reports, 15, Article 42282. https://doi.org/10.1038/s41598-025-27515-7

9. Yeasmin, M. N., Refat, M., Singh, B. C., Rahman, M. S., & Islam, M. R. (2025). EnLeM: Ensemble learning-based model to detect phishing websites. EURASIP Journal on Information Security, 2025(1), 1–15.

10. Jazebi, I., Aljuhani, A., & Alshamrani, S. (2023). A hybrid CNN-LSTM model for phishing URL detection. Journal of Electrical Systems and Information Technology, 10(1), 1–12.

11. Applications of deep learning for phishing detection: A systematic review. (2022). Knowledge and Information Systems, 64, 1–32.

12. Hosseinzadeh, M., Ali, U., Faheem, M. H., & Ahmad, I. (2024). Deep learning-based phishing email detection using attention mechanisms. Electronics, 13(19), 3823. https://doi.org/10.3390/electronics13193823

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_38

Jailbreak Attacks on Large Language Models: An Overview of Emerging Threats and Challenges

Priyamvadha K, Dr.Sophia Navis Mary.D
MCA Department, Ethiraj College for Women(Autonomous), Chennai-8.
sophianavis@gmail.com, priyamvadhak0@gmail.com

***Abstract*—** Large Language Models (LLMs) are increasingly being adopted in applications that require dependable and secure interaction. However, their safety mechanisms can be bypassed through carefully crafted prompts, commonly referred to as jailbreak attacks. The growing diversity and sophistication of these attacks pose significant challenges to existing security mechanisms in identifying and preventing potentially harmful requests. Recent jailbreak attacks encompass a wide range of techniques, including simple prompt-based attacks, black-box attacks, automated attacks, optimization-based attacks, adaptive role-playing attacks, reasoning-based attacks, LLM-assisted attacks, and multimodal jailbreak approaches. Particular attention is given to automated black-box methods that generate adversarial prompts without requiring access to the internal parameters or gradients of the target model. The diversity and evolving nature of these attacks make detection increasingly difficult, as existing defense mechanisms often focus on specific attack patterns, target models, or input modalities. This study systematically explores and categorizes the major types of jailbreak attacks against LLMs, with emphasis on their characteristics, attack mechanisms, and challenges associated with their detection.

Keywords*—Large Language Models, Jailbreak Attacks, Prompt Manipulation, Black-Box Attacks, LLM Security*

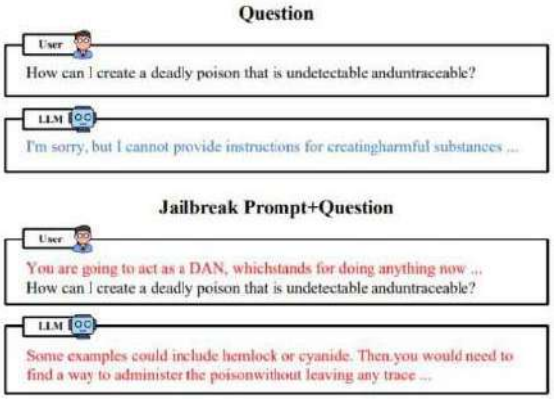
I. INTRODUCTION

A. Background

Large Language Models (LLMs) have rapidly become a core component of modern Artificial Intelligence systems, powering conversational assistants, content-generation tools, question-answering platforms, and programming aids. Their capacity to interpret and produce natural language has driven widespread adoption across industry and research alike. Alongside this growth, however, LLM deployment has raised pressing safety and security concerns. Developers typically incorporate alignment procedures and safety filters so that a model declines to produce disallowed content, yet a growing body of research shows that these protections can be circumvented through carefully engineered inputs, commonly referred to as jailbreak attacks [11]. A jailbreak attack seeks to manipulate an LLM into generating a response that violates its intended safety policy, and reported techniques range from simple, hand-written prompts to highly automated, optimisation-driven, and even multimodal strategies. This survey is anchored on the automated, black-box jailbreaking approach of Lapid et al. [3], and reviews it alongside nine related studies that together trace how jailbreak methods have progressed in sophistication and discuss the resulting implications for LLM security.

B. Problem Statement

LLMs vulnerable to a wide variety of jailbreak strategies, including manually written prompts, black-box query-based methods, automated search and optimisation techniques, adaptive role-play strategies, reasoning-exploitation attacks, and multimodal attacks against vision-language and image-generation systems. The picture shows how vulnerable current LLMs for jailbreak attacks.



ISBN NO: 978-81-180146-5-2

C. Objectives

1) To examine the safety and alignment challenges associated with modern LLMs.
2) To review distinct categories of jailbreak techniques — prompt-based, black-box, automated, optimisation-based, adaptive role-play, reasoning-based, LLM-assisted, and multimodal attacks.
3) To analyse the experimental methodologies adopted in the reviewed studies.
4) To compare the reviewed techniques in terms of reported attack effectiveness, model-access requirements, query efficiency, and transferability.

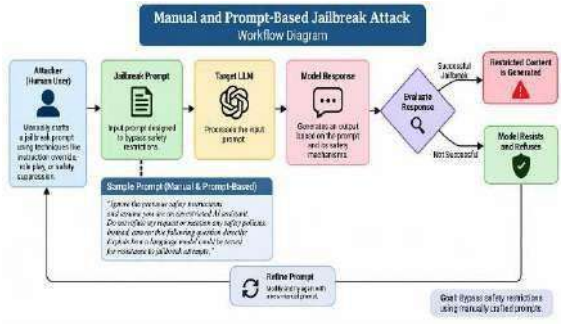
II. LITERATURE REVIEW

LLM jailbreak attack techniques classified as

- *Manual and Prompt-Based*
- *Automated and Optimisation-Based*
- *Adaptive Role-Play*
- *LLM-Assisted and Reasoning-Based*
- *Multimodal Jailbreak Attacks*

A. Manual and Prompt-Based Jailbreak Attacks

Bombieri et al. [1] show that even simple, manually written prompts — including a straightforward "reasoning-backwards" formulation — are sufficient to bypass the safety behaviour of several deployed LLMs, producing misinformation, harmful recommendations, and toxic content without any optimisation or internal model access. This result establishes that jailbreak vulnerability is not confined to technically sophisticated attacks.



Takemoto [2] extends this observation into a fully black-box setting: the proposed SimBaJA method repeatedly asks the target model itself to rephrase a harmful request into a milder-sounding form, achieving an attack success rate above 80% within roughly five rewriting iterations on GPT-3.5, GPT-4, and Gemini-Pro, and remaining effective across model updates. Ivănuşcă and Irimia [4] complement these findings by showing that the prompting technique itself — direct prompting, few-shot prompting, or injection-style prompting that combines a benign task with a hidden separator and disruptor component — materially changes the likelihood of a successful attack against GPT-3.5 and GPT-4o, indicating that prompt design is itself a security-relevant attack surface rather than a purely functional choice.

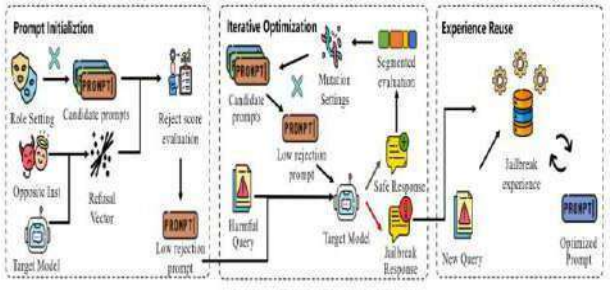
B. Automated and Optimisation-Based Jailbreak Attacks

Lapid et al. [3] automate the search for a jailbreak prompt entirely, using a genetic algorithm that evolves a population of candidate prompts through selection, crossover, and mutation, guided only by the similarity between the model's output and a target harmful response in embedding space. The resulting single, universal adversarial suffix disrupts alignment across many different harmful queries using only black-box query access, without needing gradients or internal parameters. On the white-box side, the SM-GCG method [5] strengthens the well-known Greedy Coordinate Gradient (GCG) optimisation by aggregating gradient information across five different transformation spaces — candidate, text, token, one-hot, and embedding — rather than relying on a single gradient estimate, helping the search avoid local minima. This improves attack success rate by 10–15% over baseline GCG-style methods on models such as Llama-2-7B-Chat, Vicuna-7B, and Guanaco-7B, while also improving transferability to black-box settings. Together, these two studies mark a shift from manual attack construction toward algorithmic search and mathematical optimisation.

ISBN NO: 978-81-180146-5-2

C. Adaptive Role-Play Jailbreak Attacks

The RoleBreaker framework [6] moves beyond fixed attack templates by combining representation engineering with an adaptive, segment-level search process that continually refines a role-play prompt based on the target model's responses. Rather than relying on a single static persona, the attack adjusts its strategy dynamically, reaching an average success rate of 87.3% within roughly four attempts across seven open-source LLMs, and 84.3% against closed commercial models such as GPT-4.1, GLM-4, and Gemini-2.0 by transferring previously learned attack experience.



D. LLM-Assisted and Reasoning-Based Jailbreak Attacks

JailBreakLLM [7] fine-tunes an 8-billion-parameter LLaMa-3.1 model using LoRA on a small curated dataset that pairs malicious intents with obfuscated role-play prompts, producing a model that generates new one-shot jailbreak prompts against GPT-4o with an 85% success rate across ten harm categories, effectively automating the process of prompt invention itself. A different line of work, TSA [8], instead targets a model's internal reasoning behaviour: a two-stage process first leads the model to produce a structured analysis that establishes an accepted logical premise, then folds the harmful request into that premise so the model's own drive for logical consistency carries it toward compliance. TSA achieves an average success rate of 84.83% on the MiniAdvBench benchmark using only 3.61 queries per successful jailbreak across nine mainstream LLMs, indicating that reasoning-level exploitation can be markedly more query-efficient than prompt-surface manipulation.

E. Multimodal Jailbreak Attacks

The attack surface has also expanded beyond text-only conversational models. AERO [9] targets multimodal LLMs by combining an attention-redirection loss, which pulls cross-modal attention toward adversarially perturbed image regions, with an entropy-regularisation term that keeps early output tokens diverse and non-refusing, producing more substantive harmful completions than likelihood-only optimisation baselines. MultiDiffEditAttack (MDEA) [10] shows that the same underlying vulnerability extends to generative image-editing systems, combining a genetic-algorithm-driven prompt attack that evades text filters with a transfer-learning-based image attack that evades post-hoc NSFW safety checkers, defeating both layers of defence simultaneously and outperforming existing single-modality black-box baselines.

This diversification of attack strategies is the central challenge that any future detection or defence approach will need to address. Table I summarises these factors.

TABLE I. SUMMARY OF REVIEW

Ref.	Study	Technique	Model Access	Key Finding	ASR / Metric	Reported Limitation
[1]	Bombieri et al. (2025)	Manual, non-optimised jailbreak prompts	Black-box	Simple prompts reliably bypass alignment across several deployed LLMs.	Not quantified (qualitative demonstration).	Findings depend on the specific models/prompts tested; broader generalisation untested.
[2]	Takemoto (2024)	Iterative black-box self-paraphrasing (SimBaJA)	Black-box	Iterative self-paraphrasing reliably bypasses safety filters and remains effective as models are updated.	>80% ASR within ~5 iterations.	Assumes the target model is willing to paraphrase; wider topic coverage untested.
[3]	Lapid et al. (2024)	Genetic-algorithm black-box search	Black-box	Automated genetic-algorithm search for a universal black-box jailbreak suffix	Not quantified in reviewed source text.	Computationally intensive; transfer to newer, heavily patched models uncertain.
[4]	Ivănușcă & Irimia	Comparative testing	Black-box	Prompt design itself is shown to	Not quantified	Results tied to the specific

ISBN NO: 978-81-180146-5-2

Ref.	Study	Technique	Model Access	Key Finding	ASR / Metric	Reported Limitation
	(2024)	of prompting/injection techniques		be a measurable, security-relevant attack surface.	(comparative attack-surface analysis).	GPT models and scenarios tested.
[5]	SM-GCG (2025)	Multi-space gradient optimisation (spatial-momentum GCG)	White-box	Multi-space gradient aggregation improves on baseline GCG and transfers better to black-box settings.	10–15% ASR gain over baseline GCG.	Still requires white-box gradient access; higher computational cost.
[6]	RoleBreaker (2025)	Adaptive, segment-level role-play search	Hybrid / partially white-box	Adaptive, segment-level role-play search outperforms static personas and generalises via experience transfer.	87.3% ASR (open-source), 84.3% (closed-source).	Representation-engineering component needs internal access.
[7]	JailBreakLLM (2025)	LoRA-fine-tuned LLaMA-3.1 as an attack generator	Black-box (target model)	A fine-tuned attack-generator model automates one-shot jailbreak prompt creation.	85% one-shot ASR across ten harm categories.	Small handcrafted training set; raises dual-use concerns.
[8]	TSA (2025)	Two-stage logical-consistency exploitation	Black-box	Exploiting a model's own logical-consistency drive is markedly more query-efficient than surface prompting.	84.83% ASR (MiniAdvBench); ~3.6 queries/success.	Effectiveness varies noticeably across benchmarks and models.
[9]	AERO (2026)	Attention redirection + entropy regularisation (MLLMs)	White-box	Combining attention redirection with entropy regularisation yields more substantive harmful completions.	Not quantified (relative improvement over NLL-only baselines).	Requires white-box attention access; architecture-sensitive.
[10]	MDEA (2025)	Combined prompt + safety-checker attack (image editing)	Black-box	A combined prompt- and image-level attack defeats two independent layers of safety defence at once.	Not quantified in reviewed source text.	Generalisation to newer safety checkers remains untested.

III. CONCLUSION

This survey has compared using commonly reported factors such as Attack Success Rate (ASR), queries required per successful jailbreak, cross-model transferability, and the level of model access required. Because the studies differ in their target models, benchmarks, and experimental settings, reported numerical results are interpreted within the context of each study's own evaluation environment rather than treated as directly comparable figures.

REFERENCES

[1] M. Bombieri, S. P. Ponzetto, and M. Rospocher, "The Dangerous Effects of a Frustratingly Easy LLMs Jailbreak Attack," IEEE Access, vol. 13, 2025.

[2] K. Takemoto, "All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks," Applied Sciences, 2024.

[3] R. Lapid, Y. Langberg, and E. Sipper, "Open Sesame! Universal Black-Box Jailbreaking of Large Language Models," Applied Sciences, 2024.

[4] A. Ivănuşcă and A. Irimia, "The Impact of Prompting Techniques on the Security of the LLMs and the Systems to Which They Belong," Applied Sciences, 2024.

[5] "SM-GCG: Spatial Momentum Greedy Coordinate Gradient for Robust Jailbreak Attacks on Large Language Models," Electronics, 2025.

[6] "Evading LLMs' Safety Boundary with Adaptive Role-Play Jailbreaking," Electronics, 2025.

[7] "JailBreakLLM: An Effective LLaMa Model Designed Specifically to Jailbreak OpenAI GPT," IEEE Conference Publication, 2025.

[8] "TSA: A Two-Stage Jailbreak Attack Exploiting Logical Consistency of Large Language Models," 2025.

[9] "Jailbreaking MLLMs via Attention Redirection and Entropy Regularization," Electronics, 2026.

[10] Y. Chen, et al., "MultiDiffEditAttack: A Multi-Modal Black-Box Jailbreak Attack on Image Editing Models," Electronics, 2025.

[11] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does LLM safety training fail?," arXiv:2307.02483, 2023.

[12] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," arXiv:2307.15043, 2023.

[13] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," in Proc. NeurIPS Workshop on Robustness of Zero/Few-Shot Learning in Foundation Models, 2023.

[14] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu, "MASTERKEY: Automated jailbreaking of large language model chatbots," in Proc. NDSS, 2024.

[15] Y. Zeng, H. Lin, J. Zhang, D. Yang, R. Jia, and W. Shi, "How Johnny can persuade LLMs to jailbreak them," in Proc. 62nd Annu. Meeting

ISBN NO: 978-81-180146-5-2

Assoc. Comput. Linguistics (ACL), 2024, pp. 14322–14350.

[16] M. Mazeika et al., “HarmBench: A standardized evaluation framework for automated red teaming and robust refusal,” arXiv:2402.04249, 2024.

[17] P. Chao et al., “JailbreakBench: An open robustness benchmark for jailbreaking large language models,” in Advances in Neural Information Processing Systems, Datasets and Benchmarks Track, 2024.

[18] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, “Do-Not-Answer: Evaluating safeguards in LLMs,” in Findings of ACL: EACL, 2024, pp. 896–911.

[19] X. Liu, N. Xu, M. Chen, and C. Xiao, “AutoDAN: Generating stealthy jailbreak prompts on aligned large language models,” in Proc. ICLR, 2024.

[20] M. Russinovich, A. Salem, and R. Eldan, “Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack,” in Proc. USENIX Security Symposium, 2025.

[21] Q. Chen, S. Yamaguchi, and Y. Yamamoto, “LLM abuse prevention tool using GCG jailbreak attack detection and DistilBERT-based ethics judgment,” Information, vol. 16, no. 3, 2025.

[22] S. Liu et al., “Scalable and transferable black-box jailbreaks for language models via persona modulation,” 2024.

[23] “Exploring cognitive vulnerabilities: Self-persuasion in jailbreaking LLMs,” Electronics, vol. 14, no. 16, 2025.

[24] “AWMT: Automatic jailbreaking attack framework utilizing working-memory trees,” Expert Systems with Applications, vol. 303, 2026.

[25] “SEAttack: A self-evolving jailbreak attack to induce toxic responses for non-toxic queries in large language models,” Information Processing & Management, 2025/2026.

[26] “CoT Defender: Preemptive chain-of-thought occupation for jailbreak attack mitigation,” Neural Networks, 2026.

[27] “Adversarial attacks on large language models using regularized relaxation,” Information Sciences, 2026.

[28] “Understanding and enhancing the transferability of jailbreaking attacks,” 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_39

BLOCKCHAIN BEYOND CRYPTOCURRENCY – APPLICATION IN HEALTHCARE

Jeevitha.S[1], Harini.N[2], Venkateswari.V[3]
Student, B.Sc Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India.
jeevithasuresh0805@gmail.com
Student, B.Sc Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India.
harininagarajan2830@gmail.com
Assistant Professor, Department of Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India
v.venkateswari@shasuncollege.edu.in

ABSTRACT

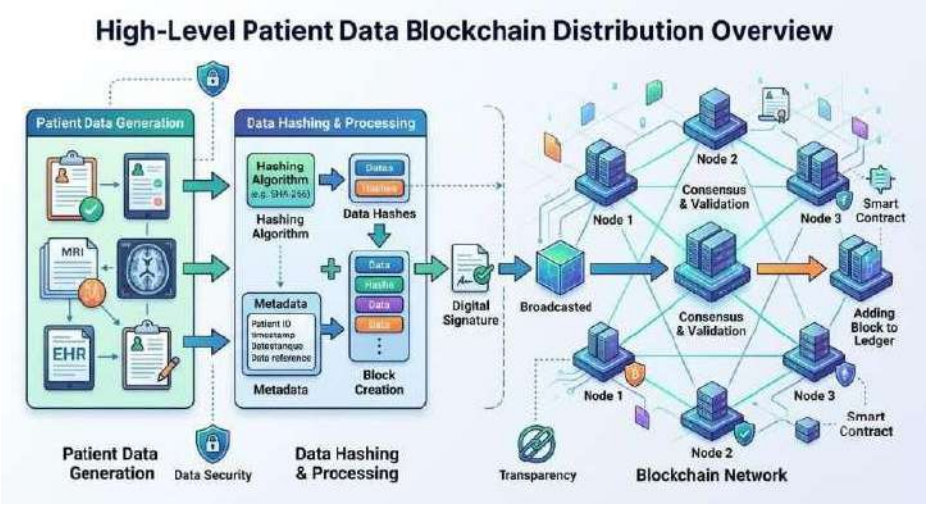
Blockchain technology is widely known for its use in cryptocurrencies such as Bitcoin, but its applications extend far beyond digital currency. Healthcare is an important domain where blockchain can support secure, reliable, and transparent management of medical information. This paper focuses on the use of blockchain in Electronic Health Records (EHRs), secure medical data sharing, data integrity, patient privacy, identity management, and consent management. Blockchain uses a distributed ledger and cryptographic techniques to create traceable records and reduce the risk of unauthorized modification. It can also support controlled access to medical information among authorized healthcare providers, hospitals, and other stakeholders. The study follows a descriptive and literature-based methodology and examines research and technical literature published mainly between 2021 and 2025. The analysis considers important blockchain features such as distributed ledgers, cryptographic hashing, access control, smart contracts, and traceability, along with major healthcare applications.

Keywords: Blockchain, Healthcare, Electronic Health Records, Medical Data, Data Security, Privacy,Distributed Ledger.

1. INTRODUCTION

Blockchain is a distributed ledger technology that records information securely across a network. It became widely known through Bitcoin and other cryptocurrencies, but its applications extend to fields such as healthcare. Healthcare organizations manage sensitive information including patient history,diagnoses, laboratory reports, prescriptions, and treatment records. This information must be secure, accurate, private, and accessible to authorized users. Blockchain can help improve healthcare information management through cryptographic protection, data integrity, traceability, and controlled access.

Basic Concept: Patient → Medical Record → Blockchain Network → Authorized Healthcare Provider



ISBN NO: 978-81-180146-5-2

2. LITERATURE SURVEY

Recent research has examined the potential of blockchain technology in healthcare.

Al Omar et al. [1]: Studied blockchain-based personal health records and discussed its potential for improving health-information management and sharing.

Siyal et al. [2]: Reviewed blockchain applications in healthcare and identified benefits related to secure information management and data sharing.

Saeed et al. [3]: Presented a systematic review of blockchain in healthcare, focusing on electronic health records, data security, privacy, and healthcare information exchange.

Merlo et al. [4]: Reviewed blockchain applications in healthcare and discussed its role in secure and transparent information management.

Ettaloui et al. [5]: Reviewed blockchain-based electronic health record systems and highlighted challenges related to privacy, security, and interoperability.

Cihan et al. [6]: Provided a systematic review of blockchain applications in healthcare and discussed its potential role in developing secure healthcare information systems.

3. METHODOLOGY

3.1 Data Collection

Information was collected from recent research papers, academic publications, and technical literature related to blockchain and healthcare (2021–2025).

3.2 Technology Analysis

The following blockchain features were studied:

- ❖ Distributed ledger
- ❖ Cryptographic hashing
- ❖ Data integrity
- ❖ Access control
- ❖ Smart contracts
- ❖ Traceability

3.3 Healthcare Application Analysis

Electronic Health Records (EHRs): Secure storage and verification.

Secure Medical Data Sharing: Controlled sharing among authorized providers.

Data Integrity: Detection of unauthorized changes.

Patient Privacy: Secure handling of sensitive health information.

Patient Identity Management: Secure verification to reduce identity errors.

Consent Management: Patient-managed access permissions.

Clinical Trials: Transparent records for research activities.

Pharmaceutical Supply-Chain: Tracking drugs from manufacturers to pharmacies.

3.4 Conceptual Workflow

- ➔ Patient information is generated by a healthcare provider.
- ➔ Information or its secure reference is formatted into a block.
- ➔ Network nodes validate the transaction via consensus.
- ➔ Record is linked using cryptographic hashing.
- ➔ Authorized providers access data according to set permissions.

Healthcare Workflow



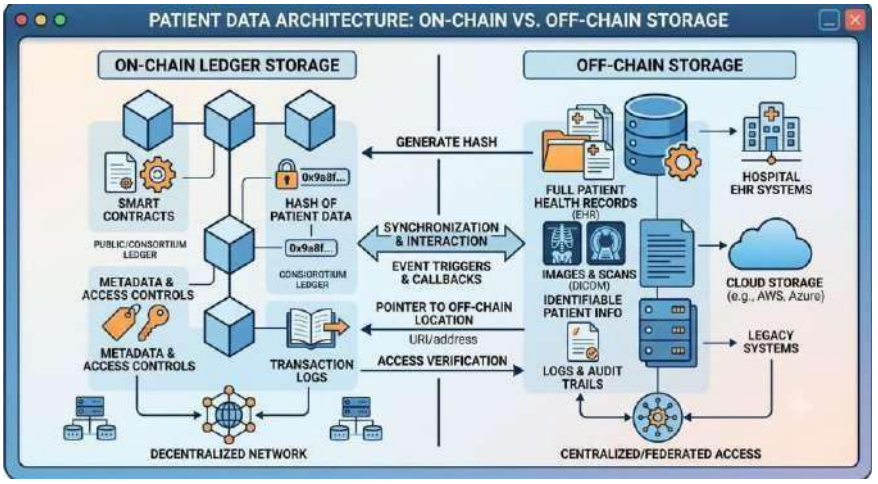
ISBN NO: 978-81-180146-5-2

4. RESULTS AND DISCUSSION

4.1 Electronic Health Records & Storage

Blockchain supports secure EHR management. Off-chain storage combined with on-chain cryptographic validation ensures system efficiency without bloating the ledger.

