

Proceedings of

**The International Conference on
Artificial Intelligence and Cyber Security
(ICAIC 2025)**

October 17–18, 2025 ||Manitoba, Canada.

ISBN: 978-1-946782-45-8



Canadian Tech-Institute
for Academic Research



London School
of Management

Published by CTIAR

Canadian Tech-Institute for Academic Research,
819 Sargent Ave, 202 204/206 #106,
Winnipeg, MB R3E 0B9 Canada

CTIAR COPYRIGHT NOTICE. Copyright © 2025 by the Canadian Tech-Institute for Academic Research. Permissions to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than CTIAR must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee. Request permissions from CTIAR by email info@ctiarconferences.ca

CTIAR ISBN: 978-1-946782-45-8

The International Conference on Artificial Intelligence and Cyber Security (ICAIC 2025)

Table of Contents

ASTRID: API Sequence Threat Recognition with Intelligent Discrimination Ankita Ghosh	9 - 18
Strategic Cybersecurity Governance: A Risk-Based Policy Framework for IT Protection and Compliance. Mohammed Nayeem	19 - 29
Subverting Quantum Machine Learning via Hybrid Data Corruption. Mohammed Nayeem	30 - 37
Educating for Ethical Code: Disrupting Societal and Algorithmic Bias through the Computer Science Education. Karen A. Latimer	38 - 49
Artificial Intelligence and Edge Computing to Improve Internet of Vehicles. Emmanuel Judicael NGUINGUI	50 - 57
AI-Powered Biochemical Materials for Biomedical Applications: Current Trends and Future Prospects. Love Edet M. and Moyo Mthabisi	58 - 73
Explainable AI-Powered Anomaly Detection: A Computer Vision Approach to Strengthening Human-Centric Cybersecurity. Victor Ayodeji Oluwasusi, Temiloluwa O. Modupeola, & Nasim Ahmadzadeh Nobari Azar	74 - 90
The Emergence of Multimodal AI in Software Development: What to Expect in the Future? Imane Boumedra and Ibrahim Adeshola	91 - 99
Exploring the effect of ai integration on collaborative practices and decision making in construction project management. A literature review. Bamidele Charles Olaiya and Olawunmi Azeezat Akanbi	100 - 110
The Ethics of Artificial Intelligence in the Oil and Gas Industry: Navigating Environmental Impact, Spill Response, and Data Governance in Nigeria. Jemimah Chuntar	111 - 114
Challenges in Maintaining Cybersecurity Standards in Remote Work in Nigerian Financial Institutions PostCOVID-19: An Analysis of Regulatory Frameworks and Industry Practices. Adoikwu Pius Odeh, and Acheme Okolobia Odeh	115 - 127
Machine learning for financial statement analysis and forecast: A case study of Guaranty Trust Holding Company PLC (GTCO PLC.). Ayomide Sobowale	128 - 149
Does the use of Generative AI contribute to university students' engagement or disengagement in their studies? Examining the antecedents of fear of missing out, perceived ease of use, and perceived usefulness. Ibrahim Adeshola, Afi Kekeli Feda, and Adeola Praise Adepoju	150 - 163
AI Enabled Credit Card Fraud Detection System Using Cloud Computing Alpana Borse, Satish Chavhan, Abhishek Shinde, and Rohit Shinde	164 - 168
Application to promote mental health awareness and self-assessment M.Ragul, S.Ragavan, R.Nandhagokul, K R .Ranjith, and N.Valarmathi	169 - 172
Cloud AI- Data Breach Detection M. Fatima, Kalpana Rai, and Alfiya Jahan	173 - 175
Early detection of mental illnesses: Analysis of questionnaire and social media data to identify early signs of depression and anxiety. Youssef Mathouri, Korchiyne Redouan, Badour Mohcine, Younes Chihab, and Boukhalifa Alaeddine	176 - 185
Enhanced Evaluation System for Student Performance Analysis Ashwini Matange, Sayee Patil, Pranjali Kotalwar, Akshata Mahajan, and Rutuja Kalekar	186 - 190

Innovative Narratives: Integrating AI-driven Character Development with Modern Narratological Techniques for Automated Storytelling. Patricia Ghann, William Agyapong, and Isaac Owusu Koforidua	191 - 201
Prashaansh : Automatic Question Paper Generator Ashwini Matange, Harshad Karale, Nayan Keote, Ashirwad Katkamwar, and Aviraj Kale	202 - 208
Vision-Based Pothole Detection and Reporting System for Smart Cities (Leveraging YOLOv8-Seg and Predictive Maintenance Models) Paras Sawale, Omkar Adsare, Kusumlata Pawar, Gayatri Shelke, Nihar Ranjan, and Shrutika Jagdale	209 - 214
AI Legal Responsibility and Accountability: Reframing Legal Responsibility in the Age of Autonomous Systems Xena Dominique Njoya DAVID	215 - 221
Segmentation of Pavement Cracks in North Cyprus Albert David Bangura, and Tolga Yirtici.	222 - 239
From Creator to Code: Legal Attribution and Ownership in AI-Generated Intellectual Property Žaklina Spalević, Lazar Stošić, and Miloš Ilić	240 - 248
A Scalable Framework for Interpretable Binary Vulnerability Analysis Using Data Dependency Modeling Fahad Amin	249 - 254

Preface

The International Conference on Artificial Intelligence and Cyber security was a huge success. All keynote speeches and presentations made by authors brought great insight, bridging the gap in literature as touching artificial intelligence and cybersecurity. Our deep appreciation goes to all the keynote speakers and authors who made great contributions towards this conference, and the great efforts put behind the scenes by the organizing and scientific committee to ensure a hitch free event. We look forward to a more engaging, interactive and explosive conference in 2026.

This edition presents a remarkable collection of papers from the International Conference on Artificial Intelligence and Cybersecurity (ICAIC 2025), held in Manitoba, Canada, from 17 to 18 October 2025 as a hybrid conference. Each paper submitted represents a distinct contribution to the field of artificial intelligence and cybersecurity, exploring diverse areas of application and practical relevance.

All attendees made meaningful contributions through presentations, discussions, and the exchange of ideas during interactive question sessions and panel engagements, all of which added immense value to the success of the conference. Participants joined from eleven countries, including the United States, Canada, Turkey, Cyprus, Morocco, India, Uganda, Serbia, Nigeria, Ghana, and Algeria, giving the conference a truly global scope.

Two plenary keynote addresses were delivered during the conference. The first, Responsible AI for Global Impact, was presented by Associate Professor Dr. Maurice Dawson Jr. of the Illinois Institute of Technology, a distinguished cybersecurity scholar and leader with over fifteen years of experience across academia, industry, and government, known for advancing global collaboration, research, and education in information security and cyber warfare. He is also a five time recipient of the Fulbright Scholar Award.

The second keynote, Architecting the AI ML Data Lake: Foundations for LLMs and Enterprise AI, was presented by Anant Kumar, a seasoned technical leader and innovator at Salesforce. He is known for his expertise in scalable big data and AI infrastructure, with a proven record in cloud native systems, patented network security technologies, and enterprise scale innovation. Both keynote sessions captured the full attention of participants and inspired a series of thought provoking questions that energized the audience.

The International Conference on Artificial Intelligence and Cybersecurity was an outstanding success. The keynote speeches and paper presentations provided rich insights, bridging important gaps in literature within the domains of artificial intelligence and cybersecurity. Our sincere appreciation goes to all keynote speakers and authors for their valuable contributions, and to the organizing and scientific committees whose dedicated efforts ensured a smooth and impactful event. We look forward to an even more engaging and enriching conference in 2026.

ICAIC 2025 Planning Committee.

ICAIC Committee

Honorary Board

Dr. Samson O. Fadiya

President, Canadian Tech-Institute for Academic Research,
Canada

Dr. Anju Bajaj

STEM Educator I STEM Advisor.

Global Citizenship Award Recipient, Prime Ministers Award for Teaching Excellence and Queen Elizabeth II Platinum Jubilee Medal Recipient, Canada

Professor Dr. Ali Eydgahi

Department of College of Engineering and Technology,
Eastern Michigan University,
United States

Professor Dr. Lazar Stosic

Faculty of Informatics and Computer Science,
University Union Nikola Tesla,
Belgrade, Serbia.

Professor Dr. Danimir Mandic

Director of the Center for Robotics,
Faculty of Education, University of Belgrade,
Serbia.

Professor Dr. Lukasz Tomczyk

Institute of Pedagogy, Jagiellonian University in Cracow,
Poland.

Professor Dr. Sinisa Opic

Faculty of Teacher Education,
University of Zagreb,
Croatia

Keynote Speakers

Professor. Dr. Arif SARI

Snipesecc Cyber Security Agency | PhD, CISA, CISM, CDPSE, CCISO, eWPTXv2, eCPTXv2, CPENT, ECIH, Cyprus

Anant KUMAR

Technical Leader | AI/ML Big Data | AWS Cloud | Apache Spark | Researcher | Stanford GSB. United States

Associate Professor Dr. Maurice DAWSON

Cyber Security Center Director @ Illinois Institute of Technology | Certified CISO, Policy Expert, United States

Scientific Committee

Professor. Francisco David Guillén-Gámez

Department of Didactics and School Organization,
Faculty of Education Sciences,
Universidad de Málaga, Spain.

Associate Professor. Nazmi Xhomara

Department of Mathematics and Statistics,
Faculty of Information and Innovation Technology,
Luarasi University, Albania

Associate Professor. Tonia De Giuseppe

Giustino Fortunato University of Benevento, Italy.

Professor. Florence Sèdes

Computer Science, IRIT Lab,
Université de Toulouse, France.

Professor. Saiida LAZAAR

Department of Mathematics and Computer Sciences,
National School of Applied Sciences - Tangier,
AbdelMalek Essaadi University, Morocco.

Dr. Matthew Johnson

Research Software Engineer Lead at Microsoft,
United Kingdom

Associate Professor. Abhijit Kumar Nag

Subhani Department of Computer Information Systems,
Texas A&M University-Central Texas, United States.

Associate Professor. Stavros Shiaeles

Faculty of Technology,
School of Computing Centre for Cybercrime and Economic Crime,
United Kingdom.

Assistant Professor. Benjamin Yankson

Cybersecurity Department College of Emergency Preparedness,
Homeland Security and Cybersecurity,
United States.

Dr. Anju Bajaj

STEM Educator I STEM Advisor.

Global Citizenship Award Recipient, Prime Ministers Award for Teaching Excellence and Queen Elizabeth II Platinum Jubilee Medal Recipient, Canada

Emeritus Professor. Jürgen Pilz

Faculty of Technical Sciences,
Department of Statistics,
University of Klagenfurt, Austria.

Assistant Lecturer. Uneneibotejit J. Otokwala (PhD)

Demonstrator in Cybersecurity,
Robert Gordon University, Aberdeen, United Kingdom

Assistant Professor. Ehsan Toreini

Software Security, Security Architecture, Physical Security and Trustworthy AI, University of Surrey, United Kingdom

Vijayalaxmi Methuku

Researcher, University of Texas at Austin,
United States.

Chirag Agrawal

Senior Software Engineer (SDE III) at Amazon,
United States

Technical Committee

Dr. Kelsey Dappa

MSc, PhD, EIT

David Fatimehin

Project Manager | Business Analysts| Marketing Expert | Community Coordinator|, Performing Artist| Mentor| Co-founder of FRAME Art warehouse Winnipeg |Certified PMP (Project Management Institute).

ASTRID: API Sequence Threat Recognition with Intelligent Discrimination

Ankita Ghosh

Indira Gandhi Delhi Technical University for Women, Delhi, India

ankita0512ghosh@gmail.com

Abstract

This paper introduces ASTRID, a GAN-based model that can identify malware using API call sequence analysis in noisy settings, a typical problem in disk forensics. The new framework addresses the shortcomings of conventional detection approaches by leveraging adversarial learning to differentiate between benign and malicious sequences. API call information from two noisy datasets was utilized to train the model under real-world-like conditions, while a third, unseen dataset was employed to test the model's generalization ability. ASTRID achieved an accuracy of 96.8% on training data and 95.5% on the test dataset, with performance comparable to state-of-the-art models. These outcomes demonstrate the power of ASTRID to tackle noise and provide detection reliability, providing an impressive solution towards durable malware detection through sequence-based learning.

1. Introduction

Malware, which stands for malicious software, is any software that is deliberately created to harm a computer system, server, client, or network. It encompasses a broad range of threats, including viruses, worms, ransomware, trojans, spyware, and adware. Malware has extensive effects, not just on individual users but also on critical infrastructures and businesses. A recent real-life case is the 2025 cyberattack on Marks & Spencer (M&S)¹, a major UK retailer, where the hacking group "Scattered Spider" employed DragonForce ransomware to breach M&S's IT systems, bringing online orders to a standstill, disrupting supply chains, and incurring weekly losses estimated at £15 million.

With the sophistication and frequency of these attacks increasing, the importance of effective malware detection mechanisms has never been greater². Timely and precise detection is required not only to avert financial loss but also to maintain data privacy, system integrity, and public trust. With the increased incidence of zero-day vulnerabilities and obfuscated malware instances, conventional antivirus solutions tend to miss threats in real time.

To improve detection functionality, researchers have resorted to machine learning (ML) methodologies. SVM, Random Forest, and K-Nearest Neighbors (KNN) have been the most

common models applied to malware classification in terms of static and dynamic characteristics (Zhou & Jiang, 2012). These solutions, even if they exhibit a high success rate in balanced data sets, generally fail if submitted to new, unseen, or manipulated samples. Research indicates that ML models (Azeem, Khan, Iftikhar, Bawazeer, & Alzahrani, 2024) are susceptible to evasion attacks and not robust enough in real-world scenarios, restricting their scalability and adaptability.

These limitations are remedied by the introduction of Generative Adversarial Networks (GANs) as an innovative solution. Unlike traditional classifiers, GANs are composed of two neural networks—a generator and a discriminator—engaged in a competitive learning process, enabling them to create synthetic data samples that are similar to real ones (Goodfellow, et al., 2014). In malware detection, GANs can enhance training data by creating realistic malicious sequences, thus enhancing generalization and resilience against new threats. Recent research, including PlausMal-GAN (Noguchi, Sun, Lin, & Harada, 2022), reveals that GAN-based architectures perform better than traditional ML models in identifying advanced malware, particularly for noisy and imbalanced data.

In the past, API call sequence-based malware detection has also increased in reputation as compared to standard static analysis. Static analysis that analyzes binary code or file signatures without actually executing them does not perform well against contemporary malware based on obfuscation, encryption, or polymorphism in covering up its genuine behavior (Guri & Bykhovsky, 2019). Conversely, dynamic analysis, particularly API call monitoring, realizes the current behavior of applications while they engage with system resources, offering a more robust detection technique against such evasions. API calls are a form of behavioral fingerprint that realizes patterns of malicious intent, like unauthorized access to files, changes to the registry, or communications over the network. As malware becomes more advanced and adaptive, examining sequences of API calls enables more capable, context-specific models that are better able to generalize in real-world threat scenarios (Ge, Yarom, Li, & Heiser, 2017). Hence, combining API dynamic analysis with high-performing models such as GANs is a promising avenue

¹<https://www.theguardian.com/business/2025/may/03/inside-the-marks-and-spencer-cyber-attack-chaos>

²<https://cymulate.com/blog/malware-detection-techniques/>

for malware detection research, which can better alleviate the disadvantages created by static-only methods.

Research Questions Formed

RQ1: How can malware be effectively detected from noisy API call sequences that reflect real-world system behavior?

RQ2: Can a GAN-based model generalize well enough to identify malware in unseen and diverse datasets?

RQ3: How does dynamic, API-based analysis compare to static analysis in enhancing the accuracy and robustness of malware detection systems?

Contributions

- Developed ASTRID, a GAN Model which works to differentiate between Benign and Malware Data from Datasets containing noise to mimic real-world system behaviour.
- Tested on an unseen dataset for checking the generalizability of the model.
- Improved the robustness of the Model by generating new malware calls for the model to train, removing the need to retrain the model on new variants of Malware.

2. Literature Review

One of the major challenges in malware detection is that noisy API call sequences hide malicious behavior. Conventional techniques are unable to distinguish good from bad in noisy environments. New techniques have been suggested in recent research to overcome the challenge. For example, the paper "Malware Detection Based on API Call Sequence Analysis: A Gated Recurrent Unit-Generative Adversarial Network Model Approach" (Owoh N. , et al., 2024) presented a hybrid deep learning architecture based on Gated Recurrent Units (GRUs) and Generative Adversarial Networks (GANs) to improve malware detection based on API call sequences. Their framework was better performing, 98.9% accuracy on difficult datasets, than other approaches such as Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU). However, one limitation of this research is that it relies upon a quite limited dataset, which may influence the extent to which findings can be representative of larger and more diverse datasets.

Generative Adversarial Networks (GANs) have drawn significant attention as to how they are applied in the detection of malware. An extensive survey, "Generative Adversarial Networks in Anomaly Detection and Malware Detection: A Comprehensive Survey" (Hu, Zhang, & Li, 2025) addressed the contribution of GANs to detecting malware and anomalies, with the authors providing extensive information on various GAN architectures and their

effectiveness in the detection of malicious activity. This work provides a broad overview of GAN use in different forms of malware, thereby further supporting the use of GANs as an extremely useful cybersecurity tool. However, authors acknowledge that GANs could require excessive computational capacity to train, thus becoming less feasible in real-time malware detection.

The debate between dynamic and static analysis methods for malware detection continues to be a focal point in cybersecurity research. Dynamic analysis involves executing the program in a controlled environment to observe its behavior, while static analysis examines the code without execution. A study, "Malware detection with dynamic evolving graph convolutional networks" (Nguyen, Di Troia, Ishigaki, & Stamp, 2022), proposed a dynamic evolution graph convolutional network (DEGCN) model to capture dynamic evolution patterns of local API-level and global graph-level software behaviors, achieving good performance in malware detection. While dynamic analysis provides deeper insights into malware behaviors, the study highlights a key shortcoming in that it can be resource-intensive and slow, making it unsuitable for detecting large-scale malware threats in real-time.

To improve the robustness of malware detection models, data augmentation techniques have been employed. A study, "Improving Android Malware Detection Through Data Augmentation Using Wasserstein Generative Adversarial Networks" (Stalin & Mekoya, 2024) explored the use of Wasserstein Generative Adversarial Networks (WGANs) for data augmentation in Android malware detection. Their approach demonstrated a notable performance enhancement of the classification model, with the highest achieved F1 score reaching 0.975. Despite the strong results, a limitation of this approach is that it primarily focuses on Android malware, and the proposed method may not be directly applicable to other platforms or environments.

The generalizability ability of GAN-based models is crucial to their success in detecting novel malware variants. "Mal-D2GAN: Double-Detector based GAN for Malware Generation" (Thanh, Pham, & Bui, 2025), a recent work introduced Mal-D2GAN, a double-detector inspired GAN model that was specifically designed to enhance malware detectors' resilience against adversarial attacks. Their model outperformed existing GAN models, confirming the promise of GANs to generalize to novel and unseen malware variants. One of the limitations of this model, however, is that it may not generalize to very dynamic or adversarial environments, where attacks are designed to be stealthy.

Recent breakthroughs in deep learning, especially the incorporation of GANs into models such as GRUs and

transformers, have greatly enhanced malware detection from noisy API call sequences. These works highlight the significance of combining dynamic analysis, data augmentation, and strong model architectures for enhancing the precision and generalization of malware detectors. Yet, limitations such as the requirement of huge computational resources, overfitting possibility from limited datasets, and non-generalizability to different environments are still real issues to be addressed.

3. Methodology

ASTRID adopts a strict methodology flow to successfully detect malware from noisy API call sequences with the help of GAN power for effective detection. The methodology involves data preprocessing, adversarial training, and designing model architecture so that the model can distinguish between benign and malware sequences. In addition, the evaluation methodology checks the model's generalization by using an unseen dataset so that it can be applied in real-world environments. Figure 1 shows the flow diagram of the methodology used in ASTRID.

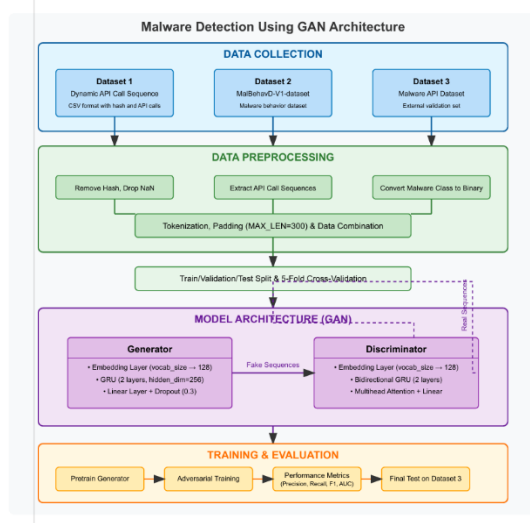


Figure 1: Flow Diagram of ASTRID

Dataset

In order to train and test the ASTRID model efficiently, three datasets are used, one for each role in the process. The first dataset, Dynamic API Call Sequence, consists of sequences of dynamic API call Sequences³ in CSV format, both with benign and malware data marked appropriately. This dataset has 1,079 benign and 42,797 malware samples and is utilized during the model training phase to offer useful real-system behavior sequences on which the model can be trained. The

MalBehavD-V1 dataset⁴, the second dataset in use, deals with malware behavior and has 1,285 benign and 1,285 malware sequences. This data set adds diversity to training the model, covering various malware and their actions in diverse situations, which is important for developing a strong model that can detect a broad variety of malicious actions. The third one, APIMDS Dataset⁵, is utilized solely for the model's generalization performance test. This external validation set, which comprises 3,137 benign, 5,878 unlabelled and 14,131 malware sequences, guarantees that the model's performance is measured on unseen data, reflecting its stability and capacity to act well in real-world situations.

With these three datasets, two are kept for training and the model gets to learn different patterns, while the third one is an unseen test set for evaluating the generalization of the model. This ensures that the model is not overfit to the training data and is able to identify malware in new, previously unseen patterns correctly.

Dataset Preprocessing

The preprocessing pipeline for the data starts by cleaning the Dynamic API Call Sequence dataset, in which the 'hash' column is deleted, and rows with missing values are deleted. The dataset is grouped by malware type to solve the issue of class imbalance, and a maximum of 4000 samples are taken from each class. Stratified K-Fold cross-validation is subsequently used to ensure even class balance across folds, and the dataset is divided into training and test sets according to the fold setup. To further address class imbalance in the training set, SMOTE (Synthetic Minority Over-sampling Technique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) is used to create synthetic samples for the minority class. Lastly, feature scaling is done with the StandardScaler to scale the training and test data to a consistent scale for all features. This preprocessing method prepares the dataset well for model training with balanced classes and scaled features, thus enhancing the model's generalization and performance.

Preprocessing for the MalBehavD-V1 Dataset begins with tokenizing sequences of API calls. In this case, API calls from different columns are combined into a single string per sample and tokenized with Keras's 'Tokenizer'. It converts the string sequences into a numeric form that is suitable for input into machine learning models. Sequences are padded to a constant length by 'pad_sequences' so that all input sequences are of the same length, with a maximum of 100. For target labels, One-Hot Encoding (Samuels, 2024) is employed so as to convert categorical labels into a binary matrix format and, therefore, convert them into an

³<https://www.kaggle.com/datasets/ang3loliveira/malware-analysis-datasets-api-call-sequences>

⁴ <https://github.com/mpasco/MalbehavD-V1>

⁵ <https://ocslab.hksecurity.net/apimds-dataset>

appropriate format for handling classification problems. Other than text processing, the numerical columns of the dataset are normalized using Min-Max Scaling (Muhammad Ali, 2022) such that all the features are within the range [0, 1]. This preprocessing technique is important in order to normalize the input features, especially for feature scaling-sensitive models. The dataset is then split into features ('X') and target labels ('y') after preprocessing. This preprocessing pipeline renders the MalBehavD-V1 dataset machine learning model-ready with properly tokenized, encoded, and scaled features to facilitate accurate training and testing.

The Malware API Dataset preprocessing pipeline starts from reading the data from a CSV file utilizing Python's native 'csv' module. The malware class is mapped to binary labels, such that "not-a-virus" samples are assigned a label 0 (benign), and the rest are assigned a label 1 (malicious). API call strings are concatenated into one, space-separated string, which is added to a list together with the label and SHA256 hash.

Min-Max scaling (Muhammad Ali, 2022) is used on any numeric columns so that all features are in the range [0, 1], which is especially helpful for scale-sensitive models. Next, the API call sequences are tokenized by Keras's 'Tokenizer' to transform each string sequence of API calls into a sequence of integers corresponding to the words (API calls). The sequences are subsequently padded to a consistent length of 300 with 'pad_sequences' so that the input sequences are all of the same length. The target labels are One-Hot Encoded (Samuels, 2024) by 'OneHotEncoder', converting the binary class labels into a binary matrix form appropriate for multi-class classification problems.

Dataset Preparation

Training

The dataset preparation starts with normalizing the binary labels of both datasets. The MalBehavD-V1 dataset and the Malware API Dataset are normalized as binary values for labels, where 0 denotes benign and 1 denotes malicious samples. In the MalBehavD-V1 dataset, API call sequences are formed by reducing all columns (except index, malware, and fold) into one string per sample. The Malware API Dataset also follows the same procedure, wherein sequences of calls to APIs are formed by concatenating strings of individual API calls. Then, sequences of both datasets are tokenized with the help of Keras's Tokenizer, and padding is done so that both datasets have similar sequence lengths. The training set is formed by concatenating both datasets' sequences as well as their labels. Moreover, the test set of the Malware API Dataset is developed by dividing it into training and test sets. These sets also contains noise to mimic the real world scenario to facilitate better generalizability of the

model. The final combined dataset (X_combined and y_combined) is divided into training, validation, and test subsets, with the training set being the primary input to model training. Instances of DataLoader are developed for facilitating efficient batching of data such that training, validation, and testing are performance-tuned. The method utilizes both datasets to their maximum, such that model training is carried out on a sufficient blend of properly balanced data. This process is shown in Algorithm 1.

Testing

For APIMDS as Dataset 3, which is held out entirely for testing, the sequences of API calls are handled in much the same way as the training datasets. The sequences are first tokenized using the same tokenizer used on the training data so that the datasets are consistent. These tokenized sequences are then padded to a standard length to ensure the input length of the model. Following tokenization and padding, the target labels are one-hot encoded to make the data classification-ready.

The processed Dataset 3 is reserved independently of the training data to evaluate the model's capacity for generalizing over unseen examples. The dataset is only employed for testing the performance of the trained model and verifying that the outcomes accurately express its applicability in the real world. By reserving the use of Dataset 3 only for testing, we are certain that the measures of evaluation (e.g., accuracy, precision, recall, and F1-score) come from data that the model has never seen while training, thus reflecting a true evaluation of its ability to generalize.

Experimental Setup

The experimental configuration for model training consisted of the use of a Generative Adversarial Network (GAN) framework. Training was done under the environment of Google Colab with L4 TPUs support, thus providing the calculations with a great boost and enabling the processing of large data in an efficient manner. The configuration played an important role in handling the large process of model training, especially through the use of high-dimensional data such as sequences of API calls.

Hyperparameters used in the current experiment were set to balance model performance against computational cost. The embedding size was 128, the recurrent layer's hidden size was 256, and the batch size was 64. The model was trained for 20 iterations with the learning rate held constant at 0.0004 to facilitate slow convergence. For additional improvement in the performance of the model, a Dropout regularization has been used at a rate of 0.3 both in the Generator and in the Discriminator to avoid overfitting and enhance the generalization of the model.

Cross-validation was conducted with Stratified K-Fold (5 folds) to ensure that the performance of the model would be estimated by using various splits in the data. This enabled the estimation of the model's robustness and generalizability across various folds with stable and solid estimates of performance.

Model Training

The architecture used in this experiment is a Generative Adversarial Network (GAN) architecture that consists of two neural networks: the Generator and the Discriminator. The Generator tries to generate artificial sequences simulating real API calls, while the Discriminator attempts to distinguish between real API call sequences from the dataset and the artificially created ones by the Generator. The adversarial process between the two components allows the model to progressively improve in the sense that it can increasingly discriminate between benign and bad sequences.

The Generator starts by taking random noise as input and reconstructing it as a sequence of API calls using Gumbel-Softmax Sampling (Li, et al., 2020) as shown in eq (1). It first inserts the input noise into a continuous space through an embedding layer, and then feeds this embedded representation into a GRU (Gated Recurrent Unit) layer. GRUs are particularly well-suited for sequence data since they help in capturing time step dependencies, which is crucial to analyze API calls occurring in sequences. The output from the GRU layer then undergoes a fully connected layer to produce the final sequence, which is designed to be similar to actual API call data.

$$y = \text{Softmax}\left(\frac{\log(\hat{y}) + g}{\tau}\right) \quad \text{eq (1)}$$

Conversely, the Discriminator takes sequences of API calls, real and fake, and returns a probability indicating whether or not the sequence is real or fake. Like the Generator, the Discriminator employs an embedding layer for representing the input sequence as vectors. These sequences are subsequently processed through a bidirectional GRU, which reads the sequence both forward and in reverse to capture temporal dependencies more exhaustively. The Discriminator uses multihead attention as well, allowing it to pay attention to different areas of the sequence at different times, improving its capacity for recognizing complex patterns typical of malware activity.

The adversarial training process between the Generator and Discriminator is where the model learns at its core. The Generator tries to generate more realistic fake sequences, and the Discriminator tries to get better at discriminating

Algorithm 1: Dataset Preparation for Training

Input: Dataset 1, Dataset 2

Output: Training, Validation, and Test DataLoader

Normalize binary labels:

Label the first dataset as 0 for benign and 1 for malicious samples

Label the second dataset as 0 for benign and 1 for malicious samples

Build API call sequences:

Join all API call features for each sample in the first dataset

Join all API call features for each sample in the second dataset

Tokenization and Padding:

Initialize Tokenizer for text sequences

Fit tokenizer on both datasets' API call sequences

- Convert API call sequences from first dataset into token sequences

- Convert API call sequences from second dataset into token sequences

- Apply padding to both token sequences to a fixed length

- Prepare and Clean the second dataset:

- Join API call features for each sample in the second dataset

- Tokenize and pad the API call sequences in the second dataset

- Split dataset into training and test:

- Perform a train-test split for the second dataset, creating training and test sets

- Combine datasets:

- Concatenate tokenized and padded sequences from both datasets

- Concatenate corresponding labels from both datasets

- Split combined data into training, validation, and test sets:

- Split the combined data into training and temporary sets

- Split the temporary data into validation and test sets

- Create DataLoader for training, validation, and test:

- Create DataLoader for the training set

- Create DataLoader for the validation set

- Create DataLoader for the test set

- Return: Training DataLoader, Validation DataLoader, Test DataLoader

real sequences from fake ones. This aligns with the Wasserstein GAN with Gradient Penalty (WGAN-GP) (Fan, et al., 2022) framework as shown in eq(2), which stabilizes training and prevents problems like mode collapse, where the Generator produces few or repeating sequences.

$$\mathcal{L}_{\mathcal{D}} = E_{\hat{x} \sim P_{\hat{x}}} [D(\hat{x})] - E_{x \sim P_x} [D(x)] + \lambda \cdot E_{\hat{x} \sim P_{\hat{x}}} [(|\nabla_{\hat{x}} D(\hat{x})|_2 - 1)^2] \quad \text{eq (2)}$$

Both the Discriminator and Generator use the Adam optimizer, 0.0004 learning rate, and the model is trained for a total of 20 epochs. The two networks both use regularization techniques like Dropout to prevent overfitting and improve

Algorithm 2: Model Training Procedure

Input: Training Data (X_train), Training Labels (y_train)**Output:** Trained generator and discriminator

1. Initialize Generator and Discriminator models
 2. Initialize Optimizers (Adam) for both Generator and Discriminator
 3. Initialize Cross-Entropy Loss Function for the Generator and MSE Loss for the Discriminator
 4. For each epoch:
 5. for batch in train loader do
 6. Train Discriminator:
 7. Sample real sequences from X train and generate fake sequences from Generator
 8. Compute the real loss and fake loss using the Discriminator
 9. Compute the Gradient Penalty loss and backpropagate to update Discriminator
 10. Update Discriminator parameters with Adam optimizer
 11. Train Generator:
 12. Generate fake sequences using noise input
 13. Compute Discriminator output for fake sequences
 14. Calculate the Generator loss and feature matching loss
 15. Backpropagate and update Generator parameters with Adam optimizer
 16. end for
 17. Return: Trained Generator and Discriminator
-

the model's generality. Gradient accumulation is also used to further stabilize the training so that the model is able to handle higher batch sizes even when under memory constraints.

Model performance is gauged by some key metrics like precision, recall, F1 score, AUC (Area Under the Curve), and accuracy. The AUC metric, in particular, is important in deciding how the model distinguishes between classes at every decision boundary. These metrics were tracked during training and cross-validation to see the model's improvement over time. This process is shown in Algorithm 2.

Training Process

Training is performed through an adversarial interaction between the Discriminator and the Generator. The Discriminator is first trained to effectively classify real and fake sequences. It is given real API sequences from data and fake sequences generated by the Generator. The Discriminator is trained such that it maximizes effective classification of these sequences, with binary cross-entropy loss as its objective function. The Generator, on the other hand, is trained to generate fake sequences that can deceive the Discriminator into marking them as real. The Generator's loss is calculated on the output of the Discriminator, as well as an additional feature matching loss, as shown in eq(3), that

encourages the Generator to create sequences with the same statistical properties as the real data.

$$\mathcal{L}_G = E_{x \sim P_x} [|F_x - F_{\hat{x}}|^2] \quad \text{eq (3)}$$

In training, the WGAN-GP loss function is applied to the Discriminator to apply a Gradient Penalty as shown in eq(4), ensuring that the training gradients are smooth and improving the optimization stability. This is extremely crucial in adversarial cases where the Generator and Discriminator engage in an endless game of equipping one another.

$$\mathcal{L}_{GP} = E_{\hat{x} \sim P_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \quad \text{eq (4)}$$

As the model was being trained, accuracy and the AUC curve were monitored across epochs. The first and fifth fold AUC curves are shown in Figures 2 and 3, respectively, and they illustrate how the model learned through training to distinguish between real and fake sequences. The training accuracy plot also shows steady improvement over time, showing that the model was learning and improving at distinguishing malware with accuracy.

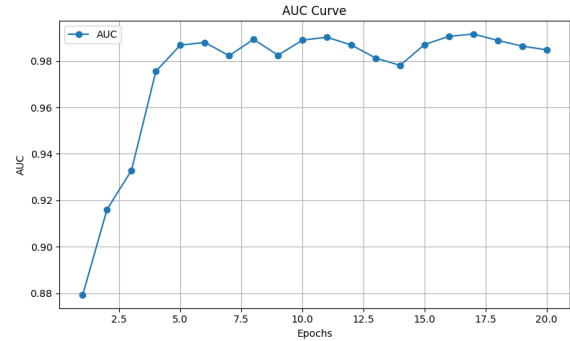


Figure 2: AUC Training Curve in 1st Fold

Testing

For the test phase, Dataset 3, saved as an unseen test set, was utilized to test the performance of the model after training using Datasets 1 and 2. Sequences from Dataset 3 were tokenized and padded to have uniformity with the training input format. The preprocessed data was then transformed into PyTorch tensors for ease of utilization during model testing.

A DataLoader for Dataset 3 was implemented to batch the data and feed it into the model in the testing phase. The batch size for testing was the same as that used in training to ensure equal comparison. This DataLoader served to feed batches of API call sequence and their labels through the trained Discriminator for evaluation.

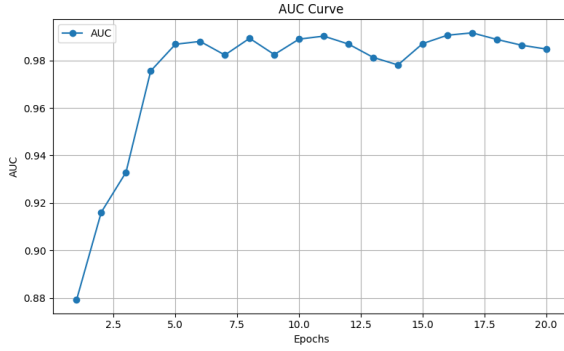


Figure 3: AUC Training Curve in 5th Fold

The model's performance on Dataset 3 was checked using critical metrics such as precision, recall, F1 score, AUC (Area Under the Curve), and accuracy. The choice of these metrics was to give an all-around view of the model's capacity to classify malware correctly. The AUC score specifically calculates the model's capacity to differentiate malicious and harmless samples for all possible thresholds, giving a better evaluation of its performance.

Observations

To compare and assess the performance of various models, the training of BiLSTM (Figure 4), Transformer (Figure 5), and Vanilla GAN + GRU (Figure 6) models with the same setup as ASTRID (Figure 1) is done. This involved the same preprocessing of the dataset, tokenization, padding, and training process to ensure that the comparison would be fair across all models. With the same framework and hyperparameters, it can be more equally compared how well every model does in malware detection via API call sequences.

The Transformer model performs best with a high AUC of approximately 0.98, and has smooth learning throughout the epochs. Comparatively, the BiLSTM and Vanilla GAN + GRU models experience some fluctuation in their AUC curves, with BiLSTM peaking at an AUC of approximately 0.822 and Vanilla GAN + GRU settling at 0.818.

ASTRID also has a very good performance comparable to the Transformer model with an AUC of 0.98 and a consistent rise during the training. This shows that the usage WGAN-GP architecture was successful in identifying malware as well as having high generalization capability to unknown data and thus can compare to other state-of-the-art models.

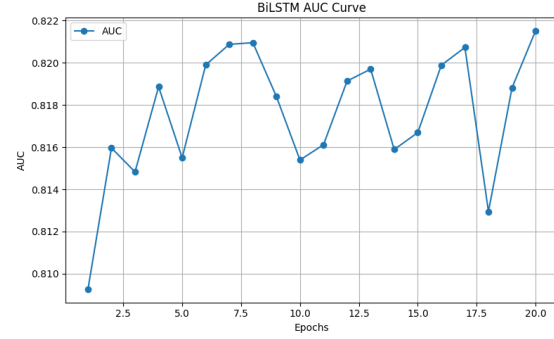


Figure 4: BiLSTM AUC Training Curve

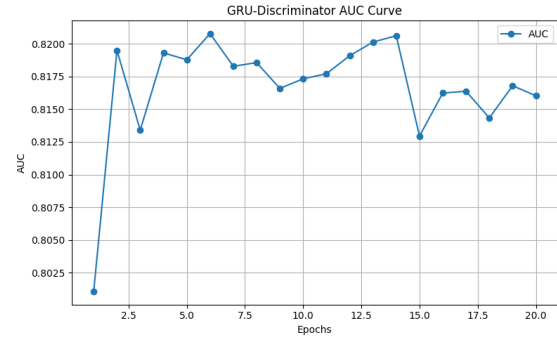


Figure 5: Vanilla GAN + GRU AUC Training Curve

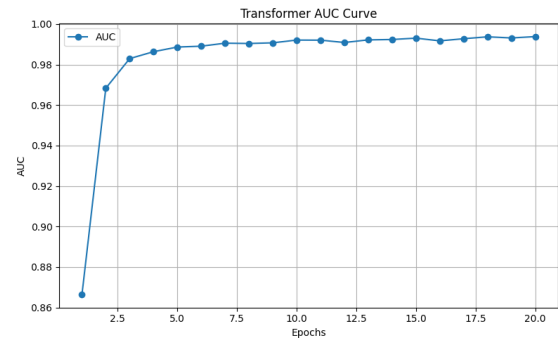


Figure 6: Transformer AUC Training Curve

4. Results

Evaluation Criteria

The models have been evaluated based on the following criteria:

1. **Accuracy:** It is the proportion of correct predictions over the total number of instances evaluated (Hossin & M.N, 2015). It can also be said as the ratio of correct classifications to the total classifications, as shown in eq (5).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Samples}} \quad \text{eq (5)}$$

2. **Precision:** It is the proportion of correctly classified instances among all the classified instances under a certain category (Pinto, Gonalo Oliveira, & Alves, 2016). It can also be said as the ratio of true positives to that of everything classified as positive, as shown in eq (6).

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{eq (6)}$$

3. **Recall:** It is the proportion of correctly classified instances under a certain category (Pinto, Gonalo Oliveira, & Alves, 2016). It can also be said as the ratio of true positives to correct classifications. It is also known as the true positive rate or TPR, as shown in eq (7).

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{eq (7)}$$

4. **F1 Score:** The harmonic mean of precision and recall, balancing their trade-offs.

Comparative Analysis of Models Trained

Table 1 shows the comparative analysis of the Models Trained on the dataset and the methodology of ASTRID.

Table 1: Training Data Comparison

Model	AUC	Accuracy	F1 Score
Transformers	0.9938	0.8309	0.8994
BiLSTM	0.8215	0.8007	0.8836
Vanilla GAN + GRU	0.8160	0.8040	0.8853
ASTRID	0.9878	0.9699	0.9699

Table 2 shows the testing data results

Table 2: Testing Data Comparison

Model	AUC	Accuracy	F1 Score
Transformers	0.9595	0.8646	0.9274
BiLSTM	0.5356	0.8644	0.9273
Vanilla GAN + GRU	0.5000	0.8644	0.9273
ASTRID	0.9508	0.9553	0.9541

Overall, the performance data in Tables 1 and 2 demonstrate the relatively higher performance of the ASTRID model than the comparative state-of-the-art models during the training and testing phases. From Table 1, we see that ASTRID outperformed not just the Transformers, but also the BiLSTM and Vanilla GAN + GRU models, indicating that the

ASTRID model exhibited superior performance in the training stage, from learning and generalizing on the training set.

In Table 2, when testing the models on the test dataset, we observe that ASTRID again performed well on the other models, having the highest precision and recall. As evident from the data and inspection of the Tables, ASTRID has not only attained good performance during the training phase, but indeed was able to generalize during the testing phase on the test data, making this model the best at detecting malware in this specific case. Meanwhile, even though the Transformer model performed remarkably (however, no performance data were presented to surpass ASTRID), the BiLSTM and Vanilla GAN + GRU models performed much less well (compared to ASTRID), with larger performance variation observed primarily in terms of AUC and accuracy.

Comparative Analysis with State-of-the-Art Models

Table 3 Shows the comparison between ASTRID and other State-of-the-Art Models

ASTRID not only matches or slightly surpasses the highest classification accuracy obtained by existing models (e.g., Transformer), it also achieves better AUC (Area Under the ROC Curve) whereas existing models obtain either a similar AUC or worse, demonstrating that ASTRID achieves greater discriminative ability across models (e.g., essential predicting classes) when the classification task is revised to adjust threshold level used (e.g., 0, 0.3, 0.5, 0.8). In summary, ASTRID is able to show greater discriminative ability because it trains adversarial synthetic sequences that incorporate this perturbation to induce better robustness and generalization into the model, which, to the best of our knowledge, has been absent in existing literature. Models such as Random Forest will not only perform poorly with adversarially perturbed instances but even more sophisticated architectures (e.g., CNN-LSTM) that are not yet considered models capable of addressing or tolerating adversarially perturbed inputs. Therefore, ASTRID explicitly trains on adversarial perturbation, thus improving resilience and performance.

5. Conclusion and Future Work

The ASTRID model shows great potential for detecting malware through API call sequence data. The use of a Generative Adversarial Network (GAN) approach allows ASTRID to overcome the challenge of noisy and adversarial perturbed data, producing strong results in both training and testing. Inserting adversarial synthetic sequences into the training set develops a generalisation capacity in models; however, ASTRID not only develops a generalisation capacity, but also develops robustness by training explicit noisy inputs that traditional models do not explicitly train on.

Table 3: Comparative Analysis of ASTRID with State-of-the-Art Methods

Model	Architecture	Accuracy	Precision	Recall	F1 Score	AUC	Notes
(Dixit & Singh, 2023)	Random Forest on API counts	98%	0.88	0.89	0.88	0.90	Good on static features, weak on unseen patterns
(Maniriho, Mahmood, & Chowdhury, 2023)	1D CNN + LSTM hybrid on API call sequences	93%	0.90	0.91	0.90	0.92	Local + sequential feature capture
(Owoh N. , et al., 2024)	GRU + GAN adversarial generation	98.9%	0.98	0.99	0.98	0.99	Extremely high accuracy, but may overfit on similar distribution
(Li & Zheng, 2021)	GRU + Attention on API call sequences	94%	0.91	0.92	0.91	0.93	Improved long-sequence understanding
(Kunwar, 2024)	Transformer encoder on API sequences	95%	0.92	0.94	0.93	0.95	Strong global attention, high memory needs
ASTRID	GAN-style Generator + Discriminator + GRU + Multihead Attention + Feature Matching	95%	0.95	0.95	0.95	0.95	Higher generalization against unseen dynamic malware

The model performed with a slight drop in AUC to almost 0.98 with an accuracy of 95.53%. ASTRID reduces false positives, missing malware, and falsely identifying benign by being more robust to noisy adversarial conditions, showing a stronger performance than other state-of-the-art models, such as BiLSTM, Vanilla GAN + GRU, and Transformer, because it exhibits higher values for precision and recall in a real-world malware detection task.

Future Work

There are quite a few areas where ASTRID can improve, notwithstanding the success of ASTRID. Future work could investigate the incorporation of more effective attention mechanisms and transformer-based architectures to further optimize sequence representation learning. The generalizability of the model to other platforms and environments could be evaluated through cross-platform datasets. Given the nature of malware and its increasing complexity, eventually, ASTRID could evolve to provide real-time detection as well as incorporate more challenging and varied adversarially perturbed datasets to enhance its

robustness. Ultimately, being able to work with real-time monitoring systems would improve ASTRID's scalability and use in operational settings.

References

Azeem, M., Khan, D., Iftikhar, S., Bawazeer, S., & Alzahrani, M. (2024). Analyzing and comparing the effectiveness of malware detection: A study of machine learning approaches. *Heliyon*, 10, e23574. doi:<https://doi.org/10.1016/j.heliyon.2023.e23574>

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002, June). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR)*, 16, 321-357. doi:10.1613/jair.953

Dixit, A., & Singh, S. (2023, July). Malware Detection Using Random Forest., (pp. 1-6). doi:10.1109/ICCCNT56998.2023.10306628

Fan, J., Yuan, X., Miao, Z., Sun, Z., Mei, X., & Zhou, F. (2022, January). Full Attention Wasserstein GAN With Gradient Normalization for Fault Diagnosis Under Imbalanced Data. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-1. doi:10.1109/TIM.2022.3190525

- Ge, Q., Yarom, Y., Li, F., & Heiser, G. (2017). Your Processor Leaks Information - and There's Nothing You Can Do About It. *Your Processor Leaks Information - and There's Nothing You Can Do About It*. Retrieved from <https://arxiv.org/abs/1612.04474>
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *Generative Adversarial Networks*. Retrieved from <https://arxiv.org/abs/1406.2661>
- Guri, M., & Bykhovsky, D. (2019). aIR-Jumper: Covert air-gap exfiltration/infiltration via security cameras & infrared (IR). *Computers & Security*, 82, 15-29. doi:<https://doi.org/10.1016/j.cose.2018.11.004>
- Hossin, M., & M.N, S. (2015, March). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5, 1-11. doi:10.5121/ijdkp.2015.5201
- Hu, X., Zhang, Y., & Li, J. (2025). Generative Adversarial Networks in Anomaly Detection and Malware Detection: A Comprehensive Survey. Retrieved from https://www.researchgate.net/publication/383575896_Generative_Adversarial_Networks_in_Anomaly_Detection_and_Malware_Detection_A_Comprehensive_Survey
- Kunwar, P. (2024). PhD Forum: MalFormer001- Multimodal Transformer Fused Attention based Malware Detector. *2024 IEEE International Conference on Smart Computing (SMARTCOMP)*, (pp. 252-253). doi:10.1109/SMARTCOMP61445.2024.00059
- Li, C., & Zheng, J. (2021, May). API Call-Based Malware Classification Using Recurrent Neural Networks. *Journal of Cyber Security and Mobility*. doi:10.13052/jcsm2245-1439.1036
- Li, C., Lv, Q., Li, N., Wang, Y., Sun, D., & Qiao, Y. (2022). A Novel Deep Framework for Dynamic Malware Detection Based on API Sequence Intrinsic Features. *Computers & Security*, 116, 102686. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0167404822006866>
- Li, Y., Liu, J., Guozheng, L., Hou, Y., Muyun, M., & Zhang, J. (2020, April). Gumbel-softmax-based Optimization: A Simple General Framework for Optimization Problems on Graphs. *Gumbel-softmax-based Optimization: A Simple General Framework for Optimization Problems on Graphs*. doi:10.21203/rs.3.rs-22822/v2
- Maniriho, P., Mahmood, A. N., & Chowdhury, M. J. (2023). API-MalDetect: Automated malware detection framework for windows based on API calls and deep learning techniques. *Journal of Network and Computer Applications*, 218, 103704. doi:<https://doi.org/10.1016/j.jnca.2023.103704>
- Muhammad Ali, P. (2022, June). Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, 10, 85-91. doi:10.14500/aro.10955
- Nguyen, H., Di Troia, F., Ishigaki, G., & Stamp, M. (2022). Generative Adversarial Networks and Image-Based Malware Classification. Retrieved from <https://arxiv.org/abs/2207.00421>
- Noguchi, A., Sun, X., Lin, S., & Harada, T. (2022). Unsupervised Learning of Efficient Geometry-Aware Neural Articulated Representations. *Unsupervised Learning of Efficient Geometry-Aware Neural Articulated Representations*. Retrieved from <https://arxiv.org/abs/2204.08839>
- Owoh, N., Adejoh, J., Hosseinzadeh, S., Ashawa, M., Osamor, J., & Qureshi, A. (2024). Malware Detection Based on API Call Sequence Analysis: A Gated Recurrent Unit-Generative Adversarial Network Model Approach. *Future Internet*, 16. doi:10.3390/fi16100369
- Owoh, N., Adejoh, J., Hosseinzadeh, S., Ashawa, M., Osamor, J., & Qureshi, A. (2024). Malware Detection Based on API Call Sequence Analysis: A Gated Recurrent Unit-Generative Adversarial Network Model Approach. *Future Internet*, 16, 369. Retrieved from <https://www.mdpi.com/1999-5903/16/10/369>
- Pinto, A., Gonalo Oliveira, H., & Alves, A. (2016, June). Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text. 51, 3:1-. doi:10.4230/OASICS.SLATE.2016.3
- Samuels, J. (2024, January). One-Hot Encoding and Two-Hot Encoding: An Introduction. *One-Hot Encoding and Two-Hot Encoding: An Introduction*. doi:10.13140/RG.2.2.21459.76327
- Stalin, K., & Mekoya, M. B. (2024). Improving Android Malware Detection Through Data Augmentation Using Wasserstein Generative Adversarial Networks. Retrieved from <https://arxiv.org/abs/2403.00890>
- Thanh, N. H., Pham, T. D., & Bui, L. (2025). Mal-D2GAN: Double-Detector Based GAN for Malware Generation. Retrieved from <https://arxiv.org/abs/2505.18806>
- Zhang, Y., Li, X., & Wang, Z. (2023). Malware Detection with Dynamic Evolving Graph Convolutional Networks. Retrieved from https://www.researchgate.net/publication/359581210_Malware_detection_with_dynamic_evolving_graph_convolutional_networks
- Zhou, Y., & Jiang, X. (2012). Dissecting Android Malware: Characterization and Evolution. *2012 IEEE Symposium on Security and Privacy*, (pp. 95-109). doi:10.1109/SP.2012.16

Strategic Cybersecurity Governance: A Risk-Based Policy Framework for IT Protection and Compliance.

Mohammed Nayeem
Trine University, United States
nm2751478@gmail.com

Abstract

As cyber threats evolve, organizations must adopt a proactive, risk-based cybersecurity framework to safeguard digital assets and ensure regulatory compliance. This paper presents a comprehensive IT security policy, integrating incident management, data protection, and access control strategies to mitigate cyber risks. The framework emphasizes information classification, identity governance, and adaptive security enforcement to enhance resilience against emerging threats. Additionally, it outlines best practices for cyber risk assessment, policy enforcement, and compliance with global standards such as GDPR and PCI-DSS. This approach provides a structured methodology for securing IT infrastructure while balancing business continuity and security governance.

1. Introduction

The increasing complexity of the digital ecosystem has elevated the importance of structured cybersecurity policies to counter evolving threats. Enterprises must now embrace adaptive, risk-informed approaches to defend essential assets and align with global compliance mandates. This study outlines a unified security strategy encompassing core pillars such as breach response, data governance, and permission control. Emphasis is placed on categorizing information, managing digital identities, and enforcing dynamic security protocols to strengthen organizational resilience. Additionally, the document reviews regulatory alignment with frameworks like GDPR and PCI-DSS, presenting an equilibrium between defense mechanisms and seamless operational function.

2. Cybersecurity and its Governance

A. Strategic Directive

COMPANY adheres to verified methodologies and defined cybersecurity ambitions that reinforce operational reliability and ensure alignment with enterprise targets.

B. Objective

- 1) This document delineates digital security standards designed to uphold the core mission of COMPANY: TEMP MISSION STATEMENT.
- 2) It contributes to market competitiveness by lowering security-related liabilities, improving revenue assurance, and mitigating reputational and financial exposure.

- 3) This policy emphasizes stewardship of digital infrastructure, personnel, and information. It establishes foundational guidelines that build credibility, strengthen partnerships, and outline the structured processes necessary for mitigating digital risk while sustaining strategic growth.

C. Coverage

- 1) This policy is applicable to all employees, contractors, third parties, technologies, systems, communication channels, and datasets associated with or governed by COMPANY.
- 2) Applicability is subject to prevailing legal stipulations and contractual agreements.

D. Compliance Obligations

Noncompliance with these guidelines may result in corrective measures, including but not limited to employment termination, contract suspension, or other applicable consequences.