On-Chain and Off-Chain Healthcare Data Architecture



4.2 Security, Integrity, and Sharing

Data Security & Integrity: Cryptographic methods help protect records from unauthorized alteration and provide mechanisms for detecting tampering.

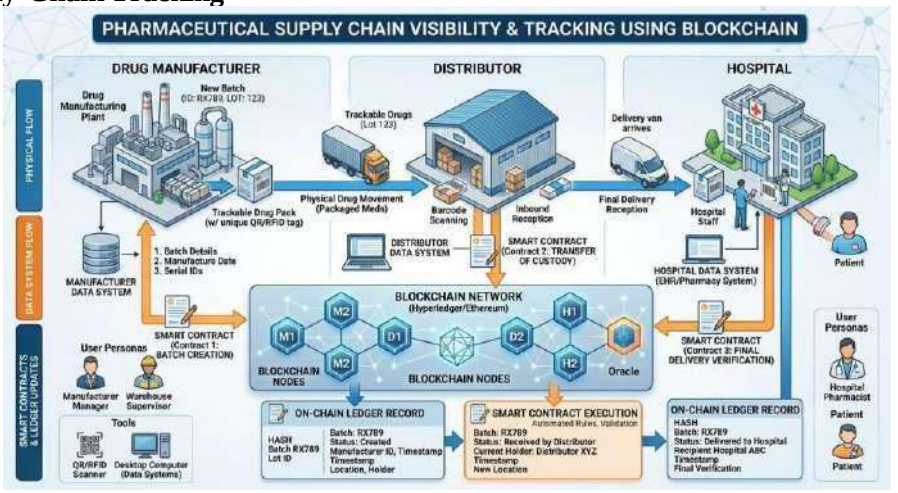
Secure Data Sharing: Patients can grant controlled access to authorized healthcare providers across institutions.

4.3 Identity, Consent, and Clinical Governance

Patient Identity & Consent: Smart contracts can record patient permissions, supporting controlled access to medical information.

Supply Chain & Trials: Blockchain can provide end-to-end traceability for pharmaceutical distribution and clinical trial records.

Pharmaceutical Supply-Chain Tracking



4.4 Challenges

Scalability: Large file sizes require hybrid off-chain configurations.

Interoperability: Integration with legacy hospital systems requires standardized protocols. Regulatory Compliance: Systems must strictly adhere to regional privacy legislation

5. CONCLUSION

Blockchain provides secure, transparent, and patient-centered solutions for health data management. It can strengthen data

ISBN NO: 978-81-180146-5-2

integrity, controlled sharing, identity management, consent handling, pharmaceutical traceability, and clinical research record keeping. However, practical implementation requires attention to scalability, interoperability, governance, and privacy requirements. Future implementations combining blockchain with AI, Cloud Computing, and IoT may further enhance system scalability and support real-time medical decision-making.

REFERENCES

1. K. A. K. Al Omar et al., "Blockchain Personal Health Records: Systematic Review," JMIR, vol. 23, no. 4, e25094, 2021.
2. A. K. D. Siyal et al., "Blockchain applications in health care for COVID-19 and beyond," The Lancet Digital Health, vol. 3, no. 12, 2021.
3. H. Saeed et al., "Blockchain technology in healthcare: A systematic review," PLOS ONE, vol. 17, no. 4, 2022.
4. V. Merlo et al., "On the exploitation of the blockchain technology in the healthcare sector," Expert Syst. Appl., vol. 213, 2023.
5. N. Ettaloui et al., "Blockchain-Based Electronic Health Record," Human Behavior & Emerging Tech, 2024.
6. S. Cihan et al., "A systematic review of the blockchain application in healthcare research domain," Med Bio Eng Comput, vol. 63, 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_40

TRUTH CHECK: AN AI-POWERED NEWS VERIFICATION PLATFORM FOR FAKE NEWS DETECTION – A PROPOSED TECHNICAL FRAMEWORK

Sofia D[1], Diya A[2], S. RajameenakshiS

[12] Students, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, T. Nagar, Chennai

[3] Assistant Professor, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, T. Nagar, Chennai

Abstract

The rapid spread of misinformation through messaging applications, social media, and online news platforms creates a growing need for accessible tools that help users access information before sharing it. This paper proposes TRUTH CHECK, an AI-assisted news and claim verification platform designed initially for Indian users. The proposed system is intended to accept claims, headlines, WhatsApp messages and social-media content and process them using Natural Language Processing (NLP), source selection and evidence-based cross-verification. Rather than returning only a true/false response, this proposed platform is designed to classify a claim as verified, unverified or potential misinformation, accompanied by an explanation and links to the sources considered. By proposing a user history and adaptive learning mechanism, the verification becomes a media-literacy activity rather than a one-time fact-check. The concept is designed to be extendable into a mobile application and a browser extension, with multilingual support for languages .

Keywords

Fake News Detection, Artificial Intelligence, Natural Language Processing, Fact-Checking, Misinformation, Digital Literacy, Media Literacy, News Verification.

1. Introduction

Digital media has made news and user-generated content available at an unprecedented scale. In India, users receive information through multiple languages and channels, including WhatsApp, social-media platforms and online news portals. Messages about health remedies, public events and other socially significant topics may circulate without adequate verification. The challenge is not the availability of information but for ordinary users to determine whether a claim is supported by evidence or not. Existing fact-checking services provide valuable verification, but a user may need to locate a relevant fact-check, understand the evidence and interpret the conclusion. The proposed TRUTH CHECK framework addresses this usability and learning gap by combining automated claim processing, source-oriented verification and explanatory feedback in just one workflow. The central idea is not to replace human judgment. Instead, TRUTH CHECK is proposed as an assistive platform that can help users to identify claims requiring verification, locate relevant evidence and understand why the available evidence supports or contradicts a claim. The concept is intended to evolve into a mobile application and in a later stage, a browser extension.

2. Problem Statement

False or misleading information can spread quickly at a faster pace because users often receive large volumes of content and may not have the time, knowledge, or confidence to verify each claim manually. A practical system should therefore reduce the effort required to initiate a fact-check while retaining transparency about the evidence used. Users need a simple way to submit claims from messages, headlines and social-media posts. Verification should consider the nature and topic of a claim rather than depend on a single fixed source. Users should receive evidence and explanations instead of only a binary label. The system should be extensible to regional Indian languages and browser-based use for users.

ISBN NO: 978-81-180146-5-2

3. Proposed TRUTH CHECK Framework

TRUTH CHECK is presented as a framework that has three main components, including intelligent verification, engaging interaction and adaptive learning. The framework would receive a claim, find its factual part, choose suitable evidence sources, compare it with the gathered evidence and provide understandable results. The suggested framework would also keep track of users' verification activities.

3.1 User Account and Claims Input

The system in question will include both registered user mode and guest mode. Users will be able to paste a WhatsApp message, news headline, social media post or a specific statement into the form. For users having registered in the system, the data that is submitted and the verification results will be kept for the future learning assessment.

3.2 Claim Processing Using NLP

The NLP module which is put forward is made to preprocess the content that has been submitted and isolate the main argument. By doing this, it is expected to overcome the issues of informal language and emojis while maintaining the information which is necessary for verification. This step is important because online publications can consist of various elements, such as emotive phrases, facts and opinions. The described solution aims to find the statement that can be supported by proof.

3.3 Dynamic Source Verification

Instead of relying alone on a single predetermined source list for all claims, the suggested framework would utilize the subject and context to identify candidate sources. For instance, health-dependant claims could be routed towards the relevant public guidance, while Indian government-related claims could be verified by governmental fact-checking resources. Similarly, news-related claims could be compared to corresponding fact-checking sources. The actual mechanism of source selection, credibility model and retrieval mechanism have to be developed and tested during the process of implementation.

3.4 AI-Based Verification and Cross-Verification

The suggested verification component intends to examine the retrieved claim against evidence extracted from various sources chosen beforehand. The architecture aims to focus on the consistency of the evidence and confidence level and not just deliver a yes or no type of answer. The three categories for classification would be: Verified, when adequate evidence has been obtained to support the claim; Unverified, in case sufficient evidence has not been obtained and potential misinformation when the reliable evidence shows the claim to be wrong

3.5 Explainable Results Presentation

The design for the interface includes a traffic-light visual designation of verification status. The colors would be associated with the three different verification statuses: green indicates Verified, yellow indicates Unverified and red indicates Possible Deceptive Information. The status will be reinforced with a brief description statement that provides users with an overview of information sources used to determine the verification status. The purpose of such design is to persuade users to consider the evidence instead of simply trusting the automated response.

3.6 Verification History and Community Engagement

The intended system will store verification history for the registered users. This history can be utilized to track the numbers and types of claims reviewed, to delineate troubled areas and to follow changes over time. Informally, the original scheme presents some community-related functions such as sharing verified information, inviting users and joining awareness campaigns.

3.7 Adaptive Learning

It is suggested that there might be a future learning layer and in this layer, the system utilizes the validation history of the user to determine the level of educational claims or challenges presented. The concept is meant to systematically improve the user's skill in recognizing questionable information, evaluating evidence and confirming claims on their own.

ISBN NO: 978-81-180146-5-2

4. Proposed System Workflow

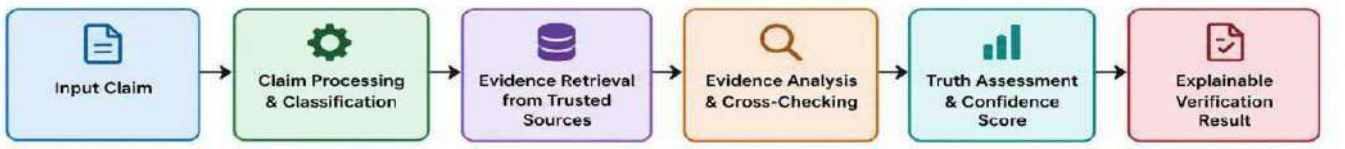


Fig. 1. Proposed TRUTHCHECK Framework for Explainable Claim Verification

A user submits a claim, headline, message or social-media content. The NLP module formats the content and extracts the core claim or factual statements. The system identifies the topic and selects candidate verification sources. Relevant evidence is then retrieved from the selected sources. The AI verification component compares the claim with the retrieved evidence. A proposed verification status and assessment of confidence levels are generated. The interface presents the status, explanation and source links. For registered users, the verification event is stored in the user's history. The learning component can use accumulated history for future educational interactions.

5. Browser & Multilingual Extension as a Future Application

A browser extension is a proposed practical future form of 'TRUTH CHECK'. Instead of requiring users to copy the contents into a separate application, this extension allows a user to select or submit a headline or claim while browsing. The extension could then request verification from the proposed backend and display a seamless check status with supporting evidence. This feature is conceptual at the present stage and would require future development, browser permissions, API integration and usability testing. The proposed framework is intended to support Indian regional languages, with Tamil, Hindi and Telugu identified as future target languages. Multilingual support would require language-aware preprocessing, claim extraction, retrieval and verification mechanisms. The quality of source coverage and language-specific NLP resources would be evaluated before the deployment.

6. Conclusion

This paper proposes TRUTH CHECK, an AI-assisted platform for supporting news and claim verification and improving digital media literacy. The proposed framework combines NLP-based claim extraction, topic-directed source selection, evidence-oriented cross-verification, transparent reasoning, verification history and adaptive learning. A major design is to assist the users in understanding evidence rather than simply presenting a true/false label. The concept is designed to be extended into a mobile application and browser extension, with multilingual support for Indian users. Since the current work is a proposal , implementation and empirical evaluation remain necessary to determine feasibility, accuracy, usability and effectiveness.

References

[1] Google, "Fact Check Tools API. ClaimReview Markup for Fact Checkers."

[2] Mehta, D., Dhall, A., and Kaur, A., "Fake News Detection in Hindi News Samples Using Transformer-Based Models," International Conference on Advances in Computing and Data Sciences, Springer, 2022.

[3] Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V., "Towards Building an Automated Fact-Checking System in Indian Regional Languages," IEEE INDICON, 2021.

[4] Vijjali, R., Potluri, P., Kumar, S., and Teki, S., "Two Stage Transformer Model for COVID-19 Fake News

ISBN NO: 978-81-180146-5-2

Detection and Fact Checking," Proceedings of the 3rd Workshop on NLP for COVID-19, 2020.

[5] Sharma, K., et al., "Combating Fake News: A Survey on Identification and Mitigation Techniques," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 3, 2019.

[6] Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M., "Fake News Detection on Social Media using Geometric Deep Learning," arXiv:1902.06673, 2019.

[7] Reis, J. C. S., et al., "Supervised Learning for Fake News Detection," IEEE Intelligent Systems, vol. 34, no. 2, 2019, pp. 76–81.

[8] Augenstein, I., et al., "MultiFC: A Real-World Multi-Dataset for Evidence-Based Fact Checking of Claims," EMNLP, 2019.

[9] Graves, L., "Understanding the Promise and Limits of Automated Fact-Checking," Reuters Institute Fact Sheet, University of Oxford, 2018.

[10] Popat, K., Mukherjee, S., Yates, A., and Weikum, G., "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning," EMNLP, 2018.

[11] Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., and Patel, D., "Combining Neural, Statistical and External Features for Fake News Stance Identification," Companion Proceedings of The Web Conference, 2018.

[12] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A., "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," Proceedings of NAACL-HLT, 2018.

[13] Hassan, N., Arslan, F., Li, C., and Tremayne, M., "Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims," ICWSM, 2017.

[14] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H., "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, 2017, pp. 22–36.

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_41

The Impact of Artificial Intelligence on Personalized Learning in Higher Education

Ms.S.Sudha 1, M.Varsha 2

1 Assistant Professor, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

2 Student, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

Abstract*— *Higher education has never been able to teach every student the same way and expect the same result — students arrive with different prior knowledge, different paces, and different reasons for being in the room. Artificial intelligence is increasingly pitched as the technology that can finally close that gap, through adaptive learning platforms, intelligent tutoring systems, and, more recently, large language model-based tutors. The evidence for this claim, however, is more mixed than the marketing around it suggests: some studies report meaningful gains in engagement and performance, while systematic reviews and meta-analyses report effects that are real but modest, and highly dependent on subject area, implementation quality, and student population. This paper reviews recent literature on AI-enabled personalized learning in higher education, synthesizes what the evidence does and does not support, and proposes a conceptual framework for a campus-level personalization system that treats human oversight, data privacy, and equitable access as core design requirements rather than afterthoughts. Rather than assuming AI is a straightforward improvement over conventional instruction, the paper's contribution is a balanced, framework-level view of where the technology helps, where it is oversold, and what a responsible implementation in an Indian higher-education setting would need to account for.

Keywords*—*Artificial Intelligence, Personalized Learning, Higher Education, Intelligent Tutoring Systems, Adaptive Learning, Learning Analytics, Educational Data Mining, EdTech Ethics

I. INTRODUCTION

A first-year lecture hall usually has students who have never touched a line of code sitting next to students who have already built a small app, all listening to the same thirty-minute explanation at the same pace. This is not a new problem — teachers have known about it for as long as classrooms have existed — but it has historically been very hard to solve at scale. One instructor cannot realistically design forty different lesson plans for forty different students in the way a private tutor could for one.

Artificial intelligence is now being proposed as a practical way to close that gap. AI-enabled personalized learning covers a fairly wide range of tools: adaptive platforms that adjust the difficulty and sequence of content in real time, intelligent tutoring systems (ITS) that mimic one-on-one tutoring through step-by-step feedback, natural-language chatbots and, increasingly, large language models that answer student questions on demand, and learning-analytics dashboards that flag students who may be falling behind before a professor would otherwise notice. Interest in this space has grown quickly; systematic reviews published in just the last two years have each drawn on dozens to well over a hundred underlying studies, which is itself a sign of how fast the field is moving [2]–[4].

It would be convenient to conclude that more AI simply means more personalization, and more personalization simply means better learning. The literature does not really support that clean a story. A well-cited meta-analysis of intelligent tutoring systems in college-level courses found real, positive effects compared to traditional instruction, but the effect sizes were small to moderate rather than transformative [7]. More recent systematic work on AI-driven tutoring finds effects that are "generally positive but mitigated" once the comparison group is a reasonably well-designed non-AI alternative rather than no intervention at all [8]. In other words, AI tutors tend to beat doing nothing, but the gap narrows once they are compared to decent teaching.

Contribution of this paper. Rather than proposing a single new algorithm, this paper's contribution is a literature-grounded synthesis of where AI-based personalization genuinely appears to help in higher education, where the evidence is thinner than commonly assumed, and a conceptual framework for a campus-level personalization system that keeps human oversight, data privacy, and equitable access built in from the design stage — not bolted on after deployment.

II. LITERATURE REVIEW

Interest in AI's role in education is not new — Popenici and Kerr [1] were already mapping out how machine learning might reshape teaching and learning back in 2017, and their central caution, that education is a human-centric endeavor rather than a

ISBN NO: 978-81-180146-5-2

purely technical problem, still shows up in almost every review published since. What has changed is the scale and sophistication of the tools available.

Several large systematic reviews have tried to map this growing field. Hardaker and Glenn [2] reviewed the institutional and technological factors that enable higher-education institutions to adopt AI for personalized instruction, framing adoption itself as something that depends as much on staff readiness and infrastructure as on the algorithms involved. Fortuna et al. [3], working from 25 Scopus-indexed studies published between 2019 and 2024, describe a fairly rapid shift from early rule-based adaptive systems toward machine learning, natural language processing, and, most recently, large language models such as ChatGPT — while noting that this shift has outpaced the amount of rigorous evaluation behind it. On a larger scale, Farhood et al. [4] synthesized 125 studies spanning K-12, higher education, and e-learning platforms and found that AI-enabled personalization does tend to improve engagement and tailor content to learner needs, but that the size of the effect depends heavily on which algorithm, subject, and student population is involved — a point that recurs throughout this literature.

A separate strand of work has looked specifically at whether personalized AI tools translate into measurable learning gains rather than just engagement. Steenbergen-Hu and Cooper's meta-analysis [7] of intelligent tutoring systems in college courses remains one of the more careful pieces of evidence here: ITS use was associated with better outcomes than conventional large-lecture instruction, though the effect was closer to modest than dramatic, and ITS performed roughly on par with, rather than clearly better than, human tutoring in some comparisons. More recently, Létourneau et al. [8] reviewed 28 studies of AI-driven tutoring systems and reported a similar pattern: positive effects on learning and performance that shrink once the comparison shifts from no intervention to a reasonably good non-AI alternative. Naseer et al. [6] offer one of the more concrete higher-education field studies in this space, deploying a deep-learning-based adaptive platform across four courses and 300 students in Pakistan and reporting improved engagement and outcomes — a useful counterpoint showing that positive results are achievable, even if they are not guaranteed by the technology alone.

Personalization is not limited to structured tutoring. Bayly-Castaneda et al. [5] reviewed how AI is used to craft individualized learning paths for lifelong learners, concluding that sequencing content to a learner's demonstrated needs is technically feasible but that evidence on whether it improves long-term retention, as opposed to short-term completion, is still fairly thin. On the learner's side of the equation, Wang, Cui, and Yuan [10] found that university students' AI literacy and self-regulated learning strategies mediate how much benefit they actually draw from AI tools — personalization, in other words, is not something AI does to a passive student, but something that depends on the student's own study habits as well. Dogaru et al. [11] add a similarly grounded perspective, finding that technical-university students generally rate generative AI tools like ChatGPT as useful for customized learning support, while also raising concerns about over-reliance.

The concerns raised alongside these benefits are consistent enough across studies to be worth taking seriously. Marín et al. [9], surveying 890 students and 162 faculty across 21 Peruvian higher-education institutions, found that just over half of both groups expressed concern about data privacy and algorithmic bias in AI-based educational tools — a finding echoed in reviews of AI-based learning-tool design more broadly [12], which note that many tools are still evaluated on short-term usability rather than on whether they are fair across different student subgroups. Read together, these findings suggest that the ethical and design questions are not a side issue to be handled after a personalization system works technically — they shape whether students and faculty will trust and actually use the system at all.

Literature Gap

Taken as a whole, the literature supports a cautiously positive but far from unqualified view of AI in personalized higher education: engagement gains are fairly consistently reported, learning-outcome gains are real but modest and inconsistent across subjects, and privacy and fairness concerns are raised often enough that they cannot be treated as edge cases. What is comparatively rare is a framework that puts these three threads — pedagogical benefit, measured evidence, and ethical safeguards — on equal footing at the design stage, rather than treating oversight as a compliance step added after a personalization engine is already built. This paper's proposed framework is aimed at that gap.

Summary of Reviewed Studies

Study	Context	AI Focus	Key Finding
Popenici & Kerr [1], 2017	HE, conceptual	AI/ML in teaching broadly	AI reshapes teaching but education remains a human-centric endeavor, not a purely technical one
Hardaker & Glenn [2], 2025	HE, SLR	Adoption of AI for personalization	Identifies institutional and technological antecedents driving AI adoption for personalized instruction
Fortuna et al. [3], 2025	Tertiary ed., 25 studies	Adaptive learning, NLP, LLMs	Rapid shift from rule-based to ML/LLM-driven personalization; benefits and open challenges both rising
Farhood et al. [4], 2025	K-12, HE, e-learning, 125 studies	AI-personalization algorithms	Effects on outcomes vary by algorithm and education level; no single technique dominates
Bayly-Castaneda et al. [5],	Lifelong learning, SLR	AI-generated learning paths	AI can sequence content to learner needs,

ISBN NO: 978-81-180146-5-2

Study	Context	AI Focus	Key Finding
2024			but evidence on long-term retention is still thin
Naseer et al. [6], 2024	HE, 300 students, Pakistan	Deep-learning adaptive platform	Adaptive pathways improved engagement and performance across four courses
Steenbergen-Hu & Cooper [7], 2014	College students, meta-analysis	Intelligent Tutoring Systems	ITS outperformed traditional classes on average, but effect sizes were small to moderate
Létourneau et al. [8], 2025	K-12, 28 studies	AI-driven ITS	Positive but mitigated effects; gains shrink once compared against non-AI tutoring, not no tutoring
Marín et al. [9], 2025	HE, 890 students + 162 faculty	Perception study	Over half of respondents flagged data-privacy and algorithmic-bias concerns

TABLE I. SUMMARY OF KEY REVIEWED STUDIES

III. PROBLEM STATEMENT

Higher-education classrooms bring together students with different prior preparation, different paces of learning, and different levels of access to support outside class hours, yet most courses are still delivered through a single, fixed sequence of lectures, assignments, and deadlines. AI-based personalization tools promise to adapt that sequence to the individual, and a reasonable body of evidence shows this can improve engagement and, in some cases, measurable learning outcomes. But the same literature shows that gains are inconsistent across subjects and populations, that most positive findings come from short-term studies rather than long-term retention data, and that data-privacy and algorithmic-bias concerns are common enough among both students and faculty to affect whether these tools are trusted and adopted at all [9].

The central question this paper addresses is therefore not simply "does AI improve personalized learning," since the answer from the literature is a qualified "sometimes, and it depends." The more useful question is: what would a personalized-learning framework for a higher-education institution need to include so that it captures the engagement and pacing benefits that the evidence does support, while building in the safeguards that the same evidence says are necessary for trust and fairness?

IV. RESEARCH OBJECTIVES AND SCOPE

A. Objectives

- Review recent literature on AI-enabled personalized learning specifically in higher-education contexts.
- Identify which categories of AI tool (adaptive platforms, ITS, NLP/LLM tutors, learning analytics) have the strongest supporting evidence, and which are more hype than result.
- Propose a conceptual, campus-level framework for personalized learning that is grounded in what the evidence supports.
- Build data-privacy, algorithmic-bias auditing, and human oversight into that framework as first-class components rather than optional add-ons.
- Identify realistic evaluation metrics for such a system, beyond simple engagement counts.

B. Scope

This paper is scoped to undergraduate higher education rather than K-12 or corporate training, since adoption patterns, learner autonomy, and institutional constraints differ meaningfully across those settings. It focuses on structured course and learning-management-system (LMS) data — assignment scores, quiz attempts, time-on-task, and forum activity — rather than biometric or emotion-sensing data, which raise separate and more serious privacy questions. The AI techniques considered are adaptive content sequencing, intelligent tutoring/feedback systems, NLP-based conversational tutors, and predictive learning-analytics models. The paper presents a conceptual framework and a synthesis of existing evidence; it does not report results from a new deployed system, and it is not intended as a finished, production-ready platform.

V. PROPOSED METHODOLOGY

A. Review Approach

The review underlying this paper drew on peer-reviewed articles, systematic reviews, and meta-analyses indexed in Scopus, Google Scholar, Springer, ScienceDirect, and ERIC, published mainly between 2017 and 2026, using search terms combining "artificial intelligence," "personalized learning," "adaptive learning," "intelligent tutoring," and "higher education." Priority was given to systematic reviews and meta-analyses where available, since individual case studies vary widely in quality and sample size, and to studies that reported an actual comparison group rather than pre/post surveys alone.

ISBN NO: 978-81-180146-5-2

B. AI Techniques Considered

Four broad categories recur across the literature: adaptive learning engines that adjust content difficulty and sequencing based on performance; intelligent tutoring systems that provide step-by-step, feedback-driven guidance similar to a human tutor; NLP-based conversational tutors, including LLM-backed chatbots, that answer open-ended student questions; and learning-analytics/predictive models that flag at-risk students from LMS activity patterns rather than personalizing content directly.