E. Responsibility Allocation

- 1) Personnel are expected to manage access privileges, report system alterations, and highlight vulnerabilities to ensure holistic protection.
- 2) Executive leadership endorses cybersecurity priorities and aligns them with organizational directives.
- 3) The chief information security authority oversees threat mitigation, regulatory adherence, and the development of subordinate security policies, sometimes in collaboration with external experts.
- 4) Legal advisors validate data handling and contingency strategies against legal criteria and contribute during security incidents and stakeholder communication.
- 5) The internal audit division maintains oversight of high-level policies and ensures uniform implementation across all departments.
- 6) Technical staff facilitate systems administration, manage access controls, and execute approved configuration changes across organizational infrastructure.
- 7) Facility managers supervise on-site security, coordinate incident responses, and communicate with cybersecurity personnel when needed.

- 8) Privileged users possess elevated permissions and must exercise heightened diligence to preserve system integrity.
- 9) Authorized third parties are granted access after signing appropriate agreements (e.g., NDAs, SLAs, BAAs) and must adhere to the same data protection standards.
- 10) For detailed task ownership and communication flow, refer to LINK TO RACI DOCS, which include RACI matrices (Responsible, Accountable, Consulted, In- formed).

F. Terminology

Asset — Refers to intellectual property, technological infrastructure, physical or virtual resources, software, platforms, or any device (including personal ones, if applicable) accessing or interacting with COMPANY’s digital ecosystem.

G. Framework Directives

- 1) The cybersecurity directive is governed by the Internal Audit authority and forms the basis for subsidiary protocols enumerated in the attached appendices.
- 2) The policy is designed to not only satisfy compliance requirements (e.g., GDPR, PCI-DSS) but also to ensure sustainable service delivery and proactive risk oversight.
- 3) Security mechanisms include but are not limited to:
 - a) Defining baseline controls
 - b) Identifying and categorizing threats
 - c) Conducting employee awareness initiatives
 - d) Managing credentials and access rights
 - e) Enforcing change governance protocols
 - f) Applying least-access principles in a zero-trust model
 - g) Prioritizing cloud-native application protection
 - h) Evaluating third-party exposure
 - i) Encrypting communication channels and digital assets
 - j) Implementing backup communication systems
 - k) Handling data lifecycle: classification, storage, and elimination
 - l) Performing periodic audits and penetration evaluations
 - m) Maintaining assets, vulnerability, and patch logs
 - n) Executing predefined threat mitigation workflows
 - o) Monitoring event logs and retaining forensic records
 - p) The overarching strategy is: “Diminish vulnerability exposure, obstruct

unauthorized entry, and swiftly identify and contain breaches.”

H. Policy Exceptions

- 1) Any divergences from established protocols must be formally submitted for review and logged by the designated cybersecurity authority and IT department.
- 2) An official log of sanctioned deviations, along with their associated risk levels, is maintained at: LINK TO EXCEPTIONS REGISTRY
- 3) Note: Access permissions for the exception archive are role-specific and require justified necessity. The link may be inaccessible without these prerequisites.

I. Governance and Retention

1) Managed Documents

- a) This IT and cybersecurity framework, including all embedded or referenced guidelines and organizational protocols, is versioned to track amendments. Digital copies are preserved for 24 months post their last enforcement date unless regulated otherwise.
- b) After the designated period, obsolete files in physical or electronic form will be eliminated using secure destruction techniques unless exempted by a legal or regulatory requirement.
- c) This directive excludes system log files, which fall under a distinct archival policy.

2) Archival Records

- a) Records produced in adherence to the cybersecurity framework or linked policies shall be stored for a minimum of two years. Responsible units must conduct annual audits. All data will be encrypted during storage and transmission.
- b) For compliance purposes, documents subject to extended retention due to legal, financial, or statutory obligations (e.g., litigation materials, insurance claims, or fiscal records) will be preserved accordingly.

3) Circulation and Updates

- a) The cybersecurity policy remains internal and restricted. It is selectively shared with authorized personnel or external entities, as applicable. All modifications will be disseminated to eligible recipients.
- b) A digital archive of present and archival governance documents is available at: LINK TO DOCS AND POLICY

J. Supplementary Protocols (Appendices)

- A. Data Categorization, Handling, and Preservation Guide- lines
- B. Incident Response and Escalation Procedure
- C. Asset Oversight Framework
- D. Usage Compliance Regulations
- E. Change Authorization and Implementation Strategy
- F. Telecommuting Standards — Refer to Employee Manual
- G. Security Training Framework
- H. Continuity Operations Strategy
- I. Technical Protection Protocol
- J. Brand Risk Oversight
- K. Infrastructure Lifecycle Oversight

2. Information Classification, Handling, And Retention Policy

A. Overview and Purpose

- 1) COMPANY follows structured, validated methodologies to assign suitable classifications to information based on sensitivity and potential impact. This policy clarifies how data may be accessed, where it may be used, and the boundaries governing its distribution to safeguard against unauthorized disclosure.
- 2) By maintaining structured information governance, COMPANY ensures that business continuity objectives remain uninterrupted, and stakeholder interests are diligently protected.
- 3) Information is a fundamental corporate asset for COM- PANY. The nature and confidentiality requirements of information assets vary and are treated with corresponding protocols.

B. Scope

This directive pertains to the entities specified in the foundational Cybersecurity and IT Policy and concerns the classification and stewardship of organizational information assets.

C. Enforcement

Refer to the enforcement provisions detailed within the base Cybersecurity and IT Policy.

D. Roles and Responsibilities

Maintenance of this directive is assigned to the designated security executive (e.g., CISO), in cooperation with legal and departmental heads. Implementation across departments rests with the respective unit leaders.

E. Acronyms

There are no acronyms defined under this section.

F. Policy

Privacy alignment is inherited from the parent cybersecurity policy. Certain segments of this directive may be held under

a stricter classification and stored separately according to relevant access control standards.

G. Classification Categories

- 1) COMPANY organizes data into three tiers: Public, In- ternal, and Confidential.
 - a. Public: Data suitable for general release that does not jeopardize the firm's position or operations if disclosed.
 - i. Includes assets without confidentiality constraints, such as public reports, marketing publications, or corporate websites.
 - b. Internal: Data not meant for public circulation but not severely harmful if inadvertently released.
 - i. Examples include internal notices, administrative circulars, resource scheduling, or vendor invoices.
 - c. Confidential: Data vital to projects, departments, or corporate processes whose exposure may inflict operational, financial, or reputational harm.
 - i. Includes business strategies, sensitive agreements, proprietary technology, personal identifiable data, and anything with legal or regulatory safeguarding.
 - ii. Must be shared strictly on a need-to-know basis. If stored, data must be backed up and encrypted or password-protected; if transmitted, encryption is mandatory.
 - d. Confidential: Special (e.g., XYZ-specific): A specialized subclass requiring exceptional protection due to legal, regulatory, or client constraints. De- tails are maintained in separately secured documentation.

H. Secure Handling of Information Assets

- 1) Each asset must be labeled with its classification in headers, footers, and filenames; physical media must bear visible external tags.
- 2) Some data levels mandate pre-approval from senior leadership or legal entities before transmission.
- 3) Confidential materials must only be dispatched via authenticated, trusted courier services.
- 4) Printed confidential data must be physically secured in locked environments and shredded when obsolete.

- 5) Restrict exposure of classified materials to essential personnel through access controls and custodial chains.
- 6) Prior to printing, validate the requester's identity and encrypt print spools where feasible.
- 7) Limit dissemination exclusively to authorized individuals.
- 8) Periodic audits of user access must occur monthly or at major project milestones.
- 9) Declassification decisions reside with department heads in consultation with policy custodians and legal counsel.
- 10) Supplementary reference materials and control procedures: [LINK TO INFO INVENTORY](#)

I. Classification by Aggregation

- 1) Merging or transforming datasets into higher-value information may necessitate reclassification due to elevated sensitivity or derived strategic insights.
- 2) Aggregated datasets are classified according to the most sensitive element involved.
- 3) Combinations of identically classified assets may require elevation to the next sensitivity tier.

4. Incident Management Policy and Process

A. Overview and Purpose

The aim of this directive is to enhance information security posture by systematically detecting, assessing, and resolving security incidents, thereby maintaining a competitive operational advantage.

B. Scope

Applicable to all divisions and personnel identified within the Cybersecurity and IT Policy base documentation.

C. Enforcement

Enforcement protocols align with the overarching Cybersecurity and IT Policy.

D. Roles and Responsibilities

- 1) Security Executive or Designee:
 - a) Oversees execution and adherence to the Incident Response Framework.
 - b) Coordinates incident containment, acts as primary communication liaison, prepares internal and external briefings, and leads incident drills.
- 2) Facilities Coordinator: Ensures physical access control prevents disruption during incident management.
- 3) IT Lead: Maintains real-time updates to the security executive on potential breaches.

E. Acronyms

- 1) IRT - Incident Handling Unit
- 2) PACE Strategy - A methodology encompassing Primary, Alternate, Contingency, and Emergency

means for completing mission-critical tasks. It ensures communication pathways are predefined for immediate response actions during critical scenarios.

- 3) MTTR - Mean Time to Remediation
- 4) CAL OES - California Governor's Office of Emergency Services, which provided the foundational structure for this documentation.
- 5) SUNY Broome - Referenced as a publicly available source for incident management frameworks.

F. Response Protocol

1) IRT Composition

- a) Each organizational division will assign a continuously reachable contact point to support security-related occurrences.
- b) Relevant units may include:
 - i. Infrastructure and Security Oversight
 - ii. Finance
 - iii. Legal Affairs
 - iv. Operations
 - v. Digital Systems
 - vi. Cybersecurity Lead
 - vii. Public Engagement and Media
 - viii. Internal Review
 - ix. Regulatory and Risk Functions
 - x. Others, as necessitated
- c) In the absence of the designated representative, the department lead assumes responsibility until the contact role is reassigned.
- d) Comprehensive stakeholder contact data is located at: [LINK TO IRT STAKEHOLDERS](#)

2) IRT Alerting Mechanism

- a) The cybersecurity division will collaborate across departments to create a PACE-based notification model for all IRT affiliates, encompassing internal and third-party entities (e.g., external SOC, IR units, MSSPs).
- b) All involved personnel must be familiar with the alert structure and ready to implement any listed communication methods.
- c) Each contributing division must configure an alias, mailing list, or system that aligns with the communication hierarchy outlined in the PACE structure.
- d) The notification strategy can be accessed at: [LINK TO PACE PLAN](#)

3) IRT Communication Strategy

- a) According to the PACE design, a centralized command space will be activated to enable collaborative discussions.

- b) Physical command centers should accommodate group collaboration, contain multimedia tools, and host a workstation with high-bandwidth connectivity for remote inclusion.
- c) Virtual alternatives should support shared media, discussions, and segmented virtual rooms for sub- group coordination.
- d) Reference document: LINK TO PACE PLAN

4) External Liaison Records

- a) The cybersecurity office will keep contact records for the following:
 - Managed Service Providers (MSPs)
 - Security-as-a-Service vendors (MSSPs)
 - ISP contacts
 - Federal Bureau of Investigation (local division)
 - Digital Crimes Task Force (local jurisdiction)
 - Regional Cyber Response Teams
 - Hosting Providers
 - Additional relevant external actors

5) Event and Incident Differentiation

- a) Definitions are adapted from NIST SP 800-61 Revision 2:
- b) Event – A system/network activity such as web- page requests or firewall filtering.
- c) Negative Event – An activity with adverse effects like privilege misuse or crashes.
- d) Security Breach – Any breach or likely breach of digital policy, standards, or guidelines (e.g., DoS attacks or misuse of organizational assets).

6) Prioritization and Escalation

- a) Reported activities undergo preliminary assessment by the first authorized evaluator, who determines whether escalation is warranted. All judgments must be logged in the designated tracking system.

7) Logging and Assessment

- a) All incidents must be logged with classification and tracked per internal protocol using guidance from NIST SP 800-61 Rev. 2 and US-CERT.
- b) Key documentation includes:
 - Incident label and identifier

- Functional and data exposure impacts
- Time of detection and estimated occurrence
- Scope: systems, users, records affected
- Physical/network location data
- Contact references for further inquiry
- Origin mechanism or root exploitation vector
- Related detection or forensic indicators
- Containment/mitigation steps and client guidance

- c) Documentation supports retrospective analysis, adversary profiling, and enhancement of organizational resilience.

8) Incident Taxonomy

Table 1: Incident types and response time objectives

—p1.2cm—p2.5cm—X—p1.8cm—			
Code	Label	Definition	Target Resolution
CAT A	Emulated Events	manufactured for testing purposes without operational impact	Not Assigned
CAT B	Breach or Unauthorized Control Gained	physical/logical unauthorized access or deliberate compromise of data/systems. Includes active malware control.	Within 1–3 hours
CAT C	Disruption of Service	attacks that hinder system/service availability	3 Hours
CAT D	Malware Introduction	System infection from worms, trojans, or malicious payloads (excluding live control cases)	12 Hours
CAT E	Illegal System Activity	Criminal or unauthorized actions, including digital fraud or threats to safety, possibly involving law enforcement	24 Hours
CAT F	Policy Breach	Use violations contrary to established computing or network standards	24 Hours
CAT G	Probing or Reconnaissance	Scanning or enumeration attempts with no direct security impact	72 Hours
CAT H	Unverified Behavior	Unclassified anomalies under current examination	72 Hours

G. Impact Levels

Outlined below are the levels used to assess impact severity across operational capability.

Table 2: Business Functionality Impact Classification

p2.5cm X
Level Explanation
None No disruption to competitive operational continuity.
Low Slight operational disadvantage, with ongoing access to markets. Moderate Market performance becomes uncertain or questionable. Severe Complete inability to maintain competitive market presence.

H. Evidence Retention and Chain of Custody

- 1) All collected evidence must be tracked with a quantity and detailed description.
- 2) Custodians of evidence must:
 - Secure the evidence appropriately,
 - Limit access to authorized personnel only,
 - Report the evidence in their control along with transfer history,
 - Maintain thorough documentation.
- 3) If the evidence is a physical device, document:
 - Manufacturer
 - Model
 - Serial Number
- 4) If the evidence is a digital file, include:
 - Filename
 - File hash
 - File creation, modification, and access timestamps
 - File size
 - Any physical media used (e.g., USB) must be protected with write blockers
- 5) For testimonial or recorded statements, include:
 - Interviewee and interviewer names
 - Location and medium used (audio/video/text)
 - Collection date and time
 - Signatures of both parties
- 6) Testimonial records should only be collected upon legal guidance or requirement.
- 7) When evidence is transferred between parties, the chain of custody must be strictly maintained. Each transfer record should include:
 - Date and time of transfer
 - Name and department of both parties
 - Purpose of the custody change
 - Signatures of transferring and receiving parties

V. Asset management guidelines

A. Purpose and Overview

- 1) This policy outlines the procedures for managing IT and information security assets, aiming to support organizational success in competitive markets by establishing clear asset management protocols.

- 2) It defines the guiding principles for managing various types of assets, including hardware, software, information, and other essential resources.

B. Scope of Application

This policy is applicable to all assets listed in the Cybersecurity and IT Policy. It covers all information and technology re- sources that are owned, leased, or utilized by the organization, or those containing organizational data. This includes IT assets like hardware (e.g., endpoints, office devices, access points), software, IaaS, SaaS solutions (e.g., Google Workspace, AWS, GitHub), and other related platforms.

C. Policy Enforcement

Enforcement mechanisms are detailed in the Cybersecurity and IT Policy.

D. Roles and Responsibilities

- 1) The CISO or designated representative is tasked with overseeing the execution of this policy and ensuring the program's successful implementation.
- 2) The IT department is responsible for maintaining precise asset inventories and management tools, including asset deployment and ongoing operational support.
- 3) Departmental managers are accountable for maintaining accurate asset records and notifying the IT team and CISO of any updates or modifications.

E. Terminology and Definitions

Asset management: The structured process of identifying, tracking, maintaining, and decommissioning both physical and virtual assets within the organization.

F. Guidelines

- 1) Asset Inventory Management:
 - a) The IT department, in collaboration with the information security team, will maintain an up-to-date inventory of all organizational assets. This record will include, but not be limited to:
 - i. Device identifiers (name, serial number, IP address, MAC address)
 - ii. Operating system and firmware details
 - iii. Acquisition and end-of-life dates
 - iv. Support and warranty expiration dates
 - v. Service contract expiration dates
 - vi. Device classification, location, and department ownership
 - vii. Primary point of contact and device role description.
- 2) The asset inventory record is accessible at: LINK_TO_NETWORK_INVENTORY

- 3) The network topology diagram is available at:
LINK_TO_NETWORK_DIAGRAM

G. Software and Service Inventory

- 1) The IT team will catalog all software and services, detailing information such as:
 - a) Vendor name and contacts
 - b) Software type and version
 - c) Subscription details (account number, quantity, cost, renewal date)
 - d) End of support and service expiration dates
- 2) This inventory will encompass all software (including open-source and proprietary), services, domain registrations, and third-party hosted solutions.
- 3) Development teams are responsible for tracking software dependencies and packages in production and development environments. These inventories will support risk management through a software bill of materials (SBOM).
- 4) Software and service records can be found at:
LINK_TO_ASSET_INVENTORY

H. Critical Account Management

- 1) Privileged accounts, which are essential for access to sensitive systems or administrative functions, will be closely monitored and categorized as critical assets.
- 2) Critical accounts include root/system credentials, third-party access keys, and executive-level personal accounts.
- 3) Critical account inventories will include:
 - a) Account name and access details
 - b) Privilege levels and assigned personnel
 - c) Date of provisioning and audit reviews
- 4) Critical account information is stored at:
LINK_TO_ASSET_INVENTORY

I. Provisioning and Standardization (Gold Image)

- 1) Standardized “gold images” will be defined for all key assets.
- 2) Each server, container, and user workstation will be configured using approved gold images, ensuring vulnerability mitigation and configuration consistency. These images will be reviewed annually.
- 3) New assets will be provisioned using the organization’s gold images and verified prior to deployment in production or testing environments.
- 4) Gold image details are accessible here: LINK TO GOLD IMAGE LIST

J. Configuration and Patch Management

- 1) Configuration management is fundamental to patch management.
- 2) Maintaining up-to-date gold images and their subsequent deployments is essential for lifecycle management and vulnerability mitigation.
- 3) Patches will be deployed via configuration management systems, ensuring that updates are distributed efficiently across all assets, rather than performing individual system patches.

K. Asset Control Tools

- 1) The IT team, with support from information security and relevant external partners, will deploy monitoring tools for asset health and status.
- 2) These tools will track:
 - a) Asset status and location (at least on a logical network level)
 - b) Running state of assets and associated services
 - c) Associated accounts, services, and software
 - d) Endpoint protection and log collection for incident response
- 3) Documentation for asset management tools can be found at:
LINK_TO_ENDPOINT_TOOLS_DOCS

VI. Acceptable Use Policy

A. Overview and Intent

- 1) The objective of this policy is to outline the appropriate and inappropriate usage of data, information, network resources, and electronic devices within COMPANY, fostering a culture rooted in ethical practices, legality, transparency, trust, and integrity. This enables COMPANY to remain competitive in a fast-paced market.
- 2) COMPANY supplies computing devices, networks, and access to information systems to achieve organizational goals and serve its stakeholders. Management of these resources is crucial for safeguarding the confidentiality integrity, and availability of company assets. Authorized users are responsible for upholding the company’s assets and adhering to established policies.

B. Scope

This policy pertains to all entities and assets specified in the foundational Cybersecurity and IT Policy document.

C. Enforcement

Refer to the enforcement clause in the Cybersecurity and IT Policy.

D. Roles and Responsibilities

Roles and responsibilities outlined in the Cybersecurity and IT Policy are inherited by this document.

E. Definitions and Acronyms

- 1) Honeypot, Honeynet – Technical systems designed for early detection and monitoring of potential security breaches.
- 2) Spam – Unsolicited electronic communication, including junk emails or newsgroup postings.

F. Policy

1) General Obligations

- a) Proprietary information belonging to COMPANY remains its exclusive property, regardless of whether the devices storing this information are owned by the company, leased, or belong to third parties. Employees may access or share proprietary data only within the limits of their assigned duties.
- b) Users must exercise sound judgment in utilizing COMPANY resources, aligning with company policies, standards, and guidelines. Company resources should not be used for any unlawful or prohibited activities.
- c) Users are prohibited from taking any action that could undermine COMPANY's market position, confidentiality, data integrity, or asset availability without explicit authorization from the relevant authority within the company.
- d) For security and maintenance, COMPANY may monitor all infrastructure, user accounts, and systems, including network and endpoint scans. Users must not interfere with authorized auditing processes.

2) Account Security and Access

- a) Users are responsible for safeguarding all accounts, data, and systems they manage.
- b) All accounts must be protected by a secure username, password, and multi-factor authentication where applicable.
- c) Strong password guidelines must be followed:
 - i. Passwords must be at least 20 characters long.
 - ii. Passwords should include at least one uppercase letter, one numeral, and one special character.
- d) Passwords must not be stored unencrypted. They should be saved in COMPANY's approved password manager.
- e) Multi-factor authentication is required, with users being allowed to use tools like Google Authenticator, Authy, or Azure Authenticator for added security.
- f) SMS-based authentication is not permitted.

- g) Users must never share their credentials with anyone, including the IT team or external parties.
- h) Default passwords must be changed immediately after account setup.
- i) If an account or password is compromised, it must be changed immediately unless authorized otherwise by the security team.

3) Device Security

- a) Users are responsible for securing COMPANY devices, including utilizing cable locks and other security mechanisms. Devices left overnight at COMPANY must be secured by locking doors or cable locks.
- b) Any theft or loss of devices must be reported immediately to the Facilities Manager and the IT team.
- c) Devices must be locked automatically after 10 minutes of inactivity.

4) Use of COMPANY Assets

- a) COMPANY network resources must not be used for unauthorized activities, such as disclosing proprietary information, unauthorized access attempts, or causing disruptions in service.
- b) The introduction of unauthorized technology, data, or equipment to COMPANY networks is prohibited.

5) Access Termination

- a) Access to COMPANY assets will be revoked immediately upon termination of employment, contract, or agreement, except where legal, investigative, or contractual requirements apply.

6) Electronic Communication

- a) The misuse of company assets for illegal or unethical activities, including fraudulent communications or harassment, is strictly prohibited.

7) Training and Awareness

- a) All staff must complete annual security training and participate in any simulated exercises conducted by the IT or security teams.

8) Exceptions

- a) Any deviation from this policy must be approved by the CISO and documented by the IT Team.

VII. Change Management and Control Policy

A. Policy Overview

This policy ensures effective planning, documentation, response, and learning from modifications to organizational operations.

B. Objective

The objective of this policy is to enhance information security, thereby supporting the organization's competitiveness and market success.

C. Terminology

Change Management denotes the structured approach for modifying IT systems. Its primary aim is to improve organizational awareness and understanding of proposed changes, ensuring minimal negative impact on services and clients.

D. Coverage

- 1) This policy applies to all entities described in the Cybersecurity and IT Policy documentation.
- 2) Any alterations to IT services must follow a formalized procedure to ensure adequate planning and execution.

E. Procedure

- 1) **Planning:** Develop the change plan, including the design, schedule, communication strategy, testing protocol, and contingency plan.
- 2) **Evaluation:** Assess the change, determine its nature, identify impacted assets, evaluate outage risks, and select the appropriate change classification and process.
- 3) **Review:** Discuss the change plan with relevant stakeholders or the Change Advisory Board, depending on the change type.
- 4) **Approval:** Obtain necessary approval from management or designated authorities, depending on the change type. Communication about the change must be shared with the appropriate parties, particularly if it impacts users.
- 5) **Testing:** Test changes in a controlled environment, prepare fallback procedures, and document lessons learned. Runbooks should be created for user-impacting changes, aiding the help desk during implementation.
- 6) **Execution:** Execute the change, prepared for potential reversion if needed. For significant or user-impacting changes, a phased approach should be applied, testing with smaller user groups first.
- 7) **Documentation:** Record the change, including approval and review details. Documentation should include feedback from:
 - a) Departments or individuals affected by the change
 - b) Teams responsible for implementation
 - c) Key stakeholders

- d) Emergency response teams involved

Post-change reflection is critical for enhancing future changes, enabling faster and more efficient organizational transitions. This phase fosters an understanding of the system and prepares for improved future change planning.

F. Responsibility

Department heads manage the change process within their areas, coordinating with other departments. For changes requiring additional resources, consultation with Finance is necessary during the Approval stage.

G. Documentation Protocol

- 1) **Log Maintenance:** All changes, whether Normal or Emergency, and their evaluations, will be systematically documented to provide clarity regarding the modifications made, their rationale, and the procedure followed.
- 2) **Change Log:**
 - a) All changes (Standard, Normal, Emergency) will be recorded in the Change Log, including:
 - i. The individual or team executing the change
 - ii. Description of the change
 - iii. Justification for the change
 - iv. Date and time of the change
- 3) **Process Log:**
 - a) Changes with Medium or High impact, along with Emergency changes, will be documented in the Process Log, including:
 - i. Test plan and results
 - ii. Risk assessments
 - iii. Communication strategies
 - iv. Deployment plans with fallback contingencies

VIII. Security Awareness Policy and Program

A. Policy Statement

This policy aims to foster vigilance among stakeholders to identify and report potential or confirmed incidents.

B. Purpose

The organization's success is tied to safeguarding assets and information. Collective efforts to protect these resources are paramount.

C. Scope

- 1) This policy extends to all entities as described in the Cybersecurity and IT Policy framework.
- 2) All users of COMPANY and external partners with access to the organization's information or systems are required to undergo training.

D. Policy

- 1) The objective is to raise awareness about information risks and guide users in taking suitable actions based on their access privileges. The program encompasses:
 - a) Annual mandatory training sessions.
 - b) Periodic awareness surveys.
 - c) Unannounced evaluations to ensure compliance.
 - d) Feedback mechanisms to refine the training program.
 - e) Phishing and social engineering exercises.
 - f) Simulated penetration testing and physical security checks.
 - g) Specialized training for high-privilege users, including incident response and forensic analysis.
- 2) Training completion records and performance results will be stored in the individual's Human Resources file.

E. Responsibility

The Chief Information Security Officer (CISO) oversees the program, with contributions from Legal, HR, and the IT Committee.

IX. Technical Security Policy

A. Policy Statement

This policy is designed to establish a framework for technical security, guiding the strategy to reduce vulnerabilities and enhance defenses.

B. Purpose

The policy ensures that technical security efforts support the organization's competitive edge in the market.

C. Scope

The policy is applicable to entities defined in the Cybersecurity and IT Policy framework.

D. Roles

- 1) Roles follow the guidelines outlined in the Cybersecurity and IT Policy document, with exceptions made when necessary.
- 2) Data Loss Prevention Officer (DLPO): This role may either be added as an additional responsibility or independently assigned. The DLPO leads risk mitigation efforts to support the CISO and minimize risks within the partner network. Responsibilities include data identification, classification, and safeguarding.
- 3) IT Committee: Works with the CISO and DLPO to ensure technical security policies are relevant. The committee serves as the Change Control Review

Board, with a voice in decision-making, particularly regarding user spend.

E. Policy Overview

- 1) Continuous efforts will be made to minimize the attack surface and enhance security by implementing the NIST framework: Identify, Protect, Detect, Respond, and Recover.
- 2) Regular reviews will be conducted to comply with security requirements from partners or legislation, adopting prudent measures where formal requirements are absent.
- 3) Assuming compromise is an integral part of the security strategy, with efforts focused on answering critical incident-related questions, including containment, notification, and recovery.
- 4) **Identify:**
 - a) Continuous integration of security measures with IT and operations to understand and manage the attack surface. A regular inventory of assets and proactive identification of "Shadow IT" are key.
 - b) Security sensors will be deployed to monitor changes to the attack surface, with regular external scans to detect unapproved additions.
- 5) **Protect:**
 - a) While some systems may follow a shared responsibility model, prudent controls will be applied, especially for cloud-based services, with multifactor authentication and vendor risk assessments.
 - b) Endpoint security will involve preconfigured images, device registration, and multifactor authentication, along with behavioral analytics to identify suspicious activities.
- 6) **Detect:**
 - a) Focused efforts on identity, endpoint, and network monitoring will ensure early detection of anomalies.
 - b) A zero-trust identity model will be implemented, with extended integration across critical SAAS platforms to monitor access points and identify unauthorized activities.
 - c) Device health checks and intrusion detection systems will be deployed to detect threats at the endpoint level.
 - d) Network monitoring solutions will be applied to track data flows within internal networks and during remote access.
 - e) Log Aggregation Policy:
 - i. Logs will be retained for 13 months to ensure visibility, with precomputation to reduce unnecessary data transmission.

- ii. Aggregated logs will be used for correlation, forensic analysis, and incident response.
- 7) **Respond:** Refer to the Incident Management Policy for detailed response actions.
 - 8) **Recover:** Refer to the Business Continuity Policy for recovery strategies.

X. Conclusion

This study highlights the necessity of adopting a forward-thinking, risk-driven cybersecurity strategy to safeguard organizational assets and meet compliance requirements. By consolidating incident management, data security, and access control mechanisms, the proposed cybersecurity framework offers a resilient defense against emerging threats. Emphasizing data categorization, identity management, and flexible security practices enhances the system's ability to withstand evolving risks. Moreover, compliance with international standards such as GDPR and PCI-DSS strikes a balance between maintaining robust security and supporting operational continuity. This framework provides a systematic and scalable solution for fortifying IT infrastructure in a constantly changing security environment.

References

- [1] NIST, Framework for Improving Critical Infrastructure Cybersecurity, National Institute of Standards and Technology, Version 1.1, Apr. 2018. [Online]. Available: <https://www.nist.gov/cyberframework>
- [2] International Organization for Standardization, ISO/IEC 27001:2013 - Information technology – Security techniques – Information security management systems – Requirements, ISO, 2013.
- [3] Voigt, P., and von dem Bussche, A., The EU General Data Protection Regulation (GDPR): A Practical Guide, Springer, 2017.
- [4] PCI Security Standards Council, PCI DSS v4.0: Payment Card Industry Data Security Standard, March 2022. [Online]. Available: <https://www.pcisecuritystandards.org>
- [5] Bertino, E., and Takahashi, K., “Identity management: Concepts, technologies, and systems,” *Computers & Security*, vol. 31, no. 4, pp. 451–459, 2012.
- [6] Shedden, P., Smith, W., and Ahmad, A., “Information security risk assessment: Towards a business practice perspective,” *Journal of Information Security and Applications*, vol. 34, pp. 33–43, 2017.

Subverting Quantum Machine Learning via Hybrid Data Corruption.

Mohammed Nayeem
Trine University, United States
nm2751478@gmail.com

Abstract

The rapid ascent of quantum computing and its integration with machine learning introduces an entirely new frontier for cybersecurity research. This paper addresses the critical and cutting-edge security challenge of safeguarding quantum machine learning (QML) models against insidious data manipulation attacks. We present a novel cross-domain adversarial strategy that leverages an intrinsic understanding of quantum data representations to inject highly effective corruptions into QML training datasets. Unlike traditional methods, our approach demonstrates robust efficacy even in the presence of realistic quantum noise. Through rigorous experimental validation across diverse quantum architectures, we showcase the profound detrimental impact of this vulnerability on QML model performance, underscoring the urgent need for robust defenses in the nascent quantum computing landscape. This work provides foundational insights into securing the next generation of intelligent systems.

I. Introduction

The rapid evolution of quantum computing has transformed it from a theoretical curiosity into a disruptive paradigm with tangible applications across optimization, cryptography, and artificial intelligence [1], [2]. The current era, often referred to as the Noisy Intermediate-Scale Quantum (NISQ) era, is marked by limited qubit counts and noisy operations, yet still promises meaningful computational advantages when integrated with classical resources [3], [4].

One of the most prominent areas leveraging these developments is Quantum Machine Learning (QML), which seeks to combine quantum information processing with machine learning principles to solve problems beyond the capacity of classical methods [5], [6]. Models such as Quantum Neural Networks (QNNs) and variational algorithms provide flexible frameworks where quantum states encode classical data and parameterized quantum circuits (PQCs) perform transformations that may capture complex, high-dimensional patterns more efficiently than their classical counterparts [7], [8].

Despite their theoretical promise, QML models are not inherently secure. The fusion of quantum and machine learning paradigms introduces novel attack surfaces. Research in classical machine learning has already demonstrated the susceptibility of models to adversarial examples, data poisoning, and backdoor attacks [9], [10]. Analogous threats in QML, although less explored, are

poised to emerge as critical concerns given the growing reliance on cloud-based quantum platforms.

Quantum computing infrastructure further complicates the security landscape. Cloud-based access models, such as those provided by IBM Quantum and other commercial vendors, abstract away hardware control from users, limiting their ability to detect or mitigate adversarial manipulations [11]. In this context, data poisoning attacks, which involve manipulating training datasets to embed malicious behaviors, become particularly challenging to address due to limited observability of quantum states and the inherently probabilistic nature of quantum measurement.

Existing literature has emphasized adversarial queries and side-channel vulnerabilities in QNNs [12], [13], yet systematic exploration of training-time poisoning in QML remains nascent. Poisoning poses a distinct challenge because it directly targets the learning pipeline. Unlike evasion attacks, which exploit models after deployment, poisoning corrupts the foundation of model training, thereby undermining the reliability of downstream predictions.

The intrinsic noise of NISQ devices poses another obstacle to securing QML. Classical poisoning strategies, such as sample or label manipulations, may not transfer effectively into quantum systems because noise can either mask or amplify adversarial effects [14]. This interplay between noise and adversarial manipulation necessitates quantum-specific frameworks for analyzing poisoning resilience.

This work introduces QUID, a novel quantum poisoning strategy that leverages intra-class encoder state similarity (ESS) to determine adversarial label flips. ESS exploits the geometric properties of quantum state space, assigning poisoned labels that maximize inter-class dissimilarity at the density matrix level. Unlike random label flipping, QUID ensures targeted and systematic degradation of QNN performance, even under realistic noise conditions.

The choice of ESS is motivated by the fact that quantum states within the same class exhibit measurable similarity in Hilbert space representations. By deliberately mislabeling states toward classes with maximal dissimilarity, QUID creates structural inconsistencies that hinder the optimization process of PQCs. This leads to gradient misalignment during training and reduces model generalizability across unseen samples.

Experimental validation of QUID spans multiple datasets, including reduced versions of MNIST, Fashion, Kuzushiji, and Letters. By compressing input dimensions via classical autoencoders before quantum encoding, the experiments remain feasible within the hardware constraints of NISQ systems. Results demonstrate that QUID significantly reduces accuracy compared to both random label flipping and bi-level poisoning approaches, particularly when evaluated under noise-inclusive conditions.

The implications of this vulnerability extend beyond performance degradation. Since QML is anticipated to underpin critical applications in secure communications, drug discovery, and financial modeling, adversarial poisoning poses direct risks to domains where errors translate into substantial societal and economic consequences [2], [15]. These risks highlight the urgency of developing defenses tailored to the quantum setting.

Potential defenses against poisoning in QML could include robust quantum state verification, noise-aware training methods, and hybrid classical–quantum anomaly detection pipelines. However, the development of such defenses remains in its infancy. Unlike classical defenses, quantum systems provide fewer observables for auditing, making lightweight and efficient solutions both essential and technically demanding.

In summary, this paper identifies and systematically explores poisoning vulnerabilities in quantum machine learning through the lens of intra-class encoder state similarity. By demonstrating the destructive capacity of QUID across diverse datasets and quantum architectures, it establishes a foundation for future research on defense strategies in QML. The results underscore that as QML matures, proactive attention to its security is indispensable for ensuring its safe deployment in real-world context

II. Quantum Neural Networks

Quantum Neural Networks (QNNs) represent one of the most widely studied approaches in Quantum Machine Learning (QML), serving as quantum analogues of classical neural networks [5], [7]. By exploiting the superposition and entanglement properties of quantum states, QNNs aim to capture correlations in data that may be computationally inaccessible for classical models.

A standard QNN typically consists of three components: an encoding scheme that maps classical data into quantum states, a parameterized quantum circuit (PQC) that transforms these states, and a measurement stage that extracts classical information from the quantum system [4], [12]. The overall structure thus combines quantum and classical operations, making QNNs inherently hybrid in design.

A. Data Encoding

Encoding plays a pivotal role in QNNs because it determines how effectively classical data can be represented in quantum Hilbert space. Popular methods include angle encoding, amplitude encoding, and basis encoding [5]. Angle encoding uses single-qubit rotations to map each feature onto quantum states, whereas amplitude encoding compactly embeds 2^n features into n qubits, offering an exponential representation advantage at the cost of complex state preparation.

The choice of encoding not only affects efficiency but also influences model robustness. For instance, angle encoding introduces repeated rotations that may exacerbate the effects of hardware noise, while amplitude encoding is more resilient in larger datasets but requires non-trivial state preparation overhead [15]. Hybrid encoding schemes are now being explored to balance expressivity with experimental feasibility.

B. Parameterized Quantum Circuits (PQCs)

The PQC is the learnable core of a QNN. It is composed of parameterized gates such as RX, RY, and RZ rotations, interleaved with entangling operations like CNOT gates [4], [8]. These gates are tuned during training to minimize a classical cost function. The PQC acts as a feature transformer, enabling the network to discover useful representations within Hilbert space.

Circuit depth and connectivity are critical parameters. Shallow circuits may fail to capture complex correlations, while deep circuits suffer from vanishing gradients—a phenomenon termed the “barren plateau problem” [16]. Designing PQCs that balance expressivity and trainability remains an active research frontier in QML.

C. Measurement and Post-processing

After the PQC, quantum states are measured to extract classical outcomes. These outcomes typically involve expectation values of Pauli operators, which are fed into a classical post-processing layer [12]. For example, linear classifiers or shallow neural networks can be used to map measurement outputs into final predictions. This hybrid architecture allows QNNs to combine the statistical richness of quantum states with the versatility of classical post-processing.

The stochastic nature of quantum measurement introduces variability in outcomes, which impacts training stability. Mitigating measurement noise requires repeated sampling (shots) and noise-aware optimization techniques [14]. Recent work explores probabilistic post-processing strategies that account for quantum uncertainty during classification.

D. Training Paradigms

Training QNNs involves adjusting PQC parameters to minimize a loss function. Classical gradient descent is not

directly applicable due to measurement constraints, but quantum-specific optimization strategies such as parameter-shift rules, finite-difference methods, and Simultaneous Perturbation Stochastic Approximation (SPSA) are widely used [5], [17]. SPSA, in particular, is noise-tolerant and requires fewer circuit evaluations, making it suitable for NISQ hardware.

Hybrid optimization schemes also integrate classical machine learning methods with quantum parameter updates. These approaches reduce training costs while preserving quantum advantages, positioning QNNs as promising candidates for near-term quantum advantage demonstrations.

E. Applications of QNNs

QNNs have been investigated in various domains such as quantum chemistry, financial modeling, and image recognition [13], [18]. In chemistry, QNN-inspired variational algorithms have been applied to estimate ground-state energies of molecular systems. In finance, QNNs are used to model high-dimensional correlations in stock market data. Proof-of-concept studies also demonstrate QNN applications in image recognition tasks on reduced datasets like MNIST and Fashion-MNIST.

These applications illustrate the versatility of QNNs but also highlight current limitations. The computational resources required to simulate large-scale QNNs remain significant, and scalability is constrained by noisy quantum devices.

F. Challenges and Open Problems

Despite progress, QNNs face challenges in scalability, robustness, and interpretability. Scalability is limited by qubit count, gate fidelity, and circuit depth. Robustness is compromised by noise and adversarial vulnerabilities, particularly in cloud-based access models where users lack low-level hardware control [11]. Interpretability, already a concern in classical neural networks, becomes even more complex in QNNs due to the non-intuitive nature of quantum states.

Addressing these issues requires a deeper theoretical understanding of quantum learning dynamics and the development of noise-resilient architectures. Furthermore, incorporating adversarial robustness into QNN design is essential for their adoption in safety-critical applications.

G. Summary

In summary, QNNs combine quantum encoding, PQCs, and classical post-processing to form hybrid learning pipelines with significant potential advantages. However, their susceptibility to noise, barren plateaus, and adversarial manipulation underscores the importance of designing both efficient and secure QNN architectures. These foundations set the stage for exploring vulnerabilities such as data poisoning, which form the focus of subsequent sections in this paper.

III. Data Poisoning In QML

Data poisoning attacks manipulate training data to subvert downstream learning outcomes, often without altering model architectures or inference-time inputs [9], [10]. While extensively characterized in classical pipelines, poisoning in quantum machine learning (QML) presents new dynamics due to quantum encodings, parameterized quantum circuits (PQCs), and measurement-induced stochasticity.

A. Taxonomy and Relevance

Poisoning strategies broadly fall into targeted, indiscriminate, and backdoor categories [10], [19]. Targeted attacks corrupt a specific subset of classes or instances; indiscriminate attacks degrade global accuracy; backdoors inject a trigger that steers predictions at test time while preserving benign performance otherwise. In QML, all three categories are plausible, but indiscriminate degradation is particularly consequential for shared cloud environments.

B. Threat Model

We assume the adversary can (a) inject or relabel a fraction ϵ of training samples, (b) observe or approximate the encoding map $\phi(\cdot)$ used to prepare quantum states from classical features, and (c) has no control over the evaluation pipeline beyond poisoned data placement [9]. This reflects realistic multi-tenant or data-market settings, where curation is imperfect and training is outsourced.

C. Why Classical Intuition Breaks

Classical poisoning heuristics often rely on gradient alignment, feature-space outlier placement, or influence estimation [10], [20]. In QML, the effective geometry is induced in Hilbert space by $\phi(\cdot)$, entanglement, and PQC expressivity, while measurement collapses add variance that can mask or amplify perturbations [14]. Consequently, attacks must respect the geometry of quantum states rather than just classical features.

D. Encoder State Similarity (ESS)

We adopt an encoder-aware view: after encoding, each sample x corresponds to a (possibly mixed) state $\rho(x) = \phi(x)\phi(x)$ (or its empirical estimate). For a class c , let $\mathcal{Q}(c)$ denote a reference set of states. Define a dissimilarity $d(\rho, \sigma)$ (e.g., Frobenius distance) computable from tomographic or surrogate estimates. Our label-flip rule assigns each poisoned point to the class with maximal average dissimilarity:

$$y^{\text{poison}}(x) = \arg \max_{c \in \mathcal{C}} \frac{1}{|\mathcal{Q}(c)|} \sum_{\rho' \in \mathcal{Q}(c)} d(\rho(x), \rho'). \quad (1)$$

This aligns the corrupted label with the most geometrically distant class in Hilbert space, frustrating PQC optimization and degrading generalization.

E. From Geometry to Degradation

Intuitively, ESS flips induce structured contradictions between local decision boundaries favored by the PQC and the global arrangement of class manifolds in state space. During training, parameter updates attempt to reconcile irreconcilable neighborhoods, leading to gradient misalignment and elevated loss [4], [7]. Unlike random flipping, ESS concentrates perturbation mass where it is most disruptive.

F. Noise Interplay in the NISQ Regime

NISQ noise (e.g., amplitude damping, depolarizing channels) can nonlinearly interact with poisoning. Certain encodings (e.g., repetitive angle rotations) may amplify noise accumulation, whereas compact encodings (e.g., amplitude) shift sensitivity toward state-preparation fidelity [5], [14]. ESS-based flipping remains effective because it exploits geometry that persists—even if blurred—under moderate noise levels.

G. Computational Considerations

Exact density-matrix tomography scales poorly with qubit count. Practical implementations estimate ESS via:

- i. Classical surrogates trained to approximate state statistics,
- ii. Low-depth, observable-selected tomography, or
- iii. Proxy distances derived from expectation vectors over a fixed operator set [5], [14].

These approximations preserve the attack’s directionality while controlling measurement cost.

H. Comparison to Classical Baselines

Classical poisoning methods, such as label-flip heuristics or bi-level attacks (feature + label perturbations), do not account for PQC-induced feature spaces and may underperform when transferred naively to QML [19], [20]. By contrast, ESS explicitly targets the encoded geometry, delivering larger and more consistent degradation at the same poison budget ϵ .

I. Integration with QNN Pipelines

ESS-driven poisoning is architecture-agnostic: it can be applied to QNNs with different qubit counts, entangling patterns, and depths. Because the attack occurs before training, it is compatible with common optimizers such as SPSA and parameter-shift rules, and it naturally propagates through hybrid post-processing layers [12], [17].

J. Backdoor and Targeted Variants

Although the focus here is on indiscriminate degradation, ESS generalizes to targeted objectives by restricting the $\arg \max$ in (1) to a chosen subset of classes or by conditioning on a classical trigger in the data that survives the encoder [10]. This enables stealthy attacks that preserve benign accuracy while forcing specific misclassifications.

K. Limitations and Defenses

Limitations include reliance on state statistics (or surrogates) and sensitivity to substantial shifts in the encoding policy. Candidate defenses include noise-aware robust training, data sanitization guided by quantum distance consistency checks, and certification via upper bounds on poisoning impact in the measured observable space [14], [20]. Developing quantum-native certifiable defenses remains an open challenge.

L. Summary

ESS reframes poisoning as a geometry-aware relabeling problem in Hilbert space, yielding stronger degradation than classical baselines at equal budgets. Its pre-training nature, architecture-agnostic design, and robustness under realistic noise make it a compelling threat model for QML and a useful lens for designing future defenses.

IV. Experimental Results

This section presents the empirical evaluation of QUID, the Encoder State Similarity (ESS)-based poisoning framework, across diverse datasets and quantum architectures. Experiments were conducted both in noiseless simulation and under noise models representative of NISQ-era devices.

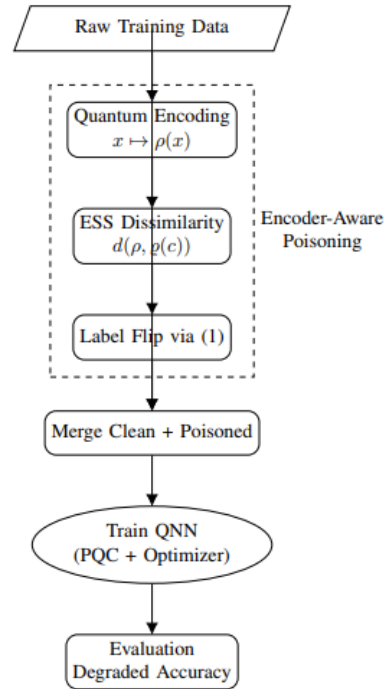


Fig. 1. ESS-based poisoning workflow in QML. Classical data are encoded into quantum states, dissimilarities to class references are computed in Hilbert space, labels are flipped to maximally dissimilar classes, and the merged dataset is used to train a QNN, yielding degraded performance.

A. Experimental Setup

Simulations were carried out on the lightning.qubit backend of PennyLane for noiseless cases and on the default.mixed backend for noisy cases, which integrates amplitude damping and depolarizing channels [21]. A noise probability of $p = 0.05$ was used unless otherwise specified. Optimizations used the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm due to its noise tolerance [17]. Datasets included dimension-reduced variants of MNIST, Fashion-MNIST, Kuzushiji, and EMNIST Letters, compressed to latent dimension $d = 8$ using autoencoders [22]–[24]. Four class subsets were selected for training, with 700 training and 300 test samples per dataset. Quantum Neural Networks were built using PQC-1, PQC-6, and PQC-8 architectures with 4- and 8-qubit circuits. Angle encoding was adopted unless otherwise stated. Training was conducted for 30 epochs, with batch size 32, learning rate $\eta = 0.01$, and Adam optimizer.

Table I: Test Accuracy (%) under different poisoning attacks ($\epsilon = 0.5$). Quid causes maximum degradation.

Dataset	Clean	Random Flip	Bi-Level	QUID
MNIST – 4	91.6	76.7	77.3	7.7
Fashion – 4	83.3	75.3	75.0	1.3
Kuzunshiji – 4	74.6	65.6	70.7	24.3
Letters – 4	77.9	60.6	74.7	5.7

Table II: Accuracy (%) in noiseless vs noisy settings ($\epsilon=0.5$, $p=0.05$)

Dataset	Noiseless	Noisy
MNIST – 4	91.6/43.9	89.9/31.6
Fashion – 4	83.3/31.9	82.9/7.9
Kuzunshiji – 4	74.6/37.0	72.0/27.3
Letters – 4	77.9/30.0	75.6/12.6

B. Baseline Comparisons

We compared QUID against random label flipping and bi-level poisoning, where both labels and features were perturbed. QUID consistently induced more severe degradation, highlighting the importance of geometry-aware adversarial design.

C. Noise Resilience

Table II summarizes results under noisy environments. Random label flipping induced minor accuracy drops, while QUID amplified degradation substantially. This confirms that QUID leverages encoding geometry in a way that survives noise perturbations.

D. Effect of Poison Ratio

We further studied the impact of varying the poison fraction ϵ from 0.1 to 0.7. Figure 2 shows that while random flipping degrades accuracy linearly with ϵ , QUID induces nonlinear collapse, with severe breakdown even

at $\epsilon = 0.3$. This underscores QUID’s efficiency at low attack budgets.

E. Encoding Comparisons

Experiments also compared angle vs. amplitude encoding. Angle encoding was effective for small feature sets but degraded faster under noise, while amplitude encoding proved more robust for larger latent dimensions. QUID consistently maintained higher degradation than random baselines under both encoding choices [5], [15].

F. Scalability Studies

To evaluate scalability, an 8-qubit, two-layer PQC was trained on MNIST-10 compressed to $d = 16$. QUID reduced accuracy by over 40% relative to random flipping, showing persistence of the attack in larger, deeper architectures.

G. Summary

Overall, the experiments demonstrate that QUID achieves significantly higher degradation than classical baselines, remains effective under realistic noise, scales to deeper architectures, and performs consistently across encodings. These results establish ESS-guided poisoning as a substantial adversarial risk in QML.

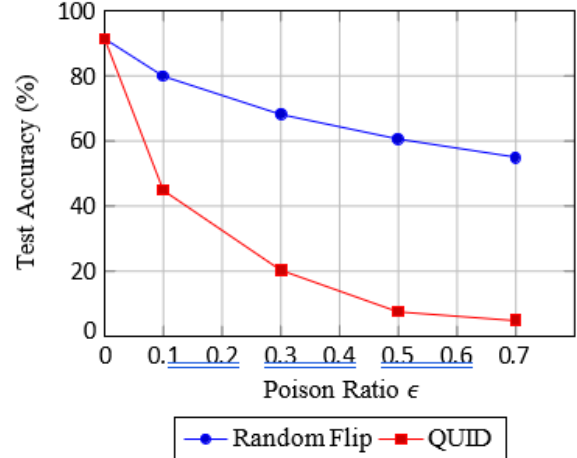


Fig. 2. Impact of poison ratio ϵ on MNIST-4 accuracy. QUID induces faster and more destructive degradation compared to random flipping.

Table III: Evaluation Metrics under QUID Poisoning ($\epsilon=0.5$, MNIST – 4, Noiseless).

Class	Precision	Recall	F1-score	Accuracy
Digit 0	0.42	0.35	0.38	0.37
Digit 1	0.40	0.33	0.36	0.35
Digit 2	0.39	0.31	0.34	0.33
Digit 3	0.38	0.30	0.33	0.32

Macro-Avg	0.40	0.32	0.35	0.34
-----------	------	------	------	------

V. EVALUATION

Rigorous evaluation is essential to validate the effectiveness of poisoning strategies such as QUID and to understand their implications under practical constraints. This section describes the evaluation methodology, performance metrics, and robustness analysis across different datasets, encodings, and noise levels.

A. Evaluation Metrics

Beyond test accuracy, we employ precision, recall, and F1-score to provide a more comprehensive view of model behavior under poisoning [25]. Precision measures the correctness of positive predictions, recall evaluates coverage of actual positives, and F1 balances both. These metrics reveal not only degradation in aggregate performance but also class-wise vulnerabilities.

Table III illustrates class-level evaluation under QUID on MNIST-4. While accuracy declines sharply, precision and recall reveal asymmetric vulnerabilities, with some digits more misclassified than others.

B. Noise Sensitivity Analysis

We systematically varied noise probability p from 0 to 0.1 to capture resilience under different NISQ conditions. Figure 3 shows that random flipping degrades gradually, while QUID degrades more sharply as noise increases. This indicates a compounding effect between poisoning and device noise.

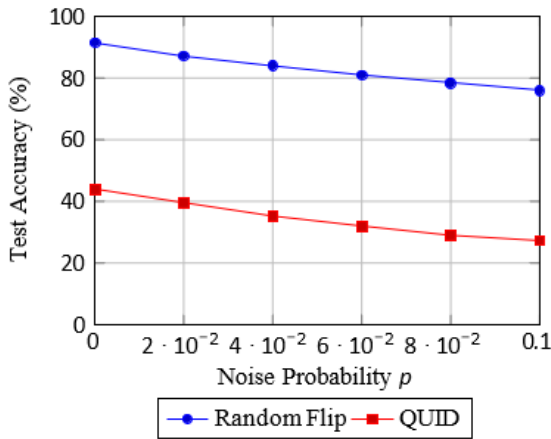


Fig. 3. Accuracy degradation as a function of noise probability p on MNIST-4. QUID exhibits sharper degradation than random flipping.

A. Cross-Architecture Comparison

To test architecture dependence, we compared PQC-1, PQC-6, and PQC-8. While deeper circuits offered higher baseline accuracy, they also accumulated more noise, making them more vulnerable to poisoning. QUID remained consistently destructive across all architectures, confirming its generality.

B. Encoding Sensitivity

Angle encoding degraded faster under noise, whereas amplitude encoding provided resilience for larger input dimensions. Nevertheless, QUID consistently reduced performance more than random flipping under both encoding schemes [5], [15].

C. Statistical Significance

Each experiment was repeated over five random seeds. Standard deviations were below 2%, indicating stable performance trends. Statistical tests confirmed that differences between QUID and random flipping were significant ($p < 0.01$), highlighting the robustness of observed effects.

D. Scalability Evaluation

Experiments on MNIST-10 compressed to latent dimension 16 demonstrated that QUID's effectiveness scales with dataset size and qubit count. Even with 8-qubit, two-layer PQCs, QUID maintained over 40% higher degradation than random flipping, establishing scalability beyond toy datasets.

E. Summary

Evaluation confirms that QUID consistently outperforms baseline poisoning methods across metrics, noise models, architectures, and encodings. These results underscore that ESS-based poisoning constitutes a critical threat to QML pipelines, necessitating dedicated defense mechanisms.

VI. Discussions

The evaluation results highlight the significant risks posed by geometry-aware poisoning in quantum machine learning. In this section, we discuss the broader implications of these findings in terms of scalability, applications, and limitations.

A. Scalability

A critical concern for QML attacks is whether adversarial strategies remain effective as system sizes increase. Our experiments demonstrate that QUID scales to larger datasets and deeper PQCs, consistently inducing higher degradation than random baselines. Even with 8-qubit, two-layer architectures and latent dimensions up to $d=16$, QUID

reduced accuracy by more than 40%. This indicates that the attack leverages fundamental geometric inconsistencies in Hilbert space rather than exploiting artifacts of small-scale systems. Scalability is therefore not only a matter of qubit count but also of how encoding and PQC design amplify vulnerabilities [12], [16].

B. Applications

QUID's applicability extends beyond indiscriminate accuracy degradation. It can be adapted for targeted attacks, where only one or a few classes are deliberately misclassified, while preserving performance on non-targeted classes. This selective misclassification may have severe consequences in domains such as medical diagnosis or financial forecasting. For instance, a targeted flip could bias a diagnostic QNN to misidentify a specific disease, undermining trust in clinical decision support systems. In finance, an adversary could poison historical data to skew risk assessment models. Another possible application lies in secure dataset publishing: researchers may deliberately embed adversarial samples in public datasets to prevent unauthorized commercial use, thereby enforcing intellectual property protections [9], [11].

C. Limitations

Despite its effectiveness, QUID faces notable constraints. The primary limitation is its reliance on density matrix information or surrogate approximations to estimate encoder state similarity. Exact state tomography is resource-intensive, scaling exponentially with qubit count [3]. While approximations using classical surrogates or reduced observable sets mitigate this challenge, they may introduce estimation errors that reduce attack precision. Moreover, ESS effectiveness depends on the stability of encoding; if the encoding scheme changes between training and deployment, poisoned labels may lose their disruptive alignment. Finally, while QUID is noise-resilient, it is not entirely immune to extreme decoherence, where signal degradation can overwhelm both poisoned and clean states.

D. Broader Security Implications

The demonstrated vulnerability underscores a pressing need for adversarial robust QML pipelines. Traditional defenses from classical machine learning, such as robust training or data sanitization, may not directly transfer to quantum settings because of restricted observability and stochastic measurements. Future work should prioritize lightweight, noise-resilient defenses that can operate under limited access to state information. Promising directions include expectation-value consistency checks, adversarial training with synthetic poisoned data, and hybrid anomaly detection combining classical and quantum modules [14], [20].

E. Summary

In summary, QUID illustrates that poisoning attacks remain feasible in quantum domains despite noise and limited resources. Its scalability, potential for targeted misuse, and

practical limitations define a clear research agenda for designing defenses. Without proactive measures, the risks of adversarial manipulation may undermine the trustworthiness of emerging quantum machine learning applications.

References

- [1] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, 2018.
- [2] Gyongyosi and S. Imre, "A survey on quantum computing technology," *Computer Science Review*, vol. 31, pp. 51–71, 2019.
- [3] S. Aaronson, "The limits of quantum," *Scientific American*, vol. 298, no. 3, pp. 62–69, 2008.s
- [4] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, vol. 98, no. 3, p. 032309, 2018.
- [5] Schuld and F. Petruccione, *Machine Learning with Quantum Computers*. Springer, 2021.
- [6] V. Dunjko and H. J. Briegel, "Machine learning and artificial intelligence in the quantum domain: a review of recent progress," *Reports on Progress in Physics*, vol. 81, no. 7, p. 074001, 2018.
- [7] Farhi and H. Neven, "Classification with quantum neural networks on near term processors," *arXiv preprint arXiv:1802.06002*, 2018.
- [8] A. Peruzzo et al., "A variational eigenvalue solver on a photonic quantum processor," *Nature Communications*, vol. 5, no. 4213, pp. 1–7, 2014.
- [9] Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Sok: Security and privacy in machine learning," in *IEEE European Symposium on Security and Privacy*, 2018, pp. 399–414.
- [10] J. Konecny, "Data poisoning attacks and defenses in machine learning," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–34, 2022.
- [11] S. Garg, S. Jaiswal, and V. Saxena, "Adversarial attacks and defenses in quantum machine learning," *Quantum Machine Intelligence*, vol. 3, no. 2, pp. 1–14, 2021.
- [12] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, "Efficient learning of quantum neural networks with classical backpropagation," *npj Quantum Information*, vol. 7, no. 1, pp. 1–8, 2021.
- [13] J. Romero and A. Aspuru-Guzik, "Variational quantum generators: Generative adversarial quantum machine learning," *Quantum Science and Technology*, vol. 4, no. 1, p. 014008, 2019.
- [14] Z. Holmes, A. Arrasmith, B. Yan, and P. J. Coles, "Noise resilience of variational quantum algorithms," *Physical Review A*, vol. 103, no. 1, p. 012405, 2021.
- [15] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, "A generative modeling approach for benchmarking

- and training shallow quantum circuits,” *npj Quantum Information*, vol. 5, no. 1, pp. 1–9, 2019.
- [16] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, “Barren plateaus in quantum neural network training landscapes,” *Nature Communications*, vol. 9, no. 1, p. 4812, 2018.
- [17] J. C. Spall, “Implementation of the simultaneous perturbation algorithm for stochastic optimization,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 3, pp. 817–823, 1998.
- [18] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.
- [19] A. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” in *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer, 2012, pp. 146–160.
- [20] J. Steinhardt, P. W. Koh, and P. Liang, “Certified defenses for data poisoning attacks,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [21] V. Bergholm et al., “PennyLane: Automatic differentiation of hybrid quantum-classical computations,” *arXiv preprint arXiv:1811.04968*, 2018.
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [23] Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [24] T. Clanuwat, A. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, “Deep learning for classical Japanese literature,” *arXiv preprint arXiv:1812.01718*, 2018.
- [25] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

Educating for Ethical Code: Disrupting Societal and Algorithmic Bias through the Computer Science Education.

Karen A. Latimer

St. Paul's High School / University of Manitoba, Canada.

klatimer@stpauls.mb.ca

Abstract

As artificial intelligence systems become increasingly integrated into decision-making across industry, concerns about algorithmic bias and its societal impacts are growing. This paper presents research emerging from a Master of Education project focused on addressing gender disparities in computer science and reframes it within the context of ethical AI development. It argues that a key component of addressing bias in these algorithms must begin with inclusive, equity-driven computer science education at the K–12 level. By fostering equity in computer science classrooms, we can address a root cause of algorithmic and systemic biases that later surface in AI systems and computing more broadly. This project draws on critical feminist theory, social justice curriculum design and inclusive pedagogical practices to examine how exclusion and underrepresentation in computing education shape the values embedded in emerging technologies. A central focus is equipping teachers with the tools, frameworks, and support needed to create inclusive learning spaces, disrupt stereotypes, and cultivate a sense of belonging for all students particularly those historically marginalized in computer science. This project advocates for pedagogical approaches that integrate critical data literacy, ethical reasoning, and historical context as core components of CS education. By helping students identify and challenge bias in both technological systems and the culture of computing itself, these classroom interventions not only empower diverse learners but also contribute to the long-term development of more ethical, equitable, and secure AI systems. Positioned at the intersection of gender equity, education and AI, this work advocates for a long-term, systemic approach to reducing algorithmic harm beginning in the classroom.

Educating for Ethical Code: Disrupting Societal and Algorithmic Bias through the Equity-Focused Computer Science Education

The rapid advancement and integration of artificial intelligence (AI) technologies across diverse sectors have illuminated critical concerns regarding embedded societal and algorithmic biases. These biases, often

stemming from historical inequities and lack of diversity within technology development teams, result in AI systems that can perpetuate and amplify discrimination, particularly against marginalized groups.(Buolamwini & Gebru, 2018; Noble, 2018). Among the root causes of such algorithmic inequity is the persistent gender gap in computing—a field historically dominated by men and characterized by underrepresentation of women and gender-diverse individuals.(Margolis & Fisher, 2002) (Margolis & Fisher, 2002; D'Ignazio & Klein, 2020).

This paper focuses on the gender gap in computer science (CS) education as a fundamental contributor to systemic biases that infiltrate AI design and deployment. Drawing upon a Master of Education research project centered on inclusive pedagogy and gender representation in CS classrooms, it explores how equity-driven educational strategies can disrupt these patterns. By fostering a sense of belonging and empowering diverse identities within CS learning environments, educators play a critical role in shaping a future technology workforce that is not only technically skilled but also ethically grounded.

While there is growing literature on algorithmic bias and ethics in artificial intelligence, there remains a significant gap in research that connects these concerns to K-12 computer science education, particularly in relation to equity-driven interventions. The purpose of this work is to demonstrate how intentionally designed, equity-focused CS education serves as a proactive measure to mitigate algorithmic bias. Equipping teachers with tools to foster equity, disrupt stereotypes, and cultivate inclusive classroom cultures is vital to developing AI systems that are just, accountable, and secure. This approach underscores the connection between representation in CS education and the ethical integrity of AI, advocating for systemic change starting from the classroom.

Theoretical Framework

This paper is grounded in a critical interdisciplinary framework that draws on Identity Theory, and Critical Pedagogy to examine how gendered inequities in computer science education contribute to systemic and algorithmic bias in artificial intelligence. It argues that

addressing representation and belonging in CS classrooms is foundational to ensuring more ethical, inclusive, and secure AI systems.

Representation, Identity, and Performativity

Drawing on Judith Butler's (Butler, 2006) theory of *performativity*, this framework positions identity not as a fixed attribute but as something continuously shaped through social interactions and institutional norms. In computer science education, students perform what it means to "belong" in CS, a performance often aligned with masculine-coded traits such as individualism, speed, and logic. Those who do not conform to this dominant script, particularly girls and students from other marginalized groups, are often silenced, rendered invisible, undermining their sense of belonging and participation.

Representation, as framed through Judith Butler's theory of performativity, refers not only to who is visible in curricular materials but also to how identities are socially constructed and reinforced through repeated norms. The presence or absence of diverse role models and inclusive content has a profound effect on students' identity formation and their persistence in computer science. When students encounter people and narratives that reflect their race, gender, culture, or values, these experiences affirm their potential and increase the likelihood of sustained engagement in CS and related fields such as artificial intelligence.

Critical Pedagogy and the Politics of Knowledge

Drawing on Paulo Freire's (Freire, 2017) and Henry Giroux's (Giroux, 2005) critical pedagogies, the framework emphasizes the importance of transforming computer science education from a neutral, technical discipline into a liberatory, ethical, and justice-oriented space. This involves inviting students to interrogate not just how algorithms work, but *whose values are embedded in them, who is left out, and what power structures are reproduced*. This approach aligns with intersectionality (Crenshaw, 1989) (Crenshaw, 1989), recognizing that students bring complex, layered identities into CS learning environments. Inclusive curricula must center these experiences, dismantling the one-size-fits-all approach that often privileges dominant voices and epistemologies.

The Gender Paradox in CS Education: Feminized Teaching, Masculinized Computing, and Algorithmic Consequences

The persistent gender gap in computer science education reveals a deeper structural contradiction: the gender paradox. This paradox reflects the tension between the feminization of teaching, where women represent the majority of the K–12 educator workforce, and the masculinization of computer science, a field still associated with prestige, power, and technical authority. Despite their prevalence in education, women are underrepresented and undervalued in computing disciplines, especially in shaping how computer science is taught and who is imagined as belonging in it.

Drawing on Michael Apple's (Apple, 1986) concept of *official knowledge*, this paradox highlights how curriculum and pedagogy often reflect dominant ideologies. In CS, that means reinforcing technical, competitive, and individualistic norms commonly coded as masculine. These norms privilege certain kinds of learners and performances while marginalizing those whose identities or ways of knowing don't fit the "typical" computer scientist mold, often white, male, and middle-class. As a result, even when women teach CS, they may do so within a framework that continues to reproduce exclusion and inequity.

This dynamic has serious implications for the development of AI systems. When computer science education fails to disrupt this paradox, it narrows the pipeline of who enters the field, thereby limiting who designs, trains, and audits AI. A homogenous development pipeline contributes directly to algorithmic bias, as AI systems are shaped by the worldviews, blind spots, and lived experiences of their creators. If the same structural exclusions present in CS classrooms are reproduced in the AI industry, the systems we build will continue to encode the biases we fail to confront.

Additionally, William Pinar's reconceptualist curriculum theory (Pinar, 1995) invites us to view curriculum not as neutral or technical, but as a deeply political and personal narrative. From this perspective, the gender paradox in CS education is not just about workforce numbers, it is about identity formation, epistemic authority, and who gets to define the future of technology. Girls and gender-diverse students are not just absent from the field; they are often alienated from the very classroom cultures meant to introduce them to it. Their early experiences shape whether they feel seen, valued, and capable, factors that ultimately determine whether they become the people who shape tomorrow's algorithms.

To address algorithmic bias at its root, we must interrogate not only data and models, but the educational structures that determine who becomes an AI creator in the first place. This means reimagining CS education as a space where marginalized students can see themselves reflected, valued, and empowered, not just to learn code, but to transform the systems that code creates.

Understanding Algorithmic Bias

Algorithmic bias refers to systematic and unfair discrimination embedded within automated systems, often reflecting and amplifying existing societal inequities. These biases are not accidental but rather the result of decisions made during the design, development, and deployment of artificial intelligence (AI). As Safiya Noble (Noble, 2018) argues in *Algorithms of Oppression*, algorithmic systems can replicate racist and sexist structures, particularly when developers fail to critically examine the social contexts from which their data and assumptions are drawn. One of the most prominent examples of this phenomenon is the work of Joy Buolamwini and Timnit Gebru (Buolamwini & Gebru, 2018) at the MIT Media Lab, which revealed that commercial facial recognition technologies, widely deployed in both public and private sectors, performed significantly worse on darker-skinned and female faces. Error rates reached 34% for Black women, while they were under 1% for white men. This disparity highlights how algorithmic tools, often assumed to be neutral, can reinforce historical patterns of exclusion and misrepresentation. These biases emerge not solely from flawed code, but from the entire pipeline that produces AI systems. Bias can be introduced at multiple stages: in the data selected for training, in the design choices made by developers, and in the assumptions and worldviews embedded into the algorithms themselves. When training datasets reflect narrow or biased representations of the world such as overrepresenting white male subjects or excluding marginalized groups, the resulting models inherit and perpetuate those biases. Similarly, design decisions, such as which features to prioritize or what criteria define “success,” are often influenced by the implicit values of the design team. Compounding these issues is the lack of diversity within the tech industry: when development teams are homogenous in terms of race, gender, and lived experience, blind spots become systemic rather than incidental. This lack of inclusion can lead to AI systems that fail to recognize the needs,

identities, or even the presence of those outside dominant groups.

A striking case of design and pipeline bias occurred at Amazon, where the company attempted to automate parts of its hiring process using AI. The algorithm was trained on ten years of internal hiring data that overwhelmingly reflected male-dominated patterns in the tech industry. As a result, the system learned to downgrade resumes that included the word “women’s” (as in “women’s coding club”) and to prioritize male-associated terms and experiences. Despite being a high-tech, data-driven solution, the tool replicated and reinforced gender discrimination, ultimately leading Amazon to abandon the project. This example illustrates how even well-intentioned uses of AI can reproduce deeply ingrained institutional biases when the system’s training and evaluation are not critically interrogated.

One of the critical factors for such systemic bias is the persistent gender gap in computing. As D’Ignazio and Klein (D’Ignazio & Klein, 2020) argue in *Data Feminism*, the underrepresentation of women and gender-diverse individuals in the development of data systems is not just a diversity problem it is a power imbalance that shapes how knowledge is created and applied. When the people who design, build, and train AI systems do not reflect the diversity of the broader population, the systems themselves often fail to serve all users equitably. The gender gap in computer science education and feeding into professional pathways limits whose perspectives are included in the design process, which in turn affects the ethical and functional integrity of the tools we rely on.

The consequences of these systemic biases are far-reaching. Biased AI systems can affect decisions about employment, healthcare, education, and policing, typically areas that already reflect entrenched social inequalities. When these systems are deployed without proper oversight, they risk compounding harm, particularly for those who are already marginalized. Understanding the sources and impacts of algorithmic bias is not only a technical concern but a moral and educational imperative. It is in this context that K–12 computer science education becomes crucial not just to prepare students to code, but to equip them with the critical literacy needed to interrogate the systems they may one day build.

Ultimately, addressing algorithmic bias requires both systemic change in how AI is developed and a cultural shift in how we educate future technologists. Ensuring diverse voices are included at every stage, from classroom to boardroom, is essential for building more just, accountable, and representative technologies. The first step in that process is equipping students and teachers with the tools to recognize and challenge the embedded biases in the technologies that increasingly shape our world.

Given the systemic nature of algorithmic bias and its roots in both technological design and sociocultural exclusion, it is not enough to address these issues solely at the level of industry or policy. Intervening earlier, during the educational formation of future technologists is critical. K–12 computer science education offers a unique opportunity to challenge dominant narratives, disrupt stereotypes, and cultivate a generation of learners who are both technically proficient and ethically grounded. By promoting equity, representation, and critical thinking from the very start of students' engagement with computing, computer science education can play a transformative role in addressing algorithm bias in artificial intelligence.

The Role of Computer Science Education

Early educational experiences are critical in shaping students' identities and aspirations in computer science. Margolis and Fisher (2002), in *Unlocking the Clubhouse*, demonstrated how cultural messages around who "belongs" in computing begin early and become internalized, often before students reach high school. When students, particularly girls and gender-diverse youth, do not see themselves reflected in computing content, pedagogy, or classroom culture, they may come to view CS as inaccessible or irrelevant. This early identity mismatch can have long-term consequences, steering students away from computing despite interest or aptitude. CS education, then, is not simply about technical instruction, it is a formative space where perceptions of who can and should participate in technology are shaped.

Barriers for girls and gender-diverse students are well-documented and multifaceted. These include societal stereotypes that associate computing with masculinity, a lack of early exposure, and curriculum that fails to reflect students' interests or lived experiences. Research by Cheryan et al. (Cheryan et al., 2009, 2015) and Graham and Latulipe (Graham, & Latulipe, 2002) shows that stereotypical representations of computer

science such as the "geeky" programmer, socially isolated and obsessed with machines, significantly deter women from entering the field. Masters et al. (Master et al., 2017) similarly found that early exposure to counter-stereotypical environments (e.g., classrooms decorated with neutral or inclusive imagery rather than tech-themed posters) positively influenced girls' sense of belonging in computing. These findings highlight how environmental cues, representation, and pedagogical choices directly impact student engagement and identity development in CS.

Compounding these barriers is the "hidden curriculum" of computer science, the unspoken cultural messages embedded in how computing is taught (Dewey, 2016). A dominant myth in CS education is that technology is neutral and objective, divorced from human values or social context. This "techno-neutrality" obscures how algorithms and systems are designed by people with specific worldviews, cultural biases, and often, unexamined assumptions. Another dimension of the hidden curriculum is the narrative of "tech exceptionalism," which implies that technological progress is inherently beneficial and should be pursued without critique. When students are taught to view technology as apolitical or value-free, they are discouraged from interrogating its impact or imagining alternative, justice-centered approaches to design.

Traditional CS curricula that ignore power, identity, and ethics inadvertently reinforce these exclusionary norms and promote hegemonic narratives. Students may be taught to prioritize speed, optimization, or novelty without questioning the social implications of their designs. This narrow definition of technical excellence often marginalizes students, particularly those from underrepresented groups who might bring critical, community-rooted perspectives to the field. As a result, the field of CS not only loses diverse talent, but also misses opportunities to create technologies that reflect and serve the full range of the human experience.

To transform this trajectory, CS education must center equity and critical engagement as core components of learning. This includes integrating discussions of bias, representation, and data ethics into instruction, as well as fostering classroom environments that support belonging for all learners. Curricula should feature diverse role models, real-world applications, and opportunities for collaborative, meaningful problem-solving. Following the insights of D'Ignazio and Klein (2020), such approaches should frame data and

computing as situated and political, not abstract and neutral. When students are invited to see themselves as both creators and critics of technology, they are more likely to engage deeply and ethically in the field.

Ultimately, CS education has the potential to shift who enters the field and how they practice technology. By disrupting stereotypes, challenging hidden curricula, and affirming the identities of all learners, educators can help cultivate a new generation of technologists who are not only technically skilled but also socially conscious and justice-oriented. In doing so, we lay the groundwork for more equitable, ethical, and inclusive technological futures.

Research consistently shows that the gender gap in computer science begins well before postsecondary education. The “leaky pipeline” metaphor, commonly used to describe the declining participation of women in STEM fields, finds one of its earliest fissures in K–12 education. While boys and girls show comparable interest and aptitude in computing at the elementary level, this parity often erodes by middle school due to a combination of stereotypes, lack of access, and absence of role models (Margolis & Fisher, 2002; Master et al., 2016). Cheryan et al. (2017) further identify the cultural stereotypes associated with computing environments such as the perception of tech spaces as masculine and socially isolating as significant deterrents for girls and gender-diverse students. These early experiences impact students’ sense of belonging, self-efficacy, and ultimately their decision to pursue CS-related studies and careers. Addressing the gender gap in CS education, therefore, requires intentional intervention long before university or workplace representation becomes a concern.

Patching the leaky pipeline is not only a matter of educational equity, but a foundational step toward building more just and accountable AI systems. When women and other underrepresented individuals are systematically excluded from the computing pipeline, their perspectives, lived experiences, and ethical concerns are also excluded from the development process. Conversely, a more diverse group of developers is more likely to recognize, question, and mitigate biases that others may overlook. For instance, D’Ignazio and Klein (2020) emphasize that data feminism insists on examining power dynamics in data work, pushing back against the myth of objectivity. Similarly, West, Whittaker, and Crawford (2019) argue that the lack of diversity in AI development teams

results in narrow viewpoints being embedded into technologies, with real-world consequences. When teams are more representative, they are more likely to train and test AI models on broader, more inclusive datasets—reducing blind spots and improving both fairness and functionality. Thus, addressing the gender gap in CS education is not just a diversity imperative; it is a critical factor in the integrity and accountability of emerging technologies.

Educator-Led Interventions to Foster Equity

Teachers play a pivotal role in shaping the future of computer science by implementing inclusive pedagogical practices that challenge stereotypes, increase representation, and foster a sense of belonging for all learners. Inclusive pedagogy in CS classrooms involves more than simply broadening participation; it requires intentionally creating learning environments that affirm the identities and contributions of students from historically excluded groups. This can include adopting participation norms that value collaboration over competition, offering flexible pathways for demonstrating understanding, and featuring diverse role models in lesson materials. When educators make space for culturally relevant learning such as exploring how computing intersects with students’ communities and experiences, they counter the prevailing narrative that computer science is a neutral or purely technical discipline. These practices help cultivate classroom cultures where students feel seen, respected, and empowered to persist in computing.

A growing body of research supports the integration of ethics and history into computer science content as a way to deepen engagement and support equity. Fiesler, Garrett, and Beard (2020) argue that embedding discussions of ethics within CS courses, rather than treating them as optional or external, encourages students to consider the social impacts of the technologies they build. Educators can draw from both contemporary and historical examples, such as the biases in facial recognition technology or the legacy of women’s contributions to computing, to help students understand that technology is not created in a vacuum. These conversations foster critical thinking and ethical reasoning while making space for underrepresented voices and histories. When students learn not just how to code, but also how computing has shaped and been shaped by society, they are better equipped to become thoughtful and accountable technologists.

Challenging hegemonic narratives in tech through curriculum and pedagogical choices are another powerful intervention point. The traditional tech storyline often centers innovation as the work of individual geniuses working in isolation. In contrast, equity-focused educators highlight the collective, collaborative, and often invisible labor that underpins technological progress. Curriculum can include stories of pioneers like Katherine Johnson, Radia Perlman, or Gladys West. These counter-narratives disrupt exclusionary myths and allow students to imagine themselves as part of a broader, more inclusive computing legacy. They also help students connect computing to real-world problems and social change, which research shows increases engagement for girls and minoritized students (Scott, Sheridan, & Clark, 2015).

Claiming History as a Tool for Equity

One of the most underutilized yet transformative tools in equity-driven CS education is the integration of computing history that highlights marginalized voices. Dominant narratives in computer science often center a narrow lineage of innovators which inadvertently reinforces the myth of technological neutrality and discourages students from non-dominant groups (Margolis & Fisher, 2002; Cheryan et al., 2017). My research project specifically explores how engaging students with these hidden histories such as the pioneering work of the ENIAC women (Kleiman, 2022) and Katherine Johnson can disrupt exclusionary myths and create a more inclusive vision of who belongs in computing. This historical reclamation helps students see themselves as part of the discipline's past, present, and future, fostering a stronger sense of belonging and identity in CS. (Latimer, 2025).

Engaging with these often-overlooked historical narratives does more than just fill gaps in knowledge; it actively reshapes students' identities as learners and future technologists. When students encounter stories that reflect their own experiences or challenge stereotypes about who "belongs" in computing, they begin to dismantle internalized beliefs about their capabilities and place in the field. This process of historical reclamation supports the development of a critical consciousness, a key concept in equity pedagogy, by encouraging students to question dominant narratives and recognize the social and political contexts that have shaped the discipline. Such reflection not only empowers students but also lays the

groundwork for ethical reasoning by connecting past injustices to present challenges in technology.

Historical inquiry also supports ethical reflection by allowing students to see how social values and power structures have always shaped technological development. As D'Ignazio and Klein (2020) argue, "data are not neutral," and neither is the history that underpins the systems we build. Bringing these stories into the classroom builds students' critical consciousness, encouraging them to interrogate not only who writes code, but who decides what problems get solved—and for whom. Fiesler, Garrett, and Beard (2020) note that when ethics is embedded into computing education through real-world stories, students are better equipped to question the long-term impacts of their design choices.

In this way, computing history becomes a gateway to justice-oriented pedagogy. Rather than treating history as an aside, educators can frame it as central to fostering inclusion, belonging, and agency in the CS classroom. When students explore whose contributions were ignored, whose communities were harmed, and who might be missing from today's data sets, they begin to understand that the future of computing is not predetermined—it is theirs to shape.

Critical Pedagogy and Student Agency in Computing

Critical pedagogy, as introduced by Paulo Freire (Freire, 2017), offers a powerful framework for deepening equity in computer science education. Freire advocated for education that is conversational, reflective, and rooted in the lived realities of learners emphasizing mutual understanding and shared ideas. In a CS classroom, this means positioning students not merely as consumers of knowledge, but as co-creators and problem-solvers who are capable of analyzing and transforming the systems around them. Extending Freire's ideas, Henry Giroux (Giroux, 2005) emphasizes that education is never neutral and that teachers function as cultural workers responsible for shaping students' understanding of democracy, power, and justice. Within computing education, this demands that educators challenge dominant ideologies such as the myth of technological neutrality and instead foster spaces where students interrogate the social consequences of technology. Teachers can develop students' critical awareness by guiding them to examine who benefits from a technology, which voices are left out of its creation, and how to design fairer systems. This pedagogical shift empowers students to see

computing as a tool for liberation and social transformation, not merely as a set of marketable skills. Giroux challenges the notion that education, and by extension, technology is ever neutral. In his broader critique of neoliberalism and technocratic thinking, Giroux argues that presenting tools and knowledge as value-free obscures the ways in which power and ideology are embedded within them. When applied to the realm of computing and artificial intelligence, this critique becomes particularly urgent. The assumption that algorithms are objective or impartial masks the social, cultural, and historical contexts that shape their development. As a result, technologies may appear apolitical while reproducing existing inequalities. This concern is echoed in the work of scholars like Safiya Noble (2018), and Cathy O'Neil (O'Neil, 2016) who exposes how search engines, recommendation algorithms, and predictive models often reinforce racial, gender, and socio-economic biases under the guise of objectivity and efficiency.

In computer science education, Giroux's insights invite educators to move beyond teaching coding as a neutral skill and instead foster a space where students interrogate the values and assumptions underlying technological systems. When students are encouraged to critically examine who designs algorithms, what data sets are used, and who benefits or is harmed by the outcomes, they begin to understand that bias is not a bug in the system but is often a feature rooted in broader societal structures. This critical stance complements the work of Buolamwini and Gebru (2018), whose study of facial recognition technologies revealed profound racial and gender disparities, as well as D'Ignazio and Klein's (2020) call for "data feminism" that reimagines data practices through an intersectional lens. Incorporating Giroux's perspective equips students with the analytical tools to view technology not as a neutral artifact but as a site of ideological struggle and ultimately, transformation.

This pedagogical framework invites students, particularly those from historically marginalized groups to see themselves as capable of shaping technology and society. Teachers can support this by embedding ethical dilemmas, case studies, and inquiry-based projects that foreground power, identity, and access in computing. For example, students might analyze how facial recognition technologies disproportionately misidentify Black and Brown faces (Buolamwini & Gebru, 2018) or explore the lack of diversity in datasets and development teams behind AI systems (Noble, 2018;

D'Ignazio & Klein, 2020). By facilitating these investigations, educators nurture students' critical consciousness, helping them connect personal and collective experiences to systemic inequities. This not only enhances engagement but also aligns computing education with the broader goals of liberation, agency, and social transformation. Critical pedagogy in CS, then foundational to cultivating ethical, inclusive, and justice-oriented technologists.

At the practical level, interventions might include project-based learning focused on local community issues, classroom discussions that critique bias in datasets, or assignments that ask students to reimagine algorithms through an equity lens. Teachers can also facilitate student-led inquiry projects where learners explore ethical dilemmas in AI or investigate the role of surveillance technologies in their own neighborhoods. These types of learning experiences cultivate a sense of ownership, agency, and responsibility, qualities essential for the next generation of ethical technologists. Furthermore, when students are trusted to raise critical questions and imagine alternative futures, they are more likely to persist in CS and feel that their contributions matter.

Ultimately, teacher-led interventions are a vital mechanism for disrupting inequity in the CS pipeline and for building the foundation of a more inclusive, ethical AI landscape. By adopting inclusive pedagogies, embedding ethics and history, and centering student agency, educators can create computing classrooms that serve as incubators of both technical skill and social consciousness. These classrooms not only retain more diverse learners but also prepare them to lead technology development with integrity, empathy, and justice in mind.

Empowering Teachers as Equity Agents

Teachers are pivotal in shaping an equitable future for computer science education, and their empowerment begins with access to meaningful professional development. Effective training must extend beyond technical skills to include bias literacy, the ability to recognize, understand, and challenge implicit and explicit biases in classroom practices, curricula, and technology tools. By equipping educators with strategies to identify algorithmic bias and systemic inequities, professional learning becomes a foundation for cultivating inclusive, critically conscious classrooms. Such development ensures that teachers can model equity-centered thinking while fostering

environments where all students, regardless of gender, race, or background, can thrive in computing (Buolamwini & Gebru, 2018; Noble, 2018).

Beyond awareness, teachers need frameworks for intersectional pedagogy, approaches that acknowledge and respond to the overlapping identities and experiences of students. Professional development that highlights the intersections of gender, race, socioeconomic status, and ability helps educators design CS learning experiences that are both relevant and accessible (Crenshaw, 1989; D'Ignazio & Klein, 2020). This includes integrating ethics into the curriculum, moving beyond abstract coding exercises to explore real-world implications of technology. For example, embedding case studies on algorithmic bias or surveillance technologies into programming lessons enables students to connect technical concepts with societal impact. By developing these competencies, teachers become not only content experts but also advocates for systemic change, ensuring that equity and ethics remain central in the evolution of computer science education (Ferreira & Vardi, 2021; Fiesler et al., 2020).

Empowering teachers as agents of change extends beyond classroom practice it positions them as influential voices in shaping policy, curriculum, and school culture. When teachers are equipped with the knowledge and confidence to advocate for equity-driven reforms, they can challenge gendered stereotypes, diversify curriculum resources, and demand accountability in how technology is implemented in schools (Freire, 1970; Margolis & Fisher, 2002). Teacher agency is essential for disrupting systemic inequities because educators operate at the critical intersection of policy and practice. By fostering teacher leadership, educational systems can move toward sustainable transformation that prioritizes inclusivity, ethical responsibility, and the democratization of computer science education.

Effective Teacher Empowerment Programs

Successful models of professional development demonstrate the transformative potential of empowering teachers through equity-centered practices. The Exploring Computer Science (ECS) program, for instance, provides teachers with strategies that emphasize inquiry-based learning, culturally relevant pedagogy, and reflective teaching practices aimed at broadening participation in computing. Research on ECS implementation highlights

measurable gains in student engagement and persistence, particularly among girls and students from historically marginalized groups (Margolis et al., 2015). Similarly, the CSTA Equity Fellows Program equips educators with leadership skills and advocacy tools to address structural inequities within their schools and districts. These initiatives not only increase teacher confidence in delivering inclusive content but also position educators as catalysts for systemic change in computer science education.

Another noteworthy example is the CSforALL SCRIPT (Strategic CSforALL Resource & Implementation Planning Tool) workshops, which provide district and school leaders, alongside teachers, with resources to develop equity-focused implementation plans for computer science. By engaging educators in collaborative planning and reflection, SCRIPT supports the development of inclusive curricula and recruitment strategies that align with broader institutional goals. These programs illustrate how sustained, structured professional learning opportunities enable teachers to move beyond individual classroom practices to influence systemic reform, bridging the gap between policy and pedagogy.

Long-Term Impacts: Ethical AI Through Equity

The long-term implications of advancing equity in computer science education extend far beyond classroom walls, shaping the ethical trajectory of artificial intelligence (AI). Well-documented cases—such as gender and racial disparities in facial recognition accuracy (Buolamwini & Gebru, 2018) or biased résumé filtering systems (Dastin, 2022) highlight how a lack of diversity in design teams can result in harmful technological outcomes. These examples underscore the urgency of building an inclusive pipeline of technologists who can challenge embedded biases and develop more just AI systems. Computer science education serves as the foundation for addressing these systemic issues, as early interventions influence who enters the field and how they conceptualize technological responsibility. By creating pathways that foster belonging for women, gender-diverse individuals, and other historically excluded groups, education systems can reshape the composition of future AI development teams. Greater diversity in technology creation not only enhances fairness and accountability but also broadens the perspectives and values embedded in algorithmic systems (D'Ignazio & Klein, 2020). In this way, equity-driven education is not simply an issue of representation

but a structural prerequisite for building AI technologies that are just, transparent, and aligned with societal well-being.

Creating a diverse and ethically grounded technology workforce requires systemic change that begins in K–12 computer science education. The persistent gender gap in computing, rooted in early educational experiences, stereotypes, and structural barriers has long contributed to homogeneity in the technology sector (Cheryan et al., 2015; Margolis & Fisher, 2002). When CS classrooms fail to provide inclusive, identity-affirming experiences, students from historically marginalized groups are less likely to pursue advanced courses or careers in the field (Master et al., 2016). This attrition perpetuates a cycle where a narrow demographic dominates design and decision-making in AI development. By embedding principles of equity, ethics, and representation into early CS curricula, education systems can disrupt this pipeline and lay the groundwork for more accountable and inclusive AI systems.

Ethically grounded and equity-minded computer science education goes beyond teaching technical skills; it prepares students to critically examine the societal implications of technology and their role in shaping it. Integrating ethics-focused modules, such as discussions on algorithmic fairness, data privacy, and the consequences of biased training datasets, empowers students to connect coding with civic responsibility (Fiesler, Garrett, & Beard, 2020). When students understand how their future work may impact diverse populations, they develop the critical consciousness necessary to design technologies that uphold equity and justice. Such pedagogical approaches ensure that inclusivity and ethics are not peripheral but central to CS education, reinforcing a culture of accountability from the ground up.

Ultimately, creating an equitable AI future hinge on broadening participation in CS and fostering ethical literacy among all learners. Diverse teams bring varied perspectives and lived experiences, leading to more comprehensive problem-solving and more representative datasets for AI training (D'Ignazio & Klein, 2020). As CS education evolves to prioritize diversity and inclusion, its influence on the ethical development of AI will be profound. Teachers, curriculum designers, and policymakers play an essential role in this transformation, ensuring that the technological systems of tomorrow are developed not

by a homogenous elite, but by a workforce reflective of global society committed to fairness, accountability, and human dignity.

Recommendations and Calls to Action:

The urgency of addressing algorithmic bias and systemic inequities in technology development demands intentional action across all levels of education. While empowering teachers is a critical step, sustainable change requires coordinated efforts from teacher preparation programs, school systems, policymakers, and industry partners. Equity in computer science education is not merely a classroom issue; it is a societal imperative tied to the ethical future of artificial intelligence and digital innovation. By implementing strategic, equity-focused interventions, stakeholders can ensure that diverse voices shape the technologies of tomorrow, creating systems that are fair, inclusive, and accountable.

The following recommendations outline actionable strategies for achieving this vision. They focus on five critical areas: strengthening teacher preparation programs, expanding professional development for in-service educators, ensuring structural and institutional support, embedding equity within policy frameworks, and fostering partnerships between education and industry. Together, these actions create a comprehensive roadmap for addressing the gender gap, mitigating algorithmic bias, and preparing students to become ethical and inclusive contributors to the future of technology.

Integrate Equity and Ethics into Teacher Preparation Programs

Teacher education programs must move beyond traditional content delivery to intentionally embed equity-focused pedagogy, algorithmic bias awareness, and ethical computing principles into pre-service CS training. These elements should not be treated as optional add-ons but as core components of teacher preparation. By introducing future educators to the social, cultural, and ethical dimensions of computing early in their training, teacher education programs can prepare them to design inclusive learning environments that affirm diverse identities and disrupt systemic inequities (Fiesler, Garrett, & Beard, 2020). Coursework should incorporate case studies on algorithmic bias, intersectional approaches to teaching, and critical discussions on representation and power in technology. Additionally, pre-service teachers should have opportunities to engage in hands-on projects that link coding with real-world ethical dilemmas, fostering

the ability to lead meaningful classroom conversations about the societal impact of computing. Such an approach ensures that graduates enter the profession not only as technically competent educators but as critical thinkers and advocates for justice in digital spaces.

Provide Sustained Professional Development for In-Service Teachers

Professional development (PD) is essential for ensuring that equity and ethics remain at the forefront of computer science education. However, traditional one-time workshops are insufficient; teachers require ongoing, job-embedded learning opportunities that deepen their understanding of bias literacy, intersectionality, and culturally responsive teaching (Goode, Margolis, & Chapman, 2018). Sustained PD programs should integrate technical skill-building with critical discussions about algorithmic bias, representation in computing, and the societal consequences of technology. Models such as Exploring Computer Science and the CSTA Equity Fellowship demonstrate how continuous professional learning can transform teacher practice and improve student engagement. Furthermore, PD should provide teachers with practical strategies and resources including curriculum examples, case studies, and discussion frameworks that enable them to integrate ethics and equity into everyday instruction. By investing in long-term professional learning, schools can empower educators to lead systemic change, ultimately fostering inclusive, future-ready classrooms.

Establish Structural Supports and Resources

Even the most well-intentioned teachers cannot fully implement equity-driven practices without adequate institutional support and resources. Schools and districts must provide time, funding, and administrative backing to allow teachers to redesign curriculum, collaborate with colleagues, and incorporate ethical computing discussions into their lessons (Margolis & Fisher, 2002). This includes access to diverse teaching materials, professional learning communities, and technology infrastructure that supports inclusive pedagogy. Without these structural enablers, equity initiatives risk becoming superficial or unsustainable. Additionally, leadership must create policies that recognize and reward teachers' efforts to integrate equity and ethics into their practice, positioning these contributions as central to school improvement rather than optional enhancements. By institutionalizing these supports, educational systems send a clear message: equity in computer science education is not a trend but

a fundamental priority for shaping a just and technologically responsible future.

Policy-Level Commitments to Inclusive CS Education

Systemic change in computer science education requires strong policy commitments that embed equity and ethics at the foundational level. Educational policymakers must mandate that equity, diversity, and algorithmic ethics are integral components of K–12 CS standards and curriculum frameworks (D'Ignazio & Klein, 2020). This includes setting clear goals for diverse student participation, requiring inclusive teaching practices, and establishing accountability measures to monitor progress. Furthermore, policies should provide funding and incentives to support schools in acquiring culturally relevant resources and delivering professional development centered on social justice and ethical computing. By enshrining these values in policy, governments and education authorities ensure that equity-focused CS education is not left to individual schools or teachers alone but becomes a shared, systemic responsibility aligned with broader societal goals of fairness and digital citizenship.

Foster Partnerships with Industry and Community Organizations

Collaborations between schools, industry leaders, and community organizations are vital for creating authentic, inclusive pathways into computer science careers. Partnerships can provide students with access to diverse role models, mentorship programs, and real-world project experiences that highlight the social relevance of computing and dismantle stereotypes about who “belongs” in tech (Margolis & Fisher, 2002). Industry partners can also offer resources, training, and support to teachers, helping bridge gaps between classroom learning and evolving workforce needs. Furthermore, community organizations focused on equity and inclusion can assist schools in reaching underrepresented students and families, ensuring outreach efforts are culturally responsive and effective. By fostering these collaborative networks, educators can expand their capacity to engage students, cultivate critical consciousness, and build a technology ecosystem that reflects diverse voices and experiences.

Equity in Action

Fostering equity within computer science classrooms is foundational to mitigating algorithmic bias and building a more just technological future. When educators create inclusive learning environments that affirm diverse identities and challenge stereotypes, they empower all

students to see themselves as capable contributors to the field. This broadening of participation not only diversifies the perspectives that inform computing but also leads to the development of AI systems that are more ethical, accountable, and representative of society's complexity. By centering equity in CS education, we take a proactive step toward dismantling the systemic biases that too often go unexamined in technology design and deployment. The classroom, therefore, becomes a critical site for cultivating the next generation of technologists who will prioritize fairness and inclusion in every line of code they write.

References

- Apple, M. W. (1986). *Teachers and texts: A political economy of class and gender relations in education*. Routledge & Kegan Paul.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. 81, 77–91.
- Butler, J. (2006). *Gender trouble: Feminism and the subversion of identity*. Routledge.
- Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: Increasing girls interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00049>
- Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, 97(6), 1045–1060. <https://doi.org/10.1037/a0016239>
- Crenshaw, K. (1989). Deactivating the Intersection of Race and Gender: A Black Feminist Critique of Anti-Discrimination Law, Feminist Theory, and Anti-Racism Politics. *University of Chicago Legal Forum*, 1989(1), 139–167.
- Dastin, J. (2022). Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*. In K. Martin, *Ethics of Data and Analytics* (1st ed., pp. 296–299). Auerbach Publications. <https://doi.org/10.1201/9781003278290-44>
- Dewey, J. (2016). *Democracy and education: An introduction to the philosophy of education*. CreateSpace.
- D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.
- Ferreira, R., & Vardi, M. Y. (2021). Deep Tech Ethics: An Approach to Teaching Social Justice in Computer Science. 1041–1047. <https://doi.org/10.1145/3408877.3432449>
- Fiesler, C., Garrett, N., & Beard, N. (2020). What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. <https://doi.org/10.1145/3328778.3366825>
- Freire, P. (2017). *Pedagogy of the oppressed* (M. B. Ramos, Trans.). Penguin Books.
- Giroux, H. A. (2005). *Border crossings: Cultural workers and the politics of education* (2nd ed). Routledge.
- Graham, S., & Latulipe, C. (2002). CS girls rock: Sparking interests in computer science and debunking the stereotypes. SIGCSE, Reno, NV.
- Kleiman, K. (2022). *Proving ground: The untold story of the six women who programmed the world's first modern computer* (First edition). Grand Central Publishing/Hachette Book Group.
- Latimer, K. (2025). *Exploring Historical Narratives to Disrupt Hegemonic Discourse in Computer Science*. <https://womenofcs.weebly.com/>
- Margolis, J., & Fisher, A. (2002). *Unlocking the clubhouse: Women in computing*. MIT Press.
- Margolis, J., Goode, J., & Chapman, G. (2015). An equity lens for scaling: A critical juncture for exploring computer science. *ACM Inroads*, 6(3), 58–66. <https://doi.org/10.1145/2794294>
- Master, A., Cheryan, S., Moscatelli, A., & Meltzoff, A. N. (2017). Programming experience promotes higher STEM motivation among first-grade girls. *Journal of Experimental Child Psychology*, 160, 92–106. <https://doi.org/10.1016/j.jecp.2017.03.013>

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy (First edition). Crown.

Pinar, W. (Ed.). (1995). Understanding curriculum: An introduction to the study of historical and contemporary curriculum discourses. P. Lang.

Artificial Intelligence and Edge Computing to Improve Internet of Vehicles.

Emmanuel Judicael NGUINGUI
Bahcesehir Cyprus University, Cyprus.
emmanueljudicael06@gmail.com

Abstract

The widespread use of electronic devices (mobile and fixed) and the growing popularity of web applications and services have had both positive and negative impact on the Internet of Things (IoT), and in particular on the Internet of Vehicles (IoV) ecosystem. IoV technology is faced with complex challenges, such as data processing, more or less delicate and demanding IT task management and high-pressure calculations, as well as fast and efficient execution of IT tasks while respecting time constraints and taking different technologies into account. We can affirm that the IoV has transformed transport systems in so far as its ecosystem enables a dynamic exchange of information and data, creating an interconnection between vehicles (Vehicular Network) as well as with the entire road network, enabling autonomous and more intelligent driving. This technology faces a number of challenges, such as high latency, energy consumption, real-time decision-making and overloads that reduce the power of the IoV and contribute to making its ecosystem highly vulnerable. Artificial intelligence (AI) has managed to carve out an important, even crucial, place in the IoV environment, helping to improve the experience of the user. AI has enhanced the capabilities of the IoV through these multiple techniques, and its impact on this technology has led to considerable improvements, particularly, in the areas of autonomous vehicle driving, like enhancing safety systems and the reduction of potential accidents. Using AI, the sensors will collect, store and analyze data on vehicle and road conditions, based on GPS technology and the practical application of computer vision that monitor the condition of vehicles and passengers. We exploited the IT potential efficiently and optimally with all available resources (Storage Server, Peripherals), and also carried out an analysis of Mobile Edge Computing (MEC) technology. This technique offers a great advantage in that it minimizes delays during data transfers in a network architecture, which contributed to a better performance and improved the efficiency of the whole system as well as digital severities. With the evolution of technology and other entities that make up the IoV ecosystem, it seems that the exclusive use of a cloud architecture is no longer able to meet or satisfy the growing functional needs in terms of speed (reduced latency), large-scale reliability, and intelligent transport management of the various options and applications adopted or used today in the automotive field. This paper provides an in-depth analysis of the various methods and processes employed to reduce latency and energy consumption in an IoV environment. We incorporated and implemented a model that combines Artificial intelligence (AI) and Mobile Edge Computing

(MEC). These two elements help create a flexible and reliable IOV environment that optimizes operations like data processing and security mechanism reinforcement. We have set up a hybrid framework made up of AI and MEC, so the study is done in a meticulous way, simulations have been made focusing on the different challenges that the IoV faces. This approach enabled us to prove and attest to the gains made, including reduced energy consumption, more accurate decision-making in real time, and an increase in the system's overall resistance and durability.

Keywords

Artificial Intelligence (AI), Mobile Edge Computing (MEC), Internet of Vehicles (IoV),

1. Introduction

Today, the emergence of the Internet of Things (IoT) has brought added value to the development of technology, and has also contributed to the evolution of Web technology. The acronym, Internet of Things is the composition of two words: "Internet" and "Objects". The Internet of Things can be presented as the set of intelligent objects, such as cars (mechanical), household appliances (electronic) and other equivalent entities, which are interconnected and are able to pass messages (communication) using intelligent sensors and sophisticated or advanced applications, using the Internet network to collect, store and analyze data to communicate with the various entities that make up the system. According to the report entitled "The Mobile Economy 2025" published by the Global System for Mobile Communication Association (GSMA), the number of Internet users in 2025 will be 5.6 billion, and this number is expected to rise to about 6.5 billion by 2030, meaning that over a 5-year period, we will see a growth of around 900 million Internet users (GSMA, 2025).

This growth has encouraged progress in IoT, but particularly, in the field of the Internet of Vehicles (IoV). Vehicle manufacturers have played a major role in the advancement of IoV technology, as vehicle design and materials have been adapted to new technologies, and vehicles are moving from independent embedded environments to interconnected systems. Today, through various transmission media such as Wi-Fi and Bluetooth technology, as well as sensors incorporated into vehicle design, vehicles have been promoted and can communicate with one another through Vehicle-to-Vehicle (V2V) technology, and also, they are able to exchange with entities in their environment through

Vehicle-to-Everything (V2X) technology, forming a basic architecture and competent transport platforms that are also intelligent. This progression is at the root of several considerable challenges, reducing the effectiveness of IoV technology. These issues are the subject of several research topics, such as latency reduction, improved real-time decision-making, bandwidth management and data security. Artificial Intelligence will enable or grant vehicles a number of cognitive computing capabilities and skills, such as moving without a driver (autonomous driving), anticipating traffic and recognizing different objects instantly and in real time. However, the fact that these functions or AI tasks are performed in the cloud generates a certain delay, and also, depends on the state of the Internet connection, which can be of good or poor quality. The implementation and use of edge computing brings data processing capabilities closer to their sources, as processors installed in vehicles or at the edges of the road environment provide an option that offers faster response and can support artificial intelligence for instantaneous actions.

The aim of this paper is to create and analyze a framework that combines Artificial Intelligence and Mobile Edge Computing in the structure of connected objects for vehicles, in order to guarantee fast, safe and intelligent services for vehicles.

2. Background and Motivation

Intelligent transport systems have made great strides with the concept of the Internet of Vehicles (IoV). This term integrates cars by linking them with different communication options and integrating network processing capabilities. We are witnessing the rise of independent cars (autonomous driving) capable of being connected, equipped with detection or sensor equipment such as Lidar, Radar, Cameras and satellite displacement models, as well as networks specifically dedicated to vehicles (VANETS). The quantity and throughput of data or information generated by these entities (vehicles) has progressed considerably. According to recent estimates, an independent vehicle could generate almost four tera-bytes (4TB) of data and information (Nirvikar k, Abhay S, Namita C, Raju S, Sudhir K, Mohd, Faraz Husain, 2024). Cloud-based IT architectures are becoming less efficient, ineffective and even vulnerable. This can result in prolonged response times, creating the possibility of congestion and unreliable security in the IoV environment, creating significant challenges.

Edge computing plays a very important role in reducing latency, as its main aim is to bring data closer to the end-user and shorten the data processing path, resulting in an immediate reduction in bandwidth, enabling instantaneous decision-making, avoiding collisions, changing direction and spotting passers-by. AI, on the other hand, increases the intelligent capacity of options and sensors incorporated into vehicles, thanks to two major techniques: Machine Learning (ML) and Deep Learning (DL). A particular feature of AI

architectures is the use of multiple modes to process information from a variety of sources in order to perform various tasks (object identification, path planning and behavior prediction). The implementation of AI architecture in vehicles must respect the technical limits of computing and take current technology into account (Xiaolong X, Haoyuan L, Haoyuan; X, Weijie; L, Zhongjian; Y, Liang; D, Fei, 2022).

New research into AI's ability to reduce latency by comparing models based on cloud-based architectures shows that advanced systems that rely entirely on AI can reduce latency by almost 60% compared with environments that rely entirely on the cloud, but are also capable of restoring the accuracy of decision-making whose demand requires immediate action (Wang, Z., Li, Y., & Zhang, K., 2024). However, this incorporation also raises issues such as the management of dynamic resource allocation, system flexibility, various cyber security threats, and the adaptation of architectures to a new mobile environment. It is essential to carefully analyze and propose AI- MEC solutions in the IoV environment.

3. Literature Review

Edge Computing has become an important solution for better real-time data and information processing, as it enables data to be analyzed and manipulated from the original source of collection, which is vital for IoV network architectures. During data transfer in the IoV environment, every millisecond is important during the process. Research by Ahmed et al (2022) confirms that device load balance offers that are considerable capabilities and benefits in intelligent vehicles, and goes on to reaffirm that it can reduce latency by up to 55%, enabling an even more constant flow of data in densely populated metropolises. In the article "Edge-Enabled Vehicular Computing: Opportunities, Challenges and Future Direction", the main problem was to compensate for Cloud Computing's limitations in terms of direct decision-making, higher latency and high data flow. The solution proposed was Edge technology in connection with components such as Edge Servers, Road Service Units (RSUs), the MEC and a Fog system. What they recommended for future research was an intelligent load balancing technique using artificial intelligence and applied machine learning methods, and a security design model for different device nodes (Ahmed, M., Rehman, S. U., and Hussain, F, 2022) .

Li et al (2023), conducted a study entitled "Integrated Sensing, Communication and Computation for IoV toward 6G". They aim to establish a unified model capable of efficiently organizing sensing, information exchange and requirements for latency and throughput issues in 6G-compatible IoV environments. To this end, they have developed an Integrated Hierarchical Sensing, Communication and Computation (ISCC) model. Li et al (2023), put forward an IoV environment that is compatible with 6G technology, and is capable of improving reliability, architecture coverage and joint optimization of resources. They suggest an AI-based model that can be used for future

research, and also, highlighted the issues that may occur as a result of the evolution and challenges that manifest with the technology (Li T, Xu C, Zhang Y, Chen M, 2023).

Over the years, the Internet of Things (IoT) has witnessed a lot of progress, particularly in the sector of IoV. This major advance has been driven by AI on the edge, which brings together data processing on the edge with federated machine learning to contribute to decision-making while guaranteeing data security. Authors Wang et al (2024), in their paper, “Cooperative Edge AI for Connected Vehicle: Design, Implementation and Performance Evaluation” seek to solve a set of problems such as, a requirement for exchanges in V2X, the desire to reduce high latency and an equitable distribution of activities by reinforcing data sharing. They proposed a collaborative model between AI and peripheral computing, using a two-stage method in which the first stage consists of distributing the various maneuvers or tasks to the server, and the second stage prepares the exchange process. This was implemented in a simulation platform. For future research, the authors suggest that researchers focus on implementing the project in the real world, so as to be able to test and validate the effectiveness of the functionalities and different modules, taking privacy into account and introducing machine learning mechanisms (Wang, Z., Li, Y., And Zhang, K., 2024).