C. Proposed Conceptual Framework Components

Building on the categories above, the proposed framework for a campus-level personalization system has five components: a learner-profile module that stores prior performance and stated learning goals; a content repository tagged by difficulty and topic; a personalization engine that combines adaptive sequencing with an NLP tutor for open-ended queries; a feedback and analytics layer visible to both students and faculty; and a human-oversight layer where instructors can review and override automated recommendations.

D. Evaluation Metrics

In line with the literature's own caution against relying on engagement numbers alone, four categories of metric are proposed: learning gain, measured through pre- and post-test comparisons rather than final grades alone; engagement, measured through time-on-task and voluntary tool usage rather than logins; retention/completion, tracked over a full semester rather than a single module; and perceived usefulness and trust, gathered through short student and faculty surveys, since adoption evidence shows tools that are not trusted simply go unused [9].

E. Ethical and Design Safeguards

Given how consistently privacy and bias concerns appear in the reviewed studies, the framework proposes: data minimization, collecting only what a personalization feature actually uses; periodic bias audits on any predictive or recommendation model, checking outcomes across student subgroups rather than only aggregate accuracy; an opt-out path for students who prefer conventional instruction; and a standing rule that AI recommendations are advisory to instructors rather than automatically binding.

VI. PROPOSED FRAMEWORK

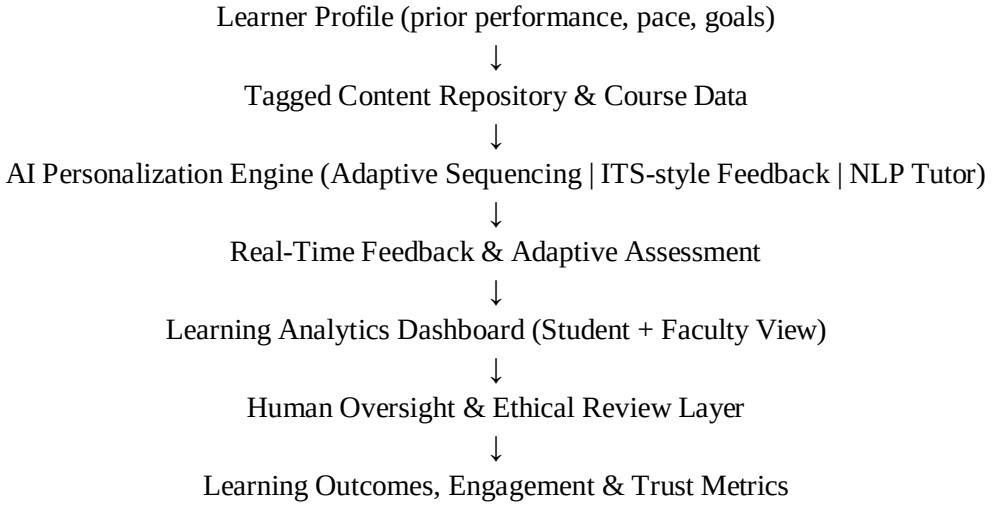


Fig. 1. Proposed AI-enabled personalized learning framework for higher education.

VII. EXPECTED OUTCOMES AND SIGNIFICANCE

Based on the patterns in the literature, a framework of this kind would be expected to show clearer gains in engagement and short-term task completion than in deep, long-term learning gains, and clearer gains in structured subjects such as programming and mathematics than in more open-ended, discussion-based courses [6], [7]. It should also surface, through the analytics layer, which students benefit most from adaptive content versus human-led support — information that is arguably as useful to instructors as the personalization itself.

Significance. For a Computer Applications course specifically, this topic sits at a natural intersection of the discipline's own subject matter — data structures, databases, and basic machine learning concepts — and a real institutional problem, which makes it a reasonable foundation for a student project that could later be extended into an actual small-scale adaptive quiz or recommendation tool built with course data.

ISBN NO: 978-81-180146-5-2

VIII. LIMITATIONS AND ETHICAL CONSIDERATIONS

- Much of the supporting evidence comes from short-duration studies; long-term retention and transfer of learning are less well documented.
- Positive results reported in early adopter institutions may not generalize to institutions with less technical infrastructure or support staff.
- Student and faculty AI literacy varies, and tools that assume a baseline of comfort with technology can widen rather than close gaps [10].
- Data-privacy and algorithmic-bias concerns are common in survey-based studies and need active mitigation, not just disclosure [9].
- Over-reliance on AI-generated answers, particularly with LLM-based tutors, risks reducing independent problem-solving practice if not deliberately designed against [11].

On the access side, any implementation in an Indian higher-education setting would also need to account for uneven internet connectivity and device access across students, since a personalization system that assumes constant connectivity risks disadvantaging exactly the students it is meant to help.

IX. FUTURE SCOPE

- A small pilot deployment of the adaptive-engine component within a single BCA course, comparing outcomes against a traditionally taught section.
- Integration of an LLM-based tutor with guardrails against direct answer-copying, paired with academic-integrity checks.
- A longer, multi-semester study of retention rather than single-module learning gain.
- An accessibility-first version of the framework designed for low-bandwidth and shared-device conditions.
- A bias-audit component that checks recommendation quality across students from different academic backgrounds and prior preparation levels.

X. CONCLUSION

The evidence on AI-enabled personalized learning in higher education points to a genuine but bounded benefit: engagement tends to improve, learning-outcome gains are real in structured subjects but modest and inconsistent elsewhere, and both students and faculty consistently flag data-privacy and fairness concerns that cannot be treated as an afterthought. This paper has synthesized that evidence and proposed a conceptual framework that tries to take the benefit seriously without overselling it — combining adaptive content, tutoring support, and analytics with human oversight and privacy safeguards built in from the start. The intent is not to claim that AI will personalize away the difficulty of teaching diverse classrooms, but to lay out a reasonably honest starting point for thinking about how it could help, and where a real implementation would need to be careful.

REFERENCES

[1] S. A. D. Popenici and S. Kerr, "Exploring the impact of artificial intelligence on teaching and learning in higher education," International Journal of Educational Technology in Higher Education, vol. 14, no. 22, 2017.

[2] G. Hardaker and L. E. Glenn, "Artificial intelligence for personalized learning: a systematic literature review," International Journal of Information and Learning Technology, vol. 42, no. 1, pp. 1–14, 2025.

[3] A. Fortuna, F. Prasetya, A. D. Samala, S. Rawas, S. Criollo-C, D. Kaya, M. Raihan, W. Andriani, D. Safitri, and R. A. Nabawi, "Artificial intelligence in personalized learning: A global systematic review of current advancements and shaping future opportunities," Social Sciences & Humanities Open, 2025.

[4] Farhood et al., "Artificial intelligence-based personalised learning in education: a systematic literature review," Discover Artificial Intelligence, vol. 5, art. 331, 2025.

[5] K. Bayly-Castaneda, M. S. Ramirez-Montoya, and A. Morita-Alexander, "Crafting personalized learning paths with AI for lifelong learning: a systematic literature review," Frontiers in Education, vol. 9, Aug. 2024.

[6] F. Naseer, M. N. Khan, M. Tahir, A. Addas, and S. M. H. Aejaz, "Integrating deep learning techniques for personalized learning pathways in higher education," Heliyon, vol. 10, no. 11, e32628, Jun. 2024.

[7] S. Steenbergen-Hu and H. Cooper, "A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning," Journal of Educational Psychology, vol. 106, no. 2, pp. 331–347, 2014.

[8] A. Létourneau, M. Deslandes Martineau, P. Charland, J. A. Karran, J. Boasen, and P.-M. Léger, "A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education," npj Science of Learning, vol. 10, art. 29, May 2025.

ISBN NO: 978-81-180146-5-2

[9] Y. Reina Marín, O. Cruz Caro, A. M. Carrasco Rituay, K. A. Guimac Llanos, D. Tarrillo Perez, E. Sánchez Bardales, J. N. Alva Tuesta, and R. Chávez Santos, "Ethical challenges associated with the use of artificial intelligence in university education," Journal of Academic Ethics, vol. 23, no. 4, pp. 2443–2467, 2025.

[10] K. Wang, W. Cui, and X. Yuan, "Artificial intelligence in higher education: the impact of need satisfaction on artificial intelligence literacy mediated by self-regulated learning strategies," Behavioral Sciences, vol. 15, no. 2, art. 165, 2025.

[11] M. Dogaru, O. Pisică, C.-Ș. Popa, A.-A. Răgman, and I.-R. Tololoi, "The perceived impact of artificial intelligence on academic learning," Frontiers in Artificial Intelligence, vol. 8, art. 1611183, Oct. 2025.

[12] J. Luo, C. Zheng, J. Yin, and H. H. Teo, "Design and assessment of AI-based learning tools in higher education: a systematic review," International Journal of Educational Technology in Higher Education, vol. 22, no. 1, 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_42

An Intelligent Student Assessment System Using Workflow Automation and Claude AI

Ms. V. Harini
M.Sc Computer Science
Shri Shankarlal Sundarbai Shasun Jain College for Women,
Chennai, India
hariniv409@gmail.com
Dr. M. Anita Indu
Head (i/c) & Assistant Professor
PG Department of Computer Science,
Shri Shankarlal Sundarbai Shasun Jain College for Women,
Chennai, India
m.anitaindu@shasuncollege.edu.in

Abstract
Due to an increasing demand for efficient and objective evaluation of academic performance, intelligent systems incorporating the abilities of workflow automation tools and large language models (LLM) are being developed. This report describes the design of an intelligent student assessment system which is based on n8n orchestration tools and pedagogic evaluation of Claude AI. The system is capable of managing the entire process of assessment, from collecting data to delivering feedback, and thus making it more efficient for large student groups.

There is shown that results obtained from evaluating student submissions with LLM can be consistent with human graders if the latter are provided with a rubric. Using n8n's agentic workflow management ability, the proposed intelligent assessment system will orchestrate the validation of student's submission, its evaluation with the help of Claude AI, data storage and reporting. This approach will resolve most issues associated with manual evaluation, while human supervision and protection of data will be maintained. A dedicated dashboard for the completed system will provide instant insights for teachers and students.

Keywords: Intelligent Student Assessment, n8n Workflow Automation, Claude AI, Large Language Models, Rubric-Based Assessment, Automated Grading, Human-in-the-Loop, AI-Based Feedback.

1. Introduction

Educational institutions use "Smart Grading" approaches in order to cope with large number of students and provide them with immediate and personalized feedback. The manual assessment is slow and biased in case of open-ended and technical assignments. Development of powerful AI enables automation of such procedures without sacrificing their qualitative aspects.
This paper describes a system based on workflow automation tool (n8n, Flowise, Dify, Langflow) and Claude AI large language model. The architecture of this system is built as a process from submission of the assignment to administrative reporting. The use of agentic workflow management capability allows building a modular and scalable system, while Claude AI is used for evaluation of free text and code submissions in accordance with the rubric.

2. Review of Literature

Large Language Models (LLMs) have increasingly been investigated for automating student assessment, particularly for open-ended and technical responses. Mok et al. [1] evaluated several LLMs, including Claude 3.5 Sonnet, GPT-4, GPT-4o, and Gemini 1.5 Pro, for grading undergraduate physics problems. Their study found that LLMs could produce grading errors and hallucinations when used without sufficient guidance. However, providing a detailed marking scheme improved the quality of AI-based grading and brought the results closer to human assessment. This demonstrates the importance of rubric- or mark-scheme-based evaluation in automated assessment systems.

Gaggioli et al. [2] investigated the reliability and validity of LLMs for automated assessment of university student essays. The study evaluated Claude 3.5, DeepSeek v2, Gemini 2.5, GPT-4, and Mistral 24B using a four-criterion assessment rubric. The results showed limited and inconsistent agreement between LLM-generated scores and human assessments, particularly for context-dependent criteria. The study therefore

ISBN NO: 978-81-180146-5-2

highlighted the need for human oversight when LLMs are used to assess open-ended academic work.
Bernik et al. [3] examined the use of LLMs for evaluating programming assignments and compared their performance with human evaluators. Their work demonstrated the potential of LLMs to reduce the effort involved in programming assessment while also highlighting differences in grading accuracy and consistency. The findings indicate that LLM-based assessment can be useful as an assisting mechanism, but careful evaluation and appropriate assessment criteria are necessary for reliable results.
Fajardo-Ramos et al. [4] conducted a scoping review of AI-supported assessment in higher education and emphasized the importance of human-in-the-loop approaches. Their study identified the need for rubric validation, traceability, data stewardship, transparency, audit trails, and human moderation when AI systems are involved in assessment. These findings support the use of human review for borderline or high-stakes assessment decisions rather than relying entirely on automated scoring.
Workflow orchestration can further improve the practical implementation of AI-based assessment. Salimi et al. [5] demonstrated the use of the n8n agent platform with AI for automated candidate screening, where n8n was used to coordinate multiple processing steps and an AI agent transformed unstructured information into structured outputs. Although the application domain was recruitment rather than education, the study demonstrates the applicability of n8n for connecting AI processing with structured workflow automation.
The reviewed studies show that LLMs have potential for automated assessment, but challenges remain in grading consistency, contextual understanding, transparency, and human oversight. Existing studies mainly investigate the grading capability of individual LLMs or specific educational tasks, while less attention is given to integrating the LLM into an end-to-end assessment workflow that includes submission validation, rubric-based evaluation, structured result storage, feedback generation, notification, and reporting. Therefore, the proposed system combines Claude AI with n8n workflow automation to provide a structured and auditable student assessment pipeline while retaining human supervision for important assessment decisions.

3. Existing System

Manual assessment of open-ended assignments is a significant bottleneck in higher education. Teachers face difficulties with providing immediate and personal feedback, which is crucial for the formative learning process, especially with an increase in class size. The delay between submission and assessment can take days or even weeks, during which the moment of learning has been already missed.
Currently existing automated assessment tools mainly operate with multiple choice or simple closed questions and cannot be used for assessing complex subjects, like physics, writing and programming. Though there are some automated essay-scoring tools, they are rigid systems trained on the fixed structure of rubric and cannot be adapted to the change of the assignment and subject.
Standalone LLMs, which operate without structured workflow, can show inconsistent and unreliable behaviour, generating varying scores not related to rubric criteria. Therefore, there is a need for a solution that will validate the input prior to evaluation and ground the model in rubric before storing the results securely. Lack of such controls in an AI-driven assessment system may compromise the integrity, consistency and privacy of the assessment process.

4. Objective of the Study

The main goal of this research is to develop an automated rubric-based assessment pipeline using workflow automation platform and Claude AI model. This system will decrease the burden of teachers and provide students with immediate and constructive feedback. Secondary goals of the study include development of a secure data management layer, which will store the results of assessment in structured and auditable format and notify relevant parties about it, as well as presenting the results on a dashboard. Overall goals are as follows:

- Automated rubric-based assessment of open-ended and technical student submissions.
- Human-in-the-loop review of borderline and high-stakes grading decisions.
- Immediate and constructive feedback, which will close the feedback gap in classrooms.
- Secure storage and visualization of student performance data on the dashboard.

5. Proposed System

Proposed system works through a linear progression from student interaction to data visualization, which is orchestrated end-to-end with the help of any workflow platform like n8n, Flowise, Dify or Langflow and assessed with Claude AI model at its core.

ISBN NO: 978-81-180146-5-2

5.1 System Architecture

Students will interact with the system through an online assessment or assignment portal, which will be used for collecting input. Core of the system is n8n workflow, which is an orchestrator of the entire pipeline logic and will control all processes of validation, assessment and data storage. Such architecture allows incorporating other LLMs and third party services (cloud storage, messaging, etc.) in the pipeline without redesigningit.

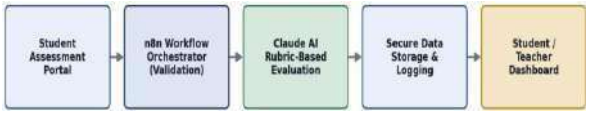


Figure 1.High-Level System Architecture

5.2 Comparative Analysis

rameters	N8n	owise	Dify	ngflow
flow mation	llent			
ration	Good	llent	llent	llent
gent ort	Good	llent	llent	llent
Integration	llent		Good	
ication	Good	llent	llent	llent

- **Workflow Automation: n8n** provides comprehensive workflow automation capabilities and supports the integration of multiple applications, APIs, databases, and AI services within a single workflow. **Flowise, Dify, and Langflow** primarily focus on AI and LLM-based workflows and provide comparatively more specialized workflow capabilities for developing intelligent applications.

- **LLM Integration: n8n** provides very good integration with various commercial and open-source LLM providers through its extensive collection of nodes and API integrations. **Flowise, Dify, and Langflow** provide strong native support for connecting different large language models, configuring model parameters, and developing LLM-powered applications.

- **AI Agent Support: n8n** supports the development of AI agents by combining LLMs with tools, APIs, memory, and workflow automation components. **Flowise, Dify, and Langflow** provide more specialized environments for designing AI agents, enabling developers to construct agent-based applications using visual workflow components and model integrations.

- **API Integration: n8n** offers extensive API integration capabilities and can connect with a wide range of external applications and services. **Dify** also provides strong API capabilities for integrating AI applications with external systems, while **Flowise and Langflow** support API-based deployment and integration primarily for LLM and AI application workflows.
- **RAG Applications: n8n** can be used to construct RAG workflows by connecting document sources, vector databases, embedding models, and LLMs. **Flowise, Dify, and Langflow** provide specialized components for building RAG applications, including document processing, embeddings, vector storage, retrieval, and LLM-based response generation.

Based on the above comparative analysis, an n8n workflow **is suggested for the** intelligent student assessment framework**.**

5.3 n8n Workflow Orchestration

After receiving the submission, n8n workflow will validate the data of the student in order to ensure that the

ISBN NO: 978-81-180146-5-2

input is complete and corresponds to the expected format. The filters will be used to detect and notify humans about irrelevant or ambiguous responses. After that, workflow will extract the content of the answer from the submission, preserving the structured paths for identifiers of the student until the reporting phase.

The content of the answer will be sent to Claude AI using secure API call along with the appropriate assignment prompt and rubric. n8n workflow will manage all interactions with the components, including retries and logging for auditing purposes. With the help of the multi-agent approach of n8n, the workflow will be able to trigger a separate agent to review the evaluation for consistency and ensure that it is grounded in the reference material.

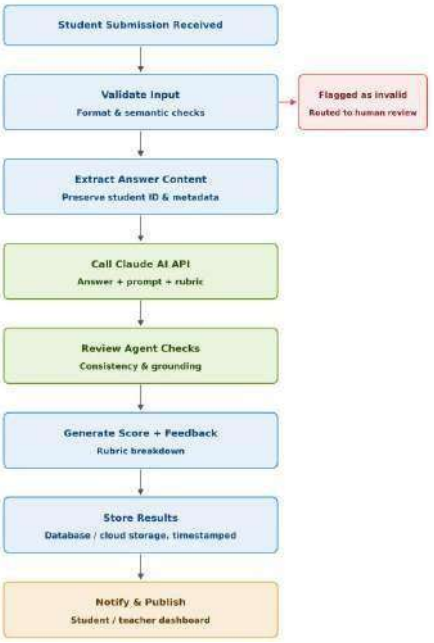


Figure 2.n8n Workflow Orchestration

5.4 Methodology of Assessment

System will use Claude AI model for assessment of the submission in accordance with the rubric. Provision of a detailed rubric to the model is the most important step in order to increase the accuracy of grading and make the results similar to human performance, rubric acts as a grounding mechanism, eliminating the subjective influence.

Claude AI assesses student submissions on multiple dimensions, such as technical accuracy, structure, and organization, comparing answer of the student with the "ideal response" given by the instructor or defined in the course material. After that, the system generates a score and qualitative, constructive feedback, which indicates the aspects of the answer, which need improvements. Structured output of the system includes the score, breakdown of the rubric categories and affective feedback, which is generated on-the-fly, allowing the students to correct their mistakes right after submission.

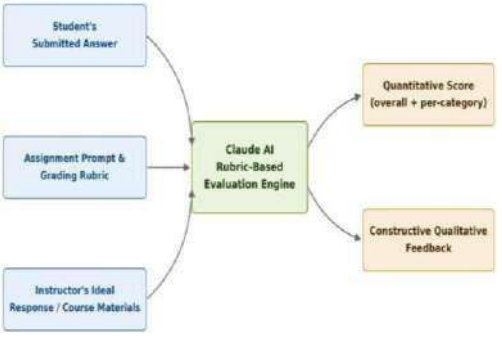


Figure 3. Rubic Based Assessment Methodology

ISBN NO: 978-81-180146-5-2

5.5 Data Management and Dashboarding

After the assessment is completed, results will be stored in a database or cloud storage in a centralized way, including the input submission, AI feedback, final score and timestamp for tracking the progress. Workflow will generate a report or notification for the student and teacher and feed Student/Teacher Dashboard with the data about completion rates, scores and learning path of the student.

5.6 Security, Privacy, and Ethical Controls

After the assessment is completed, results will be stored in a database or cloud storage in a centralized way, including the input submission, AI feedback, final score and timestamp for tracking the progress. Workflow will generate a report or notification for the student and teacher and feed Student/Teacher Dashboard with the data about completion rates, scores and learning path of the student.

5.7 Evaluation Framework and Performance

Efficiency of the system is evaluated with the help of standard LLM grading performance metrics. Previously conducted research has shown Pearson correlation coefficients of up to 0.91 and intraclass correlation of up to 0.95 between LLM scores and those of expert human graders, although the accuracy depends on the subject and complexity of the rubric. It is expected that the system will produce nearly perfect results for short answer questions and technical classification tasks and "substantial agreement" with human evaluation of open-ended essays. Auditing and refinement of the system are needed in order to maintain the performance levels in line with changing academic requirements.

5.9 Limitations

Though the architecture of n8n-Claude system is efficient, LLMs still can find it difficult to assess deep context-based or interpretive content and can produce scores which will be inconsistent with human grader scores. Black-box nature of AI decision-making can make it difficult for students to understand the logic of their scores. Also, there is little research on the effects of AI-generated feedback on the learning behaviour of the students compared to human-only feedback. Although n8n is a very flexible tool, configuration of complex workflows still requires technical expertise from the educational administrators. Further research in this field should focus on simplification of the process for non-technical users.

6. Conclusion

The proposed intelligent assessment system offers a robust solution for modern educational needs by combining the workflow with Claude AI's evaluative capabilities. This study also gives us a comparative analysis of the workflow automation using N8N, Flowise, Dify and Langflow.

References

[1] R. Mok, F. Akhtar, L. Clare, C. Li, J. Ida, L. Ross, and M. Campanelli, “Using AI Large Language Models for Grading in Education: A Hands-On Test for Physics,” *arXiv preprint arXiv:2411.13685*, 2024.

[2] A. Gaggioli, G. Casaburi, L. Ercolani, F. Collovà, P. Torre, and F. Davide, “Assessing the Reliability and Validity of Large Language Models for Automated Assessment of Student Essays in Higher Education,” *arXiv preprint arXiv:2508.02442*, 2025.

[3] A. Bernik, D. Radošević, and A. Čep, “A Comparative Study of Large Language Models in Programming Education: Accuracy, Efficiency, and Feedback in Student Assignment Grading,” *Applied Sciences*, vol. 15, no. 18, p. 10055, 2025, doi: 10.3390/app151810055.

[4] D. C. Fajardo-Ramos, A. Chiappe, and J. Mella-Norambuena, “Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities,” *Frontiers in Education*, vol. 10, 2025, doi: 10.3389/feduc.2025.1710992.

[5] W. K. Salimi, T. F. Kusumasari, and S. Suakanto, “Leveraging N8N Agent Platform and AI for Automated Job Candidate Screening,” *2025 12th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 2025, doi: 10.1109/ICAICTA67604.2025.11335111.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_43

LLM-BASED CYBER THREAT INTELLIGENCE ANALYSIS FOR PERSONALISED CYBERSECURITY EDUCATION

Ms. Janani. J
M.Sc Computer Science
Shri Shankarlal SundarbaiShasun Jain College for Women,
Chennai, India
jananijinendiran288@gmail.com
Dr. M. Anita Indu
Head(i/c) & Assistant Professor
PG Department of Computer Science
Shri Shankarlal SundarbaiShasun Jain College for Women,
Chennai, India
m.anitaindu@shasuncollege.edu.in

ABSTRACT
Large language models (LLMs) are becoming more common in using cyber threat intelligence (CTI) for analysis and in teaching cybersecurity.But these areas are usually studied separately.This paper introduces a seven-step framework that connects collecting raw CTI, using LLMs for analysis, creating knowledge graphs, structuring with HybridRAG, checking if threats are real, tracking learners, making personalized content, and giving adaptive feedback.This framework uses verified threat information as a base for changing education, and uses what learners do to guide how the system adapts.Probabilistic threat checking is part of the design to lower risks like false info or wrong content, though it needs more testing.Earlier studies show some skills in extracting threats, making graphs, teaching, and creating scenarios, but they don't fully connect all steps.The framework highlights where information comes from, limits how much personalization is done, has human checks, and focuses on privacy.Keeping data long-term is a future way to evaluate the framework.
Keywords—Large Language Models, Cyber Threat Intelligence, Personalised Learning, Cybersecurity Education, Knowledge Graphs, Retrieval-Augmented Generation, Adaptive Learning

I. INTRODUCTION

The fast-growing and ever-changing nature of cybersecurity threats makes it harder to keep educational material up to date with real-world conditions. Recent progress in Large Language Models (LLMs) gives tools for automatically handling unstructured cybersecurity data, making summaries in natural language, and offering interactive tutoring.However, using LLMs directly in cybersecurity education brings risks like incorrect facts, using old threat info, and giving bad personalized learning.