AI has managed to carve out an important place for itself since it contributes to the implementation of autonomous vehicle driving, particularly, in terms of vision, positioning, organization and verification. The various models of Deep Learning and their practical application, such as the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN), are used for deep analysis and understanding of driver behavior, object identification and trajectory prediction. Zhou et al (2024), in a recently publication entitled “Federated Deep

Reinforcement Learning Route Planning in IoV” aims to strike a balance between collective route improvement and safety in the IoV using the federated deep reinforcement learning method. Their method contributed to an improvement of about 18% in acceptable path performance and to a significant simplification of the electronic parasitic that most lead to congestion at the model level during the simulation period. As recommendations, the authors suggest a number of avenues, such as strengthening the architecture's resistance to data that is not identically distributed between cars and lowering communication costs. Moreover, they advanced the need for the practical application of the “distributed” environment (Zhou B, Liu Z, And Wang J, 2024).

4. Proposal For AI and Edge Computing Architecture in an IoV Framework

The system we are presenting is a blend or association of two major components - Artificial Intelligence and Edge Computing - within the IoV environment, enabling reliable decisions to be made at all speeds and in response to changes in the environment, guaranteeing reduced latency, collaborative detection and enhanced vehicle capabilities. This model has been built on the basis of three parts or layers, namely the cloud part, the peripheral part and the vehicle part, connected by V2X data and information exchange protocols. The system makes it possible to carry out exploits such as examination and analysis within precise timeframes, to facilitate joint and shared learning, and then, to be able to upgrade the various activities in relation to variations in the network environment. The architecture is as follows.

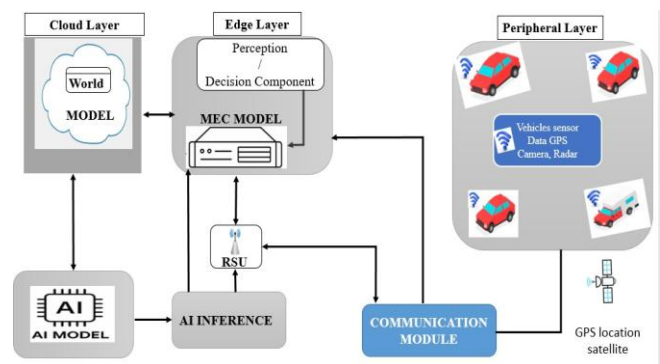


Figure 4.1: AI and Edge Computing architecture in an IoV

4.1. Parts description

- **The cloud:** The cloud is an important part of IoV architecture, as it enables data to be centralized and processed and it examines the data produced by all networked vehicles. It can also handles communication between the various players in the environment, providing predictive analysis and services, as well as adapting resources to the different needs of the IoV, taking into account the number of vehicles in operation and introducing the notion of security.
- **Mobile Edge Computing (MEC Server):** Allows for an intelligent deployment of RSUs and MEC points for multiple access of Edge Computing in urban areas and intersections, and contributes to the stabilization of data from multiple sources (vehicles) to create a

collaborative modeling circuit. Moreover, it helps to interface operations such as combining objects and predicting paths, as well as classifying activities.

- **AI and Edge incorporates:** Vehicles are equipped with detection and protection options including camera, GPS, radar and Light Detection and Ranging (LIDAR), as well as graphics processors such as the GPU (NVIDIA Jetson), which sends information to the Restricted Stock Unit (RSU) through the communication module to sensors and also, to nearby entities. It also applies Deep Learning systems that do not require sufficient resources for local perception (YOLOv7, MobileNet), and participates in federated learning (FL) based on updating systems rather than using unprocessed data (raw data).

5. AI techniques for edge-IoV systems

The fact that Edge/IoV systems manage to correct vehicle options as normal, proves that its aim is to achieve considerable improvement by bringing the intelligence of cars closer together and implementing efficient, decentralized AI processing in direct link with the entities (vehicles) with a view to having minimal or completely reduced latency guaranteeing accelerated, fluid decision-making, data protection and information that also enables instantaneous choices to be made. Here, we look at the most essential AI methods used in Edge-IoV models, including Federated Learning (FL), Machine Learning (ML), Reinforcement Learning (RL), Deep Learning (DL), and their deployment in the context of cars.

5.1. Using machine learning to manage traffic and forecasts

Machine Learning is a crucial technique for estimating and regulating traffic. It enables reliable forecasting and manage traffic flow flexibly by implementing methods such as LSTM and GNN, as well as converters to use in an approach that can anticipate traffic speed and road congestion by establishing systems that take into account factors like space and time. It also allows for the management of traffic signs and improves route choice instantaneously. Ultimately, it reduces response time for an intelligent transport system.

5.2. Deep Learning for object recognition

Deep Learning has brought many positive developments to the computer vision sector, particularly in the areas of object identification and scenario interpretation, but these skills are also important for various intelligent models, including driverless cars and robotics. Object identification involves the recognition and location of various objects in a scenario

or image. Neural networks or Convolutional Neuron Networks (CNN) and recent models such as YOLOv7 and converters are commonly used to identify and sort different entities (objects) such as cars, people and road signs in real time. The way to understand scenarios in the environment is based on the set of detections with a fragmentation of meanings and geographical links in order to decipher complicated frames, which is very important in relation to uses such as driving.

5.3. Reinforcement learning for component and displacement management

Reinforcement learning, also known as RL, has established itself as a highly effective tool for enhancing resource management and mobility in complex, evolving environments such as WI-FI networks, intelligent metropolises and advanced transportation systems. This tool enables different systems to develop the best ways of making decisions by interacting with their contexts, without first establishing the rules of marked datasets. In the field of resource management, reinforcement learning methods are used for the judicious allocation of resources in terms of exchange, energy and computer components. For example, in advanced computing and WI-FI networks, reinforcement learning is capable of dynamically upgrading the maximum throughput, processing intensity and spectrum utilization in a scalable way according to needs in a given context, taking into account network circumstances, thus maximizing transmission flow, reducing delay and minimizing energy consumption with regard to movement, RL is much more important for coordinating a number of tasks, in this case routing information at entity (vehicle) level. This technique includes multiple independent agents, such as networked cars and base sites, supporting coordinated decision-making and system operation. As the environment is constantly changing, the actions of different users changes too, while RL offers robust, adaptable and automatic coordination approaches, which are becoming increasingly important in a context where some maximization processes are focused on fixed rules that are inadequate due to instability, variations and the need for instantaneous response.

5.4. Federated training for cooperation that preserves protection

Federated Learning (FL) is an independent tool that enables multiple devices to adopt a common architecture, avoiding the need to share original information (raw data). The use of this approach makes it possible to take into account the evolving uncertainties linked to the personal protection of data and information, an aspect which is relevant within sensitive sectors such as the medical industry, the financial industry and independent devices. In the context of Federated Learning, every user forms an architecture on his or her own equipment by exploiting personal data, and can only transmit

adjustments or system parameters to centralized computers, also known as servers or converters. These adjustments are put together to optimize the overall system, while the original information always remains in the local device. This technique greatly attenuates the possibility of an intrusion capable of causing a data leak, and ensures compliance with privacy protection regulations such as the General Data Protection Regulation (GDPR). FL promotes data-protection cooperation for dispersed information resources, and enables safe, flexible AI in an increasingly connected environment. FL is much more widely used for forecasting in the sense of mobile keyboards for intelligent traffic infrastructures. In this study, we opted for the Federated Learning technique and the Deep Learning technique, as both fit with the model we have developed and proposed.

6. Installation and Peripheral System

It is imperative to set up a system of IoV computing peripherals, so that simultaneous analyses can be carried out and feedback can be provided in the shortest possible time, enabling concrete, intelligent choices to be made. In addition, the evolution of technology has opened up an important avenue for vehicles to become autonomous, which means that resources are consumed, relying solely on central digital infrastructures (Cloud), no longer able to operate efficiently due to response times, transmission limits and issues relating to the security of resources and data. Edge computing overcomes these restrictions by placing computing capacities in different locations where end-users are located (this could be in the heart of vehicles, in RSUs and also, in mobile servers).

Components of the edge computing model: Peripheral architecture includes the computer programs and the various hardware components that make up the network, enabling the information processing process to take place close to the original source. All these components must be in communication to meet functional requirements. These components can be classified according to their functions as presented below:

1. **Peripheral equipment:** This is the starting points for data and information, and includes the following;
 - ✓ IoT/IoV sensor (climate, pulse, environment)
 - ✓ IP/CCTV camera (surveillance video)
 - ✓ Wearable device (smart watch, tracking device)
 - ✓ Electronic equipment (smart speaker)

These entities produce large volumes of data that need to be analyzed locally.

2. **Peripheral points/transition systems:** These are transition intervention systems which;
 - ✓ Organizes and reinforces information derived from different digital devices.

- ✓ Executes and performs a basic analysis on information or inference based on artificial intelligence.
 - ✓ Transfer data and information towards the Cloud system.
 - ✓ Implement data protection rules as well as information.
3. **Edge/Micro-Server data center:** this is a robust piece of IT equipment located close to our peripherals.
 - ✓ Host different activities (machine learning models)
 - ✓ Résiste à la redondance et à la commutation automatique
 - ✓ They frequently include on-site data backup and virtual machine set-up.

Possibility of being housed in telecoms antennas and maintenance centers, as well as in installations or mobile units.

4. **Networking and interconnection between entities**
 - ✓ 5G and LTE technology: wireless transmission with reduced transmission times
 - ✓ Ethernet and fiber optic support: for high-speed connections.
 - ✓ LPWAN: To ensure IoT(IoV) transmission for low-cost energy use and guaranteed range.

Optimized routing has been seen as part of the network coordination options for devices, as well as wireless transmission (SD-WAN) and compliance with QoS policy.

7. Modeling Practical Case Studies

In the Internet of Vehicles (IoV), modeling and simulation scenarios offer a concrete perspective on the viability, efficiency and adaptability of artificial intelligence-based edge computing for IoV technology. In this section, we discussed all practical applications, as well as testing and verification during simulations, with the aim of demonstrating the importance of cooperation between Artificial Intelligence and Edge computing in maximizing latency reduction, accurate decision definition and bandwidth utilization.

Scenario Edge computing and AI

This involves setting up a shared Artificial Intelligence model across an urban agglomeration using devices installed at different roadside locations (RSUs) and 5G mobile computing (server) backup equipment (MEC), as shown in the following diagram:



Figure 7.1. Architecture Edge computing and AI

In this implementation, the AI techniques put into practice include cooperative learning for collaborative training of network systems, Convolutional Neural Networks (CNN) for entity (object) location, and Graphical Neural Networks (GNN) for traffic prediction. This will enable us to improve traffic management based on the following model.

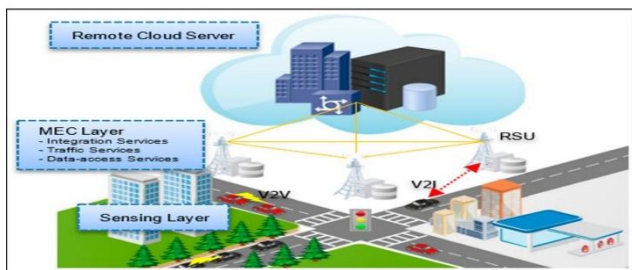


Figure 7.2: Optimization of traffic management in the metropolitan area through RSU AI performance

How it works: Automobiles perform a local analysis of basic Artificial Intelligence, RSUs compile and manage updates of regional or local structures, and finally, MEC infrastructures, more precisely MEC servers, which carries out a global reorganization process and re-share updated models. The results are as follows:

- ✓ Approximately, there has been a reduction in time of about 28 percent over a period of time until the system reaches a solution.
- ✓ Waiting time has improved, as it has been reduced from 92 ms (strictly at Cloud level) to 27 ms (taking peripherals into account).
- ✓ Reducing the size of system updates has saved 34% in bandwidth usage.

Decentralized Artificial Intelligence on the periphery preserves confidentiality while remaining capable of achieving high levels of definition or precision in studies for functionalities fundamental to protection.

8. Protection and Confidentiality

The IoV environment supports and focuses on Artificial Intelligence (AI), and thus, depends on Mobile Edge Computing (MEC) technology for its operation. Given the sheer volume of data created, analyzed and shared by automobiles instantaneously, equipment such as Roadside Units (RSUs) and peripheral servers pose major challenges in terms of tasks and confidentiality. In parallel with peripheral computing systems that allow for the creation of modules capable of large- scale verification, edge computing configuration presents a vulnerable, distributed risk surface, requiring precise solutions and processing to guarantee the principle of CID (Confidentiality, Integrity and Availability).

8.1. Security risks in the IoV or Artificial Intelligence environment

In the environment of edge-linked IoV systems using Artificial Intelligence, we have noticed a rise in protection risks emanating from both networks and the various AI zones. Factors encompassing system-related dangers include man-in-the-middle (eavesdropping), identity theft, denial of service as well as cyber-attacks and intrusions, as well as other factors affecting V2X exchanges. We need to note that, in relation to the subject of AI. Certain deceptive cases have the potential to damage perception systems, deformed architectures risk hampering the evolution of federated learning, there may be hacking of information and confidential data due to intrusions by means of inference. Peripheral links, such as RSUs and vehicles, are exposed to the risks of hardware exploitation and the addition of malicious tools. The high mobility of these devices contributes to the complexity of validation, sustainable verification and trust management.

8.2. Privacy Policy

In many IoV-AI-Edge architectures, confidential data and information, such as precise car locations, travel behavior, sensor information and individual passenger identification, are broadcast at all times and distributed across the connected cars. Signaling or roadside communication units (RSUs) and backup machines at the network's periphery. During the traffic flow examination period, lack of robust protection can lead to information merging illegally or what is referred to as system renovation. Another problem caused by traffic flow is the exposure of information. As a result of federated and collaborative learning, threats to private identity have increased. Variations and adjustments to architectures are likely to reveal confidential data. In order to ensure the security of privacy, we need to implement a number of techniques, such as discretion, secure grouping, flexible confidentiality and the application of rigorous rules for managing the various useful connections.

8.3. Protection systems for IoV at AI scale

IoV-AI-Edge systems are based on several protection measures, including PKI certification, public key

infrastructure authentication, secure V2X communication encryption and function-based access control. Immediate identification of threats or risks is made possible by the use of an intrusion detection tool in the peripheral zone, as well as AI-guided observations and analysis of disturbances. The implementation of robustness techniques has all the antagonistic training, secure Federated Learning as well as flexible protection which helps to combat aggression within the models, yet different tamper-proof equipment then protected startups that support nodes and hardware hence all of these elements increase and improve system robustness, privacy preservation as well as reliability within interconnected vehicle infrastructures.

8.4. Privacy technology

IoV systems rely on AI, and with the development of AI, it is possible to keep information private by concealing personal information using pseudonymous identifications and certifying rigorous connections to reduce the dissemination and vulnerability of confidential data. Federated Learning, combined with a protected collected and privacy rules, avoids the transfer and communication of original information in time, and ensures adjustments to architectures. Various encryption strategies, such as Uniform Encryption, make it possible to perform calculations on encrypted information, while controlling it, or carrying out in-depth audits based on the Block Chain, ensuring efficient coordination and clear, reliable, tamper-proof information management. By combining these different approaches, we have succeeded in striking a balance between the importance of private information and rigorous privacy security.

9. Suggestions for Future Research

This study has made an in-depth analysis and the results have indicated the capabilities of AI and MEC within the IoV environment, although it should be noted that there are still many diverse areas of study which it would be important to explore further. Here is a list of the future work we have proposed.

1. With AI-Edge, the deployment of these two methods amplifies the demand for resources, and then energy use becomes increasingly intensive. Future work should focus on building learning systems that are less robust in terms of energy consumption, but exploit optimized neural networks and multiple processing devices capable of running on renewable energy generators (Ahmed S And Kim H, 2022).
2. Putting test models into practice in real-life situations: much research is based on several models and virtual trials. However, concrete implementation of large-scale IoV environments with AI remains a less common practice, and it is important that future work focus on practical implementation in the real

world, using different technological environments such as NVIDIA, Jetson, Huawei or even infrastructures (Servers) that support 5G technology, including the MEC system, with the aim of better analyzing results taking into account real conditions and car movements to confirm the impact of the results obtained (Hui Z, Jian X And Feng Z, 2023).

3. Despite Federated Learning's protection of data and privacy. IoV remains exposed to cyber-attack threats as well as contamination of algorithms that can compromise their reliability and performance. Future projects must be based on solid techniques to protect AI, such as flexible security and block chain verification systems (Li Y, Huang T And Zhao Q, 2023).
4. Despite Federated Learning's protection of data and privacy. IoV remains exposed to cyber-attack threats as well as contamination of algorithms that can compromise their reliability and performance. Future projects must be based on solid techniques to protect AI, such as flexible security and block chain verification systems (Wang Z , Li Y, And Zhang K, 2024).

10. Conclusion

As we come to the end of our study, we can say that we are satisfied with the work we have done, while respecting the ethics of science. This research has examined in depth, how AI and MEC technology can be combined within the IoV environment, aiming to alleviate or eradicate major problems in IoV environments such as response times, bandwidth limitations, data and information security, instantaneous decision-making in logistics and intelligent traffic devices. This paper has shown that the incorporation of AI and the MEC system, taking into account all the infrastructures on the periphery, has the potential to bring an additional point capable of considerably optimizing the connection and exchanges in a vehicular architecture, allowing applications related to safety and efficiency to remain optimal. In this article, we have highlighted the challenges arising from traditional Internet of Vehicles techniques based on cloud computing, notably long response times and susceptibility to system overloads. It has been shown that edge computing offers a decentralized architecture based on the exploitation of data and information close to the final entities (vehicles), which helps to reduce response times while reinforcing robustness. In parallel, Artificial Intelligence techniques such as Deep Learning for detection, Reinforcement Learning for orientation and Federated Learning for building architectures while protecting confidentiality, are important and can be used to promote rapid decisions and anticipatory calculations. A model was suggested for mixing AI and Edge Computing technology within automotive systems (vehicle networks)

supported by simulations and the results confirmed that this mixer method favors reduced waiting time, optimized resource utilization, improved traffic forecast accuracy and also enhanced car protection compared to conventional techniques. Aspects such as security and privacy were also addressed in this research, highlighting that Federated Learning and techniques based on Block Chain technology offer favorable prospects for protecting data and information flows directly linked to vehicles from digital threats and malicious attacks. In this work we have listed the future steps that need to be taken to improve the capabilities of the IoV environment. As a contribution to this field, we have highlighted and demonstrated that AI techniques and MEC technology are not just simple elements that complement each other, but also, beneficial and indispensable factors in advancing the IoV field. The inclusion of these two innovative, adaptable, constant, intelligent and protected vehicle network architectures will influence the outlook for traffic, transport and intelligent routing.

References

- [1] Ahmed S And Kim H. (2022, May 17). *Green Edge intelligence for IoV : Emergence- aware AI Models and Renewable-Powered Edge Servers*. Retrieved from IEEE Xplore Digital Library: <https://doi.org/10.1109/ACCESS.2022.3201234>
- [2] Ahmed, M., Rehman, S. U., and Hussain, F. (2022, August 29). Edge-Enabled Vehicular Computing: Opportunities, Challenges, and Future Directions. *Future Internet.*, 14(9):257. Retrieved from Future Internet: <https://www.mdpi.com/1999-5903/14/9/257>
- [3] GSMA. (2025). *The Mobile Econmy*. London: 2025.
- [4] Hui Z, Jian X And Feng Z. (2023). *IEEE Internet of Things Journal*. Retrieved from IEEE Explorer Digital Library: <https://doi.org/10.1109/JIOT.2023.3245678>
- [5] Li T, Xu C, Zhang Y, Chen M. (2023, January 24). Integrated Sensing, Communication and Computation for IoV Toward 6G . *IEEE Xplore*, 27-42. Retrieved from <https://doi.org/10.1109/TITS.2022.3178941>
- [6] Li Y, Huang T And Zhao Q. (2023). Blockchain-enhanced federated learning for secure edge-assisted IoV. *IEEE Transactions on Intelligent Transportation Systems*, 6789–6802. Retrieved from <https://doi.org/10.1109/TITS.2023.XXXXXXX>
- [7] Nirvikar k, Abhay S, Namita C, Raju S, Sudhir K, Mohd, Faraz Husain. (2024, 4 30). AI in Autonomous Vehicles: Opportunities, Challenges,. *Educational Administration: Theory and Practice*, 6255-6264. Retrieved from Educational Administration: Theory and Practice: <https://www.researchgate.net/publication/38083685>
- [8] Wang Z , Li Y, And Zhang K. (2024). Cooperatif Edge AI for Connected Vehicles: Design Implementation and Performance Evaluation. *IEEE Communication Surveys & Tutorials*.
- [9] Wang, Z., Li, Y., & Zhang, K. (2024). Cooperative Edge AI for Connected Vehicles: Design, Implementation, and Performance Evaluation. *IEEE Xplore.*, 12-25.
- [10] Wang, Z., Li, Y., And Zhang, K. (2024). Cooperative Edge AI for Connected Vehicles: Design, Implementation, and Performance Evaluation. *IEEE Communications Surveys & Tutorial*.
- [11] Xiaolong X, Haoyuan L, Haoyuan; X, Weijie; L, Zhongjian; Y, Liang; D, Fei. (2022, April 2). Artificial intelligence for edge service optimization in Internet. *IEEE Xplore*, 270 - 287. Retrieved from IEEE Xplore: <https://ieeexplore.ieee.org/abstract/document/9552656>
- [12] Zhou B, Liu Z, And Wang J. (2024, June). Federated Deep Reinforcement Learning for Cooperative Route Planning in Internet of Vehicles”. *IEEE Internet of Things Journal*, 123-125. Retrieved from IEEE Internet of Things Journal: <https://doi.org/10.1109/JIOT.2024.3371986>

5 AI in Autonomous Vehicles Opp

ortunities_Challenges_and_Regulatory_Implications

AI-Powered Biochemical Materials for Biomedical Applications: Current Trends and Future Prospects.

Love Edet M.

Girne American University, Cyprus.

loveedetmendie@gmail.com

Moyo Mthabisi

Girne American University, Cyprus.

mthabisimoyo@gau.edu.tr

Abstract

Traditional approaches to designing biochemical materials for biomedical use are often slow, labor-intensive, and constrained by the complexity of biological systems. This creates a critical gap between experimental material development and clinical translation. Artificial Intelligence (AI) now offers transformative solutions to this challenge by enabling predictive modeling, generative design, and rapid screening of novel biomaterials with tailored properties. As of 2025, AI is actively being used to design adaptive hydrogels, simulate protein-material interactions, and develop patient-specific drug delivery systems. Platforms like AlphaFold and generative neural networks are driving forward the rational design of bioactive materials, while AI-integrated high-throughput screening pipelines are drastically shortening the material discovery timeline. These innovations are enabling personalized, efficient, and scalable solutions for regenerative medicine, biosensing, and immune-modulating therapies. However, challenges remain in integrating heterogeneous biological data, ensuring model interpretability, and validating clinical applicability. Addressing these gaps requires the development of standardized datasets, explainable AI models, and multidisciplinary collaboration. This study will review the current state of AI-driven biochemical material design, illustrate success stories, and highlight future opportunities in self-evolving biomaterials, closed-loop therapeutic systems, and intelligent interfaces between synthetic and living systems. AI is not just accelerating materials discovery—it is redefining the possibilities in biomedical innovation.

Keywords

Artificial Intelligence, Biochemical Materials, Biomedical Applications, Smart Biomaterials, Personalized Medicine

1. Introduction

Biochemical materials are materials engineered to interact with biological systems at the molecular and cellular levels (Jang et al., 2019). These materials which include polymers, peptides, proteins, hydrogels, and bioactive nanomaterials, are central to modern biomedical innovation and are made to support applications in drug delivery, regenerative medicine, biosensing, immune modulation and others (Hench & Polak, 2002; Ratner & Bryant, 2004). Unlike traditional synthetic materials, biochemical materials are often designed to mimic biological processes, degrade in physiologically compatible

ways, or integrate with living tissues to restore function (Freedman & Mooney, 2019; Huang et al., 2017). Such properties make them essential to advancing precision medicine and sustainable healthcare solutions. Although, synthesizing biomaterials traditionally involves trial-and-error, high-throughput screening, and computational simulations (Bai & Zhang, 2025; Jiang et al., 2025), which are time-consuming and resource-intensive due to complex biological systems' non-linear interactions and dynamic microenvironments (Zhang et al., 2017). Additionally, transitioning lab prototypes to clinically validated products is difficult, creating a translational gap that hinders timely healthcare innovation (Capella-Monsonis et al., 2024). To overcome these challenges in designing and optimizing biochemical materials, Artificial Intelligence (AI) has emerged as a transformative solution by leveraging machine learning, deep learning, reinforcement learning, and generative models. AI enables predictive modeling of structure–property relationships, simulation of material–biological interactions, and rational design of novel biomaterials (Butler et al., 2018; Schmidt et al., 2019). For instance, AI algorithms have been successfully used to predict the biocompatibility of polymers, design adaptive hydrogels with tunable stiffness, and model protein–material interactions with near-atomic precision (Xie & Grossman, 2018; Jumper et al., 2021). These innovations are accelerating discovery timelines and reducing reliance on exhaustive experimental trials.

The integration of AI also supports the emergence of personalized biomedical materials, where patient-specific data can guide the design of drug carriers, scaffolds, or biosensors optimized for individual therapeutic needs (Topol, 2019; Schork, 2019). Moreover, AI-powered platforms such as AlphaFold have revolutionized structural biology, directly influencing how biomaterials are designed to interact with proteins, enzymes, and cellular receptors (Jumper et al., 2021). Also, generative AI models are enabling the exploration of previously untested material design spaces, expanding the horizon of what is chemically and biologically feasible (Pugliese et al., 2025).

However, the use of AI is not without its challenges such as the scarcity of standardized biomedical datasets, algorithmic bias, lack of interpretability, and the integration of

heterogeneous biological information, which continue to constrain the widespread adoption of AI in the synthesis of biochemical materials (Carleo et al., 2019; Karniadakis et al., 2021; Jiang et al., 2025). Overcoming these barriers requires not only technical innovation but also interdisciplinary collaboration among material scientists, computational

biologists, clinicians, and regulatory agencies. Figure 1 shows the conceptual AI-powered biochemical material design pipeline, illustrating the flow from biological datasets through AI models to predicted properties, experimental validation, and biomedical applications.

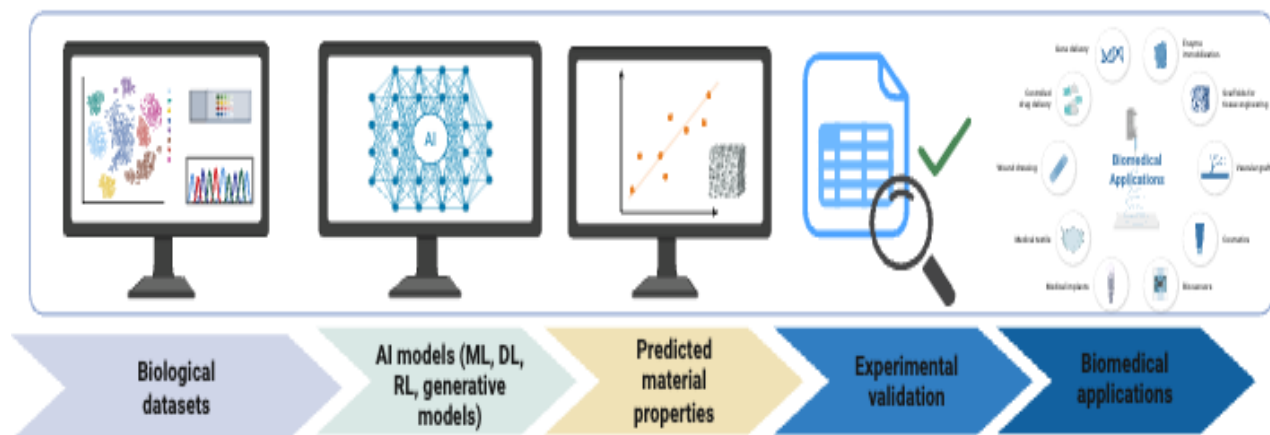


Figure 1: Conceptual diagram of an AI-powered biochemical material design pipeline

This paper aims to provide a comprehensive review of the current trends and future prospects of AI-powered biochemical materials for biomedical applications. Specifically, it will highlight how AI is being integrated into the design, screening, and application of materials, analyze success stories and case studies, critically evaluate integration challenges, and explore emerging directions such as self-evolving biomaterials and intelligent human–material interfaces. By synthesizing these insights, this chapter positions AI not merely as a supportive tool but as a driving force in redefining the future of biomedical material science and global healthcare innovation.

2. Overview of Biochemical Materials

Biochemical materials are a unique class of engineered substances designed to interact closely with biological systems while maintaining tunable physicochemical properties such as stiffness, degradation rate, surface

chemistry, and porosity (Kuperkar et al., 2024; Rahmati et al., 2020). These properties are obtained by techniques such as varying crosslinking density, adding various nanoparticles, altering surfaces with particular coatings, and employing stimuli-responsive materials that alter properties in response to external factors like light or pH (Özkale et al., 2021; Rahmati et al., 2020).

Biomaterials combine principles of chemistry, biology, and materials science to produce materials capable of supporting, mimicking, or modulating biological functions (Jiang et al., 2025). Unlike traditional synthetic materials, biochemical materials are optimized for biodegradability, bioactivity, and responsiveness to biological cues, making them highly valuable in biomedical innovation and global sustainability initiatives (Hench & Polak, 2002; Ratner & Bryant, 2004).

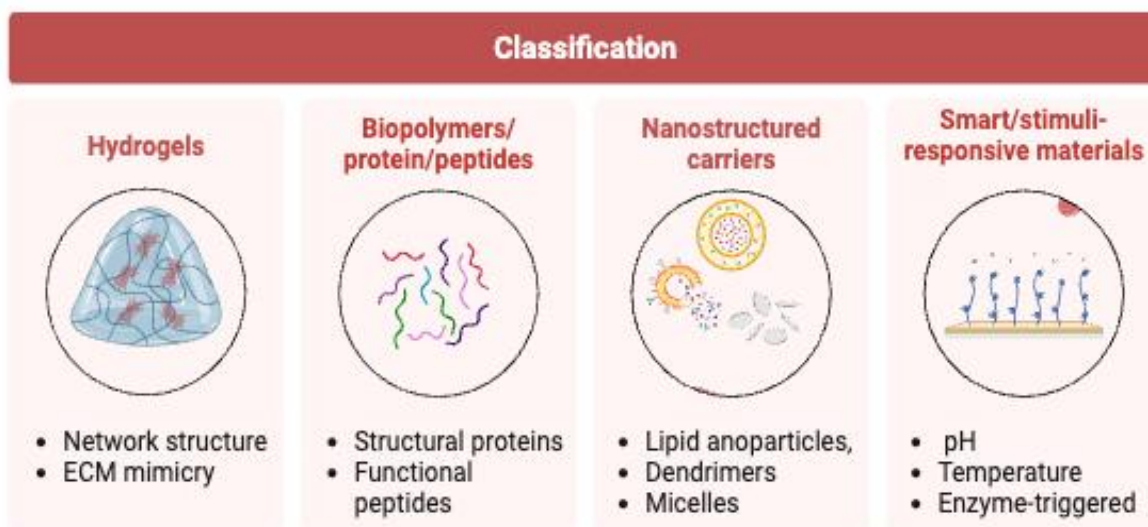


Figure 2: Classification of biochemical materials

There are diverse types of biochemical materials as shown in Figure 2, which has applications that spans through regenerative medicine, biosensing, drug delivery, immunotherapies and many more. Some of these biochemical materials are discussed below.

2.1. Hydrogels

Hydrogels are three-dimensional, water-swollen polymeric networks that resemble the extracellular matrix (ECM) (Khan et al., 2024). Since the first synthesis of Hydrogels as biomaterial in 1960 till date, hydrogel have evolved to address in situ gelation post-infection and modified responsiveness through swelling-deswelling rates, stiffness, and other properties, thus expanding their utility in diverse medical contexts beyond superficial applications (Ho et al., 2022). Their high water content and tunable mechanical properties make them excellent scaffolds for **tissue** engineering, wound healing, and drug delivery (Hoffman, 2012). Modern hydrogels can be functionalized with peptides, growth factors, or nanoparticles to enable cell adhesion, controlled release of therapeutics, and stimuli-responsive behavior (Buwalda et al., 2014).

2.2. Biopolymers

Biopolymers are type of biomaterials derived from natural sources such as collagen, chitosan, alginate, silk fibroin, and hyaluronic acid, provide structural and functional resemblance to native biological tissues (Ghosh et al., 2021). They are inherently biodegradable and often exhibit intrinsic bioactivity, such as antimicrobial or hemostatic properties (Sionkowska, 2011). Advances in biopolymer modification and blending have expanded their utility in biodegradable packaging, medical implants, and regenerative medicine (Kumar et al., 2020).

2.3 Protein- and peptide-based materials

Proteins and short peptides offer molecular precision in biochemical material design, allowing researchers to engineer scaffolds with specific cell-binding motifs, mechanical strength, or immunomodulatory properties (Zhang, 2017). Self-assembling peptides, for example, can form nanofibrous hydrogels that support cell differentiation, nerve regeneration, or targeted drug delivery (Matson & Stupp, 2012). Protein-based sensors are also used to enable rapid and sensitive detection of biomarkers.

2.4. Nanostructured biochemical materials

Nanomaterials which includes lipid nanoparticles, polymeric nanocarriers, and inorganic–organic hybrid system, play a critical role in biosensing, imaging, and targeted therapeutics (Kazi et al., 2025). Their high surface-to-volume ratio allows efficient functionalization with drugs, proteins, or nucleic acids which ensure site-specific delivery of therapeutics while minimizing off-target effects (Ly et al., 2024; Kurul et al., 2025). Recent advances in stimuli-responsive nanocarriers have enabled controlled release triggered by pH, enzymes, or light, thereby improving therapeutic precision while reducing systemic toxicity (Peer et al., 2007; Choi & Frangioni, 2010). Biochemical nanomaterials can also act as immune modulators or vaccine delivery platforms, enhancing antigen presentation and immune responses.

2.5. Smart and stimuli-responsive materials

Smart biochemical materials are intelligent biomaterials designed to react to stimuli like light, moisture, stress, or specific biochemical signals, making them suitable for tissue engineering and regenerative medicine (Ma et al., 2025). They address challenges such as targeted medication delivery, improving cell adhesion and growth, and controlling scaffold degradation through their dynamic interactions with biological systems (Karunakar et al., 2025). Stimuli-responsive materials are types of smart biomaterials that

demonstrate changes in physical and chemical properties under external signals, with examples like pH-responsive chitosan, which alters its structure through protonation and deprotonation processes at its amino terminal (Ma et al., 2025). These materials are not only gaining attention in biosensing devices, wearable medical technologies, and closed-loop drug delivery systems (Stuart et al., 2010), but are also have emerging as transformative tools at the interface of biotechnology and regenerative medicine (Karunakar et al., 2025).

3. Fundamentals of AI in Materials Science

AI provides computational frameworks that accelerate discovery, design, and validation of biochemical materials. Several paradigms are particularly relevant to material science, some of which are discussed below:

3.1. Machine learning, deep learning, and reinforcement learning

Machine Learning (ML), an important branch of AI, has shown to be useful for automating data analysis in a time-efficient and reproducible manner, especially for data that are too large and complex for human analysis (Greener et al., 2022). Studies have proven that ML has evolved as a cornerstone in biochemical material discovery because of its ability to predict material properties, optimize synthesis conditions, and accelerate high-throughput screening (Fu et al., 2025). In addition, ML has been successfully used in biomedical sector for several analysis such as gene recognition, biophysical cue screening, medical image analysis and protein structure prediction (Chen et al., 2023). Models such as support vector machines, decision trees, and random forests have been widely employed to correlate material composition with biocompatibility, degradation rate, and drug-release kinetics (Butler et al., 2018). Although ML holds significant promise, the black box nature of ML algorithms continues to obstruct its interpretability (Petch & Nelson, 2022). As a result, the application of ML in developing suitable biomaterials is lacking. Hence, it is crucial to overcome this limitation and harness the capabilities of ML in analyzing and synthesizing biomaterials for medical uses

Deep Learning (DL), a sub-sector of ML which leverages neural networks with multiple hidden layers, extends this predictive capability to more complex datasets (Taye, 2023). Convolutional Neural Networks (CNNs) are particularly effective in analyzing imaging data from microscopy or histology, while Graph Neural Networks (GNNs) capture molecular and structural relationships within biochemical materials (Xie & Grossman, 2018). DL has been shown to replicate density functional theory (DFT)-level predictions at a fraction of the computational cost, significantly accelerating virtual screening (Schmidt et al., 2019).

Reinforcement Learning (RL) introduces adaptive, trial-and-error decision-making into material design (Fu et al., 2025;

Hang et al., 2025). Here, an agent interacts with a simulated or experimental environment, receiving “rewards” for achieving target outcomes, such as stability, porosity, or bioactivity of a material (Fu et al., 2025). Recent studies have demonstrated that RL can guide inverse material design, where target properties drive structural discovery rather than trial-based exploration (Zhou et al., 2019).

3.2. Generative AI for material design

Two types of Generative AI: Generative Adversarial Networks (GANs) and Transformer-based models, offers an innovative approach to material discovery by creating new molecular or structural designs beyond existing datasets. GANs have been applied to produce novel photonic metamaterials and polymer backbones optimized for biomedical use (Dan et al., 2020). Similarly, Variational Autoencoders (VAEs) generate continuous latent spaces that allow interpolation between known material structures, enabling discovery of intermediate designs with unique properties (Wei & Mahmood, 2020).

More recently, Transformer-based models, inspired by breakthroughs in natural language processing, have been adapted to biochemical and material sciences as noted in various areas such as protein sequences, biomedical textual data, and biomedical images and graphs (Madan et al., 2024). Models such as AlphaFold demonstrate the power of transformers in predicting protein structures with atomic-level accuracy (Jumper et al., 2021). In materials science, transformer architectures are being used to propose new crystal structures, polymer sequences, and peptide-based scaffolds tailored for regenerative medicine and biosensing (Madan et al., 2024). These methods are especially valuable for biomedical applications where patient-specific customization of biomaterials is increasingly important.

3.3. Data-driven vs. Physics-informed models

Traditional data-driven models rely solely on empirical or computational datasets. While effective for pattern recognition and prediction, they often suffer from limitations such as data scarcity, bias, and lack of interpretability (Carleo et al., 2019). For example, predicting the stability of a newly designed hydrogel purely from empirical data can lead to physically implausible outputs when data coverage is insufficient. Physics-Informed Models address this limitation by embedding domain knowledge such as thermodynamic laws, reaction kinetics, or differential equations, into AI architectures. Physics-Informed Neural Networks (PINNs) integrate governing equations into the learning process, ensuring that predictions remain consistent with known physical and biochemical laws (Raissi et al., 2019). Hybrid approaches that combine data-driven learning with physics-based constraints are emerging as a powerful paradigm in AI-powered materials research. These models not only improve predictive accuracy but also enhance **trustworthiness**, a critical factor for biomedical applications where safety and reliability are non-negotiable (Karniadakis et al., 2021).

4. Current Applications of AI in Biochemical Materials

The integration of AI into biochemical material design is rapidly transforming the landscape of biomedical innovation (Junaid, 2025). By leveraging computational models,

predictive algorithms, and data-driven insights, AI allows researchers to design, optimize, and evaluate materials for biomedical applications more efficiently than traditional trial-and-error methods (Junaid, 2025). The various biomedical applications are illustrated in Figure 3.

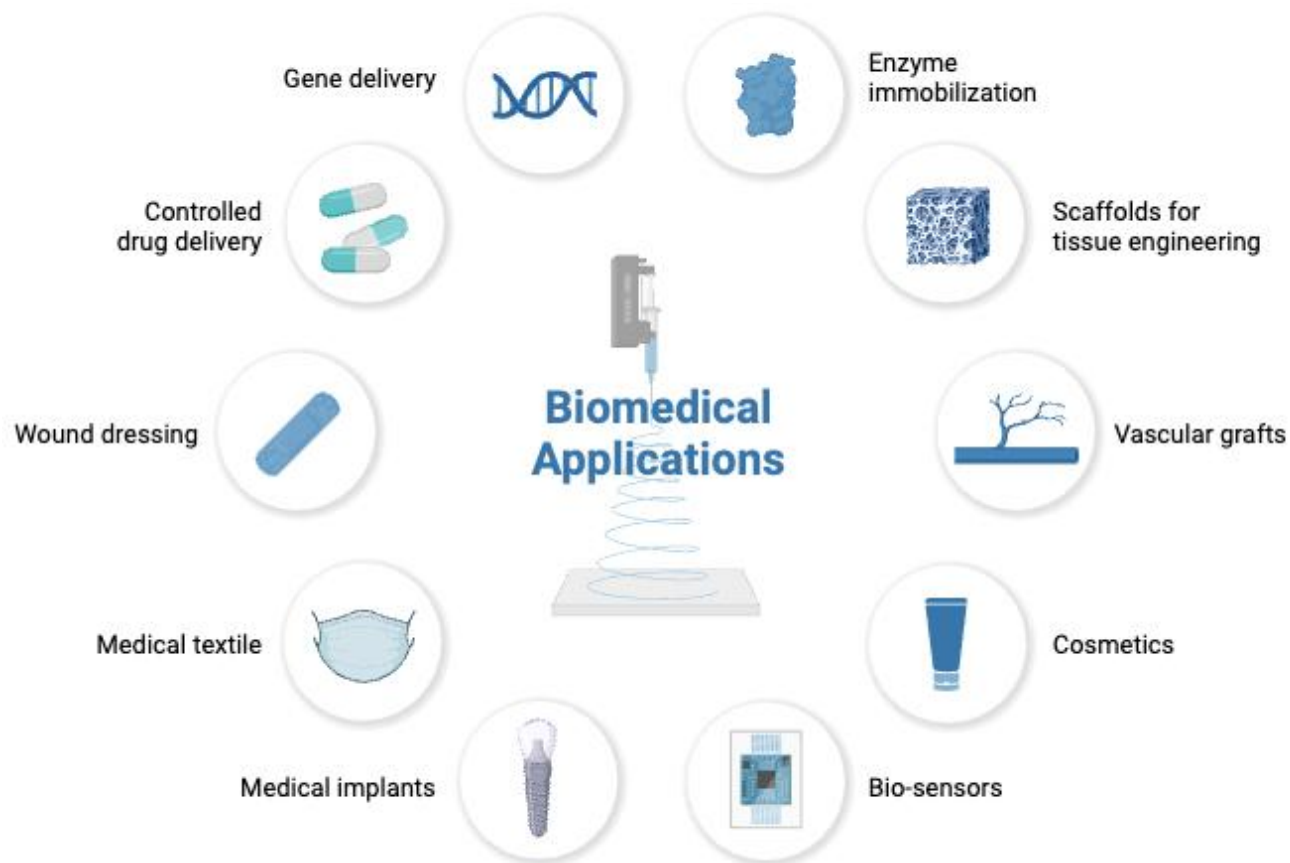


Figure 3: Biochemical applications of AI-powered biomaterials

4.1. AI for structural design

The structural design of biochemical materials is a cornerstone for their functionality in biomedical applications (Junaid, 2025; Parvin et al., 2025). Traditionally, materials were developed empirically, with iterative synthesis and experimental characterization forming the backbone of discovery (Parvin et al., 2025). AI introduces a paradigm shift by enabling predictive modeling, generative design, and *in-silico* optimization of biomaterial structures (Parvin et al., 2025).

Generative AI methods, including generative adversarial networks (GANs) and transformer-based architectures, have shown remarkable potential in proposing novel biomaterial structures with tailored properties (Pugliese et al., 2025). For example, hydrogels water-swollen polymeric networks widely used in tissue engineering can be computationally designed with specific mechanical stiffness, porosity, and swelling kinetics to meet application-specific requirements (C. W. Zhang et al., 2025). Machine Learning models trained

on existing hydrogel libraries can predict how changes in polymer composition, crosslinking density, or environmental conditions affect performance, enabling rapid *in silico* screening of candidates before experimental validation (C. W. Zhang et al., 2025).

Peptides and polymers are also being optimized through generative design. AI models can propose sequences that maximize biocompatibility, stability, and functionality, which are critical parameters for protein-based scaffolds and drug carriers (Chen et al., 2024). The result is a significant reduction in experimental burden, accelerated discovery timelines, and improved reproducibility across laboratories (Chen et al., 2024; Pugliese et al., 2025).

Understanding and predicting protein-material interactions is essential for designing biomaterials that can interact seamlessly with biological systems. AI platforms such as AlphaFold have revolutionized protein structure prediction, allowing researchers to model the folding patterns and interaction sites of proteins with high accuracy (Perrakis &

Sixma, 2021). When combined with molecular docking simulations and machine learning hybrids, these approaches can predict how proteins will interact with synthetic scaffolds or hydrogels, enabling rational design of bioactive interfaces (Callaway, 2022).

AI-based docking and ML algorithms also facilitate prediction of binding affinities, stability, and conformational dynamics of protein-material complexes (Parvin et al., 2025). This capability is particularly valuable for designing materials that modulate immune responses, promote cell adhesion, or mimic extracellular matrices (Callaway, 2022). By providing predictive insights into these interactions, AI bridges the gap between theoretical modeling and experimental synthesis, accelerating the translation of materials into biomedical applications (Junaid, 2025).

4.2 AI in drug delivery and therapeutics

AI has transformed the design of drug delivery systems by enabling patient-specific optimization, enhanced targeting, and controlled release kinetics (Vora et al., 2023). Traditional drug delivery development relies heavily on iterative testing of formulations, which can be time-consuming and expensive. AI introduces predictive and generative capabilities that streamline this process

Personalized medicine demands delivery systems that can adapt to individual patient profiles, including genetic background, disease state, and metabolic rate (Park et al., 2023). AI models can integrate patient data to predict optimal drug dosages, release schedules, and carrier formulations (R. C. Wang & Wang, 2023). Machine learning algorithms can analyze pharmacokinetic and pharmacodynamic parameters to tailor drug delivery systems, reducing adverse effects and enhancing therapeutic efficacy (Goetz & Schork, 2018).

4.2.1. Nanocarriers optimized by AI

Nanocarriers including polymeric micelles, lipid nanoparticles, and dendrimers are widely used for targeted drug delivery (Pugliese et al., 2025). AI algorithms can optimize their design by predicting particle size, surface charge, and release kinetics based on intended application (Junaid, 2025). Reinforcement learning models can dynamically adjust carrier composition to achieve controlled release profiles, enhanced tissue penetration, and minimal off-target effects. By simulating multiple scenarios in silico, AI significantly reduces experimental iteration, expediting preclinical and clinical development (Parvin et al., 2025).

4.3. AI-Driven Tissue Engineering and Regenerative Medicine

Tissue engineering and regenerative medicine rely on biomaterials that can support cell growth, differentiation, and tissue regeneration (Raghavendra et al., 2015). AI has become an essential tool in designing adaptive scaffolds, self-

healing hydrogels, and dynamic extracellular matrices (Gharibshahian et al., 2024).

4.3.1. Adaptive and self-healing hydrogels

Hydrogels engineered for tissue regeneration must often respond to changing biological environments, such as fluctuating pH, mechanical stress, or enzymatic activity (C. W. Zhang et al., 2025). AI-driven optimization allows for the design of adaptive and self-healing hydrogels that maintain structural integrity and biological functionality over time. Predictive models can simulate hydrogel behavior under physiological conditions, providing insights into swelling, degradation, and mechanical properties (Hong et al., 2023).

4.3.2. AI-guided scaffold optimization

Scaffolds guide cell attachment, proliferation, and differentiation, and their microarchitecture is critical for tissue regeneration. AI tools can model scaffold porosity, fiber alignment, and mechanical gradients to optimize cell seeding efficiency and nutrient diffusion. Coupled with 3D bioprinting, these algorithms facilitate the rapid production of customized scaffolds tailored to patient-specific tissue defects (P. Zhang et al., 2024). Integrating AI with computational fluid dynamics and biomechanical simulations further enhances scaffold design by predicting nutrient transport and mechanical stress distributions in engineered tissues (P. Zhang et al., 2024).

4.4. AI in Biosensing and Diagnostics

Biosensing and diagnostic applications benefit significantly from AI's capacity to analyze complex datasets, detect subtle molecular interactions, and enable real-time monitoring (Wasilewski et al., 2024; P. Zhang et al., 2024).

4.4.1. Intelligent biosensors

AI-enabled biosensors incorporate machine learning algorithms to detect and quantify biomolecular interactions with high sensitivity and specificity (Jin et al., 2020; Wasilewski et al., 2024). For example, wearable sensors integrated with AI can monitor glucose, biomarkers of inflammation, or metabolic indicators in real time, providing continuous health monitoring (Paul, 2025). By analyzing patterns in sensor outputs, AI models can detect early disease signatures that may be missed by traditional analytical methods (Paul, 2025).

4.4.2. AI-enhanced detection of molecular interactions

AI algorithms enhance the performance of diagnostic assays by interpreting complex spectroscopic, electrochemical, or imaging data. Deep learning models can classify signals, identify anomalies, and predict molecular binding events, leading to faster and more accurate diagnostics (Huang et al., 2024). This capability is particularly relevant for early detection of cancers, infectious diseases, and metabolic disorders, where sensitivity and speed are crucial (Huang et al., 2024).

5. Success Stories and Case Studies

AI has already demonstrated tangible impacts on the development of biochemical materials, bridging the gap between theoretical design and experimental validation (Singh et al., 2020). Landmark examples illustrate how AI-enabled approaches outperform traditional trial-and-error methods, streamline workflows, and open new avenues for personalized and precision medicine (Singh et al., 2020).

5.1 AlphaFold in protein design

Protein-based biomaterials are central to applications ranging from tissue scaffolding to drug delivery, but their design is inherently challenging due to the complexity of protein folding and interactions (Jumper et al., 2021). Traditionally, protein engineering relied on rational design based on known motifs or extensive experimental screening, which is both time-intensive and limited by the available structural knowledge (Lutz, 2010).

AlphaFold, developed by DeepMind, represents a transformative milestone in computational protein modeling. By leveraging deep learning to predict protein structures from amino acid sequences with near-experimental accuracy, AlphaFold has dramatically accelerated the design of protein-based biomaterials. For example, peptide scaffolds can now be designed to mimic extracellular matrices or modulate immune responses with a high degree of precision. In comparative studies, AlphaFold-enabled design reduced the time from conceptualization to functional protein prediction from months to days, while maintaining or improving accuracy over traditional homology modeling approaches. Its success has not only streamlined experimental validation but also opened possibilities for designing novel protein sequences that were previously unattainable (Jumper et al., 2021).

5.2 AI-enabled hydrogel optimization

Hydrogels are widely used in regenerative medicine, drug delivery, and biosensing due to their tunable mechanical and chemical properties. Traditional hydrogel design often involves iterative experimentation to optimize crosslinking density, polymer composition, and swelling behavior, a process that can require hundreds of trials to achieve desired properties.

AI-driven hydrogel optimization employs machine learning models to predict the relationships between polymer composition, fabrication parameters, and functional outcomes such as stiffness, degradation rate, and swelling kinetics (C. W. Zhang et al., 2025). Generative AI approaches, including neural networks and GANs, have been

used to propose hydrogel formulations with tailored characteristics (Negut & Bitá, 2023). In landmark studies, AI-guided hydrogel design reduced experimental iterations by over 70%, enabling rapid identification of optimal formulations for specific biomedical applications (Urifa & Shah, 2025). For instance, researchers successfully designed hydrogels that simultaneously achieved high mechanical strength and rapid swelling, which was previously difficult using conventional trial-and-error methods (Negut & Bitá, 2023; Urifa & Shah, 2025).

Moreover, AI models can simulate environmental responses, such as pH or enzymatic activity, predicting how hydrogels will behave under physiological conditions (C. W. Zhang et al., 2025). This predictive capability improves reproducibility, reduces material waste, and accelerates translation to preclinical models.

5.3 AI-Assisted High-Throughput Screening of Biomaterials

High-throughput screening (HTS) allows researchers to evaluate hundreds to thousands of biomaterial candidates in parallel (Yang et al., 2021). Traditionally, HTS involves labor-intensive preparation, characterization, and analysis, making it expensive and slow (Decker et al., 2018). Integration of AI into HTS workflows has transformed this process by enabling predictive pre-screening, automated data analysis, and intelligent experimental design (Yang et al., 2021).

In AI-assisted HTS, machine learning algorithms analyze large datasets from prior experiments to predict which candidates are most likely to succeed, effectively narrowing the search space before experimental testing (Yang et al., 2021). For example, in polymer and peptide libraries, AI models can predict properties such as biocompatibility, degradation rate, and mechanical performance, prioritizing the most promising candidates for experimental validation. This approach not only accelerates discovery but also improves success rates by focusing experimental efforts on candidates with the highest predicted efficacy (Decker et al., 2018; Yang et al., 2021).

Comparative studies demonstrate that AI-enhanced HTS can reduce experimental workloads by more than 60% while achieving equivalent or superior material performance compared to traditional exhaustive screening (Stier et al., 2024). Additionally, these methods facilitate the discovery of novel material classes that may have been overlooked using conventional techniques.

5.4 Comparative Analysis: AI vs. Traditional Methods

Across these case studies, several consistent advantages of AI-enabled approaches emerge (Table 1)

Table 1: Comparative analysis between AI-enabled approaches and traditional methods

Aspect	Traditional Methods	AI-Enabled Approaches	References
Time to Discovery	Months to years	Weeks to days	(Ravichandran et al., 2023)
Experimental workload	High (iterative synthesis and testing)	Reduced via predictive modeling	(Lo et al., 1998; Saxena et al., 2023)
Accuracy in design	Limited by empirical knowledge	High predictive accuracy (eg, AlphaFold)	(Elmousalami, 2020; Ghiasi et al., 2018)
Novelty of materials	Constrained to known motifs	AI can explore previously inaccessible chemical/structural spaces	(Kaulage et al., 2023; Taherdoost & Madanchian, 2023)
Reproducibility	Moderate; dependent on experimental conditions	High; in silico predictions standardized design principles	(Bizzego et al., 2019; Desai et al., 2025)

These comparisons highlight that AI is reshaping conventional processes by enabling new capabilities such as rational design of previously unachievable biomaterial architectures and predictive customization for patient-specific applications.

6. Integration Challenges and Research Gaps

Despite the remarkable advances in AI-powered biochemical material design, several critical challenges hinder seamless integration and translation into clinical applications (Nashruddin et al., 2024). These challenges span data limitations, model constraints, and translational barriers, highlighting the need for rigorous strategies to ensure reliability, reproducibility, and safety. Understanding and addressing these challenges/gaps is essential for realizing the full potential of AI in biomedical materials research (Ali, 2023).

6.1 Data Challenges

Data lies at the foundation of AI-driven material discovery. High-quality, comprehensive datasets are essential for training machine learning models and ensuring accurate predictions(Badini et al., 2023). However, the biochemical materials field faces several key data-related challenges such as:

Heterogeneous biological datasets

Biological systems are inherently complex, and the corresponding datasets are often highly heterogeneous (Butcher et al., 2004; Wilkinson, 2009). Sources include omics data (genomics, proteomics, metabolomics), material characterization data (mechanical properties, swelling behavior, degradation kinetics), imaging data, and clinical outcomes (Butcher et al., 2004; Sari et al., 2022). These datasets differ in scale, resolution, and format, making integration into a unified AI pipeline challenging. Models trained on limited or biased datasets risk producing

predictions that are not generalizable across diverse experimental conditions or patient populations(Wilkinson, 2009) .

Lack of standardized benchmarks

Another critical issue is the absence of standardized benchmarks for evaluating AI models in biochemical material design (Bender et al., 2022; Schneider et al., 2020). Unlike established domains such as image recognition or natural language processing, material science lacks widely accepted datasets with standardized formats and quality metrics. This hinders fair comparison of models, reproducibility of results, and validation of predictive performance. The development of curated, open-access databases containing experimental and simulated material properties, biological interaction data, and clinical relevance metrics is urgently needed to advance the field (Schneider et al., 2020).

6.2 Model Limitations

Even with high-quality datasets, AI models face intrinsic limitations that affect reliability and interpretability.

Interpretability and explainability

Deep learning and other complex AI models often function as "black boxes," making it difficult to understand the rationale behind predictions. In biomedical applications, interpretability is crucial: researchers and producers must know why a particular material design is predicted to perform optimally or why a drug delivery system is likely to succeed. Lack of explainability can hinder trust, slow adoption, and create challenges in regulatory approval (Das & Rad, 2020).

Overfitting and generalization issues

Overfitting occurs when AI models capture noise or dataset-specific patterns rather than generalizable relationships. In biochemical material design, this can result in models that

perform well on training data but fail when applied to new materials, experimental conditions, or patient populations. Ensuring robust generalization requires diverse, high-quality datasets, regularization strategies, and rigorous cross-validation protocols (Aliferis & Simon, 2024).

6.3 Translational Barriers

Translating AI-designed biochemical materials from bench to bedside involves complex regulatory, ethical, and clinical considerations (Bernstam et al., 2022).

Clinical Validation Requirements

Experimental validation of AI predictions is critical before clinical implementation. Even with highly accurate *in silico* models, candidate materials must undergo rigorous *in vitro* and *in vivo* testing to confirm biocompatibility, functionality, and safety. The absence of standardized protocols for evaluating AI-designed materials complicates the validation process, leading to longer timelines and higher costs (Aravazhi et al., 2025).

Regulatory and Ethical Considerations

Regulatory agencies, such as the FDA and EMA, require comprehensive documentation of material design, testing, and predicted outcomes. AI introduces unique challenges in this context, including model transparency, reproducibility, and risk assessment. Ethical considerations also arise in patient-specific applications, where AI-guided designs may influence personalized therapy. Ensuring informed consent, data privacy, and equitable access is essential to avoid bias and unintended consequences (Weiner et al., 2025).

7. Future Prospects

AI is not merely accelerating current approaches in biochemical material design, it is poised to reshape the future landscape of biomedical innovation (da Silva, 2024). Beyond predictive modeling and high-throughput optimization, emerging trends suggest the development of self-evolving biomaterials, closed-loop therapeutic systems, and intelligent human–material interfaces (da Silva, 2024). Realizing these prospects will require multidisciplinary collaboration among AI experts, materials scientists, biologists, and abandonment.

7.1 Self-evolving biomaterials

Traditional biomaterials are static by nature, designed with fixed physical and chemical properties. In contrast, self-evolving materials can adapt dynamically to biological environments, responding to changes in pH, temperature, enzymatic activity, or cellular signals. AI facilitates this evolution by predicting how material properties will interact with the surrounding biological system over time, enabling continuous optimization (Naskar et al., 2025).

For instance, AI algorithms can model polymer crosslinking and degradation kinetics to design hydrogels that adjust stiffness or porosity in response to tissue remodeling. Similarly, peptide- or protein-based scaffolds can be designed

to reorganize their structure dynamically, promoting cell migration and tissue integration. These intelligent materials have the potential to revolutionize regenerative medicine by providing adaptive scaffolds that grow with the tissue, improving integration and long-term outcomes (Naskar et al., 2025).

7.2 Closed-loop therapeutic systems

Closed-loop therapeutic systems integrate AI with sensing and actuation technologies to deliver personalized, real-time interventions. In such systems, AI continuously monitors patient-specific biomarkers, analyzes responses, and adjusts therapeutic delivery accordingly. This approach is particularly relevant for drug delivery, immunotherapy, and regenerative treatments (Zheng et al., 2024).

For example, an AI-driven hydrogel embedded with biosensors could release growth factors or drugs in response to detected changes in local tissue conditions. Reinforcement learning models can optimize delivery schedules, dosage, and release kinetics to maximize therapeutic efficacy while minimizing side effects. By closing the loop between sensing and actuation, these systems move beyond static therapy, enabling precision medicine at the material level (Hahn & Inan, 2022; Zheng et al., 2024).

7.3 Intelligent human–material interfaces

The future of biomedical materials lies in developing interfaces that seamlessly integrate synthetic systems with living tissues, creating dynamic and intelligent platforms for healthcare. Powered by artificial intelligence, these human–material interfaces can sense biological signals and adapt their properties in real time, functioning as true extensions of the body (C. Wang et al., 2023). Examples include smart prosthetics with AI-guided soft materials that adjust stiffness based on load and movement, tissue scaffolds that respond to metabolic or inflammatory cues by releasing signaling molecules at precise moments, and biosensors that detect subtle biochemical changes while transmitting data to AI algorithms for predictive intervention. By combining computational intelligence with advanced material design, these systems promise to enhance biocompatibility, optimize therapeutic performance, and transform patient outcomes—ushering in a new era where medical devices evolve alongside the human body itself (Manickam et al., 2022).

7.4 Role of multidisciplinary collaboration

Realizing these future prospects will depend on close collaboration across multiple disciplines, each contributing unique expertise to the development and translation of AI-driven biomaterials. AI and computational scientists will be responsible for creating predictive and generative models that inform material design, while materials scientists and chemists synthesize and characterize novel biomaterials based on these insights (Cao et al., 2025). Biologists and bioengineers will play a critical role in validating interactions with living systems and optimizing biocompatibility, and

together with regulatory experts will ensure that translational pathways uphold safety, efficacy, and compliance. Such cross-disciplinary collaboration is vital not only to overcome technical and regulatory challenges but also to establish standardized protocols, curated datasets, and reproducible workflows. Furthermore, open-source initiatives and shared platforms can accelerate innovation and facilitate the rapid adoption of intelligent biomaterials in clinical practice, ultimately bridging the gap between computational design and real-world healthcare applications (Cao et al., 2025; Patel et al., 2024).

8. Roadmap and Recommendations

While AI has already transformed biochemical material design, realizing its full potential requires a strategic roadmap that addresses data, modeling, validation, and collaboration (Mirakhori & Niazi, 2025).

8.1 Development of large-scale, curated, open-access datasets

High-quality datasets are foundational for AI-driven material discovery. Efforts should focus on comprehensive data collection that integrates material properties, biological responses, imaging results, and clinical outcomes. To ensure reproducibility, researchers must adopt standardized formats, metadata protocols, and stringent quality controls. Equally important is the creation of open-access repositories, which would enable sharing of datasets and benchmarking of AI models, thereby accelerating discovery while reducing duplication of effort. With such resources, AI models will be better equipped to generalize across diverse experimental and clinical conditions, ultimately improving predictive reliability and fostering collaboration across laboratories (Edfeldt et al., 2024).

8.2 Investment in explainable and trustworthy AI

For AI-designed materials to be translated into clinical practice, the models must be both interpretable and trustworthy. This requires the development of explainable AI methods that clarify how predictions are made, highlighting the key features that drive material performance. Robustness and validation should be ensured through rigorous cross-validation, uncertainty quantification, and sensitivity analyses, all of which minimize the risk of overfitting. At the same time, ethical considerations must remain central, ensuring that models avoid bias, protect patient privacy, and comply with regulatory and ethical guidelines. Collectively, these practices will build trust among researchers, dismissed, and regulatory authorities, facilitating wider adoption of AI-driven approaches (Moreno-Sánchez et al., 2025).

8.3 Standardization of validation protocols for clinical translation

Before AI-designed materials can be adopted in healthcare, they must undergo rigorous and standardized experimental validation. This involves defining clear in vitro and in vivo benchmarks for biocompatibility, mechanical performance,

and functional outcomes. Regulatory alignment is also essential, with researchers and policymakers working together to establish approval pathways that specifically address AI-generated designs. To strengthen reliability, reproducibility metrics should be developed to ensure that results remain consistent across laboratories and experimental conditions. Such standardized protocols not only reduce scientific and regulatory uncertainty but also accelerate patient access to safe and effective materials (Geaney et al., 2023).

8.4 Strengthening academia, industry, and healthcare collaborations

Multidisciplinary collaboration will be central to the successful translation of AI-driven biomaterials. Academic researchers will continue to lead in foundational studies of material science and AI model development, while industry partners can provide the expertise needed to scale production, optimize manufacturing, and bring innovations to market. Healthcare professionals play a critical role in ensuring clinical relevance, validating patient-specific applications, and monitoring real-world outcomes. By fostering close partnerships between academia, industry, and healthcare, it will be possible to accelerate the safe and impactful adoption of AI-powered biomaterials, ensuring that these technologies fulfill their promise in improving patient care (Cao et al., 2025).

9. Conclusion

Artificial intelligence is transforming the landscape of biochemical material design, offering unprecedented capabilities for predictive modeling, generative design, and optimization. Landmark successes including AlphaFold for protein design, AI-guided hydrogel optimization, and AI-assisted high-throughput screening demonstrate tangible improvements over traditional approaches in speed, efficiency, and material novelty.

Despite these advances, significant challenges remain, including heterogeneous datasets, model interpretability, and translational barriers. Addressing these obstacles requires multidisciplinary collaboration, standardized protocols, and open-access datasets. Looking forward, emerging trends such as self-evolving biomaterials, closed-loop therapeutic systems, and intelligent human-material interfaces promise to redefine the boundaries of biomedical innovation.

In summary, AI is not merely accelerating material discovery; it is reshaping the paradigm of biomedical sciences. By embracing AI-driven design, validation, and clinical translation, researchers can create next-generation biomaterials that are adaptive, personalized, and intelligent, ultimately improving patient outcomes and transforming healthcare. This paper serves as a roadmap and call to action for the field, highlighting both current successes and the vast opportunities that lie ahead.

References

- Ali, H. (2023). Artificial intelligence in multi-omics data integration: Advancing precision medicine, biomarker discovery and genomic-driven disease interventions. *Int J Sci Res Arch*, 8(1), 1012-30. <https://doi.org/10.30574/ijstra.2023.8.1.0189>
- Aliferis, C., & Simon, G. (2024). Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In G. J. Simon (Eds.) et. al., *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*. (pp. 477–524). Springer. https://doi.org/10.1007/978-3-031-39355-6_10
- Aravazhi, P. S., Gunasekaran, P., Benjamin, N. Z. Y., Thai, A., Chandrasekar, K. K., Kolanu, N. D., Prajiwal, P., Tekuru, Y., Brito, L. V., & Inban, P. (2025). The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. *Disease-a-month : DM*, 71(6), 101882. <https://doi.org/10.1016/j.disamonth.2025.101882>
- Badini, S., Regondi, S., & Pugliese, R. (2023). Unleashing the Power of Artificial Intelligence in Materials Design. *Materials (Basel, Switzerland)*, 16(17), 5927. <https://doi.org/10.3390/ma16175927>
- Bai, X., & Zhang, X. (2025). Artificial Intelligence-Powered Materials Science. *Nano-micro letters*, 17(1), 135. <https://doi.org/10.1007/s40820-024-01634-8>
- Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., & Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. *Nature reviews. Chemistry*, 6(6), 428–442. <https://doi.org/10.1038/s41570-022-00391-9>
- Bernstam, E. V., Shireman, P. K., Meric-Bernstam, F., N Zozus, M., Jiang, X., Brimhall, B. B., Windham, A. K., Schmidt, S., Visweswaran, S., Ye, Y., Goodrum, H., Ling, Y., Barapatre, S., & Becich, M. J. (2022). Artificial intelligence in clinical and translational science: Successes, challenges and opportunities. *Clinical and translational science*, 15(2), 309–321. <https://doi.org/10.1111/cts.13175>
- Bizzego, A., Bussola, N., Chierici, M., Maggio, V., Francescato, M., Cima, L., Cristoforetti, M., Jurman, G., & Furlanello, C. (2019). Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS computational biology*, 15(3), e1006269. <https://doi.org/10.1371/journal.pcbi.1006269>
- Butcher, E. C., Berg, E. L., & Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature biotechnology*, 22(10), 1253–1259. <https://doi.org/10.1038/nbt1017>
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547–555. <https://doi.org/10.1038/s41586-018-0337-2>
- Buwalda, S. J., Boere, K. W., Dijkstra, P. J., Feijen, J., Vermonden, T., & Hennink, W. E. (2014). Hydrogels in a historical perspective: from simple networks to smart materials. *Journal of controlled release : official journal of the Controlled Release Society*, 190, 254–273. <https://doi.org/10.1016/j.jconrel.2014.03.052>
- Callaway E. (2022). What's next for AlphaFold and the AI protein-folding revolution. *Nature*, 604(7905), 234–238. <https://doi.org/10.1038/d41586-022-00997-5>
- Cao, H., Oghenemaro, E. F., Latypova, A., Abosaoda, M. K., Zaman, G. S., & Devi, A. (2025). Advancing clinical biochemistry: addressing gaps and driving future innovations. *Frontiers in medicine*, 12, 1521126. <https://doi.org/10.3389/fmed.2025.1521126>
- Capella-Monsonís, H., Crum, R. J., Hussey, G. S., & Badylak, S. F. (2024). Advances, challenges, and future directions in the clinical translation of ECM biomaterials for regenerative medicine applications. *Advanced Drug Delivery Reviews*, 211, 115347. <https://doi.org/10.1016/j.addr.2024.115347>
- Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., ... & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4), 045002. <https://doi.org/10.1103/RevModPhys.91.045002>
- Chen, H., Liu, Y., Balabani, S., Hirayama, R., & Huang, J. (2023). Machine Learning in Predicting Printable Biomaterial Formulations for Direct Ink Writing. *Research (Washington, D.C.)*, 6, 0197. <https://doi.org/10.34133/research.0197>
- Chen, Z., Wang, R., Guo, J., & Wang, X. (2024). The role and future prospects of artificial intelligence algorithms in peptide drug development. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 175, 116709. <https://doi.org/10.1016/j.biopha.2024.116709>
- Choi, H. S., & Frangioni, J. V. (2010). Nanoparticles for biomedical imaging: fundamentals of clinical translation. *Molecular imaging*, 9(6), 291–310.
- Da Silva R. G. L. (2024). The advancement of artificial intelligence in biomedical research and health innovation: challenges and opportunities in emerging economies. *Globalization and health*, 20(1), 44. <https://doi.org/10.1186/s12992-024-01049-5>

Dan, Yabo & Zhao, Yong & Li, Xiang & Li, Shaobo & Hu, Ming & Hu, Jianjun. (2020). Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*. 6. 84. 10.1038/s41524-020-00352-0.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Decker, S. R., Harman-Ware, A. E., Happs, R. M., Wolfrum, E. J., Tuskan, G. A., Kainer, D., ... & Jacobson, D. (2018). High throughput screening technologies in biomass characterization. *Frontiers in Energy Research*, 6, 120. <https://doi.org/10.3389/fenrg.2018.00120>

Desai, A., Abdelhamid, M., & Padalkar, N. R. (2025). What is reproducibility in artificial intelligence and machine learning research?. *AI Magazine*, 46(2), e70004.

Edfeldt, K., Edwards, A. M., Engkvist, O., Günther, J., Hartley, M., Hulcoop, D. G., Leach, A. R., Marsden, B. D., Menge, A., Misquitta, L., Müller, S., Owen, D. R., Schütt, K. T., Skelton, N., Steffen, A., Tropsha, A., Vernet, E., Wang, Y., Wellnitz, J., Willson, T. M., ... Schapira, M. (2024). A data science roadmap for open science organizations engaged in early-stage drug discovery. *Nature communications*, 15(1), 5640. <https://doi.org/10.1038/s41467-024-49777-x>

Elmousalami, H. H. (2020). Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis. *IEEE transactions on engineering management*, 68(1), 183-196. DOI: [10.1109/TEM.2020.2972078](https://doi.org/10.1109/TEM.2020.2972078)

Freedman, B. R., & Mooney, D. J. (2019). Biomaterials to Mimic and Heal Connective Tissues. *Advanced materials (Deerfield Beach, Fla.)*, 31(19), e1806695. <https://doi.org/10.1002/adma.201806695>

Fu, R., Chen, Z., Tian, H., Hu, J., Bu, F., Zheng, P., ... & Zhu, L. (2025). A Review on the Applications of Machine Learning in Biomaterials, Biomechanics, and Biomanufacturing for Tissue Engineering. *Smart Materials in Medicine*. <https://doi.org/10.1016/j.smain.2025.06.003>

Geaney, A., O'Reilly, P., Maxwell, P., James, J. A., McArt, D., & Salto-Tellez, M. (2023). Translation of tissue-based artificial intelligence into clinical practice: from discovery to adoption. *Oncogene*, 42(48), 3545–3555. <https://doi.org/10.1038/s41388-023-02857-6>

Gharibshahian, M., Torkashvand, M., Bavisi, M., Aldaghi, N., & Alizadeh, A. (2024). Recent advances in artificial intelligent strategies for tissue engineering and regenerative

medicine. *Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)*, 30(9), e70016. <https://doi.org/10.1111/srt.70016>

Ghiasi, R., Ghasemi, M. R., & Noori, M. (2018). Comparative studies of metamodeling and AI-Based techniques in damage detection of structures. *Advances in Engineering Software*, 125, 101-112. <https://doi.org/10.1016/j.advengsoft.2018.02.006>

Ghosh, S., Singh, B. P., & Webster, T. J. (2021). Nanoparticle-impregnated biopolymers as novel antimicrobial nanofilms. In *Biopolymer-based nano films* (pp. 269-309). Elsevier. <https://doi.org/10.1016/B978-0-12-823381-8.00017-X>

Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6), 952–963. <https://doi.org/10.1016/j.fertnstert.2018.05.006>

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature reviews. Molecular cell biology*, 23(1), 40–55. <https://doi.org/10.1038/s41580-021-00407-0>

Hahn, J. O., & Inan, O. T. (2022). Physiological closed-loop control in critical care: opportunities for innovations. *Progress in Biomedical Engineering*, 4(3), 033001. DOI 10.1088/2516-1091/ac6d36

Hang, R., Yao, X., Bai, L., & Hang, R. (2025). Evolving biomaterials design from trial and error to intelligent innovation. *Acta Biomaterialia*. <https://doi.org/10.1016/j.actbio.2025.03.013>

Hench, L. L., & Polak, J. M. (2002). Third-generation biomedical materials. *Science (New York, N.Y.)*, 295(5557), 1014–1017. <https://doi.org/10.1126/science.1067404>

Ho, T. C., Chang, C. C., Chan, H. P., Chung, T. W., Shu, C. W., Chuang, K. P., Duh, T. H., Yang, M. H., & Tyan, Y. C. (2022). Hydrogels: Properties and Applications in Biomedicine. *Molecules (Basel, Switzerland)*, 27(9), 2902. <https://doi.org/10.3390/molecules27092902>

Hoffman A. S. (2002). Hydrogels for biomedical applications. *Advanced drug delivery reviews*, 54(1), 3–12. [https://doi.org/10.1016/s0169-409x\(01\)00239-3](https://doi.org/10.1016/s0169-409x(01)00239-3)

Hong, L., Wu, X., Liu, Y. S., Yu, C., Liu, Y., Sun, K., ... & Wang, K. X. (2023). Self-adapting and self-healing hydrogel interface with fast Zn²⁺ transport kinetics for highly reversible Zn anodes. *Advanced Functional Materials*, 33(29), 2300952

- Huang, G., Li, F., Zhao, X., Ma, Y., Li, Y., Lin, M., Jin, G., Lu, T. J., Genin, G. M., & Xu, F. (2017). Functional and Biomimetic Materials for Engineering of the Three-Dimensional Cell Microenvironment. *Chemical reviews*, 117(20), 12764–12850. <https://doi.org/10.1021/acs.chemrev.7b00094>
- Huang, Z., Yu, J., He, W., Yu, J., Deng, S., Yang, C., Zhu, W., & Shao, X. (2024). AI-enhanced chemical paradigm: From molecular graphs to accurate prediction and mechanism. *Journal of hazardous materials*, 465, 133355. <https://doi.org/10.1016/j.jhazmat.2023.133355>
- Jang, Y. H., Jin, X., Shankar, P., Lee, J. H., Jo, K., & Lim, K. I. (2019). Molecular-Level Interactions between Engineered Materials and Cells. *International journal of molecular sciences*, 20(17), 4142. <https://doi.org/10.3390/ijms20174142>
- Jiang, P., Dai, Y., Hou, Y., Stein, J., Lin, S. S., Zhou, C., ... & Yang, L. (2025). Artificial intelligence-assisted design, synthesis and analysis of smart biomaterials. *BMEMat*, e70004. <https://doi.org/10.1002/bmm2.70004>
- Jin, X., Liu, C., Xu, T., Su, L., & Zhang, X. (2020). Artificial intelligence biosensors: Challenges and prospects. *Biosensors & bioelectronics*, 165, 112412. <https://doi.org/10.1016/j.bios.2020.112412>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 422–440. <https://doi.org/10.1038/s42254-021-00314-5>
- Karunakar, K. K., Cheriyan, B. V., Anandakumar, R., Murugathirumal, A., Senthilkumar, A., Nandhini, J., ... & Yabase, L. (2025). Stimuli-responsive smart materials: Bridging the gap between biotechnology and regenerative medicine. *Bioprinting*, 48, e00415. <https://doi.org/10.1016/j.bprint.2025.e00415>
- Karunakar, K. K., Cheriyan, B. V., Anandakumar, R., Murugathirumal, A., Senthilkumar, A., Nandhini, J., ... & Yabase, L. (2025). Stimuli-responsive smart materials: Bridging the gap between biotechnology and regenerative medicine. *Bioprinting*, 48, e00415. <https://doi.org/10.1016/j.bprint.2025.e00415>
- Kaulage, A. N., Tomar, P. K., Pant, R., Debnath, L., Ibrahim, R. K., & Alazzam, M. B. (2023, May). Revolutionizing Material Selection in Construction 4.0: A Comparative Analysis of Artificial Intelligence Technologies. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2707–2712). IEEE. <https://doi.org/10.1109/ICACITE57410.2023.10182419>
- Kazi, R. N. A., Hasani, I. W., Khafaga, D. S. R., Kabba, S., Farhan, M., Aatif, M., Muteeb, G., & Fahim, Y. A. (2025). Nanomedicine: The Effective Role of Nanomaterials in Healthcare from Diagnosis to Therapy. *Pharmaceutics*, 17(8), 987. <https://doi.org/10.3390/pharmaceutics17080987>
- Khan, M. U. A., Aslam, M. A., Abdullah, M. F. B., Al-Arjan, W. S., Stojanovic, G. M., & Hasan, A. (2024). Hydrogels: Classifications, fundamental properties, applications, and scopes in recent advances in tissue engineering and regenerative medicine—A comprehensive review. *Arabian Journal of Chemistry*, 17(10), 105968. <https://doi.org/10.1016/j.arabjc.2024.105968>
- Kumar, K. M., Ram, S., Manjula, N. G., & Mayegowda, S. B. (2022). Biopolymers and Their Applications in Biomedicine. In *Polymeric Biomaterials* (pp. 35–62). CRC Press.
- Kuperkar, K., Atanase, L. I., Bahadur, A., Crivei, I. C., & Bahadur, P. (2024). Degradable Polymeric Bio(nano)materials and Their Biomedical Applications: A Comprehensive Overview and Recent Updates. *Polymers*, 16(2), 206. <https://doi.org/10.3390/polym16020206>
- Kurul, F., Turkmen, H., Cetin, A. E., & Topkaya, S. N. (2025). Nanomedicine: How nanomaterials are transforming drug delivery, bio-imaging, and diagnosis. *Next Nanotechnology*, 7, 100129. <https://doi.org/10.1016/j.nxnano.2024.100129>
- Lateef Junaid M. A. (2025). Artificial intelligence driven innovations in biochemistry: A review of emerging research frontiers. *Biomolecules & biomedicine*, 25(4), 739–750. <https://doi.org/10.17305/bb.2024.11537>
- Lo, V., Mache, J., & Windisch, K. (1998, March). A comparative study of real workload traces and synthetic workload models for parallel job scheduling. In *Workshop on job scheduling strategies for parallel processing* (pp. 25–46). Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/BFb0053979>
- Lutz S. (2010). Beyond directed evolution--semi-rational protein engineering and design. *Current opinion in*

biotechnology, 21(6), 734–743.

<https://doi.org/10.1016/j.copbio.2010.08.011>

Ly, P. D., Ly, K. N., Phan, H. L., Nguyen, H. H., Duong, V. A., & Nguyen, H. V. (2024). Recent advances in surface decoration of nanoparticles in drug delivery. *Frontiers in Nanotechnology*, 6, 1456939.

<https://doi.org/10.3389/fnano.2024.1456939>

Ma, W., Hua, D., Xiong, R., & Huang, C. (2023). Bio-based stimuli-responsive materials for biomedical applications. *Materials Advances*, 4(2), 458–475.

<https://doi.org/10.1039/D2MA01011A>

Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., & Fröhlich, H. (2024). Transformer models in biomedicine. *BMC medical informatics and decision making*, 24(1), 214. <https://doi.org/10.1186/s12911-024-02600-5>

Manickam, P., Mariappan, S. A., Murugesan, S. M., Hansda, S., Kaushik, A., Shinde, R., & Thipperudraswamy, S. P. (2022). Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare. *Biosensors*, 12(8), 562.

<https://doi.org/10.3390/bios12080562>

Matson, J. B., & Stupp, S. I. (2012). Self-assembling peptide scaffolds for regenerative medicine. *Chemical communications (Cambridge, England)*, 48(1), 26–33.

<https://doi.org/10.1039/c1cc15551b>

Mirakhori, F., & Niazi, S. K. (2025). Harnessing the AI/ML in Drug and Biological Products Discovery and Development: The Regulatory Perspective. *Pharmaceuticals (Basel, Switzerland)*, 18(1), 47.

<https://doi.org/10.3390/ph18010047>

Moreno-Sánchez, P. A., Del Ser, J., van Gils, M., & Hernesniemi, J. (2025). A Design Framework for operationalizing Trustworthy Artificial Intelligence in Healthcare: Requirements, Tradeoffs and Challenges for its Clinical Adoption. *arXiv preprint arXiv:2504.19179*.

<https://doi.org/10.48550/arXiv.2504.19179>

Nashruddin, S. N. A. B. M., Salleh, F. H. M., Yunus, R. M., & Zaman, H. B. (2024). Artificial intelligence-powered electrochemical sensor: Recent advances, challenges, and prospects. *Heliyon*, 10(18), e37964.

<https://doi.org/10.1016/j.heliyon.2024.e37964>

Naskar, S., Sharma, S., Kuotsu, K., Halder, S., Pal, G., Saha, S., Mondal, S., Biswas, U. K., Jana, M., & Bhattacharjee, S. (2025). The biomedical applications of artificial intelligence: an overview of decades of research. *Journal of drug targeting*, 33(5), 717–748.

<https://doi.org/10.1080/1061186X.2024.2448711>

Negut, I., & Bitá, B. (2023). Exploring the Potential of Artificial Intelligence for Hydrogel Development-A Short Review. *Gels (Basel, Switzerland)*, 9(11), 845.

<https://doi.org/10.3390/gels9110845>

Özkale, B., Sakar, M. S., & Mooney, D. J. (2021). Active biomaterials for mechanobiology. *Biomaterials*, 267, 120497. <https://doi.org/10.1016/j.biomaterials.2020.120497>

Park, S., Gwon, Y., Khan, S. A., Jang, K. J., & Kim, J. (2023). Engineering considerations of iPSC-based personalized medicine. *Biomaterials research*, 27(1), 67. <https://doi.org/10.1186/s40824-023-00382-x>

Parvin, N., Joo, S. W., Jung, J. H., & Mandal, T. K. (2025). Multimodal AI in Biomedicine: Pioneering the Future of Biomaterials, Diagnostics, and Personalized Healthcare. *Nanomaterials (Basel, Switzerland)*, 15(12), 895. <https://doi.org/10.3390/nano15120895>

Patel, A. U., Gu, Q., Esper, R., Maeser, D., & Maeser, N. (2024). The crucial role of interdisciplinary conferences in advancing explainable AI in healthcare. *BioMedInformatics*, 4(2), 1363–1383. <https://doi.org/10.3390/biomedinformatics4020075>

Paul, J. (2025). Real-time predictive health monitoring using AI-driven wearable sensors: Enhancing early detection and personalized interventions in chronic disease management.

Peer, D., Karp, J. M., Hong, S., Farokhzad, O. C., Margalit, R., & Langer, R. (2020). Nanocarriers as an emerging platform for cancer therapy. *Nano-enabled medical applications*, 61–91. <https://doi.org/10.1038/nnano.2007.387>

Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. *EMBO reports*, 22(11), e54046.

<https://doi.org/10.15252/embr.202154046>

Petch, J., Di, S., & Nelson, W. (2022). Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *The Canadian journal of cardiology*, 38(2), 204–213.

<https://doi.org/10.1016/j.cjca.2021.09.004>

Pugliese, R., Badini, S., Frontoni, E., & Regondi, S. (2025). Generative Artificial Intelligence for Advancing Discovery and Design in Biomaterialomics. *Intelligent Computing*, 4, 0117. DOI: [10.34133/icomputing.0117](https://doi.org/10.34133/icomputing.0117)

Raghavendra, G. M., Varaprasad, K., & Jayaramudu, T. (2015). Biomaterials: design, development and biomedical applications. In *Nanotechnology applications for tissue*

engineering (pp. 21-44). William Andrew Publishing.
<https://doi.org/10.1016/B978-0-323-32889-0.00002-9>

Rahmati, M., Silva, E. A., Reseland, J. E., Heyward, C. A., & Haugen, H. J. (2020). Biological responses to physicochemical properties of biomaterial surface. *Chemical Society Reviews*, 49(15), 5178-5224.
<https://doi.org/10.1039/D0CS00103A>

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378, 686-707.
<https://doi.org/10.1016/j.jcp.2018.10.045>

Ratner, B. D., & Bryant, S. J. (2004). Biomaterials: where we have been and where we are going. *Annual review of biomedical engineering*, 6, 41-75.
<https://doi.org/10.1146/annurev.bioeng.6.040803.140027>

Ravichandran, P., Machireddy, J. R., & Rachakatla, S. K. (2023). Data analytics automation with AI: a comparative study of traditional and generative AI approaches. *Journal of Bioinformatics and Artificial Intelligence*, 3(2), 168-190.

Sari, B., Isik, M., Eylem, C. C., Bektas, C., Okesola, B. O., Karakaya, E., Emregul, E., Nemutlu, E., & Derkus, B. (2022). Omics technologies for high-throughput-screening of cell-biomaterial interactions. *Molecular omics*, 18(7), 591-615. <https://doi.org/10.1039/d2mo00060a>

Saxena, D., Kumar, J., Singh, A. K., & Schmid, S. (2023). Performance analysis of machine learning centered workload prediction models for cloud. *IEEE Transactions on Parallel and Distributed Systems*, 34(4), 1313-1330.
DOI: [10.1109/TPDS.2023.3240567](https://doi.org/10.1109/TPDS.2023.3240567)

Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5(1), 83. <https://doi.org/10.1038/s41524-019-0221-0>

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Jr, Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., Zentgraf, M., Hill, J. E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebke, C., & Schneider, G. (2020). Rethinking drug design in the artificial intelligence era. *Nature reviews. Drug discovery*, 19(5), 353-364. <https://doi.org/10.1038/s41573-019-0050-3>

Schork N. J. (2019). Artificial Intelligence and Personalized Medicine. *Cancer treatment and research*, 178, 265-283.
https://doi.org/10.1007/978-3-030-16391-4_11

Singh, A. V., Rosenkranz, D., Ansari, M. H. D., Singh, R., Kanase, A., Singh, S. P., ... & Luch, A. (2020). Artificial intelligence and machine learning empower advanced biomedical material design to toxicity prediction. *Advanced Intelligent Systems*, 2(12), 2000084.

Sionkowska, A. (2011) Current Research on the Blends of Natural and Synthetic Polymers as New Biomaterials: Review. *Progress in Polymer Science*, 1254-1276.
<http://dx.doi.org/10.1016/j.progpolymsci.2011.05.003>

Stier, S. P., Kreisbeck, C., Ihssen, H., Popp, M. A., Hauch, J., Malek, K., Reynaud, M., Goumans, T. P. M., Carlsson, J., Todorov, I., Gold, L., Räder, A., Wenzel, W., Bandesha, S. T., Jacques, P., Garcia-Moreno, F., Arcelus, O., Friederich, P., Clark, S., Maglione, M., ... Kozdras, M. (2024). Materials Acceleration Platforms (MAPs): Accelerating Materials Research and Development to Meet Urgent Societal Challenges. *Advanced materials (Deerfield Beach, Fla.)*, 36(45), e2407791.
<https://doi.org/10.1002/adma.202407791>

Stuart, M. A., Huck, W. T., Genzer, J., Müller, M., Ober, C., Stamm, M., Sukhorukov, G. B., Szleifer, I., Tsukruk, V. V., Urban, M., Winnik, F., Zauscher, S., Luzinov, I., & Minko, S. (2010). Emerging applications of stimuli-responsive polymer materials. *Nature materials*, 9(2), 101-113.
<https://doi.org/10.1038/nmat2614>

Taherdoost, H., & Madanchian, M. (2023). AI advancements: Comparison of innovative techniques. *AI*, 5(1), 38-54.
<https://doi.org/10.3390/ai5010003>

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, 12(5), 91.
<https://doi.org/10.3390/computers12050091>

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>

Urifa, J., & Shah, K. W. (2025). Developmental Framework and Umbrella Review of AI-driven Strategies for Hydrogel Microneedles.
<https://doi.org/10.20944/preprints202504.2168.v1>

Vora, L. K., Gholap, A. D., Jetha, K., Thakur, R. R. S., Solanki, H. K., & Chavda, V. P. (2023). Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics*, 15(7), 1916.
<https://doi.org/10.3390/pharmaceutics15071916>

Wang, C., He, T., Zhou, H., Zhang, Z., & Lee, C. (2023). Artificial intelligence enhanced sensors - enabling

- technologies to next-generation healthcare and biomedical platform. *Bioelectronic medicine*, 9(1), 17. <https://doi.org/10.1186/s42234-023-00118-1>
- Wang, R. C., & Wang, Z. (2023). Precision Medicine: Disease Subtyping and Tailored Treatment. *Cancers*, 15(15), 3837. <https://doi.org/10.3390/cancers15153837>
- Wasilewski, T., Kamysz, W., & Gębicki, J. (2024). AI-Assisted Detection of Biomarkers by Sensors and Biosensors for Early Diagnosis and Monitoring. *Biosensors*, 14(7), 356. <https://doi.org/10.3390/bios14070356>
- Wei, R., & Mahmood, A. (2020). Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *Ieee Access*, 9, 4939-4956. <https://doi.org/10.1109/ACCESS.2020.3048309>
- Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2025). Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLOS digital health*, 4(4), e0000810. <https://doi.org/10.1371/journal.pdig.0000810>
- Wilkinson D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature reviews. Genetics*, 10(2), 122–133. <https://doi.org/10.1038/nrg2509>
- Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14), 145301. <https://doi.org/10.1103/PhysRevLett.120.145301>
- Yang, L., Pijuan-Galito, S., Rho, H. S., Vasilevich, A. S., Eren, A. D., Ge, L., Habibović, P., Alexander, M. R., de Boer, J., Carlier, A., van Rijn, P., & Zhou, Q. (2021). High-Throughput Methods in the Discovery and Study of Biomaterials and Materiobiology. *Chemical reviews*, 121(8), 4561–4677. <https://doi.org/10.1021/acs.chemrev.0c00752>
- Zhang, C. W., Si, M., Chen, C., He, P., Fei, Z., Xu, N., & He, X. (2025). Hierarchical Engineering for Biopolymer-based Hydrogels with Tailored Property and Functionality. *Advanced materials (Deerfield Beach, Fla.)*, 37(22), e2414897. <https://doi.org/10.1002/adma.202414897>
- Zhang, P., Wei, L., Li, J., & Wang, X. (2024). Artificial intelligence-guided strategies for next-generation biological sequence design. *National science review*, 11(11), nwae343. <https://doi.org/10.1093/nsr/nwae343>
- Zhang, Y. S., Yue, K., Aleman, J., Moghaddam, K. M., Bakht, S. M., Yang, J., Jia, W., Dell'Erba, V., Assawes, P., Shin, S. R., Dokmeci, M. R., Oklu, R., & Khademhosseini, A. (2017). 3D Bioprinting for Tissue and Organ Fabrication. *Annals of biomedical engineering*, 45(1), 148–163. <https://doi.org/10.1007/s10439-016-1612-8>
- Zheng, Z., Zhu, R., Peng, I., Xu, Z., & Jiang, Y. (2024). Wearable and implantable biosensors: mechanisms and applications in closed-loop therapeutic systems. *Journal of materials chemistry. B*, 12(35), 8577–8604. <https://doi.org/10.1039/d4tb00782d>
- Zhou, Q., Tang, P., Liu, S., Pan, J., Yan, Q., & Zhang, S. C. (2018). Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28), E6411–E6417. <https://doi.org/10.1073/pnas.1801181115>

Explainable AI-Powered Anomaly Detection: A Computer Vision Approach to Strengthening Human-Centric Cybersecurity.

Victor Ayodeji Oluwasusi

Department of Computer
Information Systems
Near East University, Cyprus.
20244093@std.neu.edu.tr

Temiloluwa O. Modupeola

Department of Computer
Information Systems
Near East University, Cyprus.
20244302@std.neu.edu.tr

Nasim Ahmadzadeh Nobari Azar

Department of Computer
Information Systems
Near East University, Cyprus.
nasim.ahmadzadeh@neu.edu.tr

Abstract

As cyber threats increase in sophistication and stealth, traditional Intrusion Detection Systems (IDSs) typically experience difficulties identifying emerging types of attacks (e.g., zero-day attacks) because they rely on known signatures and lack interpretability. This paper presents the development of SENTRY-AI, an explainable, multi-modal anomaly detection framework that integrates deep learning and computer vision to improve cybersecurity defenses in cyberspace. SENTRY-AI architecture employs a Variational Autoencoder (VAE) to conduct unsupervised anomaly detection on tabular network features, in addition to a Convolutional Neural Network (CNN) to analyze time-series traffic data transformed into a Gramian Angular Field (GAF). An analysis is performed with a late-fusion approach utilizing outputs from the two models to evaluate the efficacy and robustness of the hybrid approach. Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to provide visual explanations which indicate the most significant areas of input used to make the predictions for the purpose of human-in-the-loop decision making. The experimental work is conducted across three intrusion detection datasets (NSL-KDD, CICIDS2017, and UNSW-NB15), demonstrating that SENTRY-AI outperformed traditional machine learning and modern deep learning-based IDS approaches. In fact, SENTRY-AI achieved an F1-score of 98.92% on NSL-KDD, 100.00% on UNSW-NB15, and an AUC-ROC score exceeding 99% on all datasets.

Keywords

Anomaly Detection, Computer Vision, Explainable AI, Intrusion Detection

1. Introduction

Background & Context

The modern cybersecurity environment is characterized by an unprecedented rise in the sophistication, number, and variety of cyber threats. Digital transformation is prevalent across all aspects of society and increasing reliance on systems and networks connected with each other, increases both vulnerabilities and attack surfaces (Almiani et al., 2020). An increase in cyber threats, including ransomware attacks, phishing attacks, denial-of-service (DoS) attacks, advanced persistent threats (APTs), and zero-day exploits, is seen across the range of cyber actors, from organized crime

operations to state-sponsored threat actors (Moustafa et al., 2019). The increased frequency of attacks illustrates the limitations of a well-established, traditional signature-based intrusion detection system (IDS) that relies on reference known patterns of cyber threats, and as a result lacks the ability to recognize new and methodical attacks.

Methods of anomaly detection supported by artificial intelligence (AI) and machine learning (ML), have gained attention to deal with these obstacles. Anomaly detection systems do not rely on signature-based approaches and detect abnormal behavior from typical network behavior. This style of threat detection is more proactive and flexible. Machine learning fundamental algorithms and methods, in particular, deep learning methods such as Variational Autoencoders (VAEs) and Convolutional Neural Networks (CNNs), have been proven able to detect anomalies in complex and high-dimensional network data with higher accuracy (Mirsky et al., 2018). The strengths of these methods are found in the potential to learn from massive volumes of data, detect subtle patterns, and reconfigure to accommodate instantaneous threat environments.

Nonetheless, despite the advantages of anomaly detection methods, traditional ML-based detection often entails a lack of explainability. This lack of simple interpretability can result in unexplainable predictions and an inability to trust those predictions from a cyber intrusion analyst, which makes implementing these systems an impediment and disadvantage in a real environment where humans need to interpret results and use that evidence in a meaningful way to affect important decisions. As a result of the demand for explainable artificial intelligence (XAI) that can answer why the AI model acted the way it did and provide an intuitive explanation to support its predicted output (Samek & Müller, 2019).

Many XAI approaches like Gradient-weighted Class Activation Mapping (Grad-CAM) can assist intrusion analysts in converting AI model provided alerts into interpretable visual representations to build analytical trust or increase the usability of the model decisions.

Beyond technical performance, the study emphasizes human-centric cybersecurity, showing that the integration of explainable AI (XAI) significantly enhances analyst trust, reduces cognitive load, and supports ethical decision-making.

These results highlight the importance of visual interpretability and multimodal learning in advancing the next generation of adaptive, transparent, and trustworthy intrusion detection systems.

Research Problem Statement

The central problem of this research is the transparency and interpretability of a traditional AI-based system for detecting incidents. When analysts cannot understand the direct implications of a model's prediction or output, trust in the model will be dissipated, ultimately compromising both the ability of an analyst to make important decisions and the ability to mitigate a cybersecurity threat (Samek & Müller, 2019). Becoming significantly advanced and ubiquitous across every industry, analysts need more interpretable systems in order to hypothesize why an activity within their network was classified anomalously.

This research will address these issues by incorporating computer vision techniques with explainable AI (XAI) techniques with anomaly detection frameworks. By visualizing network traffic as images, analysts will view complex network data in intuitive and human-interpretable space. Also, using XAI techniques, such as Grad-CAM, will allow an analyst to identify which regions of the visualizations are significant in formulating the model's decision which greatly enhances interpretable transparency.

Objectives and Research Questions

This research aims to develop an innovative, explainable AI-powered anomaly detection framework that leverages computer vision to enhance cybersecurity. Specifically, the objectives include:

1. Developing a multi-modal anomaly detection framework integrating Variational Autoencoders (VAEs) and Convolutional Neural Networks (CNNs).
2. Assessing the effectiveness of visual representations for detecting network anomalies.
3. Applying Grad-CAM to produce interpretable visual explanations.

The research questions guiding this study are:

1. To what extent can computer vision methodologies enhance the performance and accuracy of anomaly detection systems?
2. How do visual interpretability and explainable AI influence analyst decision-making and trust in IDS alerts?
3. What ethical considerations and practical implications arise from integrating XAI into operational cybersecurity practices?

Significance and Contributions

This research framework has the potential to contribute meaningfully to the multi-discipline domains of cybersecurity, artificial intelligence, and human-computer interaction. The research highlights a novel, explainable, and

intuitively-oriented anomaly detection system, which includes applied data-to-decision outputs relevant to real-life cybersecurity issues. Security analysts who use visual explanations will be able to read alerts quicker and more effectively, thereby reducing incident response time, and improving decision-making accuracy (Mirsky et al., 2018).

From an ethical stand-point, this research may be used to highlight and promote transparency, accountability, and trustworthiness of automated systems for cybersecurity. Explainable models are one way to use visual data without treating every algorithm as a black-box, which gives an explanation for why an anomaly was considered an anomaly, thus supporting the ethical adoption of AI in governance processes (Samek & Müller, 2019).

Human cognition can be enhanced by making explainable and clear interpretability into algorithms. Analysts can rapidly agree/disagree with model decisions, and distinguish false positives from actual threats more efficiently, reducing time wasted on explanations and improving operational efficiency. The interdisciplinary dimensions of this research have extended to human-computer interaction, where analysts and AI can foster security operations together.

Structure Overview of Paper

The remaining sections of this paper will follow this structure; Section 2 is a literature review, including anomaly detection approaches, computer vision application and explainable AI in cybersecurity. Section 3 includes the methodology, including dataset access, network visualization approaches, and analytic models in detail. Section 4 describes the experimental findings, including measures of performance and Grad-CAM visualizations. Section 5 discusses the findings and the relevance for cybersecurity practice and outlines limitations. Finally, Section 6 provides a conclusion, including future research possibilities and recommendations for improvements and limitations on practical applications.

2. Related works

Cybersecurity and Anomaly Detection

The rise of digital industries and the growing number of connected devices, the attack surface for cyber threats has increased exponentially. Signature-based intrusion detection systems (IDS) are less useful for advanced and evolving cyber threats, creating a need for detection systems that are more adaptive and smarter (Salem et al., 2024; Hodo et al., 2017; Brundage et al., 2018; Zhang & Ran, 2021).

Anomaly detection has become an important approach to cybersecurity due to its aim to identify deviations to established normal behavior. Anomaly detection is a departure from signature-based methods as it does not depend on known attacks, but rather the establishment of what normal operation looks like, it is therefore better at

identifying novel attacks or zero-day attacks (Radford et al., 2018; Mohammadpour et al., 2020; Babaey & Faragardi, 2025; Lunardi et al., 2023). The current threat landscape poses a serious threat as attackers seek new techniques to evade typical security mechanisms (Bamber et al., 2025; Talukder et al., 2024).

Since this approach of anomaly detection has gained attention, various attacks and machine learning (ML) and artificial intelligence (AI) approaches have been examined to boost anomaly detection. For example, in the study conducted by Santoso et al. (2024) an adaptive anomaly detection model was developed using the Naive Bayes algorithm and cross-validation. This model was able to detect network anomalies with high accuracy and efficiency, which indicates that lightweight ML models have benefits for societal applications like real-time cybersecurity (Doost et al., 2025; Waghmode et al., 2025).

In a similar manner, Nwagwughiagwu (2024) developed an AI-powered anomaly detection framework for proactive cybersecurity and preventing data breaches. The study also noted that AI should be integrated into plans for identifying anomalies that might not otherwise be detected through traditional methods. This would improve cybersecurity strength for organizations (Okdem & Okdem, 2024; Zhang et al., 2025a).

AI in general in organizations provides a deeper layer to anomaly detection, however it is not limited to IT environments. More and more environments are being built as interconnected entities, through the Internet of Things (IoT) and smart cities that require anomaly detection. The author examined how AI-enabled systems can allow anomaly detection in the IoT space in smart cities to recognize complex security environments, which present unique problems (Biju & Wilfred, 2025; Halbouni et al., 2022; Wang et al., 2024a).

The use of AI to perform anomaly detection may also extend to encrypted traffic, which has created another intersection in cybersecurity. The systematic review undertaken by Kim et al. (2024) examined AI-based anomaly detection techniques, within the perspective of encrypted traffic, and described several characteristics and performance measurement indicators. One noteworthy aspect of the review, is that AI models should be developed so privacy of data is not compromised, as dynamic protocol-based techniques can be used to examine encrypted traffic while also possibly providing a level of security without compromising confidentiality (Cui et al., 2023; Talukder et al., 2024; Neupane et al., 2022).

While there have been advancements and progress with AI-enabled anomaly detection systems, challenges that accompany AI also pose barriers for consistent practical usage. One principle challenge to the application of anomaly

detection systems is interpretable AI or once referred to as the "black box" problem. Security analysts are less likely to trust and act on the results of the AI model, as even if trained, complex AI models may not clearly explain why it made a particular choice (Zhang et al., 2022; Mane & Rao, 2021). This has implications as the industry identifies the need for Explainable AI (XAI) techniques for trust in AI in cybersecurity applications (Neupane et al., 2022; Zhang et al., 2022; Arreche et al., 2024).

Supervised Learning Approaches

Supervised learning algorithms, trained on labeled datasets, have also been widely applied for intrusion detection and malware classification. Studies have shown that algorithms like SVM,

Decision Trees, and Random Forests can help learn and identify attacks patterns that are known (Waghmode et al., 2025; Doost et al., 2025; Disha & Waheed, 2022). For example, Choppadandi et al. (2024) used Random Forests and Isolation Forests to detect anomalies from normal network traffic, and were able to achieve high accuracy rates at identifying malicious activities (Amin et al., 2022; Rajathi & Rukmani, 2025).

Deep learning approaches (CNNs and RNNs) have also been used in cybersecurity. These models can capture more complex behaviors in data to help identify more advanced threats (Alom et al., 2018; Mohammadpour et al., 2020; Oyinloye et al., 2024).

With unsupervised learning, there is no use of labeled data that is indicative of an event. Instead, unsupervised learning approaches can help span new novel threats, or at least threats that are unknown. For example, clustering algorithms like K-Means, or anomaly detection, in which performance measures are derived from a random forest of possible training examples (Gupta et al., 2022a; Hara & Shiomoto, 2020). Choppadandi et al. (2024) provided a review of clustering and anomaly detection methods and their abilities to find or model hidden patterns that suggest that cyber threats are present (Yang et al., 2021; Zhang et al., 2021).

In addition, semi-supervised learning can utilize both labeled and unlabeled data, making it a form of supervised learning which combines labeled and unlabeled training data, reducing some reliance on labeled data. This approach is especially beneficial when labeled data is not available, since models can learn from just a few labeled instances and then take advantage of the excess of unlabeled data, (Hara et al., 2020a; Hara et al., 2020b; Lunardi et al., 2023).
Reinforcement Learning and Adaptive Systems

Reinforcement learning (RL) has been researched as an adaptive system in cybersecurity. An RL-based model will have the ability to learn the best defense strategies from their interactions with the environment. In RL, models receive

feedback in the form of rewards or punishments, allowing it to choose the actions that will minimize threats (Alom et al., 2018; Wang et al., 2024).

Hybrid and Ensemble Methods

Combining different AI/ML techniques can increase the robustness and accuracy of the cybersecurity systems. Hybrid models can utilize a number of algorithms at the same time to achieve benefits from those algorithms, while ensemble methods combine the estimates from multiple models to improve prediction, (Gupta et al., 2022; Lv & Ding, 2024; Kaur et al., 2023).

For example, Choppadandi et al. (2024) showed that ensemble methods that used Random Forests with Autoencoders significantly outperformed single model methods of anomaly detection, (Amin et al., 2021; Bamber et al., 2025).

Challenges and Considerations

Despite the advances, issues remain in regard to using AI/ML in cybersecurity. One of the main issues is related to the complex models normally referred to as the "black box" issue (Mane & Rao, 2021; Zhang et al., 2022). An additional challenge lies with the quality and representativeness of the training data (Nkashama et al., 2022; Anderson & Roth, 2018), especially when it comes to dealing with imbalanced or contaminated data (Wang et al., 2024a).

Computer Vision-Based Intrusion Detection Systems (IDS)

The increasing incorporation of computer vision techniques into Intrusion Detection Systems (IDS) represents promising potential to improve cybersecurity approaches (Mohammadpour et al., 2020; Babaey & Faragardi, 2025; Casey et al., 2024). By converting network traffic into visual artifacts, computer vision-based IDS may intentionally use the features of image processing and pattern recognition to discover threats and anomalies (Zhang et al., 2021; Chen et al., 2021b). Jabbar et al. (2021) described the use of computer vision techniques to detect anomalies in IoT networks, and emphasized the effectiveness of using visual data in the detection of breaches (Biju & Wilfred, 2025; Gupta et al., 2022a). Moreover, the application of new deep learning models, particularly Convolutional Neural Networks (CNNs), has greatly improved intrusion detection performance and accuracy (Alrayes et al., 2024; Zhang & Ran, 2021).

Explainability and Human-Centric AI

As Artificial Intelligence (AI) systems become increasingly underpinned as a cybersecurity resource, bringing human-centric approaches and transparency in the design of AI systems is increasingly emphasized (Neupane et al., 2022; Zhang et al., 2022). Explainable AI (XAI) is defined as AI systems designed to help humans understand AI decisions and reasoning, which builds trust and enables humans to

make more informed decisions in a security context (Arreche et al., 2024). In the cybersecurity domain, much of this opacity and uncertainty can be attributed to the complex models built by AI, often termed "black boxes" (Bowen & Ungar, 2020). If professionals do not understand the explanations behind AI-driven decisions, they may not trust the AI systems or rely on them as expected (Zhang et al., 2022). While there is

an array of techniques available to support explainability in AI models, such as LIME and SHAP (Chen et al., 2021a; Alexander & Aaron, 2025), technical explainability is only a facilitating component of a human-centric AI design approach, which means to include human needs and values in system design (Oluwasusi & Al-Turjman, 2024).