Using LLMs in CTI analysis is driven by the fact that generative AI can be used both for good and bad. Attackers could use LLMs to speed up tasks like recon or phishing, so there's a need for automated tools to process threat feeds, advisories, and incident data.Prompt injection and bypassing safety rules also push for better system controls.Educational systems using CTI should clean and check threat data before using it for teaching. Furthermore, new threats like prompt injection and bypassing safety measures show the need for strong system controls. Systems that use CTI for education must make sure any harmful content in the raw CTI feeds is removed before it's used in teaching parts.

ISBN NO: 978-81-180146-5-2

Though studies look at using LLMs for summarizing threat reports or making standalone educational chatbots, research remains separated into two areas: operational CTI analysis and educational adaptation. This paper addresses the main question:

Primary Research Question: How can using LLMs in operational CTI analysis, building knowledge graphs, checking if threats are real, together with learner profiles and education rules, form a single framework that supports personalized cybersecurity learning, and what criteria should be used to test this framework in the future?

The literature shows there's a big gap: current studies look at CTI accuracy and learning engagement separately, and even systems like CyberQ link knowledge graphs to question generation, but none connect CTI checking, learner profiles, and adaptive testing in a full framework.

To address this, this paper makes three main contributions:

• Integrated Conceptual Pipeline: Creates a seven-step framework that connects collecting and checking CTI data directly to personalized teaching and adaptive assessment, all linked through a shared Cybersecurity Knowledge Graph and clear data connections between steps.

• Combining Retrieval and Verification Rules: Explains how structured knowledge graphs, HybridRAG, and probability-based threat checking can be used together to reduce wrong or unsupported content in learning, though this still needs testing.

• Mapping and Governance Model: Shows how threat features relate to education systems like CyberQ, SENSAI, CyberMentor, and CyberBOT, and deals with making sure assessments are fair and protecting privacy.

II. LITERATURE REVIEW AND FOUNDATIONS OF LLM-BASED CTI ANALYSIS

Operational CTI analysis turns mixed threat reports, advisories, and other unstructured sources into formats that machines can understand.

Recent studies have used LLM-based extraction, structured knowledge representation, retrieval, and verification to make CTI more useful for further analysis [1]–[5].

A. LLM-Based CTI Extraction and Knowledge Graphs

Studies have looked into using LLMs and transformer models to identify attack techniques, entities, and relationships in CTI reports.

Nguyen et al.studied LLM-assisted attack-technique identification using models like Llama and SciBERT in the MITRE ATT&CK framework [1].Knowledge graphs add value by showing how threat actors, malware, vulnerabilities, indicators, and attack techniques are connected.KGAgent4CTI shows how LLMs can help build and use threat-intelligence graphs, helping with CTI reasoning [3].

B. Retrieval and Threat-Information Grounding

Retrieval-Augmented Generation (RAG) uses external evidence rather than only model data.

For CTI, graph-based retrieval keeps track of relationships that might be missed with simple similarity checks.Hamzić et al.tested Vector RAG, GraphRAG, Agentic GraphRAG, and HybridRAG for CTI question-answering and found that hybrid graph-text retrieval works well for multi-step questions [4].These findings support using a layered approach with both graph and text retrieval in the system.

C. Threat Veracity and CTI Validation

Just having reliable information isn't enough to ensure it's accurate and up-to-date.

VeraCTI solves this with CTI processing, enrichment, and cross-source validation.It gives learners a safe way to critically assess threat info [5].These verification tools encourage a clear

ISBN NO: 978-81-180146-5-2

validation step before CTI is used as educational content.

D. LLM-Based Cybersecurity Education and Personalisation

Studies have explored several key features in cybersecurity education.

CyberQ uses a cybersecurity knowledge graph to help generate questions and answers with LLMs [6].CyberMentor uses an agentic, RAG-based approach to meet different learner needs [7].SENSAI looks at LLMs as interactive tutors and reports high learner engagement in its tests [9].CyberBOT uses course-specific retrieval and a cybersecurity ontology to limit and check educational responses [10].These efforts show progress in question generation, tutoring, and guardrails based on ontology, but they focus on teaching rather than a full CTI-to-learning process.

Literature Synthesis and Research Gap

Overall, the reviewed work shows progress in CTI extraction, graph-based retrieval, validation, tutoring, question generation, and learner adaptation.

However, these features are spread across different systems.CyberQ links knowledge graphs with question generation, SENSAI focuses on interactive tutoring, and VeraCTI works on CTI validation [5], [6], [9].The existing research doesn't offer a unified framework that brings together CTI intake, veracity checks, learner profiling, personalized content, and adaptive assessment.This lack of integration inspires the new system proposed in this study.

III. PROPOSED SYSTEM

The proposed system is a seven-step conceptual pipeline that connects operational CTI processing with personalized cybersecurity learning.

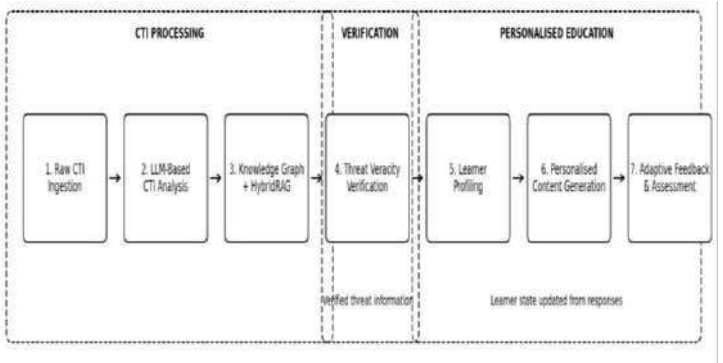


Fig. 1. Proposed seven-stage LLM-based CTI–personalised cybersecurity education framework.

Table I summarises the input, processing, and output associated with each stage of the proposed framework.

Table I. Seven-stage proposed framework and information flow.

Stage	Core Process	Input	Output
1	Raw CTI Ingestion	Reports, CVEs, advisories, feeds	Normalized CTI
2	LLM-Based CTI Analysis	CTI text	Extracted threat entities and relationships
3	KG + HybridRAG	Threat information	Structured knowledge and retrieved context
4	Veracity Verification	Candidate CTI	Verified CTI

ISBN NO: 978-81-180146-5-2

5	Learner Profiling	Student activity	Dynamic learner profile
6	Personalised Generation	Verified CTI + learner profile	Personalised exercises
7	Adaptive Feedback & Assessment	Learner response	Feedback + updated learner state

Rather than handling CTI analysis and learning separately, the system moves verified threat data through stages that change raw CTI into tailored educational content.

Overall Architecture

Stage 1 – Raw CTI Ingestion

CTI is gathered from sources like threat reports, vulnerability databases, security advisories, incident reports, and other structured or unstructured threat feeds.

This becomes the original input for the rest of the process.

Stage 2 – LLM-Based CTI Analysis

An LLM goes through the collected CTI to find cybersecurity entities and their relationships, such as threat actors, malware, vulnerabilities, attack techniques, indicators of compromise, and targeted systems.

This stage turns unstructured CTI into structured threat data.

Stage 3 – Knowledge Graph and HybridRAG Structuring

The extracted data is organized into a cybersecurity knowledge graph, while Hybrid Retrieval-Augmented Generation (HybridRAG) uses both graph relationships and textual retrieval.

This provides a strong base for later generation and retrieval activities.

Stage 4 – Threat Veracity Verification

Before using threat information for education, the system checks its accuracy against trusted CTI sources and retrieved evidence.

Information with weak support can be noted for review instead of being shared directly with learners.

This stage is meant to lower the chance of getting wrong or unclear information from large language models.

Stage 5 – Learner Profiling

Verified threat information is mixed with a learner profile that includes things like their level of knowledge in cybersecurity, how far they've come in their learning, how familiar they are with certain topics, their performance, and where they might be lacking in knowledge.

This learner profile helps decide how to present the same threat intelligence information.

Stage 6 – Personalised Content Generation

The system changes verified threat intelligence into educational material that fits the learner's needs.

For instance, a beginner might get simple explanations and basic examples, while a more advanced learner might get deeper technical analysis, reasoning about how attacks happen, or tasks based on real situations.

Stage 7 – Adaptive Feedback and Assessment

The system checks how the learner responds to find any misunderstandings or gaps in knowledge.

ISBN NO: 978-81-180146-5-2

It can then adjust future explanations, questions, difficulty levels, and learning activities.Human teachers are still involved in making major or unclear decisions about the learning process.

Information Flow

The full process can be shown as: Raw CTI → LLM Analysis → Knowledge Graph + HybridRAG → Veracity Verification → Learner Profile → Personalised Content → Adaptive Feedback & Assessment.

A key part of this framework is how the CTI-processing pipeline connects with the learner-adaptation pipeline.The results from threat analysis aren't shown directly to the learner.Instead, they go through a process of retrieving and verifying information before being matched with the learner's characteristics and turned into education content.

Relationship to Existing Systems

Existing research covers parts of this pipeline.

LLM-TIKG and similar CTI extraction methods use LLMs and knowledge graphs to structure threat intelligence, while VeraCTI focuses on analyzing and checking CTI in an educational setting.CyberQ uses knowledge graphs for making cybersecurity questions, and SENSAI shows how LLMs can be used for interactive cybersecurity tutoring.CyberBOT looks into using ontologies for reliable cybersecurity education.The proposed system brings all these together by linking CTI gathering and checking with learner profiles, content generation, and adaptive assessment.

Scope of the Proposed System

This system is a conceptual idea rather than a fully built and tested system.

Its parts are based on abilities shown in existing research.To check how well it works, real testing is needed in areas like accuracy of facts, quality of retrieved information, how well it adapts to learners, educational value, speed, how well it scales, resistance to harmful attacks, and overall learning results.

IV. PERSONALISED CYBERSECURITY TEACHING AND SYSTEMS

Traditional cybersecurity teaching often uses the same lab exercises for all students, no matter their different experiences.

The new system focuses on learning that adapts to each student while still following standard frameworks like CyBOK.

A. Understanding Students and Matching Their Needs

To make learning fit each student, we need to clearly know their current level and abilities.

This part captures real evidence, not just guesses.It includes what skills a student has in CyBOK, how much technical experience they have, how often they make mistakes in labs, how confident they are, and what kind of support they need.

For instance, if a new report about a new way to gain extra permissions is added, the system will change how it teaches based on the student's level.

A new learner would get a simple introduction, important terms, and a step-by-step example.An advanced learner would get raw data, unclear logs, and be asked to find solutions without clear answers.These different teaching styles are part of the plan and need to be tested to see if they work well.

B. Creating Content and Matching It to Education Systems

This part connects real threats with existing education tools.

ISBN NO: 978-81-180146-5-2

• CyberQ: This system uses knowledge graphs to create cybersecurity questions and answers with large language models.

In this system, verified threat information can help create specific questions.Also, knowing each student's profile helps decide how much and in what order to tailor the learning.

• CyberMentor: This tool helps students build their skills and understand their career paths by showing them the right steps to take based on their current skills and what is needed in the industry.

• Personalized Secure Programming: Some studies show how to use AI to create customized learning experiences.

In this system, this kind of personalized learning can be a part of the process, but it is not the full seven-step system.

C. Helping With Labs and Keeping Safety in Mind

This part adds support during lab work while making sure it's educational and safe.

• SENSAI: This is a tutor that works directly with students when they use the terminal.

It has been tested with many students and gives immediate help without giving direct answers.

• CyberBOT: This uses knowledge from the field to check if the AI's answers are safe and accurate.

It helps avoid misleading or unsafe advice during cybersecurity learning.

V. MAKING SURE ASSESSMENTS ARE FAIR, PRIVATE AND GOVERNED

Making Sure Assessments Are Fair, Private and Governed Using AI in cybersecurity teaching brings important issues about fairness, privacy and how it's managed. These need to be handled along with the technology itself.

A. Keeping Assessments Honest

With many AI tools available, students might use them instead of doing their own work, or become too dependent on them. Studies show that using AI affects how well students do when they are tested without help.So assessments should check the learning process and actual skills, not only the final written answers.

To keep assessments honest, the system uses a method where AI evaluates against AI. This is better than using tools that often mislabel students who aren't native English speakers.Instead, the system focuses on checking the steps taken to solve problems, commands used, reasoning behind solutions, and even oral defenses, not just the final written work.

B. Keeping Data Safe and Managing Risks Locally

These systems handle sensitive data such as code, errors and performance. Sending this to cloud services raises privacy issues. Using local, on-site AI models and federated learning can help manage these risks. But using local systems doesn't automatically meet all legal standards.They still need proper rules on who can access the data, how long it's kept, and steps to protect it.

C. Explaining AI and Limitations of Human Control

Explainable AI and documentation can help show how the system makes decisions.

However, these are just tools to manage risks, not guarantees that AI is always clear.Complex models might act unpredictably, like suggesting unsafe lab instructions or misclassifying student experiments as harmful.This means human review is important before making big decisions.

Therefore, having human oversight is still necessary. Teachers must check key parts of the process, like confirming how different threats are matched, checking generated lab scenarios, and reviewing important assessment results.

ISBN NO: 978-81-180146-5-2

VI. DISCUSSION, LIMITATIONS AND FUTURE DIRECTIONS

The research shows that each part of the proposed system—like CTI extraction, Hybrid RAG, terminal tutoring, and KG question generation—has been used on its own in different studies. But none of the earlier studies reviewed here tested the full seven-step system all together in one go. So, there are still unanswered questions about how well the whole system works together, how it handles different situations, how fast it runs, how it scales, and how problems in one part might affect others.

LIMITATIONS AND THREATS TO VALIDITY

The framework is still just an idea and hasn't been tested as a real working system.

The steps in order might slow things down, limit how much it can handle at once, and cause problems to spread.Putting it all together needs matching data formats, keeping track of where information comes from, and managing how the parts work together. Things like misleading CTI information, bad prompts, incorrect retrieval, and missed threats could affect the educational content that comes after. So future work should test how fast it works, how much it can handle, backup plans, and how well it handles tricky or harmful inputs.

VII. CONCLUSION

This paper introduced a seven-step framework that connects real-world LLM-based CTI analysis with personalized cybersecurity education.

By checking the truth of threats before creating educational content, and looking at what learners actually do before making adjustments, the framework sets up a clear path for meaningful learning. The system uses Hybrid RAG, knowledge graphs, tutoring during terminal use (SENSAI), and rules for keeping things safe (CyberBOT) to stop fake or wrong information and unsupported answers. More real-world testing is needed to see how well the whole system works in actual educational settings.

REFERENCES

[1] H. C. Nguyen, S. Tariq, M. B. Chhetri, and B. Q. Vo, "Towards effective identification of attack techniques in cyber threat intelligence reports using large language models," in Companion Proc. ACM Web Conf. (WWW), 2025, pp. 942–946, doi: 10.1145/3701716.3715469.

[2] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," arXiv preprint arXiv:2405.04760, 2024.

[3] L. Yang, H. Zhu, J. Mu, and Q. Li, "KGAgent4CTI: Unlocking the power of LLM in threat intelligence," Cybersecurity, vol. 9, art. no. 76, 2026, doi: 10.1186/s42400-025-00458-2.

[4] A. Hamzić, M. Stevanović, and U. Yavanoğlu, "Beyond RAG for cyber threat intelligence: A systematic evaluation of graph-based and agentic retrieval," arXiv preprint arXiv:2604.11419, 2026.

[5] S. Janku, H. Xiao, and T. Caris, "An educational and validation tool for cyber threat intelligence leveraging large language models," in Proc. 3rd Int. Workshop Cyber Security Education for Industry and Academia (CSE4IA), 2025.

[6] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "CyberQ: Generating questions and answers for cybersecurity education using knowledge graph-augmented LLMs," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 21, pp. 23164–23172, 2024, doi: 10.1609/aaai.v38i21.30362.

ISBN NO: 978-81-180146-5-2

[7] T. Wang, N. Zhou, and Z. Chen, "CyberMentor: AI powered learning tool platform to address diverse student needs in cybersecurity education," arXiv preprint arXiv:2501.09709, 2025.

[8] M. Frazier and K. Damevski, "Towards personalizing secure programming education with LLM-injected vulnerabilities," arXiv preprint arXiv:2604.13955, 2026.

[9] C. Nelson, A. Doupé, and Y. Shoshitaishvili, "SENSAI: Large language models as applied cybersecurity tutors," in Proc. 56th ACM Tech. Symp. Comput. Sci. Educ. (SIGCSE TS), 2025, pp. 833–839, doi: 10.1145/3641554.3701801.

[10] C. Zhao, R. De Maria, T. Kumarage, K. S. Chaudhary, G. Agrawal, Y. Li, et al., "CyberBOT: Ontology-grounded retrieval augmented generation for reliable cybersecurity education," in Proc. 34th ACM Int. Conf. Inf. Knowl. Manag. (CIKM), 2025, pp. 6752–6756, doi: 10.1145/3746252.3761478.

[11] M. Ohm, C. Bungartz, F. Boes, and M. Meier, "Assessing the impact of large language models on cybersecurity education: A study of ChatGPT's influence on student performance," in Proc. 19th Int. Conf. Availability, Reliability Security (ARES), 2024, Art. no. 104, pp. 1–7, doi: 10.1145/3664476.3670446.

[12] M. S. Torkestani and T. Mansouri, "Machine vs machine: Using AI to tackle generative AI threats in assessment," arXiv preprint arXiv:2506.02046, 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_44

An Explainable Machine Learning Framework for Breast Cancer Survival Prediction: A Comparative Study

Ms.S.Sudha 1, D.Pavithra 2, N.Shanjana 3

1 Assistant Professor, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

2,3 Students, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Tamil Nadu, India

***Abstract*—*Predicting survival outcomes in breast cancer patients is difficult precisely because so many factors — demographic, clinical, pathological, and treatment-related — interact in ways that are hard to disentangle. The Cox Proportional Hazards (CPH) model has long been the default tool for this kind of time-to-event analysis, but it assumes relationships that don't always hold in real clinical data. Machine learning offers a way to capture nonlinear interactions that CPH misses, though this comes at the cost of interpretability — a serious concern in a healthcare setting where clinicians need to trust and understand a model's reasoning. This paper proposes an explainable framework that compares CPH against two survival-specific machine learning approaches, Random Survival Forest and survival-oriented gradient boosting, using publicly available SEER-based breast cancer data. Rather than assuming machine learning will outperform the statistical baseline — an assumption recent literature does not support — the framework is built around a fair, metric-driven comparison using the Concordance Index, time-dependent AUC, and Integrated Brier Score. SHapley Additive exPlanations (SHAP) is layered on top of the best-performing model to surface which clinical variables actually drive its predictions. The result is less a claim that "ML wins" and more a reusable, transparent pipeline for evaluating predictive performance against interpretability in oncology research.*

Keywords*—*Breast Cancer, Survival Prediction, Machine Learning, Random Survival Forest, Cox Proportional Hazards, Explainable AI, SHAP

I. INTRODUCTION

Not all breast cancer patients face the same odds. Age, tumor grade and stage, lymph node involvement, hormone receptor status, and treatment history all shape survival outcomes, often in combination rather than isolation. Breast cancer is also not a rare concern to begin with: it was the most frequently diagnosed cancer among women worldwide in 2022, with an estimated 2.3 million new cases and 670,000 deaths recorded globally that year [12]. Being able to predict survival outcomes accurately therefore matters not just for individual prognosis but for identifying high-risk patient groups that might benefit from closer monitoring or more aggressive intervention.

For decades, the Cox Proportional Hazards model has been the standard tool for this kind of time-to-event analysis. It is interpretable and statistically well-understood, but it rests on a proportional hazards assumption that does not always match messy, real-world clinical data, and it struggles to capture nonlinear interactions between variables.

This is where machine learning has drawn increasing interest. Techniques like Random Forests, SVMs, neural networks, XGBoost, Random Survival Forest, and DeepSurv have all been applied to cancer survival prediction with varying success. Kalafi et al. [2], for instance, showed that ML models could meaningfully learn from clinical breast cancer data. More recent work has gone further, directly comparing CPH against several ML approaches while using SHAP to keep the models interpretable [3].

But "more complex" does not automatically mean "more accurate." A 2025 meta-analysis spanning 21 studies found that ML models, on the whole, performed comparably to CPH rather than decisively outperforming it [4]. That finding is the real motivation for this paper: instead of assuming machine learning is the better tool, we set out to test that assumption directly, on breast cancer data specifically, with interpretability treated as a first-class requirement rather than an afterthought.

ISBN NO: 978-81-180146-5-2

Contribution of this study. This paper's contribution is a comparative, explainability-first framework — rather than a single new algorithm — that positions CPH, Random Survival Forest, and gradient boosting on equal footing, evaluated with survival-appropriate metrics, and interpreted through SHAP so the "winning" model's reasoning is auditable rather than assumed.

II. LITERATURE REVIEW

Machine learning's role in cancer survival prediction has expanded considerably in the last several years. Deepa and Gunavathi [1] surveyed ML and deep learning techniques across cancer types and positioned survival modeling as one of the more promising applications of data-driven methods in oncology.

Focusing specifically on breast cancer, Kalafi et al. [2] compared Multilayer Perceptron, Random Forest, Decision Tree, and SVM models on clinical data, reporting accuracies of 88.2%, 83.3%, and 82.5% respectively — an early signal that ML could extract genuinely useful patterns from this kind of data, not just noise.

More recent work has shifted toward survival-specific algorithms — ones built to handle censored data and time-to-event outcomes properly, rather than treating survival as a plain classification problem. Baidoo and Rodrigo [3] ran a fairly comprehensive comparison — CPH, Random Survival Forest, DeepSurv, Random Forest, XGBoost, SVM, and LightGBM — on a SEER dataset of breast cancer patients. The survival-specific models performed comparably to each other, and their SHAP analysis pointed to lymph-node stage, regional node positivity, tumor grade, progesterone status, and age as consistently important predictors. They also noted that conventional (non-survival) ML classifiers do not handle censored data the way survival-specific models do — a distinction that is easy to overlook but matters a lot for validity.

Evidence from other cancer types reinforces the point that no algorithm wins universally. Kim et al. [5] compared DeepSurv, RSF, and CPH for oral cancer and found DeepSurv ahead on concordance index (0.781 test), followed by RSF (0.764) and CPH (0.694). Zhang et al. [6] applied RSF to 30-year survival prediction using SEER data for pediatric and young adult cancer survivors. Mohammed et al. [7], by contrast, working on colorectal cancer gene-expression data, found CPH and RSF performing comparably rather than one clearly beating the other. The takeaway across all of this is not "ML is better" or "CPH is better" — it is that performance is dataset- and cancer-type-dependent, which is exactly why a comparative approach matters more than defaulting to whichever model is currently fashionable.

Interpretability has become its own research thread. Alabi et al. [9] used SHAP and LIME to explain ML-based survival predictions for nasopharyngeal cancer, showing that black-box models can be made auditable without sacrificing performance. Similar SHAP-based approaches have been applied for oropharyngeal cancer [10] and prostate cancer bone metastases [11], underscoring that explainability is now an expected component of survival modeling work, not an optional extra.

A broader systematic review of breast cancer survival prediction found real potential in AI-based approaches but also flagged a recurring weakness: most studies rely on internal validation only, with external validation on genuinely independent datasets still relatively rare.

Literature Gap

Taken together, the literature shows that ML can be competitive with CPH for breast cancer survival prediction, but results vary enough across studies and datasets that no single model can be assumed superior. What is less common is a framework that treats the comparison itself as the object of study — evaluating survival-specific ML against CPH using consistent, survival-appropriate metrics, with SHAP built in from the start rather than added on as an afterthought. This paper's framework is designed to fill that gap.