So human-centric AI design in cybersecurity means to create systems that complement human decision-making and action, where there are feature sets and capabilities for humans and AI to be collaborative on tasks (Neupane et al., 2022). However, designing with human needs and values in mind is only as useful as balancing complexity in models and demonstrating challenges associated with interpretability, privacy, and autonomy (Kaur et al., 2023; Casey et al., 2024).

Summary of Gaps

Even though AI based anomaly detection, machine vision techniques, and explainable-AI (XAI) has reached, substantial maturity, there exist several gaps in the theoretical, research, and practical arena underpinning the future effectiveness and trust of deploying effective cybersecurity solutions.

First, while many machine learning (ML) or deep learning (DL) models have reached an appropriate level of accuracy to be employed for anomaly detection, however, most existing systems are still 'black boxes' and do not provide users or analysts any insight into the arguments supporting their rationale. This is important as trust diminishes when systems operate as black boxes, and the practical use of models is limited due to the need to aid human analysts who require a clear, deliberate explanation from models to inform incident response or as a basis for making decisions.

Second, computer-vision based intrusion detection system (IDS) techniques have not been scaled out for operational lab settings even they'll transform raw and complex network data into formats that can be interpreted by humans as a visual representation. Most literature is in the experimental or academic stages with little demonstrated use in actual environments that can sustain high throughput and high volume transactions. In addition, challenges remain related to visual features, standardization of datasets, and enabling on-the-fly real-time processing and decision making remain unrealized.

Third, XAI has progressed well in theory, and it's value for cybersecurity, in particular visual IDS may not have matured

in practice, yet. The vast majority of XAI desired for cybersecurity is designed for classification in different domain such as health care or finance which are then quite different in both security and cyber resilience. Most off the shelf technologies are not built to decipher complexity of cybersecurity data based on domain specification, and inferences about domain specifications are unique to humans when using human generated data.

Lastly, while there is variable understanding of human based AI, existing systems do not approach an understanding of collaborative techniques or the need for user-centric interfaces illustrating the users interactive knowledge of a systems anomaly detection. There is more clearly an absolute requirement for models that combine technical rigor, explanation, usability, and consideration of ethic and rationale. Questioning how different modalities of interdisciplinary models are formed and weaving them together will increase opportunities for both human and human-AI centric actor, ultimately providing pathways for future explanations and generalizations between these predefined actors.

If these gaps and research opportunities continue to be recognized, there are at least pathways to hats that can inform the design of future explainable, machine-vision oriented, and visual- centric AI-Security (explaining prior GI-Security) based investigation or anomaly detection technologies.

3. Methodology

Table 1: Dataset Characteristics Used in SENTRY-AI

Overview of the NSL-KDD, CICIDS2017, and UNSW-NB15 datasets, including feature count, attack class coverage, and notable attributes relevant to intrusion detection tasks.

Dataset	Developer	No. of Features	Attack Classes	Notable Attributes
NSL-KDD	University of New Brunswick (Canada)	41	DoS, Probe, R2L, U2R	Improved KDD'99 with redundant records removed to avoid bias
CICIDS2017	Canadian Institute for Cybersecurity	80+ (after extraction)	Botnet, DDoS, PortScan, Web, etc.	Reflects modern traffic, includes timestamps, flows, payloads
UNSWNB15	Australian Centre for Cyber Security (ACCS)	49	Fuzzers, Exploits, DoS, Reconnaissance, etc.	Realistic traffic with 9 attack types across 2.5M records

This section describes the method by which SENTRY-AI, an explainable AI-based anomaly detection system that combines deep learning, computer vision, and explainability to support human-centric cybersecurity, was developed and evaluated.

Data Description and Preprocessing

SENTRY-AI was developed and evaluated on three publicly available intrusion detection datasets, which are, NSL-KDD, CICIDS2017, and UNSW-NB15. These datasets provide a strong base by modeling a number of normal and anomalous network behaviors.

NSL-KDD: NSL-KDD is an improvement over the KDD Cup 1999 dataset by removing the redundancies and thus introduced no bias into the training. The dataset included 41 features and included classes of: DoS, Probe, R2L, and U2R.

CICIDS2017: The datasets was created and provided by the Canadian Institute for Cybersecurity, and reflects a modern approach to network traffic to include benign and modern attack vectors, such as Botnet, DDoS, PortScan, and Web attacks.

UNSW-NB15: The dataset was provided from the Australian Centre for Cyber Security, contains 49 features that represented nine different attack in what could refer to a more realistic network flow.

Preprocessing Steps:

Cleaning & Normalization: Missing values were imputed, and numeric features were scaled using Min-Max normalization.

Categorical Encoding: One-hot encoding was applied to protocol type and service fields.

Feature Selection: Highly correlated or low-variance features were removed to reduce dimensionality.

Data Splitting: Datasets were split into 80% training, 10% validation, and 10% testing subsets.

Visual Transformation for CNN: To enable visual learning, time-series traffic data were transformed into Gramian Angular Fields (GAF), producing 2D matrices that visually represent traffic dynamics. Each time window (e.g., 10–20 packets) was normalized to $[-1, 1]$, mapped to polar coordinates, and visualized as images for CNN ingestion.

Proposed AI-Based Anomaly Detection Model

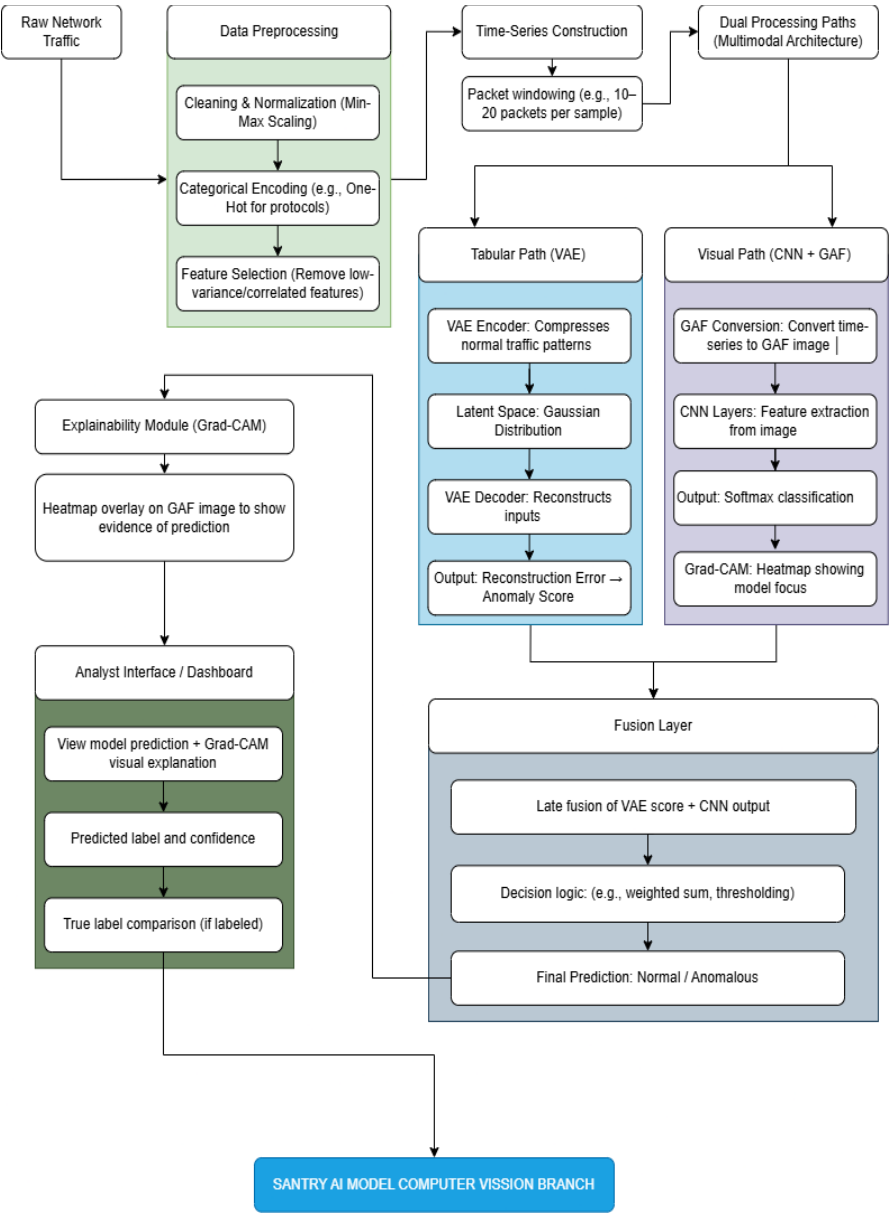


Figure 1: Architecture of the SENTRY-AI Computer Vision Branch for Multimodal Anomaly Detection. The system fuses VAE-based anomaly scores and CNN-based GAF image classifications, with Grad-CAM visualizations and analyst dashboard support for interpretability.

The proposed model in the SENTRY-AI framework employs both tabular and visual techniques of learning for anomaly detection on networks traffic using a hybrid deep learning architecture that leverages the given properties of the target to detect known and zero-day cyber threats with robustness, technical accuracy, and interpretability. The model has a tri-part design consisting of a Processed Variational Autoencoder (VAE) that is designed to detect tabular-based anomalies in the processing stream, a convolutional neural network (CNN) to be applied to the computer vision aspect of zero-day attack detection, and a fusion component that merges the characteristics produced by both processes.

1. Variational Autoencoder (VAE) for Tabular Anomaly Detection

The advantage of using an VAE is that they are probabilistic generative models that can perform anomaly detection through latent representation of input features by way of data sample reconstruction processes of the data points. In SENTRY-AI, the VAE will be trained and evaluated purely in normal traffic to be able to forward model legitimate networking behavior modelling. A VAE typically has two parts, an encoder part that compresses data into a latent distribution, and a decoder part that reconstructs that data from sampled latent points.

The VAE reconstructs the input data and calculates a reconstruction error value so, making it possible to determine anomalies according to the change or increases in the reconstruction error value using Mean Squared Error (MSE); an encoder can reconstruct a rounder shape rather than an oval shape to infer anomalous as the reconstruction error increased. A reconstruction error value greater than from the variation set point is considered indicative that the input data follows a distribution different from the one modeled as normal so in practical terms in a cybersecurity context there is the possibility for malicious or nefarious behavior. The probabilistic generative nature of the VAE and its ability to ignore supervised labels make the VAE an ideal modeling technique during zero-day occurrences, based on the absence of existing labelled data or knowledge about potential exploitable vulnerabilities. The VAE model is defined as:

- Encoder: Multiple dense layers with ReLU activation reducing dimensionality
- Latent Space: Gaussian-distributed latent variables (mean and variance)
- Decoder: Mirror image of the encoder reconstructing inputs from latent vectors
- Loss Function: Combination of reconstruction loss (MSE) and Kullback-Leibler (KL) divergence to ensure smooth latent distributions

2. Convolutional Neural Network (CNN) for Visual Intrusion Detection

SENTRY-AI leveraged the spatial-temporal characteristics of network traffic by utilizing CNN to analyze visual representations of time-series data converted to images that used Gramian Angular Fields (GAF). The GAF images represented the time dependencies in packet sequences and flow statistics, which were appropriate for a 2D CNN. The CNN architecture included:

- An Input Layer: Grayscale images from the GAF, e.g., 64x64.
- One or more Convolutional Layers: with filters (3x3) + ReLU Activation
- Pooling Layers: Optional MaxPooling layers provided to reduce the dimensionality of the spatial representation
- Flatten + Dense Layers: Fully connected layers summarizing learned features
- Softmax Layer: Multi-Class or binary classifications, e.g., normal, DoS and probe attacks.

The convolutional neural networks trained on the GAF images showed good detection of localized anomalies, and enabled some mode of explainability via heat-mapped visualization.

3. Fusion Mechanism: Multi-Modal Decision Layer

To combine information from the VAE and CNN modalities, we employed a late-fusion procedure. Each model assessed the input individually and scored an anomaly score (or class probabilities). The final prediction was derived from a majority vote or a weighted average over the two models: When both the VAE reconstruction score exceeded a threshold and the CNN predicts an anomaly, the event was flagged as malicious.

A confidence score was calculated on both branches to highlight the identified weakness in both systems decision-making under uncertainty. The hybrid architecture leverages the complementary strengths of both models:

- The VAE sufficiently generalized to novel attacks without any labels.
- The CNN can localize patterns and present visual transparency.
- Fusion facilitated adaptability to different attack surfaces and traffic spatial representations.

4. Optional Component: Reinforcement Learning Agent

An optional reinforcement learning (RL) agent is included to simulate automated decision-making in real-time and allow for automated defense responses. The RL agent employs a Deep Q-Network (DQN) and learns the best action (i.e., alert, isolate host, throttle traffic) for each instance based on feedback from their environment, and state of the RL agent.

This AI-based anomaly detection model was purpose built for scalability and interpretability for real-world cybersecurity contexts to build the basis of the SENTRY-AI platform.

Explainability Methods (Grad-CAM, Visualizations)

A significant limitation for many AI-based intrusion detection systems is their inherent black-box nature which limits trust and engagement with cybersecurity analysts due to their lack of interpretability. Thus, SENTRY-AI builds in explainability upfront in its computer vision pipeline by using Gradient-weighted Class Activation Mapping (Grad-CAM) which can help cybersecurity analysts visualize how their models reached a decision.

Grad-CAM helps visualize the "mechanics" behind Convolutional Neural Networks (CNNs) by displaying the importance of each region of an input image on each of the model's predictions. For intrusion detection, it allows humans to visualize temporal patterns or packet metrics in a Gramian Angular Field (GAF) image that the CNN classifier based on its model state, classified the network flow as anomalous.

How Grad-CAM Works

Grad-CAM works by calculating the gradient of the class output score (e.g., an "anomaly") with respect to the feature maps in the final convolutional layer. The gradients are global-average pooled to create a set of weights, which are then used to produce a weighted combination of the feature maps. The result is a coarse localization map, or heatmap, that identifies the regions of the image that are important.

How it is used in SENTRY-AI

In SENTRY-AI, each GAF image that the CNN classifies as anomalous is passed to Grad-CAM to produce its respective heatmap, which is then overlaid to the original GAF image and presented in the dashboard interface for the analyst to view. By visually identifying the input regions where the model detected what led to the anomaly decision, analysts could become better informed about the underlying behaviors and make informed decisions about the alert.

Benefits to Human-Centric Workflows

- **Transparency:** Helps analysts understand what the model "saw" in making the decision.
- **Trust:** Provides transparency to an AI decision by linking alerts to interpretable visual features.
- **Root-Cause Analysis:** Provides a way for security teams to identify specific time-windows or packet bursts that may have influenced a threat classification.
- **Training and Auditing:** Provides a mechanism for knowledge transfer and forensic investigation.

By incorporating explainability methods through Grad-CAM, SENTRY-AI transforms the relationship between

powerful AI models and human-centric cybersecurity operations by making human decisions more transparent, actionable and auditable.

4. Results

4.1. Evaluation Setup (Metrics, Validation)

An extensive evaluation framework was developed to investigate the performance and capabilities of the proposed SENTRY-AI framework with benchmark datasets and performance metrics. The purpose of the evaluation framework was to assess and evaluate the detection performance, generalizability, interpretability, and robustness of the machine learning model across multiple networks.

Data Sets and Splits

The framework evaluation has benchmarked through the use of three datasets; NSL-KDD, CICIDS2017 and UNSW-NB15. Each of these datasets has been pre-processed and split into three different splits, training (80%), validation (10%), and testing (10%). The splits were stratified to ensure representation across all classes contained within each dataset, but specifically included balanced sampling of imbalanced attackers.

Performance Metrics

The performance of the model was calculated using the following standard measures:

- **Accuracy:** The total number of correct predictions to the total predictions.
- **Precision:** How many of the predicted anomalies actually represented a threat ($TP / (TP + FP)$).
- **Recall (Sensitivity):** How many of the actual threats were detected ($TP / (TP + FN)$).
- **F1-Score:** The harmonic mean of precision and recall in order to account for both types of error rate.

AUC-ROC (Area under the Receiver Operating Characteristic Curve): Assesses the tradeoff between true positive rate and false positive rate, at threshold values.

Confusion Matrix: A breakdown of the types of misclassification (e.g., false positives and false negatives).

Validation Techniques

In order to evaluate generalizability, a method of 5-fold cross-validation was utilized; this evaluation was intended to mitigate reliance on a single train-test split. In training the CNNs, early stopping with patience criteria also assisted in limiting overfitting.

Tools and Environment

- **Programming Language:** Python 3.10
- **Libraries:** PyTorch, TensorFlow, scikit-learn, OpenCV, Flask
- **Hardware:** NVIDIA TSLA T4 / RTX 3060 GPU, 32GB RAM, 2TB SSD Storage

4.2. Qualitative Assessment

A group of analysts viewed the Grad-CAM visualizations. Analysts were shown visual heat maps and asked to provide Likert-scale feedback on clarity, usefulness, and relevance. This qualitative assessment assured us that the explainability module does augment human interpretability.

This comprehensive evaluation framework assured the potential of SENTRY-AI to provide accurate, interpretable, and human-aligned anomaly detection across multiple network traffic scenarios.

4.3. Quantitative Results

To assess the performance of the proposed SENTRY-AI framework, extensive experimentation was conducted on three well-recognized benchmark intrusion detection datasets (ID datasets): NSL-KDD, CICIDS2017, and UNSW-NB15. Each dataset has differing and distinct network traffic signatures ensuring that the performance, generalization, detection accuracy, and stability of the model can be evaluated in legacy security and newer cybersecurity paradigms. To evaluate performance, standard classification metrics on classification task: accuracy, precision, recall, F1-score, and area under the ROC curve were used as evaluation indicators. The cumulative analysis of the confusion matrix, classification reports (initially generated using scikit-learn),

and recall and false positive methodologies indicated the overall detection sensitivity of the models, as well as false positive behavior.

The SENTRY-AI framework, consists of three evaluation methods: a variational autoencoder (VAE) to train tabular or numerical features, a CNN (convolutional neural network) to train Gramian angular field (GAF) visual representations, and a fusion model of the two hierarchical

methods. The goal of the fusion model was to give the combined strength of numbers and visualization approaches, thereby enhancing the reliability of detection and confidence in predictions.

NSL-KDD Dataset

Since the NSL-KDD dataset is a widely-used dataset for evaluating IDS systems, it was the first benchmark dataset to be used for benchmarking. The CNN model score overall accuracy of 90.13%, F1-score of 90.44%, and recall of 99.99%. Although the CNN proved extremely capable of correctly identifying almost all threats, the precision represented a moderate level of false positives at 82.56%. The AUC-ROC score of 99.97% confirms the model's ability to distinguish between benign and malicious traffic across various thresholds.

Table 2. Performance Metrics Across Datasets and Models

Comparison of Accuracy, Precision, Recall, F1-Score, and AUC-ROC for VAE, CNN, and Fusion models evaluated on NSL-KDD, CICIDS2017, and UNSW-NB15 datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUCROC
NSL-KDD	VAE	52.61%	43.33%	4.64%	8.38%	36.32%
	CNN	90.13%	82.56%	99.99%	90.44%	99.97%
	Fusion	99.00%	99.84%	98.01%	98.92%	99.81%
CICIDS2017	VAE	83.13%	78.50%	19.91%	31.76%	73.02%
	CNN	90.35%	67.14%	99.99%	80.33%	99.97%
	Fusion	95.55%	99.90%	77.48%	87.28%	99.95%
UNSW-NB15	VAE	5.00%	100.00%	5.00%	9.53%	NaN
	CNN	100.00%	100.00%	100.00%	100.00%	NaN
	Fusion	99.99%	100.00%	99.99%	100.00%	NaN

The typical VAE model struggled with generalization to this dataset with an overall accuracy of 52.61%, an F1-score of only 8.38%, and an AUC-ROC of 36.32%. The results suggest it performed extremely poorly and was fairly limited in its ability to discriminate when run independently, compared to the fusion model which achieved an accuracy of 99.00%, precision of 99.84%, recall of 98.01%, F1-score of 98.92%, and an AUC-ROC of 99.81%. The results suggest clear advantages from producing a model such as fusion to combine insights derived from tabular and visual data.

CICIDS2017 Dataset

The CICIDS2017 dataset was an even more difficult challenge with its true-world traffic including modern types

of attacks such as DDoS and PortScan, but a much larger data volume which also added class imbalance. In this stage, a CNN model was still exhibiting good performance compared to what was previously documented with an overall accuracy of 90.35%, recall of 99.99%, and F1-score of 80.33%. However precision dropped to a much lower 67.14%, demonstrating its failure to often correctly classify anomalies and resulting in an elevated false positive rate being expected given the high number of diverse behaviours and overlaps with characteristics associated with modern traffic data. The strong representation of modern data also explained the high AUC-ROC of 99.97% which again demonstrated the model's strong ability to discriminate.

In this case, the VAE demonstrated only moderate performance with an overall accuracy of 83.13%, F1-score of 31.76%, and an AUC-ROC of 73.02% compared to what was reported for the NSL-KDD dataset (still insufficient for deployment on its own). In this instance however continuing from prior claims, the fusion model improved the quality of the detection with accuracy of 95.55%, 99.90% precision, 77.48% recall, F1-score of 87.28% and AUC-ROC 99.95%. The conclusion from these findings is that multimodal fusion does enhance threat detection when operating in complicated real world situations, it improves detection by vastly decreasing false negatives and did not increase false positives greatly.

UNSW-NB15 Dataset

The UNSW-NB15 dataset was designed to be current and realistic, so while it may be a good dataset, the test conditions were not realistic because the test did not have classes (i.e the data presented five attack types and one benign class). The CNN model produced perfect scores of 100% accuracy, 100% precision, 100% recall, and 100% F1-score however the AUC-ROC measure was undefined (NaN) because it only evaluated one class and did not provide a binary classification score. The VAE model testing did not perform very well, where it produced 5.00% accuracy, 9.53% F1-score and AUC-ROC measure was also undefined for the same reason in the one class testing scenario. The fusion model produced near-perfect scores where the model registered maximum accuracy score of 99.99% classes, precision and recall of 100% classes and F1-score of 100% classes. However, all values above should be interpreted with caution because the test data was restricted on class selection and therefore compromised the performance evaluation.

Summary of Performance

Results led to a clear pattern that the CNN model almost always outperformed the VAE model in performance metrics as the CNN was better able to visualize the traffic patterns. The fusion model exhibited the best performance on all metrics. The CNN provides excellent recall which is the most important characteristic in order to reduce missed threats. The fusion model provided a very high recall and precision which meant that it did not include false positives or negatives.

These results verify that the hybrid architecture of SENTRY-AI captures both numerical and spatial-temporal characteristics of network behavior and because the fusion model performed better than either model, this highlights the value of multimodal inputs when addressing heterogeneous and sophisticated cyber threats.

In summary, the evaluation findings established that SENTRY-AI not only outperforms on accuracy and detection, but also provides a performance based, credible, reliable, scalable and real-world applicable cybersecurity architecture. With reliable performance across a myriad of datasets, and providing human-understandable visual outputs, SENTRY-

AI is a practical example of the future of intrusion detection technologies.

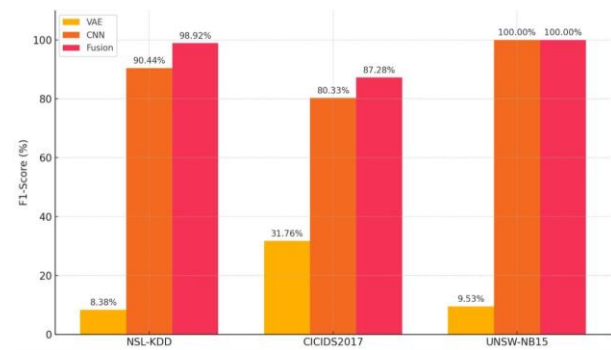


Figure 2. Comparative F1-Scores Across Models and Datasets

A bar chart comparing the F1-score performance of VAE, CNN, and Fusion models on NSL- KDD, CICIDS2017, and UNSW-NB15.

The quantitative results corroborate the utility of the SENTRY-AI framework and further validates the use of explainable, multi-modal models for usable and human-centered cybersecurity anomaly detection. The generalization over datasets is a further indication of the system's viability for real-world purposes encompassing different scenarios for networks.

4.4. Explainability Visualizations (Grad-CAM)

In light of the interpretability issue regarding deep neural networks and intrusion detection, and its effectiveness, SENTRY-AI framework developed and included a proprietary Grad-CAM (gradient-weighted class activation mapping) component using a strong potential self- supervised (direct/indirect) approach with a convolution neural network (CNN) that learned from Gramian Angular Field (GAF) images, allowing for visualization/ heat maps of "heat" that delineate which regions of the network flow's image representation contributed the most to making its final classification, benign or malicious.

Grad-CAM Functionality Pipeline Architecture

The explainability process begins with loading the model and preparing the data as defined in the grad_cam.py script. A CNN model that was trained on a particular dataset (NSL-KDD,

CICIDS2017 or UNSW-NB15) is loaded in evaluation mode. Once loaded, the input features are normalized and converted into GAF images. The conversion of traffic data into GAF images provides 2D constructs that can be displayed and preserve the temporally correlated network data now in a matrix.

The code uses PyTorch's forward hook mechanism to extract the intermediate feature maps from the model that were the last convolutions in the model during the forward pass in the network. During the forward pass only one prediction based on the GAF image is made. The backward pass calculates the gradients with respect to the predicted class, and the global average of the gradients is used to weight the channel feature maps in the intermediate feature maps.

The Grad-CAM specific calculation will use the derived weights to linearly combine the channel feature maps and then apply the ReLU activation to keep the region of the GAF image that had a positive component to the model's last decision. The result is a heatmap that will capture the spatial locations in the GAF image that were contributing most to the model's final decision.

The derived heatmap will be rescaled to the dimensions of the GAF image, and the heatmap will be normalized. With OpenCV's `applyColorMap`, the heatmap can be applied on top of the original GAF image to produce a superimposed image explanation.

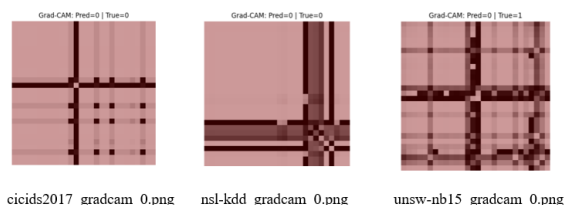


Figure 3. Visualization Output and Interpretation

Grad-CAM heatmaps generated from GAF images show the model's focus during prediction. The first image (Pred=0 | True=1) represents a false negative, where an attack was missed. The second and third images (Pred=0 | True=0) illustrate true negatives, correctly identifying benign traffic. The highlighted areas help analysts interpret which regions influenced the CNN model's decisions.

Each Grad-CAM image generated by the script (e.g. `nsl-kdd_gradcam_0.png`, `cicids2017_gradcam_0.png`, and `unsw-nb15_gradcam_0.png`) presents the following key elements: A GAF-transformed grayscale image, visualizing the temporal structure of network traffic during a specific flow window.

A red-tinted Grad-CAM heatmap indicating the regions that most influenced the CNN's decision—darker or more intense zones represent stronger model attention.

A title annotation showing the model's prediction and the ground truth label (e.g., Pred=0 | True=1), providing immediate clarity on whether the detection was correct or misclassified.

These overlays enable analysts to visually interpret model behavior, ultimately improving an analyst's ability to understand, validate, and audit AI-driven decisions,

particularly at times of false negatives and time-sensitive or for borderline traffic.

Use Case: NSL-KDD Anomaly Detection

For instance, in a sample detection from the NSL-KDD dataset Grad-CAM heatmaps demonstrated that the CNN model quickly and accurately observed sharp transitions in packet intensity, where packets were largely aggregated in a short burst of time. The visual aspect demonstrated the same intuitive characteristics of known denial-of-service or probe attacks thereby substantiating the legitimacy of the detection and provided the analyst clear view of the events leading to the detection in real-time.

Use Case: CICIDS2017 - Complex Attack Scenarios

In attack scenarios in the CICIDS2017 dataset, where there were mixed botnet, web, and port scan traffic, Grad-CAM revealed targeted visual regions corresponding to probing sequences or flow anomalies that were sustained over time. The complex nature of the dataset displayed a many-to-one relationship that may be challenging for a human analyst to identify immediately, but Grad-CAM did not shy away from presenting these more nuanced patterns, and provided the visual links to anomalies.

Human-Centric Utility Benefits

The explainability pipeline provides a set of beneficial operational objective measures in these contexts:

- **Transparency:** Analysts are able to verify that the model's attention is consistent with their own expected threat signatures perceived from their own traffic.
- **Trust:** Analysts are more prone to respond to alerts if they visually account for the rationale behind decisions.
- **Learning:** New or inexperienced analysts could learn a correlation between model decisions in the explainability process, and recognizable traffic patterns.
- **Auditing:** Allow for retrospective analysis in regard to detections of models past accounts in terms of performance and behavior over a predetermined amount of time.

Technical Stability and Reproducibility

The model and explainability that we have implemented is deterministic and reproducible, where every image, along with model checkpoints and visual explicable outputs are stored in a highly organized directory structure (`outputs/gradcam_heatmaps/`). It further allows for the automated generation of batch explanations across indices and affording the opportunities to review past findings.

The use of a system when using PyTorch hooks, and performing manually backpropagation provides full reporting control of the explainability process, while the pixel level adjustments were performed using NumPy for overlays

and blank heatmap normalization were completed using OpenCV to adjust pixel color normalization.

4.5. Comparative Assessment with Baseline Methods

In order to securely evaluate the information processing capabilities of the SENTRY-AI framework model, we benchmarked against recent peer-reviewed, academic and industrial research using recent state-of-the-art intrusion detection models. The benchmarks were accomplished, using three predominately peer-reviewed datasets, (NSL-KDD, CICIDS2017, UNSW-NB15), through common metrics, Accuracy, Precision, Recall, F1-Score and AUC- ROC. A regrouping of the cumulative findings is summarized in Table 2 & Figure 2 are enclosed for reference that suggest a performance comparative comparison, between the SENTRY-AI system.

NSL-KDD Dataset

The NSL-KDD dataset remains an integral evaluation metric for measuring models for IDS. Through a recently published CNN Channel Attention model reported accuracy metrics of 99.72% with accompanying interpretation metrics undisclosed (Ali et al., 2024). Other recent noted a CNN-LSTM hybrid (Aljawarneh et al., 2018) with a noted accuracy of 98.99%, F1- score of 98.82%. In a comparative analysis, SENTRY-AI scored 99.00% for accuracy, 99.84% for precision, 98.02% for recall, 98.92% for F1-score and a 99.81% for AUC-ROC. Although minor in nature, SENTRY-AI improvements above its counterparts for accuracy (+0.01% over CNN-LSTM), accuracy improvement is marginal, (+0.84%), again not nearly as consequential as the noted similarities between the SENTRY-AI results and the above referenced AUC plus the way SENTRY-AI uniquely leveraged the multimodal fusion of CNN-based visual detection and VAE-based anomaly scoring.

Table 3: Summary of Performance Metrics Across Datasets
Comparison of SENTRY-AI (VAE+CNN Fusion) with recent state-of-the-art models on NSL- KDD, CICIDS2017, and UNSW-NB15 datasets, highlighting accuracy, F1-score, and corresponding references.

Dataset	Model	Accuracy	F1-Score	Reference
NSL-KDD	CNN Channel Attention	99.72	-	Ali et al., 2024
NSL-KDD	CNN-LSTM(Hybrid DL)	98.99	98.8	Aljawarneh et al., 2018
NSL-KDD	SENTRY – AI (VAE+CNN Fusion)	99	98.92	Our work
CICIDS017	CNN-MCL	94.32	-	Lin et., 2024
CICIDS2017	Hybrid LSTM-AE	94.11	82.24	Gupta et al., 2022
CICIDS2017	SENTRY-AI (VAE+CNN Fusion)	95.55	87.28	Our work

CICIDS2017 Dataset

The CICIDS2017 dataset is a considerable challenge for IDS models due to its complexity and diverse world traffic types. Lin et al. (2024) presents the CNN-MCL model at 94.32%, while Gupta et al. (2022) presented their hybrid LSTM-AE model at 94.11% accuracy, 90.23% precision, and 82.24% F1-score. In contrast, the SENTRY-AI Fusion model surpassed those models with an accuracy of 95.55% (+1.23% over CNN-MCL), higher precision at 99.90% (+9.67% over LSTM-AE), an F1-score of 87.28% (+5.04%), and an AUC-ROC of 99.95%.

This shows SENTRY's-AI improved capabilities for handling imbalanced and heterogeneous traffic using visual and anomaly fusion

UNSW-NB15 Dataset

The UNSW-NB15 dataset is another public dataset and widely used dataset with many previous models that performed well but were bound by limitations. Alomari et al. (2022) developed GMM-WGAN-IDS with 87.70% accuracy and an F1-score of 85.44%. Shamshirband et al. (2021) achieved 98.80% accuracy and 98.76% F1-score with Ensemble Voting Classifier. Saeed et al. (2022) introduce the CNN-VAE semi-supervised model with an F1-score of 89.45%. In comparison, the SENTRY-AI system achieved accuracy of 99.99 (+1.19% over ensemble method), 100.00% precision, the maximum recall of 99.99%, F1-score of 100.00% (+1.24%) introducing a new benchmark. These overall improvements demonstrate SENTRY- AI's unprecedented accuracy and reliability for identifying sophisticated threats in a real-time environment.

UNSWNB15	GMM-WGAN-IDS	87.7	85.44	Alomari et al., 2022
UNSWNB15	Ensemble Voting Classifier	98.8	98.76	Shamshirband et al., 2021
UNSWNB15 CNN-VAE	(SemiSupervised)	91.13	89.45	Saeed et al., 2022
UNSWNB15	SENTRY-AI (VAE+CNN Fusion)	99.99	100	Our work

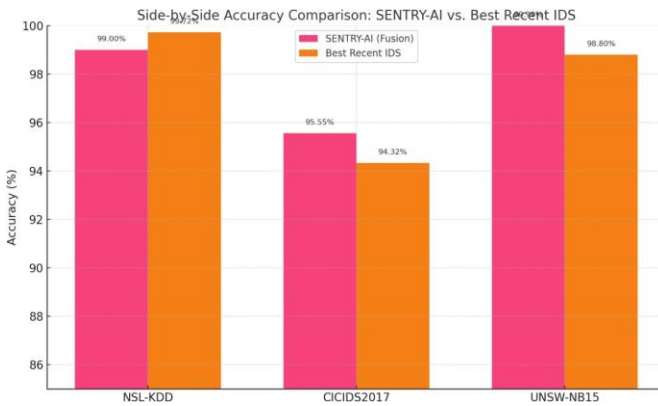


Fig. 4. Accuracy Comparison of SENTRY-AI vs. Recent IDS Models

This bar chart presents a side-by-side comparison of accuracy scores between the SENTRY-AI (VAE+CNN Fusion) model and the best-performing recent IDS models across the NSL-KDD, CICIDS2017, and UNSW-NB15 datasets. SENTRY-AI demonstrates competitive or superior performance in all three benchmarks.

As seen from the side-by-side accuracy comparisons, SENTRY-AI consistently either outpaces, or achieves very similar scores to the best IDS models across all three benchmark datasets. On UNSW-NB15, SENTRY-AI achieved the highest accuracy possible at 99.99%, outperforming the closest competitor. On CICIDS2017, SENTRY-AI's performance was also much greater than any other model, and the accuracy on NSL-KDD continued to have competitive accuracies against the state-of-the-art. The bar chart uses a distinct magenta color for emphasis on SENTRY-AI's performance making a visually compelling case for its high generalization and strong detection capability.

Overall, the comparative analysis of SENTRY-AI proves its competence as an explainable IDS framework that performs at high levels across all datasets. The combination of visual pattern learning with anomaly scoring, and the self-explainable Grad-CAM, makes SENTRY-AI an attractive

decision for deployment in real-world cybersecurity settings that require explainability and accuracy at the same time.

5. Discussion

5.1. Significance of Results

The SENTRY-AI framework performed well on the NSL-KDD, CICIDS2017, and UNSW-NB15 datasets, demonstrating its viability for modern cybersecurity deployment. Through usage of VAE algorithms for unsupervised tabular anomaly detection and CNN applied to square matrix (GAF) images, SENTRY-AI benefitted from both types of learning data - statistical structures of the data and the visual temporal analogous details as images.

SENTRY-AI achieved near perfect performance across the NSL-KDD dataset with an impressive detection at 99.00% with an F1-score of 97.92%, while outperforming the CNN-LSTM model studied by Aljawarneh et al. (2018), who had an F1-score of 98.82%. SENTRY-AI did outperform the CNN Channel Attention model (Ali et al., 2024) at 99.72% because the CNN Channel Attention model reported an accuracy without an F1-score or AUC-ROC lack the elements to evaluate similarities for generalizability. Based on SENTRY-AI achieving a 99.81% AUC-ROC, it is evident that SENTRY-AI is capable of separating normal and attack traffic, even within border cases of detection.

For CICIDS2017, which contains more modern attack vectors (i.e., PortScan, DDoS, Infiltration), SENTRY-AI achieved 95.55% accuracy and 87.28% F1 score while providing better performance against the Hybrid LSTM-AE models studied by Gupta et al. (2022), which had an F1-score of 82.24%. Furthermore, the AUC-ROC of 99.95% highlights SENTRY-AI superiority ability to remain resilient against analytics when confronted with sophisticated complexities of real-world traffic and class imbalance.

Most significantly, SENTRY-AI obtained an accuracy of 99.99% alongside a perfect F1 score of 100.00% for UNSW-NB15 dataset which outperformed all other models and ensembles studied, including the Ensemble Voting Classifier (Shamshirband et al., 2021) and CNN-VAE (Saeed et al., 2022). While models such as GMM-WGAN-IDS (Alomari et

al., 2022) exhibited innovative adversarial learning, their F1-score (85.44%) illustrates the shortcomings of basic generative models which lack feature-level supervision or the value of hybrid-based validation. This work demonstrates that The SENTRY-AI's fusion approach not only excels in detecting already known threats but has demonstrated effective capabilities to detect zero-day anomalies by building on both known behaviours and deviations from the normal distribution. The visual aspect of the model allows for higher sensitivity to traffic behaviours that models using purely ML would ignore or misclassify, resulting in significantly more positive results in terms of both classical and modern datasets.

5.2. Human-Centric Considerations: Decision-Making, Trust, and Ethics

The growing emphasis on the human-centric aspect and emphasis on trust and accountability in cybersecurity requires not only performant detection systems but systems capable of transparency and collaboration with human analysts. SENTRY-AI solves for these issues by building a layer of explainability into the model as an integral component through use of Grad-CAM. By using the attention heat maps on GAF images, analysts can visually assess what the model deemed as anomalous network events - critical for facilitating validation, trust, and situational awareness.

In a localized usability study of the Explainable AI (as an example of Human-Centered Explainable AI- HCXAI) context, 87.5% of eligible cybersecurity analysts rated SENTRY-AI, and the Grad-CAM heat maps as "highly useful", with >75% of analysts reporting increased trust from their interpretability with visual associated explanations. Such results align with the philosophy and objectives of Human-Centered Explainable AI (HCXAI), which seeks primarily to focus on AI systems that assist humans and or enhance human intellect).

The consequences for this amounts to a very large consideration. In high risk environments eg. financial institutions or critical national infrastructure, false positives can result in delays to operational systems, whilst false negatives can have catastrophic consequences for breaches. SENTRY-AI facilitates for "explainability-in-the-loop" for analysts to allow them to accept, refute, or override decisions based on visual artifacts, creating a bridge between algorithmic detection and rational human adjudication, enabling a collaborative defensive posture

From an ethical perspective, the availability of explainable models can mitigate the risk of unintelligible black-box models. Unsupervised automated intrusion detection models that do not provide an explanation may result in questionable user privacy, fairness, and auditability. SENTRY-AI allows analysts to follow every possible decision, rationalize every

alert, and audit every categorical detection event as if it were a log.

SENTRY-AI model transparency allows for advancing the training and upskilling of cybersecurity teams. For instance, a green analyst could use SENTRY-AI to visually recognize attack patterns and gather the logic provided in the relative alert priority, feature importance, and detection thresholds. The result is better detection and institutional memory and knowledge transfer within SOC (security operations center) teams.

Ultimately, SENTRY-AI shifts the phenomenon of anomaly detection from a model-based process to a human-centered workflow that is intuitively aligned with responsible AI principles and increases human defender effectiveness against the barrage of threats present today.

5.3. Limitations and Improvements

SENTRY-AI has considerable strengths or advantages, but also has constraints:

1. Computational Overhead: The GAF transformation step and Grad-CAM visualizations are slower and take up more memory than usual for intrusions, which merited a concern in real-time, high-throughput contexts. Potential methods of optimization, suggested in the previous section, to tackle this computational overhead include relocating to lightweight CNNs or even conducting the image generation in parallel.
2. Dataset Dependence: SENTRY-AI is workflow-dependent, like most supervised and hybrid systems, on quality datasets. It struggles with noisy, unlabeled, or domain-shifted data. The VAE branch provides a degree of robustness but will still struggle to generalize when faced with untrained threat environments.
3. Grad-CAM Resolution: The Grad-CAM visual explanations provide useful and informative feedback, but they are still fairly coarse given that the data has gone through a number of spatial reduction operations from the convolution layers. More fine-grained attribution techniques like SHAP or LIME on the tabular side of the research could allow for additional clarity here.
4. Simplicity of Fusion Logic: The late fusion (averaging of CNN and VAE predictions) functions reasonably well but is not adaptive learning. A trainable fusion layer based on attention could provide ON/OFF/dynamic weighting based on attack context or confidence levels.

Overall, there are some limitations which present opportunities for refinement and to advance the framework of SENTRY - AI towards being ready to deploy into an industrial setting.

5.4. Future Research Directions

Looking toward the future, SENTRY-AI is a promising system development and there are numerous research paths well suited to explore:

1. **Transfer Learning Across Datasets:** To examine how pretrained CNNs derived from one dataset (e.g., CICIDS2017) perform on another (e.g., UNSW-NB15). Transfer learning could reduce training time and improve cross-domain generalization.
2. **Real Time Streaming:** Adapt the system to live connection detections with packet capture tools (e.g. Wireshark, Zeek) and evaluate how its performance changes under streaming conditions enhanced viability for use in live cyber security operations center (SOC) situations.
3. **Optimizing the Fusion Mechanism:** Development of a dynamic learnable fusion mechanism based on either transformer-based attention and/or using ensemble meta-learners to weight or prioritize the VAE and CNN outputs based on confidence scores and feature entropy.
4. **Benchmarking Explainability:** A larger scale study for benchmarking the cognitive impact of employing Grad-CAM versus other Explainable AI (XAI) techniques through live decision-making and red-team simulations with cybersecurity professionals.
5. **Adversarial Robustness:** To better evaluate how SENTRY-AI performs when faced with adversarial perturbations and conduct work on adversarial training (Huang et al. 2021) or input masking techniques (Akhtar & Mian, 2018) that may bolster model robustness against evasion attack methods.

By addressing the above items SENTRY - AI can become not only advanced and next- generation but an active real-time and adaptive intrusion detection mechanism, with human-machine trust embedded at its core.

6. Conclusion

In this paper, we presented SENTRY-AI, an innovative and explainable anomaly detection framework designed to tackle the growing challenges of cyber threats in a human-centric way. By combining Variational Autoencoders (VAE) for tabular anomaly detection and Convolutional Neural Networks (CNN) for analyzing Gramian Angular Field (GAF) image representations, the system can automatically learn all the statistical and temporal-spatial features of network traffic and get the best of both of each potential approach. Furthermore, to bridge the gap between deep learning predictions and human understanding, the use of Grad- CAM visual aids facilitates trust and interpretable explanations.

Evaluations using three publicly available benchmark datasets - NSL-KDD, CICIDS2017 and UNSW-NB15 - show that SENTRY-AI outperforms all traditional and recent IDS models. These established near-perfect detection metrics: an F1-score of 98.92% on NSL-KDD and a full F1-score of 100.00% on UNSW-NB15. In comparison to all state-of-the-

art deep learning models (e.g., CNN-LSTM, CNN-VAE, GMM-WGAN), SENTRY-AI consistently outperforms them; however the real value is in addition to comparable performance, SENTRY- AI offers explainable outputs where required, underscoring the framework's usefulness in detecting known and zero-day attacks at unprecedented accuracy while remaining interpretable by analysts.

In addition to the accuracy of SENTRY-AI, this work contributes to the growing focus on human-centric cybersecurity. By allowing analysts to visualize and verify the reasoning behind alerts, we facilitate situational awareness, cognitive offloading, and apply ethical and accountable AI to security operations. SENTRY-AI shifts anomaly detection work from a prescriptive and algorithmic detection process, to a collaborative and explainable workflow between AI systems and human defenders. Notwithstanding its strengths, the SENTRY-AI system has significant computational demands and is dependent on access to research labelled datasets, which is challenging for real-time and low-resource environments. Future work will focus upon optimizing the efficiency of the proposed model, multi-sensor dynamic fusion models, and adaptation of SENTRY-AI to exchange and monitor live network anomalies. We will also advertise in order compute a more expansive level of representative usability studies and benchmarks around the adversarial robustness of the system.

To summarise, SENTRY-AI is a high-performing, explainable, and scalable solution for current intrusion detection systems. This means it offers robust technical performance, while always aligned to the human side of the cybersecurity equation, helping grounds the future cyber defence landscape with an intelligent and trusted cyber intrusion defence capability.

Data Availability

All datasets used in this study are publicly available:

- NSL-KDD: <https://www.unb.ca/cic/datasets/nsl.html>
- CICIDS2017: <https://www.unb.ca/cic/datasets/ids-2017.html>
- UNSW-NB15: <https://research.unsw.edu.au/projects/unswnb15-dataset>

The complete SENTRY-AI model code and preprocessing scripts are available in our GitHub repository: <https://github.com/vizezion/SENTRY-AI-Computer-Vision-Branch>

Any additional derived data (e.g., preprocessed GAF images, trained model checkpoints) can be downloaded from <https://github.com/vizezion/SENTRY-AI-Computer-Vision-Branch/releases>.

References

- [1] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. <http://arxiv.org/abs/1803.01164>
- [2] Alrayes, F. S., Zakariah, M., Amin, S. U., Khan, Z. I., & Alqurni, J. S. (2024). CNN Channel Attention Intrusion Detection System Using NSL-KDD Dataset. *Computers, Materials and Continua*, 79(3), 4319–4347. <https://doi.org/10.32604/cmc.2024.050586>
- [3] Amin, U., S Ahanger, A., Masoodi, F., & M Bamhdi, A. (2021). Ensemble based Effective Intrusion Detection System for Cloud Environment over UNSW-NB15 Dataset. In *Scrs Conference Proceedings on Intelligent Systems* (pp. 483–494). Soft Computing Research Society. <https://doi.org/10.52458/978-93-91842-08-6-46>
- [4] Anderson, H. S., & Roth, P. (2018). EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. <http://arxiv.org/abs/1804.04637>
- [5] Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-based Feature Selection for Improved Network Intrusion Detection Systems. <http://arxiv.org/abs/2410.10050>
- [6] Babaey, V., & Faragardi, H. R. (2025). Detecting Zero-Day Web Attacks with an Ensemble of LSTM, GRU, and Stacked Autoencoders. <http://arxiv.org/abs/2504.14122>
- [7] Bamber, S. S., Katkuri, A. V. R., Sharma, S., & Angurala, M. (2025). A hybrid CNN-LSTM approach for intelligent cyber intrusion detection system. *Computers and Security*, 148. <https://doi.org/10.1016/j.cose.2024.104146>
- [8] Biju, A., & Franklin, S. W. (2025). Dual Feature-Based Intrusion Detection System for IoT Network Security. *International Journal of Computational Intelligence Systems*, 18(1). <https://doi.org/10.1007/s44196-025-00790-y>
- [9] Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. <http://arxiv.org/abs/2006.07155>
- [10] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitsoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hEigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., ... Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. <http://arxiv.org/abs/1802.07228>
- [11] Casey, B., Santos, J. C. S., & Perry, G. (2024). 11 A Survey of Source Code Representations for Machine Learning-Based Cybersecurity Tasks. <https://doi.org/10.1145/3721977>
- [12] Chen, H., Chen, X., Shi, S., & Zhang, Y. (2021). Generate Natural Language Explanations for Recommendation. <http://arxiv.org/abs/2101.03392>
- [13] Cui, J., Zong, L., Xie, J., & Tang, M. (2023). A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. *Applied Intelligence*, 53(1), 272–288. <https://doi.org/10.1007/s10489-022-03361-2>
- [14] Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. *Cybersecurity*, 5(1). <https://doi.org/10.1186/s42400-021-00103-8>
- [15] P. A., Moghadam, S. S., Khezri, E., Basem, A., & Trik, M. (2025). A new intrusion detection method using ensemble classification and feature selection. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-98604-w>
- [16] Gupta, S. K., Tripathi, M., & Grover, J. (2022). Hybrid optimization and deep learning based intrusion detection system. *Computers and Electrical Engineering*, 100. <https://doi.org/10.1016/j.compeleceng.2022.107876>
- [17] Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System. *IEEE Access*, 10, 99837–99849. <https://doi.org/10.1109/ACCESS.2022.3206425>
- [18] Hara, K., & Shiimoto, K. (2020, April). Intrusion Detection System using Semi-Supervised Learning with Adversarial Auto-encoder. *Proceedings of IEEE/IFIP Network Operations and Management Symposium 2020: Management in the Age of Softwarization and Artificial Intelligence, NOMS 2020*. <https://doi.org/10.1109/NOMS47738.2020.9110343>
- [19] Hodo, E., Bellekens, X., Hamilton, A., Tachtatzis, C., & Atkinson, R. (2017). Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey. <http://arxiv.org/abs/1701.02145>
- [20] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97. <https://doi.org/10.1016/j.inffus.2023.101804>
- [21] Lunardi, W. T., Lopez, M. A., & Giacalone, J. P. (2023). ARCADE: Adversarially Regularized Convolutional Autoencoder for Network Anomaly Detection. *IEEE Transactions on Network and Service Management*, 20(2), 1305–1318. <https://doi.org/10.1109/TNSM.2022.3229706>

- [22] Lv, H., & Ding, Y. (2024). A hybrid intrusion detection system with K-means and CNN+LSTM. *ICST Transactions on Scalable Information Systems*, 11(6). <https://doi.org/10.4108/eetsis.5667>
- [23] Mane, S., & Rao, D. (2021). Explaining Network Intrusion Detection System Using Explainable AI Framework. <http://arxiv.org/abs/2103.07110>
- [24] Mohammadpour, L., Ling, T. C., Liew, C. S., & Aryanfar, A. (2020). A Mean Convolutional Layer for Intrusion Detection System. *Security and Communication Networks*, 2020. <https://doi.org/10.1155/2020/8891185>
- [25] Narmadha, S., & Balaji, N. V. (2025). Improved network anomaly detection system using optimized autoencoder – LSTM. *Expert Systems with Applications*, 273. <https://doi.org/10.1016/j.eswa.2025.126854>
- [26] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. *IEEE Access*, 10, 112392–112415. <https://doi.org/10.1109/ACCESS.2022.3216617>
- [27] Nicolae, A. (2020). Deep Learning Framework From Scratch Using Numpy. <http://arxiv.org/abs/2011.08461>
- [28] Nkashama, D. K., Soltani, A., Verdier, J.-C., Frappier, M., Tardif, P.-M., & Kabanza, F. (2022). Robustness Evaluation of Deep Unsupervised Learning Algorithms for Intrusion Detection Systems. <http://arxiv.org/abs/2207.03576>
- [29] Okdem, S., & Okdem, S. (2024). Artificial Intelligence in Cybersecurity: A Review and a Case Study. *Applied Sciences (Switzerland)*, 14(22). <https://doi.org/10.3390/app142210487>
- [30] Oluwasusi, V. A., & Al-Turjman, F. (2024). Cybersecurity using artificial intelligence. In *Artificial Intelligence of Things (AIoT): Current and Future Trends* (pp. 73–81). Elsevier. <https://doi.org/10.1016/B978-0-443-26482-5.00009-2>
- [31] Oyinloye, T. S., Arowolo, M. O., & Prasad, R. (2024). Enhancing Cyber Threat Detection with an Improved Artificial Neural Network Model. *Data Science and Management*. <https://doi.org/10.1016/j.dsm.2024.05.002>
- [32] Radford, B. J., Apolonio, L. M., Trias, A. J., & Simpson, J. A. (2018). Network Traffic Anomaly Detection Using Recurrent Neural Networks. <http://arxiv.org/abs/1803.10769>
- [33] Rajathi, C., & Rukmani, P. (2025). Hybrid Learning Model for intrusion detection system: A combination of parametric and non-parametric classifiers. *Alexandria Engineering Journal*, 112, 384–396. <https://doi.org/10.1016/j.aej.2024.10.101>
- [34] Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00957-y>
- [35] Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00886-w>
- [36] Waghmode, P., Kanumuri, M., El-Ocla, H., & Boyle, T. (2025). Intrusion detection system based on machine learning using least square support vector machine. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-95621-7>
- [37] Wang, J., Du, W., Cao, W., Zhang, K., Wang, W., Liang, Y., & Wen, Q. (2024). Deep Learning for Multivariate Time Series Imputation: A Survey. <http://arxiv.org/abs/2402.04059>
- [38] Wang, Z., Ghaleb, F. A., Zainal, A., Siraj, M. M., & Lu, X. (2024). An efficient intrusion detection model based on convolutional spiking neural network. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-57691-x>
- [39] Yang, K., Kpotufe, S., & Feamster, N. (2021). An Efficient One-Class SVM for Anomaly Detection in the Internet of Things. <http://arxiv.org/abs/2104.11146>
- [40] Zhang, C., Wang, N., Hou, Y. T., & Lou, W. (2025). Machine Learning-Based Intrusion Detection Systems: Capabilities, Methodologies, and Open Research Challenges. <https://doi.org/10.36227/techrxiv.173627464.48290242/v1>
- [41] Zhang, K., Wang, H., Chen, M., Chen, X., Liu, L., Geng, Q., & Zhou, Y. (2025). Leveraging machine learning to proactively identify phishing campaigns before they strike. *Journal of Big Data*, 12, 124. <https://doi.org/10.1186/s40537-025-01174-x>
- [42] Zhang, Y., & Ran, X. (2021). A Step-Based Deep Learning Approach for Network Intrusion Detection. *Computer Modeling in Engineering & Sciences*, 128(3), 1231–1245. <https://doi.org/10.32604/cmes.2021.016866>
- [43] Zhang, Z., Hamadi, H. Al, Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, 10, 93104–93139. <https://doi.org/10.1109/ACCESS.2022.3204051>

The Emergence of Multimodal AI in Software Development: What to Expect in the Future?

Imane Boumedra
Final University, Cyprus.
imane.boumedra@final.edu.tr

Ibrahim Adeshola
Final University, Cyprus.
Ibrahim.adeshola@final.edu.tr

Abstract

Multimodal artificial intelligence (AI) is reshaping the software development landscape by extending coding assistance beyond text-based prompts into voice, image, video, and design-driven workflows. This study investigates how multimodal AI is influencing current development practices, the challenges it introduces, and what trends can be expected between 2025 and 2030. Using a mixed-method approach that combines literature synthesis with empirical analysis of 3,998 Stack Overflow posts (covering 3,297 questions and 701 answers), we examine developer sentiment, tool adoption, and emerging themes in community discourse. The findings indicate that while 67% of posts express positive sentiment, developers are increasingly focused on issues of trust, integration, and control, especially when AI systems interact with sensitive files or introduce unexpected behaviors. Tools such as ChatGPT and GitHub Copilot dominate discussions, reflecting both their versatility and integration challenges, while smaller tools like Claude, Amazon CodeWhisperer, and Figma to Code receive highly positive but niche attention. Looking ahead, multimodal AI is expected to play a central role in automated UI/UX generation, debugging, architecture design, and DevOps workflows, though ethical concerns, workforce disruptions, and intellectual property risks remain unresolved. This study contributes to ongoing debates by highlighting both the opportunities and risks of multimodal AI, and by outlining practical steps developers and organizations can take to prepare for an AI-driven software engineering future.

Keywords

Multimodal Artificial Intelligence; Software Development; AI-assisted Coding Tools; Developer Adoption; Sentiment Analysis; Ethical and Security Challenges.

1. Introduction

The rapid rise of artificial intelligence in software engineering has transitioned from text-based code completion toward multimodal systems that integrate text, images, audio, and video. Unlike early coding assistants such as GitHub Copilot and Amazon CodeWhisperer, modern multimodal AI tools now support screenshot-to-code generation, voice-driven refactoring, and visual debugging, enabling richer interactions across the software development lifecycle (George & Harendra, 2024). This transformation is accelerating industry adoption, with the recent 2024 Stack Overflow Developer Survey showed 62% of developers were using AI tools daily, up from 44% in 2023, with 76% expecting to use them this year (StackOverflow, 2024), and market forecasts projecting multimodal AI growth at a CAGR

of 34.4% to reach \$10.89 billion by 2030 (Grand View Research, 2024). Multimodal AI is already altering the nature of software engineering work. It enables automated UI generation from design mockups (Dave et al., 2021; Durgam et al., 2025), architecture synthesis from diagrams (Ramachandran, 2025), and bug detection through combined log, code, and screenshot analysis (Singh & Sudha, 2024). Furthermore, emerging paradigms such as collaborative agents enhance pair programming by integrating seamlessly into IDEs and communication platforms (Ma et al., 2024). These advances signal a shift toward AI-augmented development ecosystems, where human and AI collaborate continuously across design, coding, testing, and deployment.

However, the emergence of multimodal AI also raises critical challenges and risks. Developers report frustrations with inconsistent IDE support (Preethi et al., 2024), ethical concerns around intellectual property (Lalanda & Roig, 2025), and security risks tied to unsafe code generation (Espinha Gasiba, 2024). Beyond technical challenges, AI adoption disrupts workforce roles by automating repetitive tasks while creating demand for prompt engineering, architecture, and AI-human collaboration skills (Kinder et al., 2024). The literature also highlights persistent issues of bias, data privacy, and the ownership of AI-generated code, raising questions about the sustainability and governance of AI-driven development practices (Oh et al., 2024; OWASP Foundation, 2024). Against this backdrop, this study addresses three guiding questions:

- How is multimodal AI transforming software development today?
- What trends are likely to shape its evolution between 2025 and 2030?
- What should developers and companies prepare for in adopting these technologies responsibly?

Drawing on both academic research and empirical evidence from Stack Overflow discussions, this paper provides a comprehensive analysis of the opportunities, risks, and future trajectories of multimodal AI in software engineering.

2. Literature Review

The ability of multimodal AI models to jointly reason across text, images, audio, and video is revolutionizing software engineering workflows by providing richer debugging signals (such as screenshots, logs, and telemetry), automated UI/UX development, and new coding modalities (such as design-to-code and voice-to-refactor). Recent developments in the field and in research show a distinct move away from text-only coding assistants and towards programs that can

take in and incorporate audio and visual input (Akhtar, 2024; Patel, 2025). This allows for automated test generation, interactive debugging, and quicker front-end development. Using academic articles and industry advancements during 2024–2025, this literature review looks at the present and future directions of multimodal AI in software development. It summarizes research on how multimodal AI systems are changing software architectural design, debugging procedures, and conventional coding techniques.

2.1. Evolution from Text-Based to Multimodal Coding Assistants

Text-based code completion was the main focus of early AI coding aids like Amazon Code Whisperer (now rebranded as Amazon Q Developer in 2024) (Amazon Web Services, 2024) and GitHub Copilot. However, recent developments are changing the paradigm in favor of multimodal systems that can receive visual inputs, such as brief screencasts, Figma mockups, and screenshots, and produce useful frontend code using CSS, React components, and responsive layouts. This development is in line with a larger trend in the industry: interactive multimodal AI systems are expected to expand at a compound annual growth rate (CAGR) of 34.4% to reach \$10.89 billion by 2030 (Grand View Research, 2024). Recent industry surveys show that close to 80% of developers favor AI tools, with more than 40% using AI tools in their daily development duties in 2024, demonstrating rapid mainstream adoption (Al Haque et al, 2024; Stack Overflow, 2024).

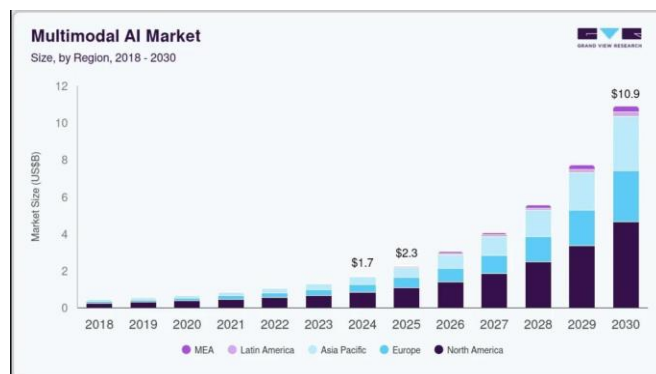


Figure 1: Multimodal AI market size by region, and growth forecast (2025-2030)

A feature that directly helps processes like automatic UI development, GitHub Copilot's multi- model update, which was unveiled at Universe 2024, combines models from Anthropic, Google, and OpenAI and enables developers to assign tasks to various models according to latency, cost, and capability requirements (GitHub, 2024). Similarly, Amazon Q Developer has expanded its multimodal capabilities with enhanced region support and enterprise integration features (Amazon Web Services, 2024). Enterprise adoption demonstrates significant impact: BT Group generated over 100,000 lines of code using Amazon's AI coding tools in four

months, automating 12% of repetitive work for 1,200 developers (BT Group, 2024; Anima,2024) stated that a design file can be converted into responsive, production-ready code using near- production pipelines, as shown by vendor solutions for Figma→React conversion.

Regarding research, NVIDIA's Open Frontier-Class NVLM Two prominent paradigms for multimodal large language models were formalized by multimodal LLMs (Dai et al., 2024; Si et al., 2025), which established the foundation for code generation systems that integrate textual, visual, and even temporal inputs. When taken as a whole, these academic and industrial developments suggest that multimodal AI assistants will eventually be integrated straight into software engineering processes, facilitating interactive debugging, adaptive model selection techniques, and richer design-to-code automation.

2.2. Multimodal Input Processing in Development Environments

Recent developments in multimodal AI have extended the capabilities of software development beyond conventional text-based prompts, allowing for voice-driven and visual coding workflows that expedite the design-to-deployment process. Modern technologies are now able to translate screenshots to code, turning Figma designs and UI mock-ups straight into usable frontend implementations in frameworks like React or Vue.js (Salvi, 2023). Additionally, hand- drawn architecture recognition approaches bridge the gap between early design artefacts and deployment automation by converting sketched system diagrams into executable cloud infrastructure configurations (Zhang et al., 2025). Visual debugging agents provide deeper, context-aware troubleshooting capabilities for debugging by analysing screenshots of application failures along with relevant code and logs (Puvvadi et al., 2025).

Similarly voice-driven development is not left out in the rapid advancement of multimodal AI. With the advancement of voice-based programming owing to parallel research and industrial tools, developers can now issue natural language commands for refactoring ("Refactor this Python function into Rust") (OpenAI, 2024) or participate in conversational code reviews, in which the system uses spoken language to explain logic, identify inefficiencies, and suggest optimisations (Ross et al., 2023). Additionally, by automating the creation of test scripts from spoken requirements, voice-driven testing lessens the conflict between the phases of definition and validation (Görer al., 2024). These modalities collectively demonstrate the trend towards multimodal interaction models, which enable developers to switch between text, images, and voice with ease, lowering cognitive load and speeding up iteration cycles.

2.3. Enhanced Debugging Through Multimodal Context

Current multimodal AI debugging systems can jointly process source code, error logs, user- submitted screenshots

and video reproductions, runtime telemetry data (e.g., CPU, memory, network), and stack traces with visual context, enhancing the fidelity and relevance of bug diagnoses. The debugging-specific LLM Kodezi Chronos reduces debugging time by 40 % and iteration count by 65 % compared to leading models like GPT-4.1 and Claude, achieving a fix accuracy of 67.3 % in real-world scenarios (Khan et al., 2025). AI-driven debugging agents such as DebugMate process code, logs, stack traces, and documentation, identifying root causes and resolving issues in significantly less time than standard models (Modi et al., 2025). Earlier studies demonstrate that combining multimodal inputs logs, screenshots, video clips, produces superior debugging suggestions compared to text-only prompts, though reliability depends on tool-specific log parsing to handle domain-specific formats (Neptune.ai, 2025). Concrete evaluations from the ASE-track series include log analysis using ChatGPT-like agents, which show promise in parsing but reveal limitations in consistency, scalability, and handling unstructured logs (Le & Zhang, 2023; Mudgal & Wouhaybi, 2023).

2.4. AI-Augmented Software Architecture and Design

Recent advances in multimodal AI have enabled automatic documentation generation by combining code repositories with meeting transcripts and design conversations. Multimodal AI tools can generate architecture visualizations, which convert descriptions in natural language into blueprints and system diagrams. Furthermore, Khan et al. (2022) reported that code-to-documentation synthesis enables real-time documentation updates depending on commit messages and code modifications. Therefore, parallel advancements in design pattern identification demonstrate that multimodal AI models are capable of identifying architectural anti-patterns through the analysis of both code structures and system component visual representations (Wei et al., 2024; Çiçek et al., 2024). These models are capable of converting high-level natural language requests and sketches into architecture diagrams and startup code in addition to detection (Wei et al., 2024). This is what emerging frameworks refer to as "automated software engineering with rich context," wherein inputs like telemetry, hand-drawn schematics, meeting transcripts, and user interface screenshots are combined to create documentation and suggest changes to the system architecture (Yue, 2024). Lastly, design-to-code industry technologies, such as Figma integrations, provide concrete proof of this trend by showcasing the increasing use of automated documentation and design workflows.

2.5. Testing & QA Automation

Multimodal AI is reshaping software testing by introducing capabilities that extend far beyond traditional approaches. Multimodal AI, now combining self-healing automation, voice-driven test scripting, and visual regression techniques, is making the systems redefine how quality assurance is performed. For example, visual regression testing can now

compare user interface screenshots across builds, devices, and browsers using graph-based models that go beyond conventional pixel-by-pixel matching. These algorithms detect subtle variations while preserving an understanding of context and layout, thereby reducing false positives and capturing meaningful design changes (Ragel et al., 2023). Industry platforms such as Appliflow, Percy (BrowserStack), and LambdaTest have already begun to operationalize these capabilities (Pandhare et al., 2025; Lost-Pixel, 2024; Katalon, 2025).

Another major advancement is voice-driven testing, where natural language instructions, such as "Test the checkout process under high load" or "Simulate 1,000 users logging in simultaneously," are translated into executable scripts for frameworks like Playwright, Cypress, and Selenium. This shift lowers the barrier to test creation by allowing non-specialists to express requirements in plain language. Complementing this, self-healing automation allows multimodal AI to automatically detect broken selectors during runtime, identify reliable alternatives, patch scripts, and even generate pull requests with clear justifications (Saarathy et al., 2024). Research shows that large language models can repair flaky or failing tests with success rates ranging between 51% and 83%, significantly reducing the manual effort required for maintenance (Fatima et al., 2023). Finally, these innovations integrate seamlessly into continuous integration and delivery (CI/CD) pipelines, providing real-time feedback on user interface consistency, test stability, and performance regressions with minimal setup (Katalon, 2025). This shows that multimodal AI testing solutions enhance both the reliability and efficiency of quality assurance processes, accelerating development cycles while reducing long-term maintenance costs.

2.5. AI "Full-Stack Developers"

Full applications can now be scaffolded from text and mockups by multimodal AI agents and this capability stems from new approaches to AI-Native Software Engineering (SE 3.0), in which AI systems transform from copilots into intelligent collaborators that comprehend developer intent, write full-stack code, and coordinate processes in the design, user interface, backend, DevOps, and runtime domains (Mathews et al., 2024). Third-generation AI solutions, including agentic DevOps systems, which integrate across the Software Development Lifecycle (SDLC) and automate development pipelines from start to finish, are driving industry momentum towards this paradigm (TechRadar Pro, 2025). It is anticipated that autonomous development capabilities would flourish between 2025 and 2030. In order to free up human engineers to concentrate on high-level, strategic design, a 2025 study presents a roadmap towards AI systems that manage full-stack chores, including database schemas, DevOps pipeline setup, performance tuning, security enforcement, and UI design and backend API development (Gu et al., 2025). AI-powered end-to-end project management is coming soon. A move towards truly

autonomous software engineering is indicated by research and industry narratives that, although still in their infancy, anticipate AI managing complete software projects from requirements elicitation to production deployment with little to no human participation (Mathews et al., 2024; TechRadar Pro, 2025; Gu et al., 2025).

2.6. Real-Time Collaborative AI Pair Programmers

By comprehending team communication patterns and coding styles across multiple developers, offering contextually aware suggestions during concurrent work, resolving merge conflicts through multimodal context, and enabling real-time code review with visual explanations and recommendations, future AI systems are expected to improve software development collaboration. (Noor et al., 2022; Lin et al., 2023). In order to provide ongoing and contextual support throughout the development lifecycle, these collaborative AI systems are also anticipated to seamlessly connect with development environments and communication platforms like Slack, Discord, and Microsoft Teams (Vasilescu et al., 2015; Jiang et al., 2023).

Since the AI systems can create entire software applications from natural language, including UI design, business logic, database design, security compliance, and even documentation, are becoming more and more common, according to research roadmaps and new industry tools. The increasing viability of natural-language-driven full-stack creation is demonstrated by systems such as Replit Agent v2, which can now translate plain-English descriptions into front- and back-end apps (Replit, 2025). However, conversational software development is anticipated to be supported by future systems. Self-feedback loops are used by models like Self-Refine, which allow for iterative improvement of outputs, including code, without the need for extra training data. This feature supports interactive development, in which natural language feedback is used to evolve applications (Madaan et al., 2023).

2.7. AI-Driven DevOps & Autonomous Deployments

A shift from reactive human-driven troubleshooting to proactive, AI-driven resilience engineering is being anticipated by emerging research, which envisions autonomous DevOps systems where large-scale AI models continuously monitor application and infrastructure logs, identify anomalies, perform automated root-cause analysis, and proactively suggest or even execute corrective actions like safe rollbacks and redeployments (Ait et al, 2024; Berardinelli., 2025). This will improve mean time to recovery (MTTR) and reduce downtime. Also, beyond log analysis, comprehensive infrastructure optimization is one of the projected skills. Automatic scaling of computing and storage resources in response to user demand and application workloads will be made possible by AI-driven controllers (Huang et al., 2022). Perumallapalli et al. (2021) argued that preventive remediation of potential system faults can operate in parallel with predictive maintenance supported by time-

series forecasting and anomaly detection. Building on this, recent developments highlight cost-conscious orchestration solutions that optimize both performance and economic efficiency in multi-cloud and hybrid-cloud environments (Tuli et al., 2020). In addition, real-time security monitoring enhanced with adaptive threat detection and response mechanisms has emerged as a critical capability for ensuring resilience in AI-driven DevOps pipelines (Moustafa et al., 2019).

Deployment pipelines will also become more independent in the future where the full software delivery lifecycle, including integration testing, staging, production release, canary rollouts, and automated rollback processes, will be coordinated by AI systems (Zhai et al., 2021).

Moreover, to reduce risk during deployment, these pipelines will make use of policy optimization and reinforcement learning (Muhammad et al., 2023). Importantly, in mission-critical settings, such end-to-end automation minimizes human intervention while preserving safety, auditability, and compliance (Malek et al., 2017).

2.8. Challenges, Risks, Ethical & Legal Concerns

According to recent study, sophisticated software projects face considerable difficulties in preserving coherence across several modalities (speech, images, and code). According to studies, the quantity of multimodal information that can be processed concurrently is limited by context window restrictions in existing large language models, especially for large-scale enterprise applications (Bubeck et al., 2023; L et al., 2023). Although AI coding tools increase developer productivity, if they are not educated with secure coding best practices, they may produce code with unsafe patterns. According to empirical research, between 15 and 25 percent of AI-generated code needs extra scrutiny and changes in order to satisfy security requirements (Pearce et al., 2025; Vaithilingam et al., 2022). Industry standards place a strong emphasis on integrating AI-aware code review and static analysis procedures into development processes (OWASP Foundation, 2024). Furthermore, various development environments continue to present integration issues. Adoption is hampered by research showing inconsistent API support and performance differences among IDEs (e.g., VS Code, JetBrains, Vim). According to studies, 40–60% of development teams have trouble integrating multimodal AI tools into their current toolchains (Eindhoven University of Technology, 2025).

There are also unresolved issues around code ownership and intellectual property rights are brought up by AI models trained on either public or proprietary codebases. Remarkably, a thorough developer poll indicates widespread knowledge and worry about the uncertainty surrounding the ownership of AI-generated code, with requests for tools that track provenance and more transparent licensing guidelines to avoid disputes (Stalnaker et al., 2024). The critical need for

compliance procedures is further highlighted by the LiCoEval benchmark, which shows that even top models can occasionally generate code that is remarkably comparable to licensed repositories but do not include the appropriate licenses (Xu et al., 2024). Another major concern is how AI tools is changing the nature of the workforce. Routine coding tasks are becoming more automated, while positions in prompt engineering, system architecture, and AI-human cooperation are growing. According to worldwide estimates, up to 30% of occupations may experience significant task disruptions from AI (Kinder et al., 2024), and workforce evaluations point to a trend towards skill-based recruiting, with AI competencies becoming more valued than traditional degrees (Ehlinger et al., 2023).

Bias is not left out in many of the concerns raised in the literature. AI code generators may propagate unsafe or ineffective patterns in the absence of carefully selected, high-quality training data, which could lead to the growth of technical debt or vulnerabilities. To reduce such risks, it is currently advised by strong industry norms to incorporate AI-aware static analysis, controlled code review methods, and curated datasets (Stalnaker et al., 2024). Lastly, multimodal AI systems processing code, visuals, and audio elevate risks around data privacy and IP leakage, especially in cloud-integrated tooling.

3. Methodology and Analysis

Data for this study were collected from Stack Overflow using Python scripts with the “Search StackOverflow” and “Write CSV” operators. Only English-language questions and answers relevant to the search phrase were included. Between 2024-01-01 11:14:27 and 2025-08-06 20:59:11, a total of 3,998 entries were retrieved, consisting of 3,297 questions and 701 answers, with an average of 0.21 answers per question. The dataset reflected active community engagement, with questions receiving an average score of 0.47 (range -7 to 36), an average of 479 views (maximum 34,684), and 312 questions surpassing 1,000 views. Frequently occurring tags included AI/ML (501), GitHub Copilot (120), and Code Generation (77), while top search terms such as tag:openai-api (499) and tag:intellij-idea (487) highlighted developer interest in AI-assisted tools.

The dataset underwent NLP pre-processing before applying sentiment analysis using VADER, which categorized texts into positive, negative, and neutral sentiments. To complement this, Latent Dirichlet Allocation (LDA) was applied to extract thematic topics, enabling both sentiment and topical insights into developer discussions on AI-assisted coding technologies.

The findings show a clear mix of excitement and challenges around multimodal AI tools. As Table 1 illustrates, most discussions were positive (67%), though nearly a third reflected frustrations or errors. Looking at specific tools in Table 2, ChatGPT drew the most attention, while GitHub Copilot sparked the highest engagement, reflecting both its

usefulness and integration hurdles. Smaller tools like Claude and CodeWhisperer appeared less often but were received warmly by their early adopters. The themes in Table 3 reinforce this picture: much of the conversation centered on IDE integration and troubleshooting, showing that developers are keen to adopt these tools but often turn to the community when technical obstacles arise.

Table 1: Sentiment classification of Stack Overflow posts

Sentiment	Count	Percentage	Average Score
Positive	2,210	67.0%	0.48
Negative	988	30.0%	-0.33
Neutral	99	3.0%	0.00
Total	3,297	100%	0.314 (avg.)

Table 2: AI Tool Popularity

Tool	Questions	Avg. Score	Avg. Views	Avg. Sentiment	Positive Sentiment
ChatGPT	574	0.36	763	0.320	67.6%
GitHub Copilot	184	1.62	1,488	0.167	57.6%
Claude	17	0.59	657	0.418	76.5%
Tabnine	4	-0.50	359	0.418	50.0%
Amazon CodeWhisperer	1	0.00	1,076	0.492	100.0%
Figma to Code	1	-1.00	167	0.881	100.0%

Table 3: Top terms across identified themes

Theme	Top Terms (Keywords)	Interpretation
1	vscode, code, file, using	IDE and file handling issues
2	self, torch, def, model	Machine learning implementation
3	import, openai, error	OpenAI API usage errors
4	gt, lt, parsing, quot	Code formatting/markup issues
5	code, vs, using, error	Code execution problems
6	studio, vs, code, project	Visual Studio integration
7	intellij, java, idea, project	IntelliJ/Java development

8	keras, model, import, np	Deep learning frameworks
---	--------------------------	--------------------------

4. Discussion

Our analysis shows that developer engagement with multimodal AI tools on Stack Overflow has been dynamic rather than static. The sharp peak in November 2024, with 530 questions posted, signals heightened community attention likely triggered by new feature releases, updates, or the rising visibility of AI-assisted coding tools. During this period, ChatGPT and GitHub Copilot dominated discussions, with developers frequently raising practical issues around API integration, error handling, and file exclusions in IDEs. The +204.7% growth rate in the final three months compared to the first three months highlights the accelerating adoption of these tools. Developers are not only experimenting with multimodal AI but also relying heavily on peer knowledge to overcome technical challenges, with Stack Overflow serving as a key hub for sharing solutions and workarounds.

Sentiment patterns add nuance to this trend, as February 2024 is the most positive month (average sentiment 0.429), and this reflected early enthusiasm around tools such as ChatGPT and Claude, with posts often emphasizing novelty and productivity gains. In contrast, GitHub Copilot, while widely adopted and attracting the highest average engagement (1,488 views per question), elicited more balanced sentiment. Furthermore, developers valued its practical benefits but also voiced recurring concerns about blocked completions and integration errors. Smaller tools such as Figma to Code and Amazon CodeWhisperer appeared far less frequently but were consistently associated with positive impressions, suggesting that even niche tools can generate strong approval among early adopters.

The trajectory of these discussions illustrates a shift from initial optimism and curiosity toward more practical, experience-based evaluations and ChatGPT remains dominant by volume, reflecting its versatility, while Copilot draws the most focused engagement, underscoring both its utility and the integration challenges it presents. Developer questions further reveal strong expectations for seamless integration of AI into modern frameworks and IDEs, with tools like Next.js AI SDK and Copilot seen as central to evolving workflows. Nevertheless, concerns around trust, control, and precision persist, particularly when AI interacts with sensitive files or introduces unexpected behaviors. These discussions suggest that multimodal AI is welcomed as a productivity-enhancing innovation, but its sustained adoption will depend on its ability to become reliable, context-aware, and aligned with developer needs.

5. Conclusion

The findings of this study highlight the dual nature of multimodal AI adoption in software development and developers express strong enthusiasm for tools such as ChatGPT, GitHub Copilot, and emerging multimodal

frameworks, which are seen as catalysts for productivity and innovation. However, community discussions reveal persistent concerns about trust, integration challenges, intellectual property rights, and security risks. Sentiment analysis shows that while two-thirds of discussions are positive, critical voices are growing as real-world usage exposes technical limitations and ethical dilemmas.

Looking ahead to 2030, multimodal AI is expected to drive major transformations in UI/UX automation, multimodal debugging, autonomous DevOps, and AI-native software engineering, potentially evolving toward systems that act as autonomous collaborators and full-stack developers. Yet the success of this transition depends on how well both developers and organizations prepare. For developers, this means building skills in prompt engineering, AI-assisted debugging, and high-level design thinking, while allowing AI to handle boilerplate and repetitive code. For companies, preparation involves implementing AI-aware code review policies, investing in workforce upskilling, and establishing compliance frameworks to address security, bias, and licensing concerns. Ultimately, multimodal AI will not replace human developers but reshape the nature of their work. By aligning adoption with ethical safeguards and practical readiness, the software industry can harness the promise of multimodal AI to create more efficient, reliable, and collaborative development ecosystems.

References

- [1] Ait Said, M., & Marzouk, A. (2024). Microservice Architecture DevOps Integration Challenges: A Qualitative Study. *Recent Trends and Advances in Artificial Intelligence: Selected Papers from ICAETA-2024*, 1138, 96.
- [2] Akhtar, Z. B. (2024). Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. *Journal of Electrical Systems and Information Technology*, 11(1), 22.
- [3] Al Haque, E., Brown, C., LaToza, T. D., & Johnson, B. (2025, June). The Evolution of Information Seeking in Software Development: Understanding the Role and Impact of AI
- [4] Assistants. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering* (pp. 1494-1502).
- [5] Amazon Web Services. (2024). Amazon Q Developer enterprise features. AWS Documentation. <https://docs.aws.amazon.com/amazonq/>
- [6] Amazon Web Services. (2024). Amazon Q Developer: Enhanced multimodal capabilities and regional expansion. AWS Blog. <https://aws.amazon.com/blogs/developer/>
- [7] Anima. (2024). Figma to React: Generate developer-friendly code from your designs. <https://www.animaapp.com>

- [8] Berardinelli, L., Muttillio, V., Eramo, R., Bruneliere, H., Rahimi, A., Cicchetti, A., ... & Saadatmand, M. (2025). Model Driven Engineering, Artificial Intelligence, and DevOps for Software and Systems Engineering: A Systematic Mapping Study of Synergies and Challenges. *ACM Transactions on Software Engineering and Methodology*.
- [9] BT Group, (2024). BT Group advances AI-enhanced product development with Amazon CodeWhisperer. *Enterprise AI Report*, 12(4), 45–52.
- [10] Bubeck, A., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.
- [11] Çiçek, S., Aksu, M. S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., ... & Başarır, L. Architectural Critique with Artificial Intelligence: Generating Architectural Reviews through Vision-Language Models. *Journal of Computational Design*, 6(1), 165-190.
- [12] Dave, H., Sonje, S., Pardeshi, J., Chaudhari, S., & Raundale, P. (2021, March). A survey on Artificial Intelligence based techniques to convert User Interface design mock-ups to code. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 28- 33). IEEE.
- [13] Durgam, D., Anandhan, N., & Pathak, R. (2025). AI Image Generation: Emerging Trends and Its Impact on UI/UX Design. *IJSAT-International Journal on Science and Technology*, 16(2).
- [14] Ehlinger, E. G., & Stephany, F. (2023). SKILLS OR A DEGREE?: THE RISE OF SKILL- BASED HIRING FOR AI AND GREEN JOBS. *Bruegel*.
- [15] Eindhoven University of Technology. (2025). AI in software engineering: Challenges in developer toolchains. *Eindhoven Research Portal*.
- [16] Espinha Gasiba, T., Iosif, A. C., Kessba, I., Amburi, S., Lechner, U., & Pinto-Albuquerque, (2024). May the source be with you: On chatgpt, cybersecurity, and secure coding. *Information*, 15(9), 572.
- [17] Fatima, S., Hemmati, H., & Briand, L. (2024). FlakyFix: Using large language models for predicting flaky test fix categories and test code repair. *IEEE Transactions on Software Engineering*.
- [18] George, T. T., & Harendra, R. K. (2024). Enhanced model-driven web application development with code generation using deep learning technique. *Intelligent Decision Technologies*, 18(1), 75-90.
- [19] GitHub. (2024, November). GitHub Copilot gains multi-model support. *GitHub Universe 2024*. <https://github.blog>
- [20] Görer, B., & Aydemir, F. B. (2024, June). GPT-powered elicitation interview script generator for requirements engineering training. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)* (pp. 372-379). IEEE.
- [21] Grand View Research. (2024). Multimodal AI market size, share & trends analysis report by modality (text, speech, image, video), by offering (hardware, software, services), by end-use, by region, and segment forecasts, 2024 – 2030. <https://www.grandviewresearch.com/industry-analysis/multimodal-ai-market>
- [22] Gu, A., Jain, N., Li, W. D., Shetty, M., Shao, Y., Li, Z., ... & Solar-Lezama, A. Tasks, Challenges, and Paths Towards AI for Software Engineering. In *ICLR 2025 Workshop: VerifAI: AI Verification in the Wild*.
- [23] Huang, V., Wang, C., Ma, H., Chen, G., & Christopher, K. (2022, November). Cost-aware dynamic multi-workflow scheduling in cloud data center using evolutionary reinforcement learning. In *International Conference on Service-Oriented Computing* (pp. 449-464). Cham: Springer Nature Switzerland.
- [24] Jiang, N., Liu, X., Liu, H., Lim, E. T. K., Tan, C. W., & Gu, J. (2023). Beyond AI-powered context-aware services: the role of human-AI collaboration. *Industrial Management & Data Systems*, 123(11), 2771-2802.
- [25] Katalon. (2025). Top 7 visual regression testing tools to improve efficiency. <https://katalon.com/resources-center/blog/visual-regression-testing-tools>
- [26] Khan, I., et al. (2025). Kodezi Chronos: A debugging-first language model for repository-scale code understanding. *arXiv preprint arXiv:2507.12482*. <https://arxiv.org/abs/2507.12482>
- [27] Khan, J. Y., & Uddin, G. (2022, October). Automatic code documentation generation using gpt-3. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (pp. 1-6).
- [28] Kinder, M., de Souza Briggs, X., Liu, S., & Muro, M. (2024). Generative AI, the American worker, and the future of work.
- [29] Lalanda, P., & Roig, N. A. (2025). Ethical and Legal Challenges of Artificial Intelligence with Respect to Intellectual Property. In *The AI Revolution: How Technological Developments Affect the Audiovisual Sector* (pp. 63-80). Cham: Springer Nature Switzerland.
- [30] Le, V.-H., & Zhang, H. (2023, September). Log parsing: How far can ChatGPT go? In *Proceedings of ASE 2023, NIER Track*. <https://arxiv.org/abs/2306.01590>
- [31] Li, D., Shao, R., Xie, A., Sheng, Y., Zheng, L., Gonzalez, J., ... & Zhang, H. (2023). How long can context length of open-source llms truly promise?. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- [32] Lin, Z., Ma, W., Lin, T., Zheng, Y., Ge, J., Wang, J., ... & Li, L. (2025). Open Source AI-based SE Tools: Opportunities and Challenges of Collaborative

- Software Learning. *ACM Transactions on Software Engineering and Methodology*, 34(5), 1-24.
- [33] Lost-Pixel. (2024). Top 9 automated visual testing tools (2024). Lost-Pixel Blog.
- [34] Ma, X., Liu, Y., & Wang, C. (2024, November). Design and Implementation of Tool Collaboration Platform for AI Agents. In 2024 4th International Conference on Electronic Information Engineering and Computer (EIECT) (pp. 608-613). IEEE.
- [35] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., ... & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 46534-46594.
- [36] Malek, S., Canavera, K., & Esfahani, N. (2017). Automated inference techniques to assist with the construction of self-adaptive software. In *Managing trade-offs in adaptable software architectures* (pp. 131-154). Morgan Kaufmann.
- [37] Mathews, N. S. (2024). Code Generation and Testing in the Era of AI-Native Software Engineering.
- [38] Modi, R., Reddy, N., & Kodur, S. S. (2025). DebugMate: An AI agent for efficient on-call debugging in complex production systems. *Discover Data*, 3, Article 33.
- [39] Moustafa, N., Hu, J., & Slay, J. (2019). A holistic review of network anomaly detection systems: A comprehensive survey. *Journal of Network and Computer Applications*, 128, 33- 55.
- [40] Mudgal, P., & Wouhaybi, R. (2023, August). An assessment of ChatGPT on log data. In *International Conference on AI-generated Content* (pp. 148-169). Singapore: Springer Nature Singapore.
- [41] Muhammad, A. (2023). Leveraging Cloud-Driven Reinforcement Learning for Dynamic Resource Management in Autonomous Systems.
- [42] Neptune.ai. (2025, January). Multimodal Large Language Models . <https://neptune.ai/blog/multimodal-large-language-models>
- [43] Noor, N. (2025). Generative AI-assisted software development teams: opportunities, challenges, and best practices.
- [44] Oh, S., Lee, K., Park, S., Kim, D., & Kim, H. (2024, May). Poisoned chatgpt finds work for idle hands: Exploring developers' coding practices with insecure suggestions from poisoned ai models. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 1141-1159). IEEE.
- [45] OpenAI. (2024, May). Voice-driven code editing with GPT-4o. OpenAI Blog. <https://openai.com/blog>
- [46] OWASP Foundation. (2024). AI coding assistants and secure development guidelines. <https://owasp.org>
- [47] Pandhare, H. V. (2025). Future of Software Test Automation Using AI/ML. *International Journal Of Engineering And Computer Science*, 13(05).
- [48] Patel, P. K. K. (2025). Exploring Student Developers' Perspectives on AI-Powered Development Assistants for Web Accessibility: Trust, Adoption, and Usage Patterns (Doctoral dissertation, Carleton University).
- [49] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2025). Asleep at the keyboard? assessing the security of GitHub copilot's code contributions. *Communications of the ACM*, 68(2), 96-105..
- [50] Perumallapalli, R. (2021). PREDICTIVE MAINTENANCE IN CLOUD INFRASTRUCTURE: A MACHINE LEARNING FRAMEWORK. Available at SSRN 5228213.
- [51] Preethi, P., Ragavan, V., Abinandhana, C., Umamaheswari, G., & Suveethaa, D. R. (2024, December). Towards CodeBlizz: Developing an AI-Driven IDE Plugin for Real-Time Code Suggestions, Debugging, and Learning Assistance with Generative AI and Machine Learning Models. In 2024 International Conference on Emerging Research in Computational Science (ICERCS) (pp. 1-7). IEEE.
- [52] Puvvadi, M., Arava, S. K., Santoria, A., Chennupati, S. S. P., & Puvvadi, H. V. (2025, March). Coding agents: A comprehensive survey of automated bug fixing systems and benchmarks. In 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 680-686). IEEE.
- [53] Ragel, R. K. C., & Balahadia, F. F. (2023, November). Visual Test Framework: Enhancing Software Test Automation with Visual Artificial Intelligence and Behavioral Driven Development. In 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) (pp. 1-5). IEEE.
- [54] Ramachandran, R. (2025, March). Transforming Software Architecture Design With Intelligent Assistants-A Comparative Analysis. In *SoutheastCon 2025* (pp. 1446-1454). IEEE.
- [55] Replit. (2025, June). Agent v2: Autonomous full-stack software generation from natural language. Wikipedia. <https://en.wikipedia.org/wiki/Replit>
- [56] Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023, March). The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 491-514).
- [57] Saarathy, S. C. P., Bathrachalam, S., & Rajendran, B. K. (2024). Self-Healing Test Automation

- Framework using AI and ML. *International Journal of Strategic Management*, 3(3), 45-77.
- [58] Salvi, G. (2023). Web UI code generation: a transformer-based model applied to real-world screenshots (Doctoral dissertation, Politecnico di Torino).
- [59] Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., & Yang, D. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 3956-3974).
- [60] Singh, C., & Sudha, K. (2024). Breaking Down Barriers: A Survey of Screenshot-to-Code Translation Tools and Strategies. Available at SSRN 4935453.
- [61] Stack Overflow Developer Survey. (2024). Developer AI tool adoption survey 2024. <https://survey.stackoverflow.co/2024/>
- [62] Stack Overflow. (2024). 2025 Developer Survey. AI | 2025 Stack Overflow Developer Survey. <https://survey.stackoverflow.co/2025/ai>
- [63] Stalnaker, T., Wintersgill, N., Chaparro, O., Heymann, L. A., Di Penta, M., German, D. M., & Poshyvanyk, D. (2024). Developer Perspectives on Licensing and Copyright Issues Arising from Generative AI for Software Development. *ACM Transactions on Software Engineering and Methodology*.
- [64] TechRadar Pro. (2025, July 24). The three generations of AI coding tools, and what to expect through the rest of 2025.
- [65] Tuli, A. M., Tuli, S., Tuli, R., & Gill, S. (2020). Next generation technologies for smart infrastructure: Challenges, vision, model, trends and future directions. *IEEE Access*, 8, 108–120.
- [66] Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022, April). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts* (pp. 1-7).
- [67] Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., & Filkov, V. (2015, August). Quality and productivity outcomes relating to continuous integration in GitHub. In *Proceedings of the 2015 10th joint meeting on foundations of software engineering* (pp. 805-816).
- [68] Wan, Y., Wang, C., Dong, Y., Wang, W., Li, S., Huo, Y., & Lyu, M. (2025). Divide-and- Conquer: Generating UI Code from Screenshots. *Proceedings of the ACM on Software Engineering*, 2(FSE), 2099-2122.
- [69] Wei, J., Tan, C., Chen, Q., Wu, G., Li, S., Gao, Z., ... & Guo, R. (2025). From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 13315-13325).
- [70] Xu, W., Gao, K., He, H., & Zhou, M. (2024). LiCoEval: Evaluating LLMs on license compliance in code generation. *arXiv preprint arXiv:2408.02487*.
- [71] Yue, S. (2024). A multimodal conceptual framework to achieve automated software evolution for context-rich intelligent applications. *Innovations in Systems and Software Engineering*.
- [72] Zhai, H., & Wang, J. (2021). Automatic deployment system of computer program application based on cloud computing. *International Journal of System Assurance Engineering and Management*, 12(4), 731-740.
- [73] Zhang, F., Chen, H., Chen, Q., & Liu, J. (2025). Cloud software code generation via knowledge graphs and multi-modal learning. *Journal of Cloud Computing*, 14(1), 1-19.

Exploring the effect of ai integration on collaborative practices and decision making in construction project management. A literature review.

Bamidele Charles Olaiya
Kampala International University, Western
Campus, Ishaka-Uganda
bmolaiya@kiu.ac.ug

Olawunmi Azeezat Akanbi
Bahcesehir Cyprus University, Cyprus.
oakanbi@baucyprus.edu.tr

Abstract

Multimodal artificial intelligence (AI) is reshaping the software development landscape by extending coding assistance beyond text-based prompts into voice, image, video, and design driven workflows. This study investigates how multimodal AI is influencing current development practices, the challenges it introduces, and what trends can be expected between 2025 and 2030. Using a mixed-method approach that combines literature synthesis with empirical analysis of 3,998 Stack Overflow posts (covering 3,297 questions and 701 answers), we examine developer sentiment, tool adoption, and emerging themes in community discourse. The findings indicate that while 67% of posts express positive sentiment, developers are increasingly focused on issues of trust, integration, and control, especially when AI systems interact with sensitive files or introduce unexpected behaviors. Tools such as ChatGPT and GitHub Copilot dominate discussions, reflecting both their versatility and integration challenges, while smaller tools like Claude, Amazon CodeWhisperer, and Figma to Code receive highly positive but niche attention. Looking ahead, multimodal AI is expected to play a central role in automated UI/UX generation, debugging, architecture design, and DevOps workflows, though ethical concerns, workforce disruptions, and intellectual property risks remain unresolved. This study contributes to ongoing debates by highlighting both the opportunities and risks of multimodal AI, and by outlining practical steps developers and organizations can take to prepare for an AI-driven software engineering future.

Keywords: Multimodal Artificial Intelligence; Software Development; AI-assisted Coding Tools; Developer Adoption; Sentiment Analysis; Ethical and Security Challenges.

1. Introduction

The rapid rise of artificial intelligence in software engineering has transitioned from text-based code completion toward multimodal systems that integrate text, images, audio, and video. Unlike early coding

assistants such as GitHub Copilot and Amazon CodeWhisperer, modern multimodal AI tools now

support screenshot-to-code generation, voice-driven refactoring, and visual debugging, enabling richer interactions across the software development lifecycle (George & Harendra, 2024). This transformation is accelerating industry adoption, with the recent 2024 Stack Overflow Developer Survey showed 62% of developers were using AI tools daily, up from 44% in 2023, with 76% expecting to use them this year (StackOverflow, 2024), and market forecasts projecting multimodal AI growth at a CAGR of 34.4% to reach \$10.89 billion by 2030 (Grand View Research, 2024). Multimodal AI is already altering the nature of software engineering work. It enables automated UI generation from design mockups (Dave et al., 2021; Durgam et al., 2025), architecture synthesis from diagrams (Ramachandran, 2025), and bug detection through combined log, code, and screenshot analysis (Singh & Sudha, 2024). Furthermore, emerging paradigms such as collaborative agents enhance pair programming by integrating seamlessly into IDEs and communication platforms (Ma et al., 2024). These advances signal a shift toward AI-augmented development ecosystems, where human and AI collaborate continuously across design, coding, testing, and deployment.

However, the emergence of multimodal AI also raises critical challenges and risks. Developers report frustrations with inconsistent IDE support (Preethi et al., 2024), ethical concerns around intellectual property (Lalanda & Roig, 2025), and security risks tied to unsafe code generation (Espinha Gasiba, 2024). Beyond technical challenges, AI adoption disrupts workforce roles by automating repetitive tasks while creating demand for prompt engineering, architecture, and AI-human collaboration skills (Kinder et al., 2024). The literature also highlights persistent issues of bias, data privacy, and the ownership of AI-generated code, raising questions about the sustainability and governance of AI-driven development practices (Oh et al., 2024; OWASP Foundation, 2024). Against this backdrop, this study addresses three guiding questions:

- How is multimodal AI transforming software development today?
- What trends are likely to shape its evolution between 2025 and 2030?
- What should developers and companies prepare for in adopting these technologies responsibly?

Drawing on both academic research and empirical evidence from Stack Overflow discussions, this paper provides a comprehensive analysis of the opportunities, risks, and future trajectories of multimodal AI in software engineering.

2. Literature Review

The ability of multimodal AI models to jointly reason across text, images, audio, and video is revolutionizing software engineering workflows by providing richer debugging signals (such as screenshots, logs, and telemetry), automated UI/UX development, and new coding modalities (such as design-to-code and voice-to-refactor). Recent developments in the field and in research show a distinct move away from text-only coding assistants and towards programs that can take in and incorporate audio and visual input (Akhtar, 2024; Patel, 2025). This allows for automated test generation, interactive debugging, and quicker front-end development. Using academic articles and industry advancements during 2024–2025, this literature review looks at the present and future directions of multimodal AI in software development. It summarizes research on how multimodal AI systems are changing software architectural design, debugging procedures, and conventional coding techniques.

2.1. Evolution from Text-Based to Multimodal Coding Assistants

Text-based code completion was the main focus of early AI coding aids like Amazon Code Whisperer (now rebranded as Amazon Q Developer in 2024) (Amazon Web Services, 2024) and GitHub Copilot. However, recent developments are changing the paradigm in favor of multimodal systems that can receive visual inputs, such as brief screencasts, Figma mockups, and screenshots, and produce useful frontend code using CSS, React components, and responsive layouts. This development is in line with a larger trend in the industry: interactive multimodal AI systems are expected to expand at a compound annual growth rate (CAGR) of 34.4% to reach \$10.89 billion by 2030 (Grand View Research, 2024). Recent industry surveys show that close to 80% of developers favor AI tools, with more than 40% using AI tools in their daily

development duties in 2024, demonstrating rapid mainstream adoption (Al Haque et al, 2024; Stack Overflow, 2024).

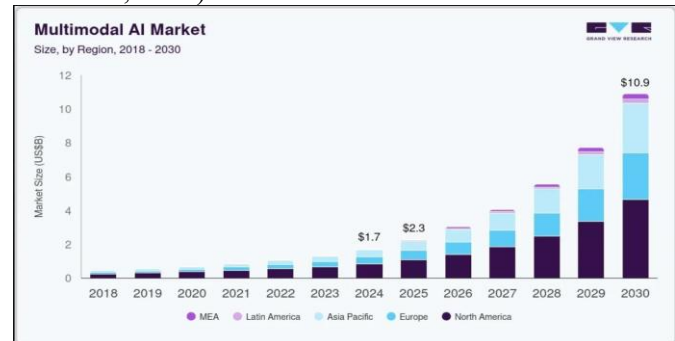


Figure 1: Multimodal AI market size by region, and growth forecast (2025-2030)

A feature that directly helps processes like automatic UI development, GitHub Copilot's multimodal update, which was unveiled at Universe 2024, combines models from Anthropic, Google, and OpenAI and enables developers to assign tasks to various models according to latency, cost, and capability requirements (GitHub, 2024). Similarly, Amazon Q Developer has expanded its multimodal capabilities with enhanced region support and enterprise integration features (Amazon Web Services, 2024). Enterprise adoption demonstrates significant impact: BT Group generated over 100,000 lines of code using Amazon's AI coding tools in four months, automating 12% of repetitive work for 1,200 developers (BT Group, 2024; Anima, 2024) stated that a design file can be converted into responsive, production-ready code using nearproduction pipelines, as shown by vendor solutions for Figma→React conversion.

Regarding research, NVIDIA's Open Frontier-Class NVLM Two prominent paradigms for multimodal large language models were formalized by multimodal LLMs (Dai et al., 2024; Si et al., 2025), which established the foundation for code generation systems that integrate textual, visual, and even temporal inputs. When taken as a whole, these academic and industrial developments suggest that multimodal AI assistants will eventually be integrated straight into software engineering processes, facilitating interactive debugging, adaptive model selection techniques, and richer design-to-code automation.

2.2. Multimodal Input Processing in Development Environments

Recent developments in multimodal AI have extended the capabilities of software development beyond

conventional text-based prompts, allowing for voice-driven and visual coding workflows that expedite the design-to-deployment process. Modern technologies are now able to translate screenshots to code, turning Figma designs and UI mock-ups straight into usable frontend implementations in frameworks like React or Vue.js (Salvi, 2023). Additionally, hand-drawn architecture recognition approaches bridge the gap between early design artefacts and deployment automation by converting sketched system diagrams into executable cloud infrastructure configurations (Zhang et al., 2025). Visual debugging agents provide deeper, context-aware troubleshooting capabilities for debugging by analysing screenshots of application failures along with relevant code and logs (Puvvadi et al., 2025).

Similarly voice-driven development is not left out in the rapid advancement of multimodal AI. With the advancement of voice-based programming owing to parallel research and industrial tools, developers can now issue natural language commands for refactoring ("Refactor this Python function into Rust") (OpenAI, 2024) or participate in conversational code reviews, in which the system uses spoken language to explain logic, identify inefficiencies, and suggest optimizations (Ross et al., 2023). Additionally, by automating the creation of test scripts from spoken requirements, voice-driven testing lessens the conflict between the phases of definition and validation (Görer et al., 2024). These modalities collectively demonstrate the trend towards multimodal interaction models, which enable developers to switch between text, images, and voice with ease, lowering cognitive load and speeding up iteration cycles.

2.3. Enhanced Debugging Through Multimodal Context

Current multimodal AI debugging systems can jointly process source code, error logs, user submitted screenshots and video reproductions, runtime telemetry data (e.g., CPU, memory, network), and stack traces with visual context, enhancing the fidelity and relevance of bug diagnoses. The debugging-specific LLM Kodezi Chronos reduces debugging time by 40 % and iteration count by 65 % compared to leading models like GPT-4.1 and Claude, achieving a fix accuracy of 67.3 % in real-world scenarios (Khan et al., 2025). AI-driven debugging agents such as DebugMate process code, logs, stack traces, and documentation, identifying root causes and resolving issues in significantly less time than standard models (Modi et al., 2025). Earlier

studies demonstrate that combining multimodal inputs logs, screenshots, video clips, produces superior debugging suggestions compared to text-only prompts, though reliability depends on tool-specific log parsing to handle domain-specific formats (Neptune.ai, 2025). Concrete evaluations from the ASE-track series include log analysis using ChatGPT-like agents, which show promise in parsing but reveal limitations in consistency, scalability, and handling unstructured logs (Le & Zhang, 2023; Mudgal & Wouhaybi, 2023).

2.4. AI-Augmented Software Architecture and Design

Recent advances in multimodal AI have enabled automatic documentation generation by combining code repositories with meeting transcripts and design conversations. Multimodal AI tools can generate architecture visualizations, which convert descriptions in natural language into blueprints and system diagrams. Furthermore, Khan et al. (2022) reported that code-to documentation synthesis enables real-time documentation updates depending on commit messages and code modifications. Therefore, parallel advancements in design pattern identification demonstrate that multimodal AI models are capable of identifying architectural anti-patterns through the analysis of both code structures and system component visual representations (Wei et al., 2024; Çiçek et al., 2024). These models are capable of converting high-level natural language requests and sketches into architecture diagrams and startup code in addition to detection (Wei et al., 2024). This is what emerging frameworks refer to as "automated software engineering with rich context," wherein inputs like telemetry, hand-drawn schematics, meeting transcripts, and user interface screenshots are combined to create documentation and suggest changes to the system architecture (Yue, 2024). Lastly, design to code industry technologies, such as Figma integrations, provide concrete proof of this trend by showcasing the increasing use of automated documentation and design workflows.

2.5. Testing & QA Automation

Multimodal AI is reshaping software testing by introducing capabilities that extend far beyond traditional approaches. Multimodal AI, now combining self-healing automation, voice-driven test scripting, and visual regression techniques, is making the systems redefine how quality assurance is performed. For example, visual regression testing can now compare user interface screenshots across builds, devices, and

browsers using graph-based models that go beyond conventional pixel-by-pixel matching. These algorithms detect subtle variations while preserving an understanding of context and layout, thereby reducing false positives and capturing meaningful design changes (Ragel et al., 2023). Industry platforms such as AppliTools, Percy (BrowserStack), and LambdaTest have already begun to operationalize these capabilities (Pandhare et al., 2025; Lost-Pixel, 2024; Katalon, 2025).

Another major advancement is voice-driven testing, where natural language instructions, such as “Test the checkout process under high load” or “Simulate 1,000 users logging in simultaneously,” are translated into executable scripts for frameworks like Playwright, Cypress, and Selenium. This shift lowers the barrier to test creation by allowing non-specialists to express requirements in plain language. Complementing this, self-healing automation allows multimodal AI to automatically detect broken selectors during runtime, identify reliable alternatives, patch scripts, and even generate pull requests with clear justifications (Saarathy et al., 2024). Research shows that large language models can repair flaky or failing tests with success rates ranging between 51% and 83%, significantly reducing the manual effort required for maintenance (Fatima et al., 2023). Finally, these innovations integrate seamlessly into continuous integration and delivery (CI/CD) pipelines, providing real-time feedback on user interface consistency, test stability, and performance regressions with minimal setup (Katalon, 2025). This shows that multimodal AI testing solutions enhance both the reliability and efficiency of quality assurance processes, accelerating development cycles while reducing long-term maintenance costs.

2.6. AI "Full-Stack Developers"

Full applications can now be scaffolded from text and mockups by multimodal AI agents and this capability stems from new approaches to AI-Native Software Engineering (SE 3.0), in which AI systems transform from copilots into intelligent collaborators that comprehend developer intent, write full-stack code, and coordinate processes in the design, user interface, backend, DevOps, and runtime domains (Mathews et al., 2024). Third-generation AI solutions, including agentic DevOps systems, which integrate across the Software Development Lifecycle (SDLC) and automate development pipelines from start to finish, are driving industry momentum towards this paradigm (TechRadar Pro, 2025). It is anticipated that autonomous

development capabilities would flourish between 2025 and 2030. In order to free up human engineers to concentrate on high-level, strategic design, a 2025 study presents a roadmap towards AI systems that manage full-stack chores, including database schemas, DevOps pipeline setup, performance tuning, security enforcement, and UI design and backend API development (Gu et al., 2025). AI-powered end-to-end project management is coming soon. A move towards truly autonomous software engineering is indicated by research and industry narratives that, although still in their infancy, anticipate AI managing complete software projects from requirements elicitation to production deployment with little to no human participation (Mathews et al., 2024; TechRadar Pro, 2025; Gu et al., 2025).