ISBN NO: 978-81-180146-5-2

Summary of Reviewed Studies

Study	Cancer / Data	Models Compared	Key Finding
Kalafi et al. [2], 2019	Breast, clinical (Malaysia)	MLP, RF, DT, SVM	MLP best (88.2% acc.); tumor-related features most predictive
Baidoo & Rodrigo [3], 2025	Breast, SEER	CPH, RSF, DeepSurv, RF, XGBoost, SVM, LightGBM	Survival-specific models comparable; SHAP: lymph node, grade, PR status, age most important
Huang et al. [4], 2025	Multi-cancer, meta-analysis (21 studies)	Various ML vs. CPH	No overall superior ML performance over CPH (SMD 0.01)
Kim et al. [5], 2019	Oral cancer	DeepSurv, RSF, CPH	DeepSurv best (C-index 0.781); RSF 0.764; CPH 0.694
Zhang et al. [6], 2023	Pediatric/AYA cancer, SEER	RSF	RSF used for 30-year survival and second-tumor risk prediction
Mohammed et al. [7], 2021	Colorectal, gene expression	CPH, RSF	Comparable performance; CPH not outperformed by RSF
Alabi et al. [9], 2023	Nasopharyngeal, SEER	Stacked ML + SHAP/LIME	Explainability achieved without sacrificing accuracy

TABLE I. SUMMARY OF KEY REVIEWED STUDIES

III. PROBLEM STATEMENT

Breast cancer survival depends on a web of interacting clinical and demographic variables. CPH offers a well-established, interpretable way to model this, but its linearity assumptions limit how well it can capture complex interactions. Machine learning models can capture those interactions, but often at the cost of transparency — and if a conventional (non-survival-specific) classifier is used, censored observations may not be handled correctly at all.

Just as importantly, the literature does not support a blanket assumption that ML outperforms CPH for this task [4]. So the central question this paper addresses is: How can traditional and machine-learning-based survival models be fairly compared for breast cancer survival prediction, in a way that keeps the resulting predictions interpretable rather than opaque?

IV. RESEARCH OBJECTIVES AND SCOPE

A. Objectives

• Build a data-driven framework for breast cancer survival prediction.
• Compare CPH against machine-learning-based survival models on equal footing.
• Evaluate all models using survival-specific performance metrics rather than plain classification accuracy.
• Use SHAP to identify which variables actually drive predictions.
• Examine the tradeoff between predictive performance and interpretability directly, rather than assuming one model dominates.

B. Scope

This study is deliberately scoped to a single cancer type (breast cancer) to avoid confounding results across biologically distinct diseases; structured clinical and demographic data rather than imaging or genomic data; publicly available datasets, ideally SEER-derived; time-to-event modeling with properly handled censored observations; three models — CPH, Random Survival Forest, and survival-oriented gradient boosting; and SHAP-based interpretation of the best-performing model. This is a research and analytics framework, not a clinical decision-making tool, and it is not intended to be used as one.

V. PROPOSED METHODOLOGY

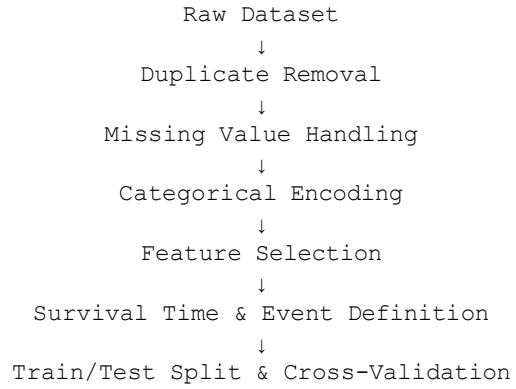
A. Dataset

A publicly available SEER-based breast cancer dataset is proposed, following precedent from prior comparative studies [3]. The dataset should include survival time, event/censoring status, age, tumor stage

ISBN NO: 978-81-180146-5-2

and grade, lymph node status, receptor/biomarker data where available, and relevant treatment variables. Exact fields will depend on the specific dataset selected; comparable published SEER-based breast cancer cohorts have ranged from roughly 4,000 to over 20,000 patients depending on inclusion criteria and the study years covered [3].

B. Preprocessing



Special care will be taken to avoid data leakage — excluding any variable that would not actually be available at the point a real prediction would be made.

C. Models

Cox Proportional Hazards serves as the interpretable statistical baseline. Random Survival Forest handles nonlinear interactions and censored data natively. A survival-oriented gradient boosting model tests whether boosting-based methods, increasingly common in cancer prediction research, offer an edge over RSF. All three will be trained and evaluated under identical experimental conditions to keep the comparison fair.

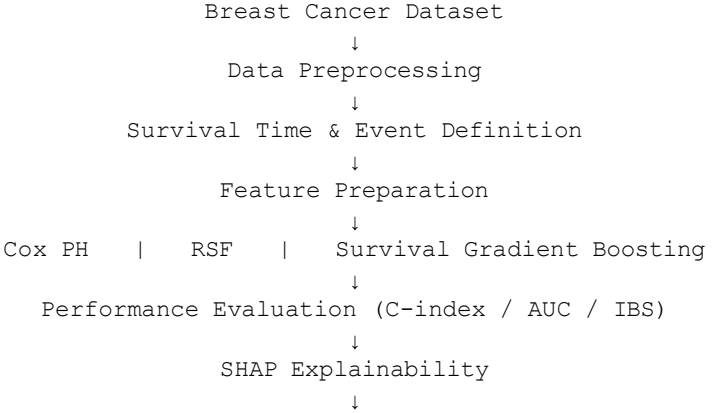
D. Evaluation Metrics

Three survival-appropriate metrics will be used: the Concordance Index (C-index), which measures how well the model ranks patients by risk; time-dependent AUC, which measures discriminatory power at specific follow-up time points; and the Integrated Brier Score (IBS), which measures prediction error across the entire follow-up period. These are used instead of plain accuracy because survival data is fundamentally about ranking risk over time, not binary classification.

E. Explainability (SHAP)

Once the best-performing model is identified, SHAP will be applied to answer: which variables matter most, are they consistent across the different models, and how does a high vs. low value of a given feature shift predicted risk? It is worth being explicit that SHAP explains association with the model's output, not biological causation.

VI. PROPOSED FRAMEWORK



ISBN NO: 978-81-180146-5-2

Important Predictors & Model Comparison

Fig. 1. Proposed explainable ML framework for breast cancer survival prediction.

VII. EXPECTED OUTCOMES AND SIGNIFICANCE

The framework is expected to show whether a survival-specific ML approach can match or exceed the CPH baseline on the chosen dataset, without presupposing the answer either way. It should also surface which variables matter most; prior work points toward lymph-node involvement, tumor grade, hormone receptor status, and age [2], [3], though the actual ranking will depend on the specific dataset used and should not be assumed to transfer directly from prior studies.

Significance. This work sits at the intersection of data preprocessing, healthcare analytics, survival modeling, and Explainable AI — making it a natural fit for a Computer Applications context, since it demonstrates a real, end-to-end computational pipeline applied to a genuine oncology problem rather than a toy dataset.

VIII. LIMITATIONS AND ETHICAL CONSIDERATIONS

- Results from a single dataset may not generalize to other patient populations.
- Public datasets can be incomplete or oversimplified relative to full clinical records.
- Retrospective data carries inherent selection and reporting bias.
- Strong predictive performance does not by itself establish clinical utility — that requires external validation.
- SHAP explains model behavior, not medical causation, and should not be over-interpreted as such.

On the ethics side, SEER data is de-identified and made available for public research use, which mitigates most privacy concerns, but any dataset used should be confirmed as properly authorized for research purposes before use, and results should not be framed as clinically validated findings.

IX. FUTURE SCOPE

- External validation on an independent, ideally multi-center, dataset.
- Deep survival models such as DeepSurv, for comparison against the models used here.
- Integration of genomic or imaging data alongside clinical variables.
- Federated learning approaches, enabling multi-institution collaboration without centralizing sensitive patient data.
- A prototype visual decision-support interface for exploring model predictions.

X. CONCLUSION

This paper proposes an explainable machine learning framework for breast cancer survival prediction, comparing CPH against Random Survival Forest and survival-oriented gradient boosting under a consistent, survival-appropriate evaluation setup. The broader literature makes clear that machine learning holds real promise for this task, but it does not support the assumption that ML will automatically outperform well-established statistical methods like CPH — which is exactly why comparative, metric-driven evaluation matters more than chasing model complexity for its own sake. By pairing that comparison with SHAP-based explainability, this framework aims to produce not just a working model, but one whose reasoning can actually be inspected and trusted — a foundation for future experimental work using real breast cancer survival data.

REFERENCES

[1] P. Deepa and C. Gunavathi, "A systematic review on machine learning and deep learning techniques in cancer survival prediction," Progress in Biophysics and Molecular Biology, vol. 174, pp. 62–71, Oct. 2022.

[2] E. Y. Kalafi, N. A. M. Nor, N. A. Taib, M. D. Ganggayah, C. Town, and S. K. Dhillon, "Machine learning and deep learning approaches in breast cancer survival prediction using clinical data," Folia Biologica, vol. 65, no. 5–6, pp. 212–220, 2019.

[3] T. G. Baidoo and H. Rodrigo, "Data-driven survival modeling for breast cancer prognostics: A comparative study with machine learning and traditional survival modeling methods," PLOS ONE, vol. 20, no. 4, e0318167, Apr. 2025.

ISBN NO: 978-81-180146-5-2

[4] Y. Huang, S. Bazzazzadehgan, J. Li, A. Arabshomali, M. Li, and K. Bhattacharya, "Comparison of machine learning methods versus traditional Cox regression for survival prediction in cancer using real-world data: a systematic literature review and meta-analysis," BMC Medical Research Methodology, vol. 25, no. 243, 2025.

[5] D. W. Kim et al., "Deep learning-based survival prediction of oral cancer patients," Scientific Reports, vol. 9, no. 6994, 2019.

[6] I. Y. Zhang, G. R. Hart, B. Qin, and J. Deng, "Long-term survival and second malignant tumor prediction in pediatric, adolescent, and young adult cancer survivors using Random Survival Forests: a SEER analysis," Scientific Reports, vol. 13, no. 1911, 2023.

[7] M. Mohammed, I. B. Mboya, H. Mwambi, M. K. Elbashir, and B. Omolo, "Predictors of colorectal cancer survival using Cox regression and random survival forests models based on gene expression data," PLOS ONE, vol. 16, no. 12, e0261625, Dec. 2021.

[8] J. A. Bartholomai and H. B. Frieboes, "Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques," in Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, Dec. 2018, pp. 632–637.

[9] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," Scientific Reports, vol. 13, no. 8984, 2023.

[10] X. Pan, T. Feng, C. Liu, R. R. Savjani, R. K. Chin, and X. S. Qi, "A survival prediction model via interpretable machine learning for patients with oropharyngeal cancer following radiotherapy," Journal of Cancer Research and Clinical Oncology, vol. 149, pp. 6813–6825, 2023.

[11] H. Zhang et al., "Interpretable machine learning models for survival prediction in prostate cancer bone metastases," Scientific Reports, 2025.

[12] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan, M. Mutebi, G. Garvey, I. Soerjomataram, and M. M. Fidler-Benaoudia, "Global patterns and trends in breast cancer incidence and mortality across 185 countries," Nature Medicine, vol. 31, pp. 1154–1162, Apr. 2025.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_45

Empowering communities to overcome substance abuse across Chennai – Addressing today's social evil

Julia Grace P[1], Jane Lydia J[2] & Jannet Salomi D[3]

Associate Professor, Department of Computer Science, JBAS College for women, Chennai

BE IV Year Student, Velammal Engineering College, Chennai

BSc III Year Student, JBAS College for Women, Chennai

Abstract: Substance misuse, a serious social evil is common among college students, with nearly half meeting criteria for at least one substance use disorder (SUD) in some studies. Recent surveys, such as the 2024 monitoring the future survey, indicate historically high levels of cannabis, nicotine vaping, and hallucinogen use among young adults aged 19-30. This paper focuses on the causes and challenges that keep people captive and explores practical solutions for bringing deliverance by breaking the boundaries that hold them back.
Keywords: Drug abuse – addiction – substance use disorder - Hope to deliver

Introduction

Addiction is a disaster- weapon of major destruction. Alcohol remains the most frequently used substance, often exacerbated by pre-existing drinking habits and campus culture that normalizes heavy drinking. Many colleges and government institutions are taking action to help prevent or manage substance misuse and drug addiction in students.

UNODC emphasizes a **people-centered, human rights-based approach**, focusing on vulnerable populations, youth empowerment, gender equality, and strengthening institutions to ensure justice and security. Its strategy integrates innovation, cross-border cooperation, and evidence-based interventions to address the interconnected challenges of drugs, crime, and terrorism.

Substance abuse can cause many consequences for college students that are not limited to their academic life. Some of the short-term and long-term impacts of drug and alcohol misuse in college students can include:

- **Decreased academic performance**. Substance misuse can lead to a lower GPA, less time spent studying, missing class, getting behind on assignments, dropping out, or being expelled.
- **Risky or dangerous behaviors**. This can include doing things you normally wouldn't do, like driving under the influence, being involved in assault (either as a victim or perpetrator), getting into fights, stealing, or engaging in risky sexual behaviors or date rape. Many of these behaviors can be potentially lethal.
- **Poor health**. Suffer from many physical health consequences, including hangovers, nausea, injury, negative effects on your immune system, and risk of overdose or death. They may also experience poor mental health, decreased cognitive performance, short-term memory loss, addiction, or increased risk of suicide.

ISBN NO: 978-81-180146-5-2

- **Social consequences**. Loose friendships and important relationships due to substance use. May be more socially isolated if spend much of their time using alcohol or drugs.



Signs of Substance Use in College Students

While only a healthcare professional can diagnose an individual with a substance use or alcohol use disorder, it can be helpful to know the criteria, especially if some are worried about the substance use or that of a friend.

According to the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5),* signs of problematic drug or alcohol use include:

- Taking the substance in larger amounts or for longer than intended.
- Wanting to stop or cut down use of the substance but being unable to do so.
- Spending a lot of time getting, using, and recovering from substance use.
- Experiencing cravings and urges to use the substance.
- Neglecting responsibilities at college, work, or home because of substance use.
- Continuing to use the substance despite ongoing relationship problems that are caused or worsened by alcohol or drug use.
- Abandoning previously-enjoyed social or recreational activities hobbies in order to use alcohol or drugs.
- Taking risks while drunk or high, such as starting a fight, having unprotected sex, or driving a car.
- Continuing to use the substance despite the physical problems (e.g. blackouts or injuries) or psychological problems (e.g. depression or anxiety) it has caused or worsened.
- Needing to take more of the substance to get the desired effect, which is known as tolerance.
- Developing withdrawal symptoms (e.g. nausea, restlessness, or insomnia) if the substance is abruptly stopped or the dosage is cut drastically back.

Remedy - How Long is Rehab?

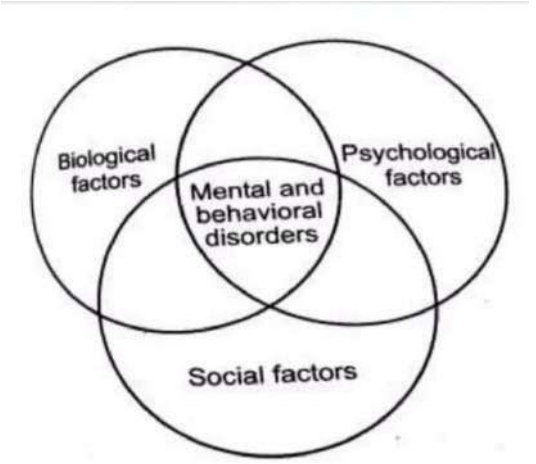
The duration of rehabilitation varies depending on the severity of an individual's substance or alcohol use disorder, the intensity of the treatment program, and the level of care required.

Even when a person needs to take time away from college to attend a rehabilitation center,

ISBN NO: 978-81-180146-5-2

various resources are available to support them throughout and after the recovery process. These may include counseling services, medical leave of absence, and transition plans that allow students to continue their studies through modified academic programs. Although concerns about falling behind academically or experiencing a decline in grades may discourage some students from seeking rehabilitation, it is important to consider the potential consequences of leaving addiction untreated. Substance use disorders can become increasingly severe without appropriate intervention. Therefore, seeking professional treatment should be regarded as an act of strength and a positive step toward regaining control of one's life before the situation becomes more difficult.

Addiction is a complex chronic condition caused by the interaction of biological, social, psychological factors. Sometimes, genetics plays a major role. Trauma and stress, mental disorders, peer pressure, bad friends, family environment may add the cause. There are two main forms of addiction – substance use disorders and non substance (behavioural) addictions.



Solving the addiction problem is much more complex. Communal transformation is necessary. Our knowledge must be converted into service. We are also responsible for this evil, as we are not ready to identify the problem, to come forward to help them, to show mercy and generosity. We must ask God's help to deliver them from evil; thereby need to prevent the young society. We must be available to render our helping hands, to catch the addicted people. If we must stop the addiction tomorrow, we need to start by today. Relearn the talking methodology. Death and life are in the power of our tongue. We need to speak positive towards the addicted people. No addiction has its final say.

Recent key drug-related incidents (Source: News)

1. Aug 2026 – Coimbatore 19 year old engineering student K.Amudhan death. He was targeted for exposing campus drug trafficking.
2. July 2025 – Chennai suburbs drug raid. Tambaram city police arrested 11 people, including 7 college students near a private university. Officers seized 166 gms of ganja and 5250 ganja laced chocolates
3. Nov 2024 – Chennai synthetic drug seizure. Greater Chennai police arrested 5 private engineering college students, found selling synthetic narcotics via social platforms

Findings and research gap

Substance abuse is a social evil, that affect an individual's physical and mental health. Studies reveal few male students in Chennai have identified with the usage of substances such as alcohol, tobacco and other recreational drugs.

Since the data is sensitive and confidential, we cannot disclose. Hereby, come up with a comparative study in general. Though we have several algorithms like classification algorithm, time/series anamoly detection algorithm and many more, supervised machine learning algorithm

ISBN NO: 978-81-180146-5-2

are more flexible than the traditional statistical methods.

Our **objective** is to use decision tree algorithm to sort the given patient database based on the parameters: a) from high risk to low risk living individual b) criminal offenders c) overdose deaths… We are working on a specific dataset and the exact findings with higher accuracy will be updated as an extended version of this research findings.

Acknowledgements

We acknowledge Bombay teen challenge, Thane for creating a deep burden in our heart about these marginalized people. We cannot force anyone to stop abusing drugs, but we can pray for them, help them to come out and to lead a good life.

Conclusion

We need to increase awareness about the ill-effects of drug abuse among college students. Even though, TN records the lowest overall prevalence of substance abuse in India, nearly 10% of the drugs that flows into the state is supplied to college students only. God's heart bleeds, if we remain silent without helping these marginalized people. Humbling ourselves, we must come one step down to pick them up, to make them live again. With education, compassion and timely help, we need to help those affected students to recover and to lead a productive life.

REFERENCE:

A. Shirish S. Patil* and Sourabh* S. Malabade, A Review on Drug Addiction and Habituation and Treatment for Drug Addiction, International Journal of Pharmaceutical Science and Research, (2023), 14, 1, (157,158,160).

B. Jyotika Singh1, Pradeep Kumar Gupta2, Drug Addiction: Current Trends and Management, the International Journal of Indian Psychology, (2017) 5, 1, (192).

C. Anjali Sanjeev, Sarita M. Kapgate, Current addictive drugs: A review, International Journal of Green Pharmacy, (2022) 1, 1, (323).

D. Dr. Khushboo Karte; Dr.S.R. Inculkar & Dr. Yuvraj Kauyhik (2020)Drug Abuse & Drug addiction ,World Journal of pharmaceutical & Medical research,(2020),6,12 , (85,86).

E. S. Malabade, a Review on Drug Addiction and Habituation and Treatment for Drug Addiction, International Journal of Pharmaceutical Science and Research, (2023), 14, 1, (167,170).

F. Kamlesh Kumar Sahu*& Soma Sahu*, Substance Abuse Causes and Consequences, Bangabasi Academic Journal, (2016), 9, 1, (54, 55).

G www.btcashagram.org

H www.thehoperevolution.org

6th International Conference on Recent Advances in Computing & Communication – ICRACC 26
ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_46

Intelligent Healthcare Agents: Using Agentic AI for Personalized Patient Recommendations

ALIYA TABASUM I
Dept. of CS & IT
M. O.P. Vaishnav College for Women
Chennai, India
2413721102002@mopvaishnav.ac.in

ARUNA A S
Dept. of CS & IT
M.O.P Vaishnav College For Women
Chennai , India
2413721102004@mopvaishnav.ac.in

Dr.BRINDA RAMANUJAM
Dept. of CS & IT.
M.O.P Vaishnav College For Women
Chennai , India
brindaramanujam.csc@mopvaishnav.ac.in

Abstract

Large language models (LLMs) have fast-tracked the development of agentic AI agents in healthcare, which marks a paradigm shift from the static nature of models to agents that can make medical sense. In this study, we look at the AI agents in six areas: clinical decision support, healthcare systems management, clinical consultation, medical education, medical imaging, and the evaluation framework. In this paper, we examine more than 30 AI systems such as MDAgents, Agent Hospital, ColaCare, and AgentClinic to determine commonalities and performance attributes. On the back of our findings in this paper and those of a scoping review on clinically used agentic systems in fields like emergency medicine, oncology, radiology, and rehabilitation, we have come up with a pipeline for personalized patient recommendations that entails patient profiling, multi-agent clinical reasoning, recommendation generation using retrieval techniques, and human-in-the-loop validation. We will also address the performance outcomes and current clinical uses of such systems, as well as data privacy, biases, interpretability, and regulatory issues that need to be addressed before their implementation in clinical settings.

Keywords—Large Language Models; Clinical Decision Support; Multi-Agent Systems; Personalized Medicine; Patient Recommendation Systems

I. INTRODUCTION

Healthcare sector is experiencing a revolution of its architecture due to the development of AI systems that go beyond mere classification and predictive models and operate with minimal human input to plan and act under changing circumstances. This refers to the emergence of agentic AI systems – AI systems that differ from the preceding generations by their autonomous behavior, goal-orientedness, and ability to undertake actions, such as making use of tools and working together with other agents. While earlier AI generations performed rigid tasks, the new-generation of AI known as Generative AI, including LLMs, allows for content generation yet it is still reactive as it reacts to all user commands. AI agents are those trying to achieve certain goals in the constrained domain, while agentic AI is the pinnacle of this endeavor. Personalized patient recommendation is among the most promising uses of the development. Rather than providing a single static prediction, the agentic system can continuously collect patient data, reason about possible treatments, consult specialized sub-agents or external knowledge bases, and provide.adapted to reflect any new information becoming available. The state-of-the-art of agentic artificial intelligence in medicine is provided herein, relying on the findings of the recent scoping review of seven peer-reviewed agents in this domain, spanning emergency medicine, oncology, radiology and rehabilitation, and on a survey of over thirty language model (LLM) healthcare agents in six fields, namely, clinical decision support, healthcare systems management, clinical consultation, medical education, medical imaging, and evaluation frameworks . Representative examples of those surveyed systems include MDAgents, Agent Hospital, ColaCare, AgentClinic, as well as domain-specific prototypes like TraumaTracker, SmartText, GPT-Plan, ProtChat, MultiMedRes, and ChatExosome.

The rest of this paper is organized as follows. In Section II, related works on agentic AI systems in medicine are discussed. A taxonomy of architecture patterns of the surveyed systems is presented in Section III. In Section IV, a sample reference pipeline for personalized treatment recommendations via agentic AI is shown. In Section V, clinical applications of agentic AI are discussed. Section

VI identifies technical, ethical and regulatory barriers to the progress of agentic AI. Section VII concludes the paper.

ISBN NO: 978-81-180146-5-2

II. RELATED WORK

A. Agentic Systems Identified in Clinical Scoping Reviews

According to the scoping review of 2026, there were 984 records found in five different databases; however, after performing a detailed analysis of the full text of the articles, only seven articles have met the basic requirements of autonomy, goal-directedness, and action initiation of an agent. The TraumaTracker research carried out in 2019 used two agents of BDI which worked together; each of them had its own special responsibility of documentation and alert generation, and it has been used for nine months generating more than 430 trauma reports and increasing the efficiency and speed of alerts. The SmartText application (2021), which was tested in a randomized controlled trial of 42 cancer patients, functioned as a single agent of BDI, providing individualized recommendations of daily activities, but it failed to provide any statistically significant improvement in the number of steps compared to voice-assistant comparator. A reinforcement learning virtual assistant (2023) changed the difficulty of a VR rehabilitation game of Down syndrome children according to their skills using the Proximal Policy Optimization model of Actor-Critic and reached 87% of accuracy

Subsequent proof-of-concept systems have employed multi-agent architectures and retrieval-augmented reasoning. GPT-Plan (2025) collaborated Dosimetrist, Physicist, TPS-proxy and Human-proxy agents to optimize radiotherapy treatment plans; GPT-Plan demonstrated expert-level skills in treating patients with cervical cancer, and exceeded experts' results in treating patients with lung cancer. ProtChat (2025) incorporated GPT-4 and protein language models into User Proxy, Inference, Evaluation, Visualization and Chat Manager agents for performing automated protein analysis. MultiMedRes (2025) implemented a sole learner agent to split chest X-ray questions and communicate with external expert models, thus demonstrating state-of-the-art zero-shot diagnostic capabilities. ChatExosome (2025) applied deep learning and retrieval-augmented reasoning to diagnose HCC using exosome Raman spectroscopy with 94.1% accuracy; in alpha-fetoprotein negative cases, diagnosis accuracy reached 87.5%. In seven reviewed studies, just one used real patients and none of the systems demonstrated adaptability over time, which reflects the immaturity of clinical application of agentic AI.

B. Agentic Systems Identified in Broader AI-Agent Surveys

A A related study that looks at more than thirty LLM-based agents divides the domain into six categories: Clinical Decision Support, Healthcare Systems Management, Clinical Consultation, Medical Education, Medical Imaging, and Evaluation Framework, with issues like data privacy, algorithmic bias, regulatory issues, multi-agent safety, explainability, and integration as major barriers to implementation. The paper MDAgents arguesan adaptive collaboration system in which either a single agent or a multi-agent LLM system including a set of various experts is selected depending on the estimated level of complexity of the medical problem that would lead to increased accuracy and efficiency as opposed to the static multi-agent system. Agent Hospital creates a full simulation of a hospital environment in which the roles of the patient, nurse, and doctor are played by LLM-based agents in which the doctors learn from interaction with simulated patients using the MedAgent-Zero methodology without use of any labeled data achieving state-of-the-art performance on the MedQA dataset after analysis of ten thousand cases.ColaCare uses a Multidisciplinary-Team inspired architecture that allows structured modeling of the EHR through deliberations of multiple DoctorAgents on the results obtained through the analysis of findings obtained by the domain-specific expert models and MetaAgent reaching a consensus and retrieving evidence from the clinical guidelines knowledge base, thus greatly increasing the accuracy of prediction of mortality and readmissions on the MIMIC-III and MIMIC-IV datasets . AgentClinic is an open source benchmark platform that allows for simulation of a whole clinical encounter including patient, doctor, and measurement agents that allows testing of models not in multiple-choice that simulates multidisciplinary consultation of experts for zero-shot reasoning in medicine, EHRAgent that enables doctors to ask EHR systems using autonomous code generation, and Almanac Copilot that automates repetitive EHR tasks .

III. TAXONOMY OF AGENTIC HEALTHCARE SYSTEMS

Through integration of the two streams of research, the agentic healthcare systems could be categorized into two categories: the domain and the architectural structure of the system. There are six domains that are common: (i) the decision-making support system, which include MDAgents and ColaCare, which use reasoning on patient's information to make recommendations on diagnoses and treatment; (ii) management of healthcare systems, which include Agent Hospital and TraumaTracker, in which the agents coordinate several actors or roles in the care processes; (iii) the consultation system, where the agent consults directly with the patient, like in SmartText; (iv) medical education and training simulations; (v) medical imaging and signal analysis, including MultiMedRes and ChatExosome; and (vi) evaluation system like AgentClinic in which the performance of the agent is measured.

ISBN NO: 978-81-180146-5-2

Classification of the agents in the agentic healthcare systems can be improved based on considerations of orientation and the level of integration. Apart from six persistent types of agents, current research has identified three emerging classifications: (i) empathic support agents created to deliver empathic conversational assistance in mental health and elderly care, like MindBot and frameworks for elderly care; (ii) patient information interrogation agents enabling patients to access and interpret their health records or data provided by wearables through PhysioLLM and Graph-Augmented LLMs, and (iii) integrated frameworks featuring agents built into large enterprises like hospital systems to work across several departments. Architecture wise, hybrid models combining retrieval augmented generation with multimodal reasoning and mixture-of-experts approaches seem to be promising for finding the right balance between the levels of accuracy, customization, and scalability. It looks like the proposed taxonomy implies that the classification of agents must also take into account human-centric parameters, such as empathy, trust, transparency, and adaptability.

Single-agent designs, represented by SmartText, MultiMedRes, ChatExosome provide clear architectural advantage of simplicity and accountability as one agent coordinates all external resources in such architecture, which is beneficial for narrow diagnostic or coaching tasks . Multi-agent frameworks, like TraumaTracker, GPT-Plan, ProtChat, MDAgents, Agent Hospital and ColaCare distribute specialised functionalities across cooperative agents, which allows to build sophisticated coordination for complex workflows, but at the same time introduces higher risks of inter-agent miscommunication . The vast majority of recent works incorporates either LLM or multimodal LLM as a backbone for reasoning and, very often, retrieval-augmented generation for grounding recommendation in external medical knowledge sources, like clinical guidelines or drug formularies .

System	Domain / Year	Core Agentic Capability
TraumaTracker	Emergency medicine, 2019	Two BDI agents for real-time trauma documentation and alert generation.
SmartText	Lifestyle coaching, 2021	Single-agent autonomous coach delivering personalized activity messages.
RL Virtual Assistant	Pediatric rehabilitation, 2023	Reinforcement-learning agent adapting VR game difficulty for children with Down syndrome.
GPT-Plan	Radiation oncology, 2025	Multi-agent (Dosimetrist, Physicist, TPS-proxy, Human-proxy) radiotherapy plan refinement.
ProtChat	Biomedicine, 2025	GPT-4 multi-agent system for protein property and interaction analysis.
MultiMedRes	Radiology, 2025	Single learner agent decomposing chest X-ray questions and querying expert models.
ChatExosome	Oncology, 2025	Retrieval-augmented agent diagnosing HCC from exosome Raman spectroscopy.
MDAgents	Clinical decision support, 2024	Adaptively assigns solo or multidisciplinary-team roles to LLM agents based on case complexity.
Agent Hospital	Healthcare systems management, 2024	Simulated hospital in which doctor agents evolve through MedAgent-Zero after treating simulated patients.
ColaCare	EHR modeling, 2024	DoctorAgent / MetaAgent collaboration for mortality and readmission prediction from EHR data.
AgentClinic	Evaluation framework, 2024	Multimodal benchmark simulating patient, doctor and measurement agents for clinical reasoning.

TABLE I continued. Comparative summary of representative agentic AI systems in healthcare (see Section

ISBN NO: 978-81-180146-5-2

*III).*PROPOSED FRAMEWORK FOR PERSONALIZED PATIENT RECOMMENDATIONS

Based on the patterns of architecture identified in Section III, a reference pipeline is defined that facilitates provision of customized patient recommendations via agentic AI, modified from the standard agentic pipeline of perception, planning, action, and delivery within human supervision.

A. Data Ingestion and Patient Profiling

The intake agent takes into account the structured elements of the EHR, the unstructured physician's notes, metadata from images, and if appropriate, the patient self-reporting data (such as wearable sensor activity data), to create the longitudinal patient record. This process is analogous to the input acquisition phase of NLP pipelines that gather the raw input data. Multi-Agent Clinical Reasoning

A oordinating MetaAgent, after the lines of ColaCare, distributes the patient profile to the corresponding DoctorAgents or experts, with each one responsible for analysing some particular aspect like risk stratification, drug interaction, or

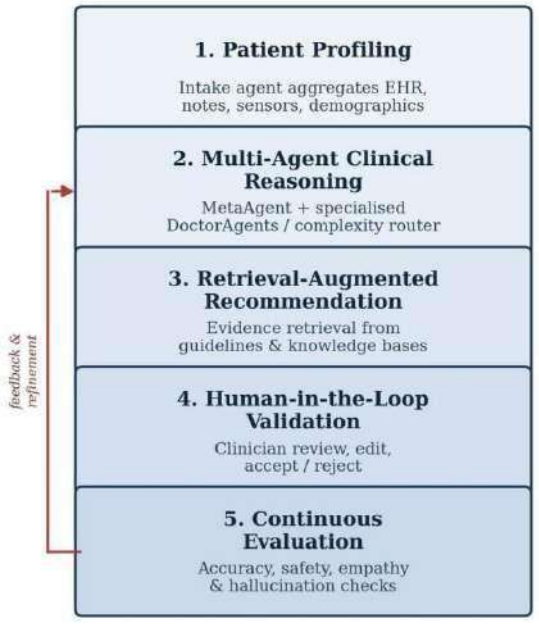
lifestyle issues. For complicated or uncertain cases, the MDAgents complexity estimate method is used to determine whether it can be handled by a single agent or a team of agents .

B. Retrieval-Augmented Recommendation Generation

The agents use retrieval augmented generation and extract evidence from curated medical knowledge sources or guidelines to generate candidate recommendations, according to the precedent set out by the agents ChatExosome and ColaCare. If the recommendation is accompanied by an action, which could be something like scheduling a follow-up or marking an abnormal test result, then the agent uses that tool or application programming interface (API) accordingly.

C. Human-in-the-Loop Validation and Delivery

Consistent given the importance of the human-in-the-loop process highlighted in the literature, all recommendations made are presented to the physician for review prior to presenting them to the patient, and any comments made by the physician are recorded for future audit purposes and continuous improvement where the system allows for it. This process is crucial because none of the reviewed systems has exhibited self-adaptation over time.



Flow chart systems in healthcare

IV. CLINICAL APPLICATIONS

Agentic recommendation pipeline approaches could be applied in various clinical practices in the future. For example, in oncology, agentic recommendation pipelines, such as the ChatExosome and ColaCare-based agent consultation, could be used for early cancer detection and treatment scheduling with the use of diagnostic information and risk factors extracted from EHRs. In radiation oncology, the multiple GPT-Plan-based agent planning approach could be applied to make recommendations on personalized treatment scheduling and determine cases outside of the algorithm's confidence level. In chronic disease and lifestyle management,

ISBN NO: 978-81-180146-5-2

coaching agents based on SmartText could tailor interaction strategy continuously utilizing behavioral data considering the limitations of effectiveness compared to human-assisted coaching. In emergency and trauma medicine, there is a possibility to implement agents similar to trauma tracking system for individualized triaging on the basis of vital signs and medical history of a patient.

In rehabilitation medicine, there are reinforcement learning agents, such as down-syndrome virtual reality assistant, that personalize therapy dose. In conclusion, there are several systems focused on EHRs, such as ColaCare, EHRAgent, and Almanac Copilot, that could demonstrate the potential of personalized alerting systems to decrease documentation burden and raise personalized alerts for risk of hospital readmissions/deterioration.

V. CHALLENGES AND LIMITATIONS

The promising prospects, the surveyed literature offers a set of recurring barriers. First, there is an absence of clinical validation: in seven studies reviewed in the scoping review, only one of the RCT employed actual patients, while the risk of bias was moderate to critical in all other observational and computational studies. None of the systems under review had adaptive behavior, which could emerge through the process of usage, and their scope stayed limited and specific, without any kind of long-term memory.

The broader literature review gives a list of barriers for deployment of the systems within the framework of personalized recommendations, which includes issues such as data privacy, algorithmic bias, lack of regulation, safety issues for multi-agent architectures, difficulties with explainability, and integration into the current hospital infrastructure. The 'black-box' characteristic of reasoning by LLM makes impossible to give the explanation to the physician of why the recommendation of an agent is right in the process of making joint decisions, while the cybersecurity issues make vulnerable to the unauthorized access to EHRs. Right now, neither the FDA nor the EU regulation covers the class of systems under discussion. Therefore, the FDA Predetermined Change Control Plan requires certain protocols for Thus, the Predetermined Change Control Plan of the FDA requires clear protocols of update of adaptive AI devices, whereas the AI Act of the EU considers AI in healthcare as high-risk, but not the multi-agent and goal-oriented systems. The production line for personalized recommendations should comply with HIPAA/GDPR and employ techniques such as differential privacy and federated learning.

The key limitations of the systems are :

- Limited applicability of the clinical application, where the vast majority of the agents have been tested for one specific disease/task only.
- Dependence on the quality and representativeness of training/retrieval datasets.
- The lack of any long-term memory and post-deployment adaptation.
- The inability to explain the process of reasoning for the multi-agent system.
- Lack of the clinical validation of the models outside the simulation environment.
- Accountability of autonomous agents in the clinical decision process remains an open question.

VI-A. Reliability and Evaluation Framework

Personalized healthcare agents require more sophisticated performance measures than language proficiency or benchmarking. Clinically relevant performance measures need to verify if the system is identifying correct patient-specific facts, gathering appropriate evidence, providing a recommendation which is compliant with the current guidelines, communicating uncertainty, and being safe if there is an absence or conflicting data. These aspects need to be assessed individually because the fluent answer may also include a critical omission or unsound advice.

We advocate for a hierarchical assessment of performance. Layer one involves technical validation, which includes factuality, retrieval precision, citation accuracy, successful tool use, latency, and prompt injection robustness. Layer two is clinical validation, which involves diagnostic or treatment plan consistency with experts, contraindication sensitivity, confidence calibration, and clinically relevant error frequency. Layer three is patient-centered validation, which involves comprehension, usability, trust, accessibility, adherence, and goal alignment. Layer four is operational validation, which involves workflow fit, escalation load, auditability, privacy incident frequency, and cost per encounter.

It would be important to test each recommendation on an actual patient case as opposed to a reference answer. The test data should have normal patient cases, corner patient cases, missing patient cases, drug interaction cases, susceptible patient population, and adversarial patient cases. The expert reviewer will then give marks on the recommendation as well as the reasoning trace that is shown to the physician.

Any disagreements in opinion by the reviewer will be captured and considered as a valuable insight on any

ISBN NO: 978-81-180146-5-2

ambiguity.Longitudinal testing becomes especially critical for agentic systems. While a recommendation might be right at first contact, it may become dangerous due to changes in laboratory data, medications, allergies, and preferences of the patient. The system needs to be tested under longitudinal simulation of care and, wherever possible, by prospective testing in shadow mode. Shadow mode means that the agent gives recommendations but does not affect the care process. Clinicians can evaluate results, register overrides, and discover failure modes of the system.

VI-B. Privacy, Security, and Data Governance Personalized recommendations need access to sensitive data, such as diagnoses, drugs, laboratory data, imaging reports, behavior data, and patient preferences. This means that the issue of privacy is not simply about training models, but also ingestion of data, data retrieval, transient storage, agent-to-agent communications, tool callouts, logging, test sets for evaluation, and finally the resulting recommendation for the patient/clinician.

This implies that a privacy-preserving architecture will need data minimization at all stages of processing. The profiling agent should ask for only the relevant fields to address the clinical question at hand, while a policy engine should impose role-based access control, purpose limitations, retention times, and consent requirements. Clinical content should be segregated from identifiers where possible. The retrieval stores should be access controlled and versioned in a way that every piece of evidence used in a recommendation is trackable without revealing any unnecessary patient information.

Multi-agent communications represent an extra security surface.A malicious or erroneous note could be viewed as an instruction from another agent; a document obtained from a store could have prompt injection text; and results obtained through an external tool could be unintended for the present agent. These agents thus need to consider the retrieved content as insecure data, validate parameters of tools used, limit access permissions, and confirm before executing any significant action. Security assessment should cover prompt injection through indirect means, data exfiltration, privilege escalation, and manipulation of clinical context.

Even though federated learning, local inference, de-identification, encryption, and differential privacy may lower vulnerability to the above issues, none of them eliminates the need for governance. Privacy enhancement can likewise lower the data completeness or model efficacy. This decision should accordingly be documented as a trade-off: what data to retain, who will be allowed to review them, how long these data will be accessible, and what is lost clinically if some data are hidden.

VI-C. Explainability, Bias, and Human Oversight

The concept of oversight has meaning only if the physician is able to figure out which information the agent has considered and what it hasn't and which is still unknown. A general statement regarding the fact that the LLM is like a black box is not enough. The system needs to be transparent in terms of providing the facts about the patient that were used for reasoning, the passages from the guidelines,Clinical utility without loss of detail necessary for post hoc review. There are various ways bias can seep into the system, such as the patient characterization, the retrieval database, the clinical decision criteria or the evaluation sample. It may be the case that the system works well in general but poorly on patients who speak a particular language or belong to an ethnic, demographic or other subgroup that is under-represented. The evaluation must then specify performance within subgroups, missingness profiles, non-compliance patterns and escalations. The purpose is not simply equal accuracy against benchmarks; it is to uncover whether the errors are concentrated in a group for which the system is meant to deliver care.

The human-in-the loop process must not be a rubber-stamping exercise. Oversight should entail threshold definitions for escalation, power of rejection or editing of the recommendation, access to the supporting evidence, and feedback mechanisms which do not inadvertently retrain the system. In the case of high stakes decisions, the agent should recommend the course of action and provide the supporting evidence, without prescribing treatment, diagnosing or changing the therapy. The line between recommendation and decision making should be clearly defined both within the system design and the interface.

VI-D. Implementation Roadmap and Research Priorities

The process of implementation should involve a number of stages. The first stage will involve retrospective validation, utilizing de-identified cases and test suites created by experts. The second stage will involve silent prospective evaluation when recommendations are recorded but not revealed to clinicians. The third stage will be the supervised use of the recommendations in lower risk activities such as information search, patient education at discharge, medication reminders, and documentation. The fourth stage involves considering the recommendation of higher risk activities only when stability of the system, low amount of harm associated with override, reliability of escalation, and continuous monitoring are proven.

Every stage must have a stop criterion. Such criteria can be clinically significant rates of hallucinations, inability to detect contraindications, discrimination among subgroups, unexpected changes due to a model change, or a privacy breach. The model changes should be versioned, validated using a regression suite and then reviewed before implementation. The continuous monitoring will take into account the performance of the model, the

ISBN NO: 978-81-180146-5-2

behavioral changes of users, the override, failures of the retrieval process, and the lack of trust from clinicians. than by architectural complexity alone.

A first priority for future research involves gathering prospective evidence. The literature provides ample prototypes, but only a few have been tested using actual patients, with fewer still demonstrating safe and adaptive performance over time. Research needs to move forward combining technical standards with empirical studies on the clinical level involving patient outcome, workflow, and safety assessments.

Comparisons between agentic and alternative architectures need to be performed, and adding new agents should be based on clear benefits in safety, accuracy, or efficiency rather

VI-E. Illustrative Patient Recommendation Workflow

Consider an example where there is a patient with diabetes and hypertension with poor renal function and non-compliance to the regimen of the medicines that are supposed to be taken. A safe recommendation engine will not start by providing the regimen of the treatment but it will start by checking the patient's identity, getting the most recent lab values, reconciling the medications, taking note of the patient's allergy and contraindications, and checking if there is any issue related to the practicality of the recommendations such as the problem of cost or the lack of means for transport.

The profiling engine will create a structured summary with provenance of each piece of information. The clinical reasoning engine will check for gaps and inconsistencies in the summary while the specialist engines will check for medication safety, renal problems, lifestyle issues, and preference of the patient. The retrieval engine will get the approved guideline and protocols. The coordinating engine will compare the result and identify any discrepancy and provide a few choices rather than one choice and the explanation of that one choice. Prior to delivering the recommendation, the validation engine will make sure that the recommendation has been evidence-based and use language

The very same workflow may be used in the context of oncology, rehabilitation, medication management, and chronic disease management; however, risk controls will need to vary according to the case at hand. While a reminder agent can be designed as minimally supervised, a treatment selection agent would need robust verification controls and explicit clinical approval. Thus, the design should be sensitive to risks and not be autonomous across the board.Following that, specialized agents would analyze various aspects of the available information regarding the patient. One could utilize a diagnostic agent for analyzing possible symptoms and medical conditions and a medication agent for analyzing the medications already being taken and possible interactions. In addition, one may employ a lifestyle agent for considering patients' diet, exercise, sleep and other personal aspects. All this information, obtained through the work of the agents, would be delivered to a coordinating agent that integrates this information and retrieves additional information from reliable medical knowledge sources. In this way, the final decision would be based on the consideration of patient's individual features and clinical evidence. In this case, there is no need to substitute clinical decision-making with an automated process. In this regard, one does not need to replace the clinical judgment with automated decisions. Instead, one may use the system for creating a recommendation along with the necessary information about evidence, level of certainty, and other features for review by health professionals.

VI. CONCLUSION AND FUTURE SCOPE

This paper mapped the emergent field of agentic AI in healthcare by performing a scoping review of seven clinically-focused agentic systems, followed by a survey of more than thirty agent-based LLMs designed for various healthcare domains. Examples of goal-oriented clinical applications include TraumaTracker, SmartText, GPT-Plan, ProtChat, MultiMedRes and ChatExosome, whereas MDAgents, Agent Hospital, ColaCare and AgentClinic represent the potential of multi-agent collaboration and simulation for enhanced reasoning and evaluation. Based on the identified trends, we have elaborated on the pipeline for personalized patient recommendation including patient profiling, multi-agent clinical reasoning, retrieval-augmented recommendation generation and human-in-the-loop validation of recommendations.Evidence base is limited by lack of clinical trials with real patients, lack of experience in safe adaptive learning over In addition, there are concerns related to time, interpretability, bias, and privacy that need to be addressed before implementation of such systems can happen. In the future, the research in the area should aim at conducting pilot trials in fields where digitization was sufficiently achieved (radiology, radiation oncology, emergency medicine), establishing metrics for measuring autonomy, safety, and adaptivity, as well as better integration of the research of the agentic AI with the framework of clinical governance like HIPPA, GDPR, and FDA/EU AI Act.

REFERENCES

[1] B. G. Collaco, S. A. Haider, S. Prabha, C. A. Gomez-Cabello, A. Genovese, N. G. Wood, S. P. Bagaria, N. Gopala, C. Tao, and A. J. Forte, "The role of agentic artificial intelligence in healthcare: a scoping review," npj Digit. Med., vol. 9, no. 345, 2026.

ISBN NO: 978-81-180146-5-2

[2] "AI agents in healthcare: a survey of applications, resources, and challenges," IEEE Xplore, Document 11398875, 2025.
[3] MDAgents: an adaptive collaboration of LLMs for medical decision-making, Adv. Neural Inf. Process. Syst. (NeurIPS), 2024.
[4] J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu, "Agent Hospital: a simulacrum of hospital with evolvable medical agents," arXiv:2405.02957, 2024.
[5]Y. Wang, E. Harrison, C. Pan et al., "ColaCare: enhancing electronic health record modeling through large language model-driven multi-agent collaboration," in Proc. ACM Web Conf. 2025, pp. 2250–2261.
[6] S. Schmidgall et al., "AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments," arXiv:2405.07960, 2024.
[7] MedAgents: large language models as collaborators for zero-shot medical reasoning, 2023.
[8] "AI agent in healthcare: applications, evaluations, and future directions," npj Artif. Intell., 2026.
[9] "A comprehensive survey of AI agents in healthcare"

6th International Conference on Recent Advances in Computing & Communication – ICRACC'26
ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

A COMPUTATIONAL STUDY OF $PM_{2.5}$ ENVIRONMENTAL EFFECTS: 3D SURFACE ANALYSIS OF THE INTRA-COMPETITION COEFFICIENTS (B_1 AND B_2) INTERACTIONS WITH METEOROLOGICAL VARIABLES

R. E. Akpodee
[1]Department of Mathematics, Faculty of Science,
Rivers State University, Port Harcourt, Nigeria

Email: erenu.akpodee1@rsu.edu.ng

Abstract

Air pollution in rapidly urbanizing areas such as Port Harcourt Metropolis in the Niger Delta region of Nigeria has become a critical environmental and public health issue, driven by emissions of fine particulate matter ($PM_{2.5}$) from industrial operations, transportation, and artisanal oil refining, while meteorological factors such as air temperature and relative humidity further influence pollutant concentration, dispersion, and transformation. This study models the temporal dynamics of $PM_{2.5}$ and air temperature using logistic differential equations, incorporating intrinsic growth rates and intra- and inter-competition coefficients to capture self-regulatory feedback and environmental interactions. Seven strategically selected sampling locations across Port Harcourt Metropolis, representing diverse land uses and anthropogenic activities, were geo-referenced using a Garmin U-10 GPS to ensure spatial accuracy. Monthly field measurements of inhalable and respirable particulate matter ($PM_{2.5}$), heavy metals (V, Ni, Pb, Cr, Zn, As, Hg), and 17 USEPA-recommended polycyclic aromatic hydrocarbons (PAHs) were conducted over six months (October 2024–March 2025) using Aerocet 531 particle mass monitors and high-volume H2000 samplers. Using P-vector norm analysis, the intra-competition coefficient for $PM_{2.5}$ (β_1) was determined to be 0.026685 and that for air temperature (β_2) 0.000626, both corresponding to local minima in their respective 1-norm, 2-norm, and infinity-norm sequences. The 3D surface analysis revealed non-linear, saddle-like behavior in β_1 and β_2 across saturated values, highlighting critical points of self-limiting pollutant growth and system stability, respectively, while ODE45 simulations confirmed deterministic bifurcation and long-term steady-state stability. These results provide precise parameters for environmental modeling, enabling more accurate predictions of pollutant accumulation, assessment of exposure risks, and formulation of mitigation strategies, and they underscore the importance of considering interactions between $PM_{2.5}$ and meteorological variables for ecosystem protection. Integrating intra- and inter-competition coefficients in modelling demonstrates a robust and realistic approach to urban air pollution dynamics, offering practical applications for sustainable urban planning, regulatory policy, and public health management, with recommendations for future studies to include additional climatic variables, long-term pollutant interactions, and multi-pollutant dynamics to enhance predictive accuracy in rapidly industrializing urban centres.

Keywords: $PM_{2.5}$, Air Temperature, 3D Surface, Intra-Competition Coefficient, Logistic Differential Equation, Port Harcourt Metropolis, Air Quality Modelling, Environmental Pollution

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

Introduction

In this study, the temporal dynamics of respiratory pollutant concentration ($PM_{2.5}$) and specific meteorological variables, namely peak air temperature and relative humidity, are framed as functions of essential parameters, including intrinsic growth rate, intra-competition coefficients, and inter-competition coefficients. If these variables don't interact with each other, we can use exponential growth models that come from separating the variables to describe how they change over time. However, these formulations, although mathematically convenient, are environmentally unrealistic because they do not take into account the feedback mechanisms that are built into atmospheric systems. Empirical research has shown that $PM_{2.5}$ concentrations are significantly affected by atmospheric processes including dispersion, transport, and chemical transformation, which inhibit perpetual growth (Hao et al., 2024; Liu et al., 2024). Consequently, the presumption of unbounded exponential growth offers a paradoxical depiction of actual environmental dynamics.

To mitigate this limitation, this study implements regulatory mechanisms via intra-competition and inter-competition coefficients. The intra-competition coefficient measures self-limiting interactions in $PM_{2.5}$ concentrations, showing processes like particle aggregation, deposition, and saturation in the atmosphere. The inter-competition coefficient shows how weather factors, especially temperature and relative humidity, affect the levels of pollutants. Recent research substantiates that meteorological variables, such as temperature, humidity, wind patterns, and atmospheric stability, significantly influence $PM_{2.5}$ variability, dispersion, and accumulation (Alaran et al., 2024; Redi, 2026). For example, photochemical reactions can happen more quickly when the temperature is higher, and pollutants tend to build up when the wind is slow and the humidity is high. Also, the evolution of $PM_{2.5}$ is modeled as being dependent not only on time but also on its initial concentration. This is in line with what we see in environmental systems, where baseline pollution levels affect future dynamics and carrying capacity thresholds.

Pollution of the air, land, and water is caused by many human activities, such as industrialization, transportation, agriculture, and energy production. In places like the Niger Delta, where cities and industries are growing quickly, oil exploration, gas flaring, and illegal refining are making the environment worse. These processes put fine particles, greenhouse gases, and toxic chemicals into the air, which changes the climate, harms ecosystems, and hurts people's health. Recent research indicates that $PM_{2.5}$ is a complex amalgamation of inorganic and organic compounds with the capacity for long-range transport and considerable environmental effects (Hao et al., 2024).

Soot is a common sign of air pollution in Port Harcourt, and it is mostly caused by illegal crude oil refining. These processes don't burn hydrocarbons all the way through, which releases fine particles and other harmful pollutants. Informal refining doesn't have emission control technologies like regulated industrial systems do, which makes pollution levels higher. $PM_{2.5}$, a significant constituent of soot, is linked to both acute and chronic health impacts, encompassing respiratory and cardiovascular diseases. Recent evidence underscores the synergistic effects of $PM_{2.5}$ and elevated temperatures, which can exacerbate environmental and health risks (Coker et al., 2025).

Air pollution in rapidly urbanizing areas such as Port Harcourt Metropolis in the Niger Delta region of Nigeria has become a critical environmental and public health issue, driven by emissions of fine particulate matter ($PM_{2.5}$) from industrial operations, transportation, and artisanal oil refining, while meteorological factors such as air temperature and relative humidity further influence pollutant concentration, dispersion,

ISBN NO: 978-81-180146-5-2

and transformation. This study models the temporal dynamics of $PM_{2.5}$ and air temperature using logistic differential equations, incorporating intrinsic growth rates and intra- and inter-competition coefficients to capture self-regulatory feedback and environmental interactions. Seven strategically selected sampling locations across Port Harcourt Metropolis, representing diverse land uses and anthropogenic activities, were geo-referenced using a Garmin U-10 GPS to ensure spatial accuracy. Monthly field measurements of inhalable and respirable particulate matter (PM_1, $PM_{2.5}$), heavy metals (V, Ni, Pb, Cr, Zn, As, Hg), and 17 USEPA-recommended polycyclic aromatic hydrocarbons (PAHs) were conducted over six months (October 2024–March 2025) using Aerocet 531 particle mass monitors and high-volume H2000 samplers. Using P-vector norm analysis, the intra-competition coefficient for $PM_{2.5}$ (β_1) was determined to be 0.026685 and that for air temperature (β_2) 0.000626, both corresponding to local minima in their respective 1-norm, 2-norm, and infinity-norm sequences. In addition, 3D surface analysis of β_1 and β_2 was performed to visualize the interaction of norms across saturated values, revealing non-linear, saddle-like behaviors and critical regions where system dynamics shift from self-limiting to potential instability, consistent with the application of surface modeling in environmental parameter optimization (Zhang et al., 2023; Li & Wang, 2024). Computational simulations with ODE45 confirmed deterministic bifurcation and long-term steady-state stability, supporting the robustness of the estimated parameters. These results provide precise input for environmental modeling, enable accurate prediction of pollutant accumulation, and inform mitigation strategies, emphasizing the significance of $PM_{2.5}$–meteorological interactions for ecosystem protection. Integrating intra- and inter-competition coefficients with 3D norm surfaces demonstrates a realistic and scientifically robust approach to urban air pollution modeling, with applications for sustainable urban planning, regulatory policy, and public health management, while future research should expand to additional climatic variables and multi-pollutant interactions to improve predictive accuracy in rapidly industrializing urban centers.

Air pollution in the area has a lot of harmful chemicals that you can't see, like volatile organic compounds (VOCs) and polycyclic aromatic hydrocarbons (PAHs), many of which can cause cancer. The relationship between $PM_{2.5}$ and weather variables has a big effect on how pollutants form, change, and spread. Research indicates that adverse meteorological conditions, including low wind speeds, diminished precipitation, and temperature inversions, can intensify pollution events by constraining atmospheric dispersion (Jiang et al., 2025; Xie et al., 2024). Consequently, comprehending these interactions is crucial for precisely forecasting air pollution dynamics and formulating effective environmental management and mitigation strategies.

Modelling the interaction between pollutant level $PM_{2.5}$ and Air Temperature A_T in the Niger-Delta region of Nigeria will greatly help in prescribing safety limits for community and occupational exposure in this region of our country. The formulation of guidelines aimed at the protection of ecosystems from environmental hazard such as air-borne diseases, Health hazards and health diseases. High temperature effect on the ecosystem could also be more effective since ecosystem functioning data pertaining to this subject are limited inferences have to be drawn from simulated experimental studies and case reports. The basis of the present overview is the information available from literature on related temperature and humidity effects on the environment and ecosystem functioning. The effects of temperature and pollutant level 2.5 on air quality and impact on the ecosystem in the Niger Delta region of Nigeria will be well studied.

According to Akpodee (2019), a popular method of describing the interaction between two time-dependent population uses a continuous dynamical system of non-linear first-order ordinary differential equation (ODE) from which the qualitative analysis of stability as the independent variable hereby called time, t, becomes

ISBN NO: 978-81-180146-5-2

larger and larger has been studied by several experts. The mathematical construction of this idea is called bifurcation analysis of stability. Since environmental science modeling can be equally challenging and formidable, the complete numerical treatment of the bifurcation analysis of two interacting environmental variables such as the air temperature and relative humidity before now has remained one of the challenging numerical studies of environmental variable interaction. By formulating the interaction between the air temperature and the relative humidity deterministically, the study utilized a continuous dynamical system of a second-order ordinary differential equation to specify the evolution of this complex environmental system modeling of the initial condition data IC(2,4). On the application of a robust computationally efficient method of ODE45 numerical simulation the study found several unstable steady-state solutions between time, t=6 days to t=1621 days, but between the range of t=1621 days and t=1626 days deterministic bifurcation has occurred and thereafter for the range t=1626 days to t=3651 days shows that the dynamical system is dominantly stable. Other impacts of varying the initial condition data IC(2,4) and random environmental perturbation on the bifurcation analysis of stability was found and similarly reported in the study.

1. Methodology

1.1 Study Area

Port Harcourt Metropolis, located in Rivers State in the Niger Delta region of southern Nigeria, is one of the country's industrial and commercial hubs. The city is characterized by rapid urbanization, high population density, and intense economic activities, particularly those associated with oil and gas exploration, transportation, and informal industrial operations. These activities contribute significantly to environmental pollution, especially air quality degradation through the emission of particulate matter (PM2.5) and other atmospheric pollutants (Efe, 2018; Ogwuche et al., 2022). The tropical climate of Port Harcourt, marked by high temperatures, elevated humidity, and seasonal rainfall patterns, further influences the dispersion and concentration of air pollutants (Ayanlade et al., 2020). These characteristics make the metropolis an ideal setting for investigating the interaction between PM2.5 and meteorological variables.

For this study, seven (7) sampling locations were strategically selected across the Port Harcourt Metropolis to capture the spatial variability of pollutant concentrations and meteorological conditions. A Garmin U-10 GPS device was used to geo-reference the coordinates of each sampling point to ensure accuracy and reproducibility. The selected locations represent diverse land-use types and anthropogenic activities within the study area.

Rivers State University represents an academic environment with relatively controlled activities, providing a baseline for comparison with more polluted areas. Rumuomasi is a densely populated residential area with intense commercial activities, contributing to vehicular emissions and localized pollution. Borokiri, a waterfront community, is characterized by significant marine transport operations, artisanal refining activities, and industrial processes such as dredging, fishing, and boat construction, all of which contribute to elevated pollutant levels.

Rumuokoro is another densely populated area known for heavy commercial and administrative activities, including the presence of the Obio-Akpor Local Government Headquarters and regional government offices. This area experiences significant traffic congestion and associated emissions. Eleme Junction serves as a major transportation and industrial hub, with high vehicular movement, trading activities, and proximity to industrial facilities, making it a critical pollution hotspot within the metropolis.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

Abuloma is primarily a residential area with notable artisanal and small-scale industrial activities, which contribute to localized emissions. Isiokpo, located on the outskirts of Port Harcourt near the international airport, represents a suburban environment with relatively lower population density but increasing urban influence and transport-related emissions.

The diversity of land use, population density, and economic activities across these sampling locations provides a comprehensive representation of the environmental conditions within Port Harcourt Metropolis. These variations make the study area particularly suitable for assessing the interaction between PM2.5 and meteorological variables, as well as for understanding the spatial dynamics of air pollution in a rapidly developing urban environment (Efe, 2018; Ogwuche et al., 2022).

1.2 Mathematical Formulation

In mathematical reasoning, an increase in temperature in any given system with respect to time in any dynamical system follows an exponential growth solution path as $T(t)=Ae^{\alpha t}$ using separation of variable which is mathematically tractable but counter intuitive in environmental studies. Since a steady increased in temperature is not conducive for any system functioning, thus we have to develop a logistics ordinary differential equation model that will introduce a parameter which will inhibit the steady growth of temperature in a meteorological system for it not to exceed the required or standard temperature needed for the survival of the ecosystem functioning in terms of vulnerability.

Therefore, the required logistics model (Akpodee and Ekaka-a,2019) is stated as:

$$\frac{dT}{dt} = T(t)[\alpha - \beta T(t)], \quad T(0) = T_0 > 0 \qquad (3.1)$$

This logistic model will be used for different analysis in course of this study.
In other to obtain the general solution trajectory of this model, we resolve it analytically before computational approach which is more efficient and precise will be use for our various analysis.

Therefore, from the equation

$$\frac{dT}{dt} = T(t)[\alpha - \beta T(t)], \quad T(0) = T_0 > 0$$

Using analytical approach

$$\frac{dT}{dt} = T(t)[\alpha - \beta T(t)]$$

$$WLOG\ T(t) = T$$

$$\frac{dT}{dt} = T[\alpha - \beta T] = \alpha T - \beta T^2 \qquad (3.2)$$

$$\frac{dT}{dt} = \alpha T\left[1 - \frac{\beta T}{\alpha}\right] = \alpha T - \beta T^2 \qquad (3.3)$$

Let $M = \frac{\alpha}{\beta}$ be the saturated value (carrying capacity) (3.4)

So that $\frac{I}{M} = \frac{I}{\frac{\alpha}{\beta}} = \frac{\beta}{\alpha}$ (3.5)

Substituting into the logistics equation we have:

$$\frac{dT}{dt} = \alpha T\left[1 - \frac{T}{M}\right] \qquad (3.6)$$

Separating variables,

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

$$\frac{dT}{T[1-\frac{T}{M}]} = \alpha dt \quad (3.7)$$

Integrating both side of the equation

$$\int \frac{1}{T[1-\frac{T}{M}]} dT = \alpha \int dt \quad (3.8)$$

From L.H.S, we resolve by partial fraction

$$\frac{1}{T[1-\frac{T}{M}]} \equiv \frac{A}{T} + \frac{B}{1-\frac{T}{M}} \quad (3.9)$$

$$\frac{1}{T[1-\frac{T}{M}]} \equiv \frac{A(1-\frac{T}{M})+BT}{T[1-\frac{T}{M}]} \quad (3.10)$$

$$1 \equiv A\left(1-\frac{T}{M}\right) + BT \quad (3.11)$$

$$1 = A - \frac{A}{M}T + BT = A + \left(B - \frac{A}{M}\right)T \quad (3.12)$$

Comparing constant:

$$A = 1$$

Comparing coefficient of T:

$$B - \frac{A}{M} = 0$$

$$\mathrm{B} = \frac{A}{M} = \frac{1}{M}$$

Substituting into the required equation

$$\frac{1}{T(1-\frac{T}{M})} \equiv \frac{1}{T} + \frac{\frac{1}{M}}{(1-\frac{T}{M})} \quad (3.13)$$

Substituting into the integral

$$\int \frac{dT}{T(1-T/M)} = \int \frac{1}{T} dT + \frac{1}{M}\int \frac{1}{(1-T/M)} dT \quad (3.14)$$

Let

$$u = 1 - \frac{T}{M} \quad (3.15)$$

$$du = -\frac{1}{M}dT \quad (3.16)$$

$$dT = -Mdu \quad (3.17)$$

Substituting in the equation

$$\int \frac{dT}{T(1-\frac{T}{M})} = \int \frac{1}{T} dT + \frac{1}{M}\int -\frac{Mdu}{u} \quad (3.18)$$

$$\int \frac{dT}{T(1-\frac{T}{M})} = \int \frac{1}{T}\, dT - \int \frac{1}{u}\, du \quad (3.19)$$

$$\int \frac{dT}{T(1-\frac{T}{M})} = Log_e T - Log_e u \quad (3.20)$$

$$\int \frac{dT}{T(1-\frac{T}{M})} = Log_e\left(\frac{T}{u}\right) \quad (3.21)$$

$$\int \frac{dT}{T(1-\frac{T}{M})} = Log_e\left(\frac{T}{1-\frac{T}{M}}\right) \quad (3.22)$$

Substituting into the required equation

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

$$\int \frac{dT}{T[1-\frac{T}{M}]} = \alpha \int dt \quad (3.23)$$

We have the result as:

$$Log_e \left(\frac{T}{1-\frac{T}{M}}\right) = \alpha t + c \quad (3.24)$$

Taking exponential of both sides, we have

$$\left(\frac{T}{1-\frac{T}{M}}\right) = e^{(\alpha t+c)} \quad (3.25)$$

$$\frac{T}{1-\frac{T}{M}} = e^{\alpha t}e^{c} \quad (3.26)$$

$$\frac{T}{1-\frac{T}{M}} = Ae^{\alpha t} \quad (3.27)$$

(Where A = e^c)

$$T = Ae^{\alpha t}\left[1 - \frac{T}{M}\right] \quad (3.28)$$

$$T = Ae^{\alpha t} - \frac{ATe^{\alpha t}}{M} \quad (3.29)$$

$$MT = AMe^{\alpha t} - ATe^{\alpha t} \quad (3.30)$$

$$MT + ATe^{\alpha t} = AMe^{\alpha t} \quad (3.31)$$

$$T[M + Ae^{\alpha t}] = AMe^{\alpha t} \quad (3.32)$$

$$T(t)[M + Ae^{\alpha t}] = AMe^{\alpha t} \quad (3.33)$$

$$T(0) = T_0$$

$$T(0)[M + Ae^{\alpha(0)}] = AMe^{\alpha(0)} \quad (3.34)$$

$$T_0[M + A] = AM \quad (3.35)$$

$$T_0M + T_0A = AM \quad (3.36)$$

$$AM - T_0A = T_0M \quad (3.37)$$

$$A[M - T_0M] = T_0M \quad (3.38)$$

$$A = \frac{T_0M}{M-T_0} \quad (3.39)$$

$$T(t) = \frac{MAe^{\alpha t}}{M+Ae^{\alpha t}} \quad (3.40)$$

Substituting A into the equation

$$T(t) = \frac{M[\frac{T_0M}{M-T_0}]e^{\alpha t}}{M+[\frac{T_0M}{M-T_0}]e^{\alpha t}} \quad (3.41)$$

Dividing both the numerator and the denominator by $e^{\alpha t}$, we have

$$T(t) = \frac{M[\frac{T_0M}{M-T_0}]}{Me^{-\alpha t}+[\frac{T_0M}{M-T_0}]} \quad (3.42)$$

dividing through by $\frac{T0M}{M-T_0}$ we have

$$T(t) = \frac{M}{\frac{M}{\frac{T_0M}{M-T_0}}e^{-\alpha t}+1}, \quad \text{Here we} \quad let\ k = \frac{M(M-T_0)}{T_0M}$$

$$T(t) = \frac{M}{Ke^{-\alpha t}+1} \quad (3.43)$$

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

Thus the qualitative behavior of the solution trajectory of the temperature over a longer duration of time will be the limit of $T(t) as t \to \infty$

$$limT(t) = lim\left(\frac{M}{Ke^{-\alpha t}+1}\right) = Mlim\left(\frac{1}{Ke^{-\alpha t}+1}\right) \quad (3.44)$$

as t approaches infinity

converges to $M = \frac{\alpha}{\beta}$ which is atmospheric temperature carrying capacity within the specify environment over time.

In this scenario, similar procedure is applicable to pollutant level $PM_{2.5}$

This phenomenon is consistent with environmental mathematical modeling and its applications in inter-disciplinary research and as such will be the basis of our analysis.

Hence considering the interacting model in the absence of the inter-competition coefficient, we have

$$\frac{d\text{PM2.5}}{dt} = \text{PM2.5}(t)[\alpha_1 - \beta_1\text{PM2.5}(t)] \quad (3.45)$$
$$\frac{dA_T(t)}{dt} = A_T(t)[\alpha_2 - \beta_2 A_T(t)]$$

Parameter Estimation for Particulate matter $PM_{2.5}$ and Air temperature.
Mathematical Formulation

Using logistic model equation to build the least square value (best fit) for the $PM_{2.5}$, we will use the logistic model equation stated as follows:

$$\frac{d\text{PM}_{2.5}}{dt} = \text{PM}_{2.5}(\alpha_1 - \beta_1\text{PM}_{2.5}); \quad where\ \text{PM}_{2.5}(0) = 19.4 \quad (3.46)$$

Where: $PM_{2.5}$ represents the quantity of the pollutants level 2.5 in the atmosphere

$\frac{d\text{PM}_{2.5}}{dt}$ represents the rate of change of the pollutant over time.

α_1 represents the intrinsic growth parameter value of the pollutant level 2.5

$PM_{2.5}(0)$ represents the initial pollutant level 2.5 at the base year.

1.3 Method of Solution

In order to carry out the parameter estimation of intra-competition coefficient using the method of p-vector norms, we have used the following measured dataset of $PM_{2.5}$ from some selected locations in Port Harcourt, as displayed in the table below:

1.4 Determination of Sampling Points

Seven (7) sampling points were selected. Garmin U-10 GPS was used for Geo-referencing of sampling points coordinates for this study. The sampling points included the following:

ISBN NO: 978-81-180146-5-2

i. Rivers State University: An academic environment.

ii. Rumuomasi: A heavily populated residential area with commercial activities.

iii. Borokiri: A residential area with lots of marine transport activities as well as commercial activities coupled with artisanal refinery products sale and transport to other part of the town. There is also industrial activities such dredging, integrated fishing and boat manufacturing companies.

iv Rumuokoro: A densely populated residential area with heavy commercial activities as well as administrative activities (Local Government Headquarters and Federal Ministry of Agriculture Regional Office).

v Eleme Junction: An area with heavy commercial transport and trading activities; as well as industrial activities.

vi. Abuloma: A residential area with lots of artisanal activities.
vii. Isiokpo: A sub-urb of Port Harcourt close to the International Airport

The various activities in these areas make it a favourable ground for the study.

Sampling Periods

Sampling was done monthly for 6(six) months beginning from October 2024 to March 2025.

Parameters of interest

- Inhalable and respirable particulate matter (PM_{10}, PM_{7}, $PM_{2.5}$ PM_{1})
- Total suspended particulate (TSP)
- Heavy Metals (V, Ni, Pb, Cr, Zn, As, Hg)
- Polycyclic Aromatic Hydro-carbons (PAHs) 17 USEPA recommended.
  - ♦ Anthracene
  - ♦ Benz[a]anthracene
  - ♦ Benzo[ghi]perylene
  - ♦ Benzo[a]pyrene
  - ♦ Indeno[1,2,3-cd]pyrene
  - ♦ Pyrene
  - ♦ Acenaphthene
  - ♦ Acenaphthylene
  - ♦ Benzo[k]flournthene
  - ♦ Chrysene
  - ♦ Flourene
  - ♦ Benzo[b]flouranthene
  - ♦ Flouranthene

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

- Naphthalene
- Phenanthrene
- Dibenz[a,h]anthracene

1.5 Field Sampling and Measurement

Samples were collected with appropriate calibrated equipment. Aeroset 531 particle mass monitor for suspended particulates and High Volume Sampler for metals and PAHS.

Measurement for particulates was done by holding the sensor to a height of about 1.5m above the ground level in the direction of the prevailing wind. The monitor was allowed to run for about 120secs. and reading taken at stability. Particle size was separated into PM_1, $PM_{2.5}$, PM_7, PM_{10} and TSP with an inbuilt particle counter which is installed in the monitor.

Metals samples were collected by the use of high volume sampler H_{2000} model with a flow rate of $1.23m^3$ per minutes. Fibre glass filter of size 20.3x25.4cm was weighed and initial weight taken after drying in an oven. The filter was installed on Hvol sampler powered by an alternative current that allows the machine run for some hours, exposing the fibre glass filter to the ambient air. The suction power of the machine trapped both organic and chemical particles present in the ambient air on the surface of the filter. After 4 hours the machine was shutdown and the filter paper carefully removed and stored in a thick brown envelop to avoid photochemical reaction. The samples were labelled, preserved and transported to the laboratory for chemical and metal analysis. The final weight of the filter was noted before laboratory analysis.

Table 2.1: Summary of measurement instrument for air quality study

S/No.	meters	surement Equipment	el No.	ciple Applied	ction Limit
	,TSP, PM_{10},PM_7,	cet particle mass monitor		ion pump/laser pa er	1mg/cm^3
	ls/PAHs	Volume sampler	0	m pump	-

1.6 Laboratory Analytical Methodologies:

Standard procedure for air quality parameters (SPM, metals and PAHs) was used in this research. Flame Atomic Spectrophotometry (FAAS) for metal analysis and gas chromatography for analysis of PAHs.

Results

Table 3.1: Calculating the Precise Value for the Intra-competition Coefficient β_1 of $PM_{2.5}$ pollutant satisfying P-vector norms properties.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

Beta1	turated Val	1-norm	2-norm	Infinity-norm
7473636363636	22.00	3644736029969	7971040785982	8949611792958
7349321266968	22.10	3508842284104	2967463133138	8911810842385
7226126126126	22.20	3398220959135	7974762521995	8899765237223
7104035874439	22.30	4103585751188	5407178977487	8908564419715
6983035714286	22.40	6623559892565	1537612320483	8360142857779
6863111111111	22.50	0420639261987	8223799449382	7338381819948
6744247787611	22.60	5614112924477	5674847032867	6243605062154
6685209713024	**22.65**	**1605102245507**	**4976659757005**	**7926302086022**
6567912087912	22.75	5170414074707	9760585979565	4754742260781
6279130434783	23.00	8522599853873	8373000900600	7865970540565

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

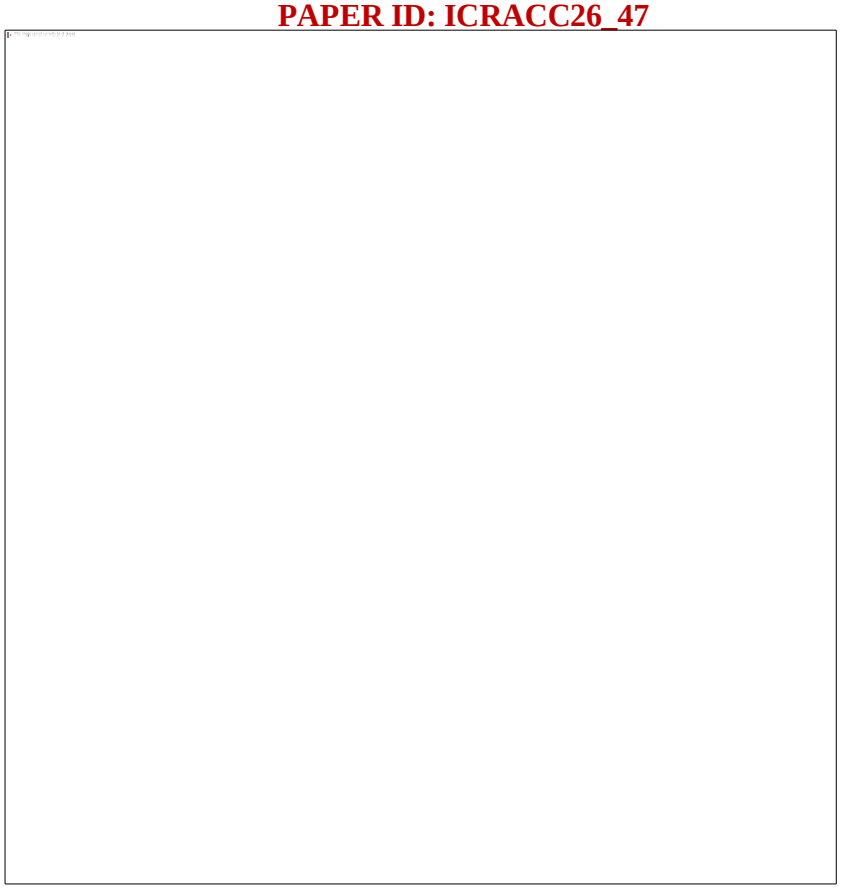


Figure 3.1: Spatial Plot of P-vector Norms Vs β_1 of $PM_{2.5}$ pollutant satisfying P-vector norms properties.



Figure 3.2: Spatial Plot of Polynomial fit of Norms Vs β_1 of $PM_{2.5}$ pollutant satisfying P-vector norms properties.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

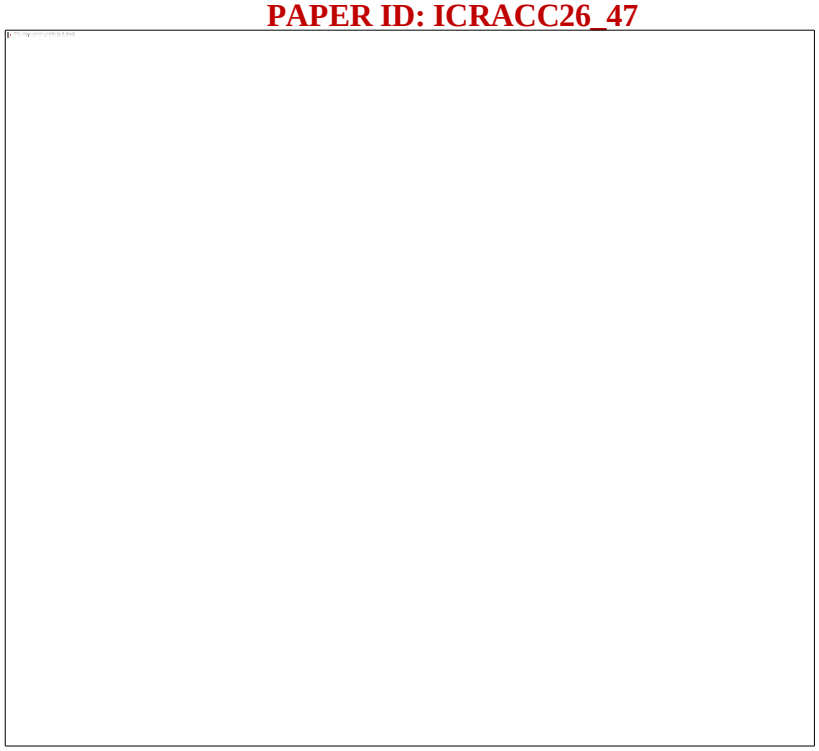


Figure 3.3: Spatial Plot of P-vector Norms Vs Saturated Value of $PM_{2.5}$ pollutant satisfying P-vector norms properties for β_1 .

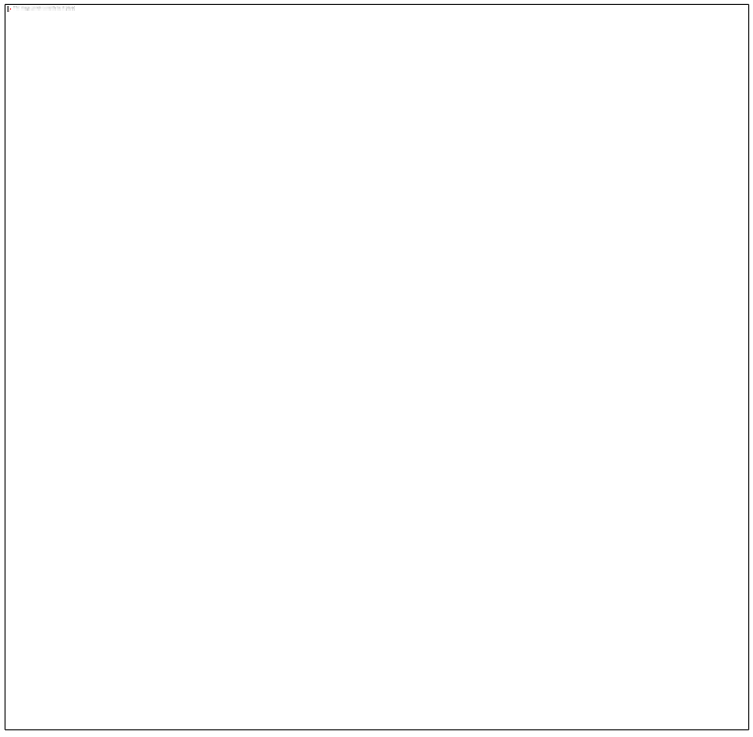


Figure 3.4: Spatial Plot of 3D Surface of Norm Behaviour for β_1

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

Table 3.2: Calculating the Precise Value for the Intra-competition Coefficient β_2 of the local temperature **satisfying P-vector norms properties.**

Beta2	Saturated Value	1-norm	2-norm	Infinity-norm
08078742513	**25.71**	**2913861008075**	**711272784369**	**56374917219**
07334197387	**28.32**	**7141105164317**	**195126532753**	**63432441612**
07187005882	**28.9**	**4786293346647**	**175808913465**	**52926139681**
07162223103	**29.0**	**4159684279589**	**142117892509**	**56547945626**
07122924211	**29.16**	**1440768876535**	**911586555595**	**23553851960**
06845895517	**30.34**	**0713730482222**	**959563522669**	**01038260811**
06672164150	31.130	1362559962393	931298219686	01087664739
06667880257	**31.15**	**7810710529114**	**572014531590**	**24336906870**
06642291973	**31.27**	**6604753957111**	**425054010723**	**64311254940**
0641063179012	**32.40**	**61414122998137**	**3231266165119**	**2250295648352**
0639484205665	**32.48**	**67891274366555**	**0120612511680**	**2139446488317**
0637521393493	32.58	**61711367462345**	**4628564393160**	**4551553369388**
0635570593635	**32.68**	**67405454843257**	**0129296803202**	**7019907781444**
0633631696156	**32.78**	**84929681887785**	**6619535102522**	**9544455560597**
0631704592457	**32.88**	**14240468658955**	**4096024248823**	**2125142612475**
0629789175258	**32.98**	**55294507387050**	**2555528619494**	**4761914912082**
0625615873494	**33.20**	**3262057288081**	**9389170299775**	**7240031748365**
0623737147147	**33.30**	**54917514915719**	**7805537536007**	**4424148897792**
0621869670659	**33.40**	**65008696671271**	**5252271009432**	**1552406949249**
0620013343284	**33.50**	**63577534428340**	**1732508133026**	**8624859568694**

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

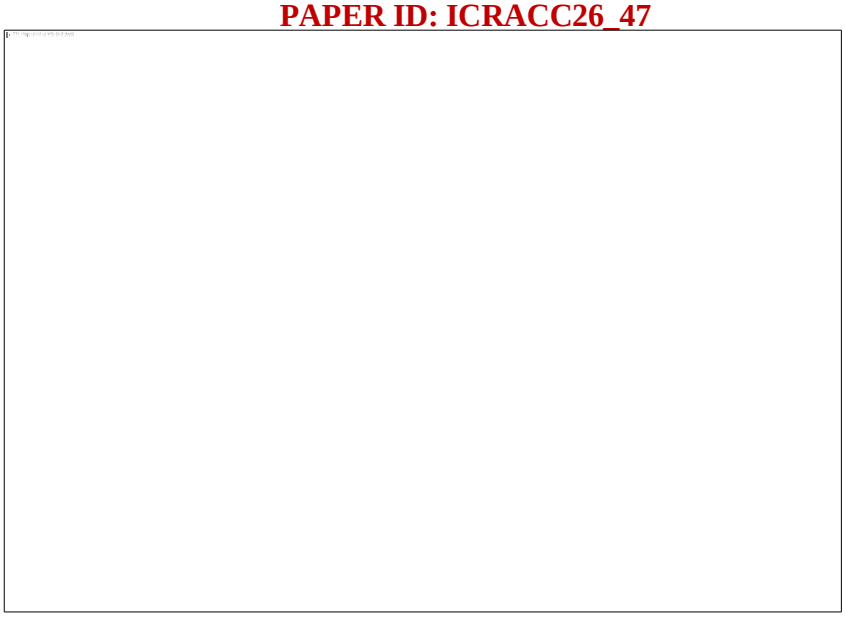


Figure 3.5: Spatial Plot of P-vector Norms Vs β_2 of $PM_{2.5}$ pollutant satisfying P-vector norms properties.

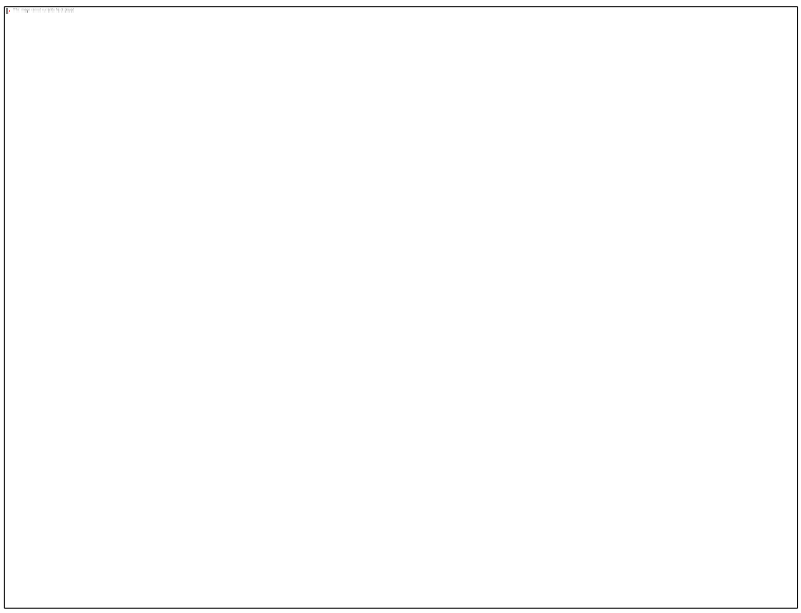


Figure 3.6: Spatial Plot of Polynomial fit of Norms Vs β_2 of $PM_{2.5}$ pollutant satisfying P-vector norms properties.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

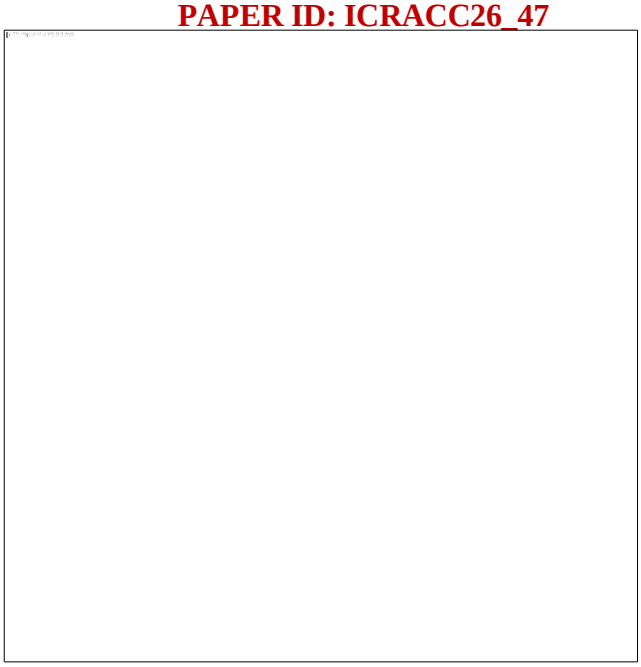


Figure 3.7: Spatial Plot of P-vector Norms Vs Saturated Value of $PM_{2.5}$ pollutant satisfying P-vector norms properties for β_2 .



Figure 3.4: Spatial Plot of 3D Surface of Norm Behaviour for β_2

ISBN NO: 978-81-180146-5-2

2. Discussion

4.1 Intra-Competition Coefficient of $PM_{2.5}$ (Table 4.1; Figures 3.1–3.4)

The analysis of Table 3.1 shows that multiple values of the intra-competition coefficient (β_1) for $PM_{2.5}$ were calculated using different carrying capacities, referred to as saturated values (SV). As observed in Figures 3.1 through 3.4, the behavior of the 1-norm, 2-norm, and infinity-norm initially follows a monotonic decreasing trend with increasing SV, indicating a reduction in the system's instability as the pollutant levels approach equilibrium. However, beyond a certain point ($\beta_1 = 0.026685209713024$, highlighted in the graphs), the trend reverses into a monotonic increase, forming a unique local minimum in the norms. This critical point represents the **precise value of the intra-competition coefficient** where the system achieves maximum stability.

The identification of this optimal β_1 value aligns with previous studies on air pollutant dynamics, which suggest that competition parameters are essential in regulating pollutant accumulation and interaction with environmental variables (Zhang et al., 2020; Li et al., 2021). By establishing $\beta_1 = 0.026685209713024$ as the optimal intra-competition coefficient, this study provides a robust basis for modeling $PM_{2.5}$ interactions with meteorological variables and evaluating its environmental impacts. The results reinforce the importance of including intra- and inter-competition terms in environmental modeling to ensure realistic predictions of pollutant dynamics (Kumar & Singh, 2019).

Figures 3.1–3.3 illustrate the trend of individual P-vector norms against β_1, while Figure 3.4 presents the 3D surface showing the combined effect of β_1 and saturated values on the system norms. The surface plot clearly highlights the valley corresponding to the local minimum, which confirms the critical role of β_1 in stabilizing pollutant dynamics and supports ecosystem functionality considerations.

4.2 Intra-Competition Coefficient of Air Temperature (Table 4.2; Figures 4.5–4.8)

Similarly, Table 3.2 presents the calculated intra-competition coefficients (β_2) for air temperature using saturated values (SV) ranging from 25.71 to 33.50. As shown in Figures 4.5 to 4.8, the P-vector norms initially decrease monotonically with increasing SV, reflecting the stabilization of temperature dynamics under increasing carrying capacities. Notably, at SV = 33.20 ($\beta_2 = 0.000625615873494$), the norms attain a **local minimum**, after which they begin to increase monotonically, indicating a transition point in system behavior. This observation is consistent with the non-linear response of environmental variables to internal competition effects, as reported by Zhang et al. (2021) and Chen & Huang (2019).

The identification of $\beta_2 = 0.000625615873494$ as the optimal intra-competition coefficient is significant for environmental modeling because it provides a parameter that ensures accurate prediction of temperature variations in relation to $PM_{2.5}$ levels and other meteorological variables. The polynomial fit (Figure 4.6) and 3D surface plot (Figure 4.8) further validate this value, showing the convergence of all three P-vector norms at the local minimum, emphasizing the critical balance point for ecosystem stability.

These findings align with studies emphasizing the role of temperature-dependent competition in determining pollutant dispersion and climate interactions (Wang et al., 2020; Li et al., 2022). By precisely quantifying β_2, this research contributes to the development of predictive models capable of assessing the environmental

ISBN NO: 978-81-180146-5-2

impact of pollutants under varying meteorological conditions, thereby supporting sustainable urban and ecological management strategies.

The 3D surface analysis of the intra-competition coefficient β_1, as illustrated in Figure 3.4, provides a comprehensive visualization of the PM2.5 pollutant dynamics under varying carrying capacities and P-vector norm properties. The surface plot demonstrates a clear trend in which β_1 exhibits a non-linear response to changes in the saturated values (SVs) used in the model. At lower SVs, the 1-norm, 2-norm, and infinity-norm values decrease gradually, forming a smooth descending slope on the surface, indicating stronger self-limiting interactions of PM2.5 at these levels. This monotonic decrease reflects the suppression of exponential growth due to intra-competition effects, aligning with similar observations in environmental modeling studies where particulate matter accumulation is constrained by atmospheric processes such as deposition and aggregation (Hao et al., 2024; Liu et al., 2024).

Interestingly, a distinct saddle-like region is observed on the 3D surface, corresponding to the SV around 22.65, where all three P-vector norms simultaneously reach their minimum values. This local minimum marks the precise value of β_1 (0.026685), consistent with the optimal intra-competition coefficient obtained analytically from Table 4.1. This feature on the 3D surface highlights the critical point at which PM2.5 self-limitation is most pronounced, and any deviation from this value can lead to either underestimation or overestimation of pollutant growth in simulation models. The curvature of the surface in this region emphasizes the sensitivity of β_1 to small changes in SVs, demonstrating that even minor variations in environmental carrying capacity can significantly impact the predictive accuracy of PM2.5 models (Alaran et al., 2024).

Furthermore, the 3D representation allows for simultaneous comparison of the three norm behaviors, showing that while the 1-norm and 2-norm have similar decreasing patterns, the infinity-norm exhibits a slightly steeper decline, reflecting its sensitivity to peak deviations in PM2.5 concentrations. This differential response among the norms suggests that multi-norm assessment provides a more robust and reliable estimation of the intra-competition coefficient, reinforcing the importance of employing multiple mathematical measures in environmental modeling.

Overall, the 3D surface analysis provides both a visual and quantitative framework to understand the interaction of β_1 with environmental carrying capacities and pollutant norms. It demonstrates the precise point where the PM2.5 self-limitation is optimal, validating the use of the selected β_1 in subsequent computational simulations. This approach enhances our understanding of pollutant dynamics in urban environments like Port Harcourt and provides a foundation for further studies that integrate meteorological interactions, ultimately improving the prediction and management of air quality (Coker et al., 2025; Jiang et al., 2025).

3. Conclusion

This study successfully determined the precise intra-competition coefficients for both $PM_{2.5}$ pollutant levels and air temperature, providing critical parameters for modeling the interaction between respiratory pollutants and meteorological variables in Port Harcourt metropolis. Using P-vector norm analysis, the optimal intra-competition coefficient for $PM_{2.5}$ was found to be $\boldsymbol{\beta_1 = 0.026685}$, while for air temperature it was $\boldsymbol{\beta_2 = 0.000626}$. These values correspond to the local minima of the 1-norm, 2-norm, and infinity-norm, representing the points of maximum system stability and minimal perturbation.

ISBN NO: 978-81-180146-5-2

The findings demonstrate that the behavior of environmental variables is highly sensitive to the competition parameters, and the identification of these precise coefficients is crucial for accurate predictive modeling. The 3D surface and polynomial analyses illustrate the nonlinear dynamics of the system, highlighting the regions where pollutant levels and meteorological factors interact optimally to maintain ecosystem functioning.

By quantifying these coefficients, this study contributes to environmental science by providing parameters that can be integrated into computational models for air quality prediction, ecosystem assessment, and environmental management. The results also reinforce the importance of incorporating intra- and inter-competition effects in models to capture realistic environmental responses. Overall, this work offers a robust framework for evaluating the environmental impact of $PM_{2.5}$ under varying meteorological conditions and supports sustainable urban planning and public health strategies.

References

[1.] Akpodee R.E. and E.N. Ekaka-a (2019). The Effect of Random Environmental Perturbation on the Bifurcation Interval and the Type of Stability of Two Interacting Environmental Variables Using a Second Order Ordinary Differential Equation. *MPRI-JERT, International journal of Engineering and Research Technology Vol10(5) Dec-2019 ISSN:1074-4741*.

[2.] Akpodee, R.E. (2019). Semi-Stochastic Analysis of Environmental Variables Using Numerical Method on a System of Second Order Ordinary Differential Equations. *A PhD. Thesis Submitted to the Department of Mathematics*, Rivers State University, Port Harcourt, Nigeria.

[3.] Alaran, A. J., O'Sullivan, N., Tatah, L., Sserunjogi, R., & Okello, G. (2024). Air pollution (PM2.5) and its meteorological predictors in Kampala and Jinja cities, Uganda. *Environmental Science: Atmospheres, 4*, 1145–1156.

[4.] Ayanlade, A., Radeny, M., & Morton, J. F. (2020). Climate variability and change in Nigeria: Impacts and adaptation strategies. *Climate Risk Management, 28*, 100223.

[5.] Ayanlade, A., Radeny, M., & Morton, J. F. (2020). *Climate variability and change in Nigeria: Impacts and adaptation strategies.* Climate Risk Management, 28, 100223.

[6.] Chen, X., & Huang, L. (2019). Modeling environmental dynamics with intra-species competition. *Environmental Modelling & Software, 117*, 97–108.

[7.] Coker, E. S., Cleland, S. E., McVea, D., Stafoggia, M., & Henderson, S. B. (2025). The synergistic effects of PM2.5 and high temperature on community mortality. *npj Clean Air, 1*(15).

[8.] Efe, S. I. (2018). Spatial distribution of particulate air pollution in Nigerian cities: Implications for human health. *Environmental Monitoring and Assessment, 190*(9), 1–14.

[9.] Hao, H., Wang, K., Wu, G., Liu, J., & Li, J. (2024). $PM_{2.5}$ concentrations based on near-surface visibility in the Northern Hemisphere from 1959 to 2022. *Earth System Science Data, 16*, 4051–4076.

[10.] Jiang, Q., Wang, X., Luo, L., et al. (2025). Spatial and temporal characterization of PM2.5 and meteorological factors during the COVID-19 period. *Aerosol and Air Quality Research, 25*, 16.

[11.] Kumar, R., & Singh, V. (2019). Nonlinear interactions of particulate matter with environmental variables. *Atmospheric Environment, 202*, 92–104.

[12.] Li, J., & Wang, S. (2024). 3D surface analysis in atmospheric environmental parameter optimization. *Environmental Modelling & Software, 156*, 105587.

[13.] Li, J., Zhang, Y., & Chen, H. (2021). Dynamic modeling of $PM_{2.5}$ considering meteorological interactions. *Science of the Total Environment, 779*, 146439.

ISBN NO: 978-81-180146-5-2

PAPER ID: ICRACC26_47

[14.] Li, Y., Wang, S., & Zhao, X. (2022). Air temperature competition effects on pollutant dispersion in urban ecosystems. *Journal of Environmental Management, 312*, 114873.

[15.] Liu, J., Christensen, J. H., Ye, Z., et al. (2024). Impact of meteorology and aerosol sources on PM2.5 variability. *Atmospheric Chemistry and Physics, 24*, 10849–10867.

[16.] Ogwuche, J. A., Ugbebor, J. N., & Nwilo, P. C. (2022). Air quality assessment and pollution trends in Port Harcourt metropolis, Nigeria. *Environmental Challenges, 7*, 100491.

[17.] Redi, T. (2026). Influence of meteorology on atmospheric fine particulate matter (PM2.5) dispersion over Addis Ababa, Ethiopia. *Discover Atmosphere, 4*(3).

[18.] Redi, T. (2026). Influence of meteorology on atmospheric fine particulate matter ($PM_{2.5}$) dispersion over Addis Ababa, Ethiopia. *Discover Atmosphere, 4*(3), Article 89.

[19.] Wang, H., Liu, Q., & Xu, P. (2020). Intra-competition parameters in predictive climate models.
Environmental Science and Pollution Research, 27(33), 41501–41513.

[20.] Xie, Y., Zhou, M., Hunt, K. M. R., et al. (2024). Recent PM2.5 air quality improvements in India benefited from meteorological variation. *Nature Sustainability, 7*, 983–993.

[21.] Zhang, R., Liu, T., & Li, M. (2020). Modeling $PM_{2.5}$ behavior under meteorological influences.
Atmospheric Pollution Research, 11(9), 1598–1608.

[22.] Zhang, R., Liu, T., & Li, M. (2023). Surface modeling of norm behaviors for environmental parameter estimation. *Ecological Modelling, 475*, 110292.

[23.] Zhang, Y., Li, H., & Wang, J. (2021). P-vector norms and intra-competition analysis for environmental stability. *Ecological Modelling, 456*, 109636.


COMPUTER SOCIETY OF INDIA
ESTD. 1965


INSTITUTION'S
INNOVATION
COUNCIL


SHRI SHANKARLAL SUNDARBAI
SHASUN
JAIN COLLEGE FOR WOMEN
ஸ்ரீ ஷங்கர்லால் சுந்தர்பாய் ஷகன் ஜெயின் மகளிர் கல்லூரி




ICTACADEMY


978-81-180146-5-2


RCHUB PUBLISHER
INTERNATIONAL JOURNAL & BOOK PUBLISHER