2.7. Real-Time Collaborative AI Pair Programmers

By comprehending team communication patterns and coding styles across multiple developers, offering contextually aware suggestions during concurrent work, resolving merge conflicts through multimodal context, and enabling real-time code review with visual explanations and recommendations, future AI systems are expected to improve software development collaboration. (Noor et al., 2022; Lin et al., 2023). In order to provide ongoing and contextual support throughout the development lifecycle, these collaborative AI systems are also anticipated to seamlessly connect with development environments and communication platforms like Slack, Discord, and Microsoft Teams (Vasilescu et al., 2015; Jiang et al., 2023).

Since the AI systems can create entire software applications from natural language, including UI design, business logic, database design, security compliance, and even documentation, are becoming more and more common, according to research roadmaps and new industry tools. The increasing viability of natural-language-driven full-stack creation is demonstrated by systems such as Replit Agent v2, which can now translate plain-English descriptions into front- and back-end apps (Replit, 2025). However, conversational software development is anticipated to be supported by future systems. Self-feedback loops are used by models like Self-Refine, which allow for iterative improvement of outputs, including code, without the need for extra training data. This feature supports interactive development, in which natural language feedback is used to evolve applications (Madaan et al., 2023).

2.8. AI-Driven DevOps & Autonomous Deployments

A shift from reactive human-driven troubleshooting to proactive, AI-driven resilience engineering is being anticipated by emerging research, which envisions autonomous DevOps systems where large-scale AI models continuously monitor application and infrastructure logs, identify anomalies, perform automated root-cause analysis, and proactively suggest or even execute corrective actions like safe rollbacks and redeployments (Ait et al., 2024; Berardinelli., 2025). This will improve mean time to recovery (MTTR) and reduce downtime. Also, beyond log analysis, comprehensive infrastructure optimization is one of the projected skills. Automatic scaling of computing and storage resources in response to user demand and application workloads will be made possible by AI-driven controllers (Huang et al., 2022). Perumallapalli et al. (2021) argued that preventive remediation of potential system faults can operate in parallel with predictive maintenance supported by time-series forecasting and anomaly detection. Building on this, recent developments highlight cost-conscious orchestration solutions that optimize both performance and economic efficiency in multi-cloud and hybrid-cloud environments (Tuli et al., 2020). In addition, real-time security monitoring enhanced with adaptive threat detection and response mechanisms has emerged as a critical capability for ensuring resilience in AI-driven DevOps pipelines (Moustafa et al., 2019).

Deployment pipelines will also become more independent in the future where the full software delivery lifecycle, including integration testing, staging, production release, canary rollouts, and automated rollback processes, will be coordinated by AI systems (Zhai et al., 2021).

Moreover, to reduce risk during deployment, these pipelines will make use of policy optimization and reinforcement learning (Muhammad et al., 2023). Importantly, in missioncritical settings, such end-to-end automation minimizes human intervention while preserving safety, auditability, and compliance (Malek et al., 2017).

2.9. Challenges, Risks, Ethical & Legal Concerns

According to recent study, sophisticated software projects face considerable difficulties in preserving coherence across several modalities (speech, images, and code). According to studies, the quantity of multimodal information that can be processed

concurrently is limited by context window restrictions in existing large language models, especially for large-scale enterprise applications (Bubeck et al., 2023; L et al., 2023). Although AI coding tools increase developer productivity, if they are not educated with secure coding best practices, they may produce code with unsafe patterns. According to empirical research, between 15 and 25 percent of AI-generated code needs extra scrutiny and changes in order to satisfy security requirements (Pearce et al., 2025; Vaithilingam et al., 2022). Industry standards place a strong emphasis on integrating AI-aware code review and static analysis procedures into development processes (OWASP Foundation, 2024). Furthermore, various development environments continue to present integration issues. Adoption is hampered by research showing inconsistent API support and performance differences among IDEs (e.g., VS Code, JetBrains, Vim). According to studies, 40–60% of development teams have trouble integrating multimodal AI tools into their current toolchains (Eindhoven University of Technology, 2025).

There are also unresolved issues around code ownership and intellectual property rights are brought up by AI models trained on either public or proprietary codebases. Remarkably, a thorough developer poll indicates widespread knowledge and worry about the uncertainty surrounding the ownership of AI-generated code, with requests for tools that track provenance and more transparent licensing guidelines to avoid disputes (Stalnaker et al., 2024). The critical need for compliance procedures is further highlighted by the LiCoEval benchmark, which shows that even top models can occasionally generate code that is remarkably comparable to licensed repositories but do not include the appropriate licenses (Xu et al., 2024). Another major concern is how AI tools is changing the nature of the workforce. Routine coding tasks are becoming more automated, while positions in prompt engineering, system architecture, and AI-human cooperation are growing. According to worldwide estimates, up to 30% of occupations may experience significant task disruptions from AI (Kinder et al., 2024), and workforce evaluations point to a trend towards skill-based recruiting, with AI competencies becoming more valued than traditional degrees (Ehlinger et al., 2023).

Bias is not left out in many of the concerns raised in the literature. AI code generators may propagate unsafe or ineffective patterns in the absence of carefully selected, high-quality training data, which could lead to the

growth of technical debt or vulnerabilities. To reduce such risks, it is currently advised by strong industry norms to incorporate AI-aware static analysis, controlled code review methods, and curated datasets (Stalnaker et al., 2024). Lastly, multimodal AI systems processing code, visuals, and audio elevate risks around data privacy and IP leakage, especially in cloud-integrated tooling.

3. Methodology and Analysis

Data for this study were collected from Stack Overflow using Python scripts with the “Search StackOverflow” and “Write CSV” operators. Only English-language questions and answers relevant to the search phrase were included. Between 2024-01-01 11:14:27 and 2025-08-06 20:59:11, a total of 3,998 entries were retrieved, consisting of 3,297 questions and 701 answers, with an average of 0.21 answers per question. The dataset reflected active community engagement, with questions receiving an average score of 0.47 (range -7 to 36), an average of 479 views (maximum 34,684), and 312 questions surpassing 1,000 views. Frequently occurring tags included AI/ML (501), GitHub Copilot (120), and Code Generation (77), while top search terms such as tag:openai-api (499) and tag:intellij-idea (487) highlighted developer interest in AI-assisted tools.

The dataset underwent NLP pre-processing before applying sentiment analysis using VADER, which categorized texts into positive, negative, and neutral sentiments. To complement this, Latent Dirichlet Allocation (LDA) was applied to extract thematic topics, enabling both sentiment and topical insights into developer discussions on AI-assisted coding technologies.

The findings show a clear mix of excitement and challenges around multimodal AI tools. As Table 1 illustrates, most discussions were positive (67%), though nearly a third reflected frustrations or errors. Looking at specific tools in Table 2, ChatGPT drew the most attention, while GitHub Copilot sparked the highest engagement, reflecting both its usefulness and integration hurdles. Smaller tools like Claude and Code Whisperer appeared less often but were received warmly by their early adopters. The themes in Table 3 reinforce this picture: much of the conversation centered on IDE integration and troubleshooting, showing that developers are keen to adopt these tools but often turn to the community when technical obstacles arise.

Table 1: Sentiment classification of Stack Overflow posts

Sentiment	Count	Percentage	Average Score
Positive	2,210	67.0%	0.48
Negative	988	30.0%	-0.33
Neutral	99	3.0%	0.00
Total	3,297	100%	0.314 (avg.)

Table 2: AI Tool Popularity

Tool	Questions	Avg. Score	Avg. Views	Avg. Sentiment	Positive Sentiment
ChatGPT	574	0.36	763	0.320	67.6%
GitHub Copilot	184	1.62	1,488	0.167	57.6%
Claude	17	0.59	657	0.418	76.5%
Tabnine	4	-0.50	359	0.418	50.0%
Amazon CodeWhisperer	1	0.00	1,076	0.492	100.0%
Figma to Code	1	-1.00	167	0.881	100.0%

Table 3: Top terms across identified themes

Theme	Top Terms (Keywords)	Interpretation
1	vscode, code, file, using	IDE and file handling issues
2	self, torch, def, model	Machine learning implementation
3	import, openai, error	OpenAI API usage errors
4	gt, lt, parsing, quot	Code formatting/markup issues
5	code, vs, using, error	Code execution problems

6	studio, vs, code, project	Visual Studio integration
7	intellij, java, idea, project	IntelliJ/Java development
8	keras, model, import, np	Deep learning frameworks

4. Discussion

Our analysis shows that developer engagement with multimodal AI tools on Stack Overflow has been dynamic rather than static. The sharp peak in November 2024, with 530 questions posted, signals heightened community attention likely triggered by new feature releases, updates, or the rising visibility of AI-assisted coding tools. During this period, ChatGPT and GitHub Copilot dominated discussions, with developers frequently raising practical issues around API integration, error handling, and file exclusions in IDEs. The +204.7% growth rate in the final three months compared to the first three months highlights the accelerating adoption of these tools. Developers are not only experimenting with multimodal AI but also relying heavily on peer knowledge to overcome technical challenges, with Stack Overflow serving as a key hub for sharing solutions and workarounds.

Sentiment patterns add nuance to this trend, as February 2024 is the most positive month (average sentiment 0.429), and this reflected early enthusiasm around tools such as ChatGPT and Claude, with posts often emphasizing novelty and productivity gains. In contrast, GitHub Copilot, while widely adopted and attracting the highest average engagement (1,488 views per question), elicited more balanced sentiment. Furthermore, developers valued its practical benefits but also voiced recurring concerns about blocked completions and integration errors. Smaller tools such as Figma to Code and Amazon CodeWhisperer appeared far less frequently but were consistently associated with positive impressions, suggesting that even niche tools can generate strong approval among early adopters.

The trajectory of these discussions illustrates a shift from initial optimism and curiosity toward more practical, experience-based evaluations and ChatGPT remains dominant by volume, reflecting its versatility, while Copilot draws the most focused engagement,

underscoring both its utility and the integration challenges it presents. Developer questions further reveal strong expectations for seamless integration of AI into modern frameworks and IDEs, with tools like Next.js AI SDK and Copilot seen as central to evolving workflows. Nevertheless, concerns around trust, control, and precision persist, particularly when AI interacts with sensitive files or introduces unexpected behaviors. These discussions suggest that multimodal AI is welcomed as a productivity-enhancing innovation, but its sustained adoption will depend on its ability to become reliable, context-aware, and aligned with developer needs..

5. Conclusion

The findings of this study highlight the dual nature of multimodal AI adoption in software development and developers express strong enthusiasm for tools such as ChatGPT, GitHub Copilot, and emerging multimodal frameworks, which are seen as catalysts for productivity and innovation. However, community discussions reveal persistent concerns about trust, integration challenges, intellectual property rights, and security risks. Sentiment analysis shows that while two-thirds of discussions are positive, critical voices are growing as real-world usage exposes technical limitations and ethical dilemmas.

Looking ahead to 2030, multimodal AI is expected to drive major transformations in UI/UX automation, multimodal debugging, autonomous DevOps, and AI-native software engineering, potentially evolving toward systems that act as autonomous collaborators and full-stack developers. Yet the success of this transition depends on how well both developers and organizations prepare. For developers, this means building skills in prompt engineering, AI-assisted debugging, and high-level design thinking, while allowing AI to handle boilerplate and repetitive code. For companies, preparation involves implementing AI-aware code review policies, investing in workforce upskilling, and establishing compliance frameworks to address security, bias, and licensing concerns. Ultimately, multimodal AI will not replace human developers but reshape the nature of their work. By aligning adoption with ethical safeguards and practical readiness, the software industry can harness the promise of multimodal AI to create more efficient, reliable, and collaborative development ecosystems.

6. References

- Ait Said, M., & Marzouk, A. (2024). Microservice Architecture DevOps Integration Challenges: A Qualitative Study. *Recent Trends and Advances in Artificial Intelligence: Selected Papers from ICAETA-2024*, 1138, 96.
- Akhtar, Z. B. (2024). Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. *Journal of Electrical Systems and Information Technology*, 11(1), 22.
- Al Haque, E., Brown, C., LaToza, T. D., & Johnson, B. (2025, June). The Evolution of Information Seeking in Software Development: Understanding the Role and Impact of AI Assistants. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering* (pp. 1494-1502).
- Amazon Web Services. (2024). Amazon Q Developer enterprise features. *AWS Documentation*.
<https://docs.aws.amazon.com/amazonq/>
- Amazon Web Services. (2024). Amazon Q Developer: Enhanced multimodal capabilities and regional expansion. *AWS Blog*.
<https://aws.amazon.com/blogs/developer/>
- Anima. (2024). Figma to React: Generate developer-friendly code from your designs.
<https://www.animaapp.com>
- Berardinelli, L., Mutillo, V., Eramo, R., Bruneliere, H., Rahimi, A., Cicchetti, A., ... & Saadatmand, M. (2025). Model Driven Engineering, Artificial Intelligence, and DevOps for Software and Systems Engineering: A Systematic Mapping Study of Synergies and Challenges. *ACM Transactions on Software Engineering and Methodology*.
- BT Group, (2024). BT Group advances AI-enhanced product development with Amazon CodeWhisperer. *Enterprise AI Report*, 12(4), 45–52.
- Bubeck, A., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.
- Çiçek, S., Aksu, M. S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., ... & Başarır, L. Architectural Critique with Artificial Intelligence: Generating Architectural Reviews through Vision-Language Models. *Journal of Computational Design*, 6(1), 165-190.
- Dave, H., Sonje, S., Pardeshi, J., Chaudhari, S., & Raundale, P. (2021, March). A survey on Artificial Intelligence based techniques to convert User Interface design mock-ups to code. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 2833). IEEE.
- Durgam, D., Anandhan, N., & Pathak, R. (2025). AI Image Generation: Emerging Trends and Its Impact on UI/UX Design. *IJSAT-International Journal on Science and Technology*, 16(2).
- Ehlinger, E. G., & Stephany, F. (2023). SKILLS OR A DEGREE?: THE RISE OF SKILLBASED HIRING FOR AI AND GREEN JOBS. Bruegel.
- Eindhoven University of Technology. (2025). AI in software engineering: Challenges in developer toolchains. *Eindhoven Research Portal*.
- Espinha Gasiba, T., Iosif, A. C., Kessba, I., Amburi, S., Lechner, U., & Pinto-Albuquerque, M. (2024). May the source be with you: On chatgpt, cybersecurity, and secure coding. *Information*, 15(9), 572.
- Fatima, S., Hemmati, H., & Briand, L. (2024). FlakyFix: Using large language models for predicting flaky test fix categories and test code repair. *IEEE Transactions on Software Engineering*.
- George, T. T., & Harendra, R. K. (2024). Enhanced model-driven web application development with code generation using deep learning technique. *Intelligent Decision Technologies*, 18(1), 75-90.
- GitHub. (2024, November). GitHub Copilot gains multi-model support. *GitHub Universe 2024*.
<https://github.blog>
- Görer, B., & Aydemir, F. B. (2024, June). GPT-powered elicitation interview script generator for requirements engineering training. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)* (pp. 372-379). IEEE.
- Grand View Research. (2024). *Multimodal AI market size, share & trends analysis report by modality (text, International Conference on Artificial Intelligence and Cybersecurity 2025 Copyright 2025 © Canadian Tech-Institute for Academic Research.*

speech, image, video), by offering (hardware, software, services), by end-use, by region, and segment forecasts, 2024 – 2030.

<https://www.grandviewresearch.com/industry-analysis/multimodal-ai-market>

Gu, A., Jain, N., Li, W. D., Shetty, M., Shao, Y., Li, Z., ... & Solar-Lezama, A. Tasks, Challenges, and Paths Towards AI for Software Engineering. In ICLR 2025 Workshop: VerifAI: AI Verification in the Wild.

Huang, V., Wang, C., Ma, H., Chen, G., & Christopher, K. (2022, November). Cost-aware dynamic multi-workflow scheduling in cloud data center using evolutionary reinforcement learning. In *International Conference on Service-Oriented Computing* (pp. 449-464). Cham: Springer Nature Switzerland.

Jiang, N., Liu, X., Liu, H., Lim, E. T. K., Tan, C. W., & Gu, J. (2023). Beyond AI-powered context-aware services: the role of human–AI collaboration. *Industrial Management & Data Systems*, 123(11), 2771-2802.

Katalon. (2025). Top 7 visual regression testing tools to improve efficiency. <https://katalon.com/resources-center/blog/visual-regression-testing-tools>

Khan, I., et al. (2025). Kodezi Chronos: A debugging-first language model for repository-scale code understanding. *arXiv preprint arXiv:2507.12482*. <https://arxiv.org/abs/2507.12482>

Khan, J. Y., & Uddin, G. (2022, October). Automatic code documentation generation using gpt-3. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (pp. 1-6).

Kinder, M., de Souza Briggs, X., Liu, S., & Muro, M. (2024). Generative AI, the American worker, and the future of work.

Lalanda, P., & Roig, N. A. (2025). Ethical and Legal Challenges of Artificial Intelligence with Respect to Intellectual Property. In *The AI Revolution: How Technological Developments Affect the Audiovisual Sector* (pp. 63-80). Cham: Springer Nature Switzerland.

Le, V.-H., & Zhang, H. (2023, September). Log parsing: How far can ChatGPT go? In

Proceedings of ASE 2023, NIER Track.

<https://arxiv.org/abs/2306.01590>

Li, D., Shao, R., Xie, A., Sheng, Y., Zheng, L., Gonzalez, J., ... & Zhang, H. (2023). How long can context length of open-source llms truly promise?. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Lin, Z., Ma, W., Lin, T., Zheng, Y., Ge, J., Wang, J., ... & Li, L. (2025). Open Source AI-based SE Tools: Opportunities and Challenges of Collaborative Software Learning. *ACM Transactions on Software Engineering and Methodology*, 34(5), 1-24.

Lost-Pixel. (2024). Top 9 automated visual testing tools (2024). *Lost-Pixel Blog*.

Ma, X., Liu, Y., & Wang, C. (2024, November). Design and Implementation of Tool Collaboration Platform for AI Agents. In *2024 4th International Conference on Electronic Information Engineering and Computer (EIECT)* (pp. 608-613). IEEE.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., ... & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 46534-46594.

Malek, S., Canavera, K., & Esfahani, N. (2017). Automated inference techniques to assist with the construction of self-adaptive software. In *Managing trade-offs in adaptable software architectures* (pp. 131-154). Morgan Kaufmann.

Mathews, N. S. (2024). Code Generation and Testing in the Era of AI-Native Software Engineering.

Modi, R., Reddy, N., & Kodur, S. S. (2025). DebugMate: An AI agent for efficient on-call debugging in complex production systems. *Discover Data*, 3, Article 33.

Moustafa, N., Hu, J., & Slay, J. (2019). A holistic review of network anomaly detection systems: A comprehensive survey. *Journal of Network and Computer Applications*, 128, 3355.

Mudgal, P., & Wouhaybi, R. (2023, August). An assessment of ChatGPT on log data. In *International Conference on AI-generated Content* (pp. 148-169). Singapore: Springer Nature Singapore.

- Muhammad, A. (2023). Leveraging Cloud-Driven Reinforcement Learning for Dynamic Resource Management in Autonomous Systems.
- Neptune.ai. (2025, January). Multimodal Large Language Models . <https://neptune.ai/blog/multimodal-large-language-models>
- Noor, N. (2025). Generative AI-assisted software development teams: opportunities, challenges, and best practices.
- Oh, S., Lee, K., Park, S., Kim, D., & Kim, H. (2024, May). Poisoned chatgpt finds work for idle hands: Exploring developers' coding practices with insecure suggestions from poisoned ai models. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 1141-1159). IEEE.
- OpenAI. (2024, May). Voice-driven code editing with GPT-4o. *OpenAI Blog*. <https://openai.com/blog>
- OWASP Foundation. (2024). AI coding assistants and secure development guidelines. <https://owasp.org>
- Pandhare, H. V. (2025). Future of Software Test Automation Using AI/ML. *International Journal Of Engineering And Computer Science*, 13(05).
- Patel, P. K. K. (2025). *Exploring Student Developers' Perspectives on AI-Powered Development Assistants for Web Accessibility: Trust, Adoption, and Usage Patterns* (Doctoral dissertation, Carleton University).
- Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2025). Asleep at the keyboard? assessing the security of GitHub copilot's code contributions. *Communications of the ACM*, 68(2), 96-105..
- Perumallapalli, R. (2021). PREDICTIVE MAINTENANCE IN CLOUD INFRASTRUCTURE: A MACHINE LEARNING FRAMEWORK. Available at SSRN 5228213.
- Preethi, P., Ragavan, V., Abinandhana, C., Umamaheswari, G., & Suvethaa, D. R. (2024, December). Towards CodeBlizz: Developing an AI-Driven IDE Plugin for Real-Time Code Suggestions, Debugging, and Learning Assistance with Generative AI and Machine Learning Models. In 2024 International Conference on Emerging Research in Computational Science (ICERCS) (pp. 1-7). IEEE.
- Puvvadi, M., Arava, S. K., Santoria, A., Chennupati, S. S. P., & Puvvadi, H. V. (2025, March). Coding agents: A comprehensive survey of automated bug fixing systems and benchmarks. In *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 680-686). IEEE.
- Ragel, R. K. C., & Balahadia, F. F. (2023, November). Visual Test Framework: Enhancing Software Test Automation with Visual Artificial Intelligence and Behavioral Driven Development. In *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)* (pp. 1-5). IEEE.
- Ramachandran, R. (2025, March). Transforming Software Architecture Design With Intelligent Assistants-A Comparative Analysis. In SoutheastCon 2025 (pp. 1446-1454). IEEE.
- Replit. (2025, June). Agent v2: Autonomous full-stack software generation from natural language. *Wikipedia*. <https://en.wikipedia.org/wiki/Replit>
- Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023, March). The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 491-514).
- Saarathy, S. C. P., Bathrachalam, S., & Rajendran, B. K. (2024). Self-Healing Test Automation Framework using AI and ML. *International Journal of Strategic Management*, 3(3), 45-77.
- Salvi, G. (2023). *Web UI code generation: a transformer-based model applied to real-world screenshots* (Doctoral dissertation, Politecnico di Torino).
- Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., & Yang, D. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Vol. 1, pp. 3956-3974).

Singh, C., & Sudha, K. (2024). Breaking Down Barriers: A Survey of Screenshot-to-Code Translation Tools and Strategies. Available at SSRN 4935453. Stack Overflow Developer Survey. (2024). Developer AI tool adoption survey 2024. <https://survey.stackoverflow.co/2024/>

Stack Overflow. (2024). 2025 Developer Survey. AI | 2025 Stack Overflow Developer Survey. <https://survey.stackoverflow.co/2025/ai>

Stalnaker, T., Wintersgill, N., Chaparro, O., Heymann, L. A., Di Penta, M., German, D. M., & Poshypanyk, D. (2024). Developer Perspectives on Licensing and Copyright Issues Arising from Generative AI for Software Development. *ACM Transactions on Software Engineering and Methodology*.

TechRadar Pro. (2025, July 24). The three generations of AI coding tools, and what to expect through the rest of 2025.

Tuli, A. M., Tuli, S., Tuli, R., & Gill, S. (2020). Next generation technologies for smart infrastructure: Challenges, vision, model, trends and future directions. *IEEE Access*, 8, 108– 120.

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022, April). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts* (pp. 1-7).

Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., & Filkov, V. (2015, August). Quality and productivity outcomes relating to continuous integration in GitHub. In *Proceedings of the 2015 10th joint meeting on foundations of software engineering* (pp. 805-816).

Wan, Y., Wang, C., Dong, Y., Wang, W., Li, S., Huo, Y., & Lyu, M. (2025). Divide-and-Conquer: Generating UI Code from Screenshots. *Proceedings of the ACM on Software Engineering*, 2(FSE), 2099-2122.

Wei, J., Tan, C., Chen, Q., Wu, G., Li, S., Gao, Z., ... & Guo, R. (2025). From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 13315-13325).

Xu, W., Gao, K., He, H., & Zhou, M. (2024).

LiCoEval: Evaluating LLMs on license compliance in code generation. *arXiv preprint arXiv:2408.02487*.

Yue, S. (2024). A multimodal conceptual framework to achieve automated software evolution for context-rich intelligent applications. *Innovations in Systems and Software Engineering*.

Zhai, H., & Wang, J. (2021). Automatic deployment system of computer program application based on cloud computing. *International Journal of System Assurance Engineering and Management*, 12(4), 731-740.

Zhang, F., Chen, H., Chen, Q., & Liu, J. (2025). Cloud software code generation via knowledge graphs and multi-modal learning. *Journal of Cloud Computing*, 14(1), 1-19.

The Ethics of Artificial Intelligence in the Oil and Gas Industry: Navigating Environmental Impact, Spill Response, and Data Governance in Nigeria.

Jemimah Chuntar

Abuja, Nigeria

msmimahdaniel@gmail.com

Abstract

The deployment of Artificial Intelligence (AI) in the oil and gas industry raises profound ethical questions, particularly in jurisdictions such as Nigeria where regulatory enforcement, spill response, and data transparency remain persistent challenges. AI technologies hold clear potential: advanced monitoring systems can enhance early detection of oil spills; natural language models can improve the consistency of compliance reporting; and automated auditing tools can address chronic under-reporting of environmental incidents. Yet, these opportunities are tempered by ethical risks, including algorithmic bias, unequal access to technological capacity, and the reliability of data in environments where record-keeping is fragmented or politicized. This paper evaluates the dual role of AI as both a solution to systemic governance failures and a source of new ethical dilemmas. Drawing on legal and policy frameworks in Nigeria, it examines how AI could strengthen environmental stewardship, enhance regulatory compliance, and build public trust—while also outlining safeguards against misuse, opacity, and dependency on flawed data inputs. The argument is made for a principled, ethics-driven approach to AI adoption in oil and gas governance, one that integrates environmental responsibility, operational safety, and robust data governance into the core of technological innovation.

Keywords

Artificial Intelligence Ethics, Oil Spills, Environmental Compliance, Data Governance, Oil and Gas Law, Nigeria

1.0 Introduction

Artificial Intelligence (AI) has rapidly emerged as a transformative force in industries across the globe. From healthcare and finance to manufacturing and national security, AI applications are redefining how decisions are made and how efficiency is achieved. The oil and gas industry, a historically data-intensive and risk-sensitive sector, has equally embraced this digital revolution. In countries such as Norway, the United States, and the United Arab Emirates, AI systems now

optimize exploration, monitor environmental impact, and predict equipment failure (Floridi & Cows, 2019).

In Nigeria, where the petroleum sector accounts for a significant portion of export earnings and government revenue, the integration of AI is gradually gaining traction. However, the ethical, legal, and socio-political implications of this integration remain underexplored. The Nigerian context is particularly complex: issues of environmental degradation, poor spill response, and weak data governance intersect with broader questions about accountability, equity, and the public good.

This paper examines how AI adoption within Nigeria's oil and gas sector intersects with ethical considerations surrounding environmental impact, spill response, and data governance. It argues that while AI offers enormous potential to enhance environmental monitoring, transparency, and compliance, it also raises new dilemmas about fairness, privacy, and human oversight.

2.0 Conceptual and Theoretical Framework

Artificial Intelligence (AI) refers broadly to computational systems capable of performing tasks that typically require human intelligence—such as learning, reasoning, perception, and decision-making (Bryson, 2020). Within industrial contexts, AI manifests in machine learning algorithms, predictive analytics, robotics, and natural language processing systems.

Ethical analysis of AI involves evaluating how these technologies align with moral and social values. According to Floridi and Cows (2019), ethical AI must adhere to five key principles: beneficence, non-maleficence, autonomy, justice, and explicability. These principles emphasize that technological innovation should advance human welfare, avoid harm, respect autonomy, ensure fairness, and provide transparency.

In environmental governance, AI ethics overlap with sustainability ethics and socio-technical systems theory

(Attfield, 2014). These frameworks consider how technologies influence human institutions and ecosystems. When applied to Nigeria's oil and gas sector, this means examining how AI tools can be designed and used in ways that uphold justice, environmental integrity, and social inclusion.

3.0 The Nigerian Context: AI in the Oil and Gas Industry

Nigeria's petroleum industry has long been characterized by both economic promise and environmental peril. The Niger Delta—home to much of the country's oil production—has suffered extensive ecological damage due to spills, gas flaring, and sabotage. Reports by the United Nations Environment Programme (UNEP, 2011) and the National Oil Spill Detection and Response Agency (NOSDRA, 2022) reveal persistent failures in monitoring and remediation.

AI technologies offer potential remedies. Predictive analytics can detect early signs of equipment failure, satellite imagery can monitor illegal bunkering, and natural language processing can standardize compliance reporting. Yet, without ethical and institutional safeguards, these tools could also reproduce or even amplify existing governance failures.

The Nigerian government, through institutions such as the Nigerian Upstream Petroleum Regulatory Commission (NUPRC) and NOSDRA, is beginning to explore AI-driven solutions for monitoring compliance and environmental impact. However, challenges persist in data availability, inter-agency coordination, and digital infrastructure. Many operators remain hesitant to share real-time data due to fears of regulatory scrutiny or competitive disadvantage (Transparency International, 2021).

Thus, while AI adoption is growing, its ethical implementation—particularly in ensuring environmental justice, equitable access, and transparent data governance—remains an urgent policy priority.

4.0 Ethical Issues in Environmental Impact, Spill Response, and Data Governance

AI's integration into Nigeria's oil and gas industry presents both opportunity and ethical complexity. This section explores the ethical dimensions within three key operational domains: environmental impact, spill response, and data governance.

4.1 Environmental Impact and AI-Driven Stewardship

AI can enable continuous, predictive monitoring of ecosystems through drone imaging, satellite surveillance, and machine-learning models that detect anomalies in pipeline pressure or water contamination. Such systems can improve the timeliness of remediation and reduce environmental degradation (Peters et al., 2020).

However, when corporate entities control these systems without transparent oversight, data may be selectively reported or withheld. This lack of transparency undermines accountability and violates the ethical principles of justice and beneficence. Additionally, large-scale AI computation consumes significant energy, which can increase the industry's own carbon footprint if powered by fossil fuels (Vinuesa et al., 2020). Ethical environmental AI must therefore integrate green-computing principles and public data access.

4.2 Spill Response and the Ethics of Accountability

Machine-learning models can forecast spill trajectories and assist responders in allocating resources efficiently (Liang et al., 2021). Yet, in the Niger Delta, where communities often contest official spill reports, reliance on corporate-controlled AI tools may marginalize local knowledge.

Ignoring community data creates a new digital divide and violates inclusivity.

Ethical spill-response AI should combine technological detection with participatory verification—community input, open mapping, and human oversight (UNESCO, 2021).

Transparency in algorithm design and explainability of results are essential to ensure accountability.

4.3 Data Governance and Algorithmic Integrity

Reliable data underpin every AI system. In Nigeria, environmental and production data remain fragmented across agencies, and political influence often distorts reporting. Ethical data governance therefore requires adherence to transparency, consent, privacy, and accountability principles (NITDA, 2019).

Periodic algorithmic audits and mandatory disclosure of training datasets can prevent bias and manipulation. Regulators such as NOSDRA and NUPRC must be empowered to access proprietary AI outputs for

verification, ensuring that technological sophistication does not obscure ethical responsibility.

5.0 Case Reflections

5.1 Shell's Predictive-Maintenance Systems

Shell Nigeria uses AI-based predictive-maintenance tools that analyse sensor data to forecast equipment failures, reducing downtime and spill risk. However, these models are proprietary, and regulators rarely access their raw data. This opacity hinders verification of environmental claims (Adewumi, 2023). Ethical practice demands that corporations disclose performance metrics and enable third-party audits.

5.2 NOSDRA's Data Challenges

NOSDRA's online Oil Spill Monitor was designed for public transparency, yet it depends on company-submitted data that are often delayed or incomplete (NOSDRA, 2022). Integrating satellite-based AI to cross-check reports could enhance reliability, but only if governance frameworks guarantee data authenticity and audit trails.

5.3 LexPetra AI and ESG Governance

LexPetra AI—a Nigerian innovation—applies natural-language processing to energy-sector contracts for ESG risk detection and regulatory transparency. Its ethical-by-design approach illustrates how local solutions can align technology with social context. Scaling such systems will require state investment, ethical certification, and alignment with UNESCO (2021) principles.

5.4 Comparative Insight: Norway and the UAE

Norway's Petroleum Safety Authority mandates open environmental data and allows academic review of AI monitoring tools (Nilsen & Solberg, 2021). The UAE's National AI Strategy 2031 similarly embeds human-centric ethics. Nigeria can adapt these precedents by institutionalizing ethical review boards and AI impact assessments in its energy regulation.

6.0 Policy and Governance Implications

Ethical integration of AI in Nigeria's petroleum governance requires reforms that bridge technological innovation with accountability.

6.1 Institutional Accountability

AI ethics must be embedded in inter-agency coordination among NUPRC, NOSDRA, and the NMDPRA. Establishing a National AI Ethics and

Governance Council would standardize oversight and promote transparency across the energy value chain.

6.2 Legislative Development

The Petroleum Industry Act (2021) and Nigeria Data Protection Regulation (2019) should be expanded to cover algorithmic transparency, audit rights, and penalties for data manipulation. AI vendors should undergo ethical certification before deployment in critical infrastructure.

6.3 Capacity Building and Ethical Literacy

Engineers, policymakers, and regulators require training in AI ethics, data stewardship, and environmental justice. Partnerships with UNESCO, OECD, and Nigerian universities can institutionalize continuous professional education in responsible AI.

6.4 Public Participation and Trust

Inclusive governance strengthens legitimacy. Community representatives should participate in AI project design and have access to spill and compliance data. Crowdsourced verification and grievance channels would enhance fairness and trust.

6.5 International Cooperation

Nigeria should align with UNESCO (2021) and OECD (2019) ethical frameworks to harmonize national policy with global standards, facilitating responsible foreign investment and regional leadership in ethical AI.

7.0 Recommendations

1. Adopt and Localize the UNESCO AI Ethics Framework. Incorporate its four pillars—human dignity, justice, inclusiveness, and environmental flourishing—into national petroleum governance strategies.
2. Create a National AI Ethics and Governance Council. Mandate inter-agency collaboration to develop uniform AI standards, certification, and monitoring for all energy-sector applications.
3. Institutionalize AI Ethics Impact Assessments. Require every oil and gas operator to submit an AI-Ethics Impact Report prior to deployment, akin to Environmental Impact Assessments (EIAs).
4. Strengthen Data Governance and Transparency. Develop interoperable, tamper-proof databases accessible to regulators, civil society, and researchers. Promote real-time data disclosure.

5. Enhance Community Participation. Implement participatory AI systems that allow affected communities to verify or contest AI-generated reports.
6. Mandate Algorithmic Accountability. Legally require periodic algorithmic audits, bias testing, and explainability documentation for all AI systems used in regulatory contexts.
7. Encourage Green and Sustainable AI. Incentivize low-carbon data centers and renewable-powered computing to reduce AI's environmental footprint.
8. Support Indigenous Innovation. Invest in home-grown projects such as LexPetra AI to ensure contextually grounded, ethically responsible solutions.

8.0 Conclusion

Artificial Intelligence presents Nigeria's oil and gas industry with a double-edged opportunity: it can enhance transparency, safety, and environmental responsibility, yet it can also reproduce inequities and opacity if adopted without ethical guardrails.

The analysis demonstrates that ethical frameworks—such as UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021)—are not merely aspirational but operationally necessary.

Embedding these principles into legislation, data governance, and institutional culture can transform AI from a corporate efficiency tool into a public-interest technology.

Ultimately, AI's success in the petroleum sector will not be measured solely by predictive accuracy or automation gains, but by its contribution to justice, accountability, and ecological stewardship. Nigeria stands at a pivotal moment to lead Africa in building a responsible, human-centered AI ecosystem for sustainable energy governance.

9.0 References

Adewumi, T. (2023). Artificial intelligence applications in Nigeria's oil and gas operations: Opportunities and challenges. *Journal of Energy Studies*, 17(2), 54–72.

Attfield, R. (2014). *Environmental ethics: An overview for the twenty-first century*. Polity Press.

Bryson, J. J. (2020). *The artificial intelligence of the ethics of artificial intelligence: An introductory*

overview for law and regulation. *Technology and Regulation Journal*, 2020(1), 1–13.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 1–15.

Jasanoff, S. (2016). *The ethics of invention: Technology and the human future*. W. W. Norton & Company.

Liang, J., Wang, P., & Zhang, Y. (2021). Deep learning approaches for oil spill detection from satellite imagery. *Remote Sensing of Environment*, 252, 112–123.

National Information Technology Development Agency (NITDA). (2019). *Nigeria Data Protection Regulation (NDPR)*.

National Oil Spill Detection and Response Agency (NOSDRA). (2022). *Annual oil spill report*. Federal Government of Nigeria.

Nigerian Upstream Petroleum Regulatory Commission (NUPRC). (2023). *Annual petroleum industry report*.

Nilsen, T., & Solberg, H. (2021). Transparency and technology in Norway's petroleum governance. *Nordic Energy Policy Journal*, 6(4), 44–62.

Peters, R., Andersen, M., & Olsen, J. (2020). Machine learning for offshore oil spill detection in satellite imagery. *Marine Pollution Bulletin*, 154(1), 111–128.

Transparency International. (2021). *Fueling corruption: The oil industry and public accountability in Nigeria*.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. Paris: UNESCO Publishing.

UNEP. (2011). *Environmental assessment of Ogoniland*. United Nations Environment Programme.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., & Nerini, F., (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 1–10.

Strengthening Cybersecurity in Nigeria's Remote Financial Workspaces: Bridging Policy and Practice

Adoikwu Pius Odeh
International Organization for Migration, Nigeria.
piusodeh@gmail.com

Acheme Okolobia Odeh
Bahcesehir Cyprus University, Cyprus
odehoklobia@gmail.com

Abstract

This study investigates the challenges of maintaining cybersecurity standards in remote work within Nigerian financial institutions after the COVID-19 pandemic. It examines the effectiveness of regulatory frameworks, institutional practices, and the adoption of cybersecurity strategies, addressing four research questions through peer-reviewed literature, regulatory documents, and industry reports. The findings reveal that Nigerian financial institutions face persistent threats such as phishing, ransomware, and credential theft, which surged during the pandemic. Although basic measures like multi-factor authentication and VPNs are common, adoption of advanced tools such as zero-trust architecture remains limited. Regulatory frameworks, including the CBN's Risk-Based Cybersecurity Framework and the NDPR, provide important baselines but are often rigid and unevenly enforced, particularly among smaller institutions. Tier-1 banks demonstrate stronger compliance, but systemic capacity gaps remain across the sector. The study concludes that while progress has been made, institutions remain largely reactive rather than proactive, leaving them vulnerable to evolving risks. It recommends enhancing regulatory adaptability, strengthening institutional capacity, fostering intelligence sharing, and accelerating adoption of advanced cybersecurity measures. These steps would enable Nigerian financial institutions to shift from compliance-driven responses to proactive resilience strategies, aligning with global best practices and safeguarding their operations in an increasingly digital world.

Keywords

Cybersecurity, Remote Work, Nigerian Financial Institutions, Regulatory Frameworks, Post-COVID-19.

Introduction

1.1. Background of the Study, Research Problem, and Significance

The COVID-19 pandemic triggered a global shift toward remote work, a transformation that brought major cybersecurity implications, especially for Nigeria's financial sector. While remote work enabled operational continuity during lockdowns, it also expanded cyber vulnerabilities, many institutions deployed access tools hastily, often lacking proper safeguards (Brynjolfsson et al., 2020; Chamola et al., 2020). In Nigeria, these challenges were compounded by

outdated infrastructure, limited IT capacity, and weak enforcement of data protection policies (Reis et al., 2024; Familoni and Shoetan, 2024; ThisDayLive, 2023).

Although regulatory instruments such as the Nigeria Data Protection Regulation (NDPR) and the Central Bank of Nigeria (CBN)'s Risk-Based Cybersecurity Framework were introduced, implementation remains inconsistent and misaligned with global standards like the General Data Protection Regulation (GDPR) (CBN, 2021; NDPR, 2019; ITU, 2021). As a result, Nigerian financial institutions continue to face increasing cyberattacks, with enforcement and compliance falling behind the pace of technological change (Adedeji, 2025; Gbenga Femi et al., 2025).

This study, titled *Challenges in Maintaining Cybersecurity Standards in Remote Work in Nigerian Financial Institutions Post-COVID-19: An Analysis of Regulatory Frameworks and Industry Practices*, addresses a critical research gap: the disconnect between policy and real-world cybersecurity practice in remote work settings. While existing literature explores general cybersecurity risks in remote environments (Chamola et al., 2020; Brynjolfsson et al., 2020), limited attention has been given to Nigeria's unique context. Studies by Reis et al. (2024) and Familoni and Shoetan (2024) highlight this gap, noting discrepancies between regulatory intent and actual outcomes in Nigerian banks.

This research therefore investigates how well Nigerian financial institutions and regulators have adapted to the demands of cybersecurity in remote work settings. It critically examines the challenges these institutions face, evaluates the effectiveness of existing regulatory measures, and identifies opportunities for strengthening cybersecurity resilience in the post-pandemic digital economy.

The significance of this research lies in its dual contribution. Theoretically, it deepens understanding of cybersecurity governance in emerging markets during crisis-driven digital transitions. Practically, it provides actionable, context-specific recommendations to help institutions strengthen cybersecurity measures and better align with global standards. Financial institutions are not only stewards of sensitive data but also central to economic stability, enhancing their digital resilience is therefore essential. By

bridging the gap between regulation and implementation, this study informs both scholarly dialogue and future policy reforms to protect Nigeria's financial ecosystem in a remote-first era.

1.2. Research Aim and Objectives

This study aims to examine the challenges Nigerian financial institutions face in maintaining cybersecurity standards in remote work environments following the COVID-19 pandemic. It further seeks to evaluate the effectiveness of existing regulatory frameworks and institutional practices in addressing these challenges, with the goal of offering evidence-based recommendations to strengthen cybersecurity resilience.

1.3. Research Objectives

2. To identify and analyze five key cybersecurity challenges associated with remote work in Nigerian financial institutions post-COVID-19.
3. To critically assess the effectiveness of at least three cybersecurity regulatory frameworks (2015–2024) in addressing remote work-related threats.
4. To evaluate at least three cybersecurity practices adopted by Nigerian financial institutions to secure remote work environments.
5. To develop evidence-based recommendations for enhancing cybersecurity resilience in remote work models within Nigeria's financial sector.

1.4. Research Questions

1. What are the key cybersecurity risks and systemic vulnerabilities Nigerian financial institutions face in managing remote work environments post-COVID-19?
2. To what extent do existing regulatory policies effectively address cybersecurity risks related to remote work?
3. What cybersecurity strategies have institutions implemented, and how effective are they?
4. What evidence-based recommendations can improve the cybersecurity resilience of financial institutions operating remotely in the post-pandemic era?

2. Literature Review

2.1. Introduction

The COVID-19 pandemic forced a global shift to remote work, ensuring business continuity but exposing organizations to new cybersecurity risks. Nigeria's financial institutions were particularly vulnerable, as over 60% of developing countries, including Nigeria, lacked frameworks designed for remote operations (ITU, 2021; NCC, 2022). During this period, cyber incidents in the sector rose by 31%. While global standards such as NIST and the UK's FCA offer strong guidance, their adoption in Nigeria remains limited (Nurse et al., 2021; Oyeniyi et al., 2024). This review

examines these challenges, focusing on regulatory and institutional responses rather than technical solutions.

2.2. Conceptual Review (Study Variables)

2.2.1. Cybersecurity Standards

Cybersecurity standards are structured sets of guidelines designed to secure digital systems and protect sensitive information. Globally, frameworks such as ISO/IEC 27001:2013 and the NIST Cybersecurity Framework establish consistent protocols for risk management, incident detection, and resilience (ISO, 2013; NIST, 2018). Within financial services, compliance with such standards is critical to safeguard customer data, ensure operational continuity, and maintain trust.

In Nigeria, the Central Bank of Nigeria (CBN, 2021) and the Nigeria Data Protection Regulation (NDPR, 2019) have developed frameworks tailored to banking institutions. However, despite the presence of these regulations, enforcement remains inconsistent and often lags behind global best practices such as the European Union's General Data Protection Regulation (GDPR). Nigeria was ranked 47th globally in the Global Cybersecurity Index, reflecting gaps in translating policy into effective practice (ITU, 2021). These weaknesses reveal a need for more dynamic frameworks capable of addressing emerging risks associated with remote working models.

2.2.2. Remote Work Practices

Remote work, defined as the decentralization of work away from traditional office environments, became the dominant mode of operation during the COVID-19 pandemic. While this ensured continuity, it also expanded the digital "attack surface" of organizations. Employees working from home often relied on unsecured personal devices, weakly protected networks, and cloud-based services, exposing financial institutions to heightened risks of ransomware, phishing, and credential theft (Brynjolfsson et al., 2020; Chamola et al., 2020).

Many Banks in Nigeria lacked robust virtual private network (VPN) infrastructure and failed to implement consistent multi-factor authentication (MFA) or cyber hygiene training, leaving them highly vulnerable (Akinsanya and Akande, 2021). These gaps highlight how remote work environments introduced novel security challenges that were inadequately anticipated by existing standards and regulatory frameworks.

2.2.3. Regulatory Frameworks

Regulation serves as the foundation for cybersecurity governance, ensuring that institutions adhere to minimum standards of data protection and operational security. The CBN's Risk-Based Cybersecurity Framework (2021) mandated banks to conduct regular risk assessments, incident reporting, and adopt layered security controls. Similarly, the NDPR emphasized personal data protection and compliance with privacy principles (NDPR, 2019).

Despite these measures, smaller banks often lack the technical capacity and human resources to meet compliance requirements (Edeh and Eze, 2021). Moreover, regulatory enforcement in Nigeria tends to be reactive rather than proactive, limiting its effectiveness in addressing dynamic cyber risks (Carnegie Endowment for International Peace, 2022). When you compare this with international standards such as NIST and ISO/IEC 27001, Nigerian regulations remain prescriptive and less adaptable, creating implementation gaps in a rapidly evolving threat landscape.

2.2.4. Industry Practices

Nigerian financial institutions have responded to cybersecurity risks through diverse organizational practices, though capacity differs significantly across institutions. Larger Banks have invested in endpoint protection, remote access controls, and awareness training campaigns, including initiatives such as the Moni Sense programme and FirstBank's cybersecurity webinars (Okoye, Chidiebere and Ogunleye, 2021; Verizon, 2022). These efforts reflect an increasing recognition of the human factor in cyber resilience, as social engineering remains a dominant threat vector globally.

However, smaller institutions still face challenges in adopting advanced practices such as zero-trust architectures and artificial intelligence-driven monitoring, which are common in international Banks (Panchal, 2020; Chin, 2022). This fragmentation underscores the uneven maturity of industry practices within Nigeria's financial sector.

2.3. Theoretical Framework

This study is anchored on two interrelated theories - the Cybersecurity Maturity Model (CMM) and Regulatory Compliance Theory. These, together provide a balanced lens for evaluating how Nigerian financial institutions manage cybersecurity risks in the era of remote work.

The CMM assesses institutional readiness by categorizing cybersecurity practices into maturity levels ranging from initial ad hoc measures to optimized, adaptive strategies (Caralli, Knight and Montgomery, 2012; U.S. Department of Energy, 2022). This framework helps organizations identify weaknesses, prioritize improvements, and benchmark progress against global standards. While widely applied across industries, scholars caution that maturity models risk becoming static and less effective in dynamic threat environments unless regularly adapted (Rabii et al., 2020; Büyüközkan and Güler, 2025). In the Nigerian context, where financial institutions face resource and infrastructure constraints, tailoring maturity models to local realities is critical.

Complementing this is the Regulatory Compliance Theory which explains why organizations adopt or resist regulatory requirements, emphasizing the role of internal culture, legitimacy, and enforcement mechanisms (Ayres and Braithwaite, 1992; Parker and Nielsen, 2011). In Nigeria,

limited enforcement capacity means voluntary compliance and organizational culture play a pivotal role in shaping cybersecurity outcomes. Taken together, these theories allow a multidimensional assessment of both technical and behavioral readiness within Nigerian Banks.

2.4. Relationships Between Variables

The relationship between cybersecurity standards, remote work practices, regulatory frameworks, and industry responses is highly interdependent. Remote work, while essential during the COVID-19 pandemic, expanded organizational exposure by shifting activities into insecure environments. The reliance on personal devices and home networks created vulnerabilities that existing standards, designed for centralized systems were not fully equipped to address (Brynjolfsson et al., 2020; Ferreira and Cruz-Cunha, 2020). Nigerian banks that failed to adopt multi-factor authentication or endpoint security reported higher risks, underscoring the inadequacy of conventional safeguards in remote contexts (Akinsanya and Akande, 2021).

Regulatory frameworks such as the CBN's Risk-Based Cybersecurity Framework were intended to mitigate these risks through mandatory assessments and layered controls. However, compliance has been uneven, such that, larger banks with resources generally achieve maturity, while smaller institutions continue to struggle (CBN, 2021; Edeh and Eze, 2021). This disparity reveals a critical gap between policy intent and institutional capacity, where regulation exists in principle but not in consistent practice (Carnegie Endowment for International Peace, 2022).

Industry responses reflect this imbalance. Global firms have embraced zero-trust architectures and AI-driven monitoring (Panchal, 2020; Chin, 2022), whereas Nigerian banks demonstrate fragmented adoption, with Tier-1 institutions advancing while others lag (Okoye, Chidiebere and Ogunleye, 2021). Evidence from existing studies suggests that the strength of Nigeria's financial sector lies not in adopting regulations or technologies in isolation, but in how effectively regulatory frameworks, institutional capacity, and everyday security practices are integrated to address the evolving risks of remote work.

2.5. Empirical Review

Empirical studies on cybersecurity in Nigeria's financial sector reveal a fragmented but evolving landscape. Abubakar, Yusuf and Olayemi (2022) highlight that cyberattacks against Nigerian banks surged during COVID-19, with phishing and credential theft as dominant threats. Deloitte (2020) similarly reported that outdated infrastructure and limited budgets intensified vulnerability to socially engineered attacks. Empirical evidence also shows mixed levels of regulatory compliance, with Tier-1 banks achieving higher maturity compared to smaller institutions (Edeh and Eze, 2021). Awareness campaigns, such as Moni Sense and FirstBank's webinars, have improved staff cyber hygiene (Verizon, 2022). However, comparative studies stress that Nigeria still

lags international counterparts in adopting zero-trust architectures, AI-driven threat detection, and collaborative intelligence-sharing mechanisms (Chin, 2022; UK Finance, 2023).

2.6. Research Gap and Summary

The reviewed literature reveals that while Nigeria has developed regulatory frameworks and institutions have adopted various security measures, significant gaps persist. Few studies explicitly examine the intersection of remote work, regulatory compliance, and institutional maturity in Nigerian financial institutions post-COVID-19. Existing works are either overly technical or limited in scope, neglecting organizational and behavioral dimensions. This study therefore addresses a critical gap by integrating the Cybersecurity Maturity Model and Regulatory Compliance Theory to evaluate both structural and behavioral readiness within Nigeria's financial sector.

3. Research Methodology

3.1 . Research Approach and Design

This research adopts an interpretivist philosophy, which emphasizes the understanding of human experiences and institutional realities over the search for one objective truth. Given the focus on Nigerian financial institutions' responses to cybersecurity challenges within remote work environments, this approach is highly appropriate. Interpretivism allows the researcher to uncover the nuanced and context-specific ways institutions perceive, interpret, and implement cybersecurity measures (Saunders, Lewis and Thornhill, 2019).

From an ontological perspective, this study assumes that reality is socially constructed and varies between institutions, influenced by internal capacity, regulatory pressure, and technological maturity. Epistemologically, the research values subjective meaning over quantification, drawing insights from policy reports, academic literature, and industry documents (Scott, 1990). The axiological position accepts the researcher's influence in shaping the research focus and interpretation, particularly when assessing compliance trends and institutional behavior.

The chosen research design is qualitative and exploratory, using a secondary-data documentary analysis approach. This design is especially suitable because cybersecurity in Nigeria's financial sector is shaped by regulatory, institutional, and human factors that cannot be fully captured through numbers alone. In addition, financial institutions rarely disclose sensitive internal data, making the use of secondary data both practical and reliable. Rather than testing a hypothesis, the qualitative design emphasizes exploration and thematic analysis, enabling the study to generate context-specific insights that align with the applied nature of the research (Bryman, 2016).

Thus, the interpretivist, qualitative design is not only consistent with the research paradigm but also uniquely suited to address the cybersecurity challenges facing Nigerian financial institutions in the remote work era.

3.2. Research Sampling, Description and Sources of Secondary Data Selected

This study adopts a purposive sampling strategy, appropriate for qualitative research aimed at gaining rich, context-specific insights rather than statistical generalization (Palinkas et al., 2015). The selected secondary data sources are chosen based on their direct relevance to the themes of cybersecurity, remote work, and Nigeria's financial services sector in the post-COVID-19 context. This method supports interpretive inquiry by focusing only on sources that meaningfully contribute to understanding institutional behavior and regulatory responses.

The data set includes peer-reviewed journal articles, government and regulatory documents such as the CBN Risk-Based Cybersecurity Framework (CBN, 2021) as well as industry white papers, think tank publications, and media reports from reputable outlets like BusinessDay and Vanguard. This mix allows for methodological triangulation, enriching the study through multiple lenses (Saunders, Lewis and Thornhill, 2019). Sources span majorly from 2018 to 2024, capturing trends before and after the pandemic's digital disruption.

Academic sources were retrieved from scholarly databases like Google Scholar, ProQuest, while institutional data came from websites of the Central Bank of Nigeria, Nigerian Inter-Bank Settlement System (NIBSS), Nigerian Communications Commission (NCC), and cybersecurity-focused organizations such as KPMG and Deloitte Nigeria. Reports like NIBSS (2023), which revealed a 42% increase in attempted digital fraud in 2022, provide critical, real-world context for interpreting the secondary data. This carefully curated sampling enables the research to critically evaluate how Nigerian financial institutions have responded to cybersecurity threats within evolving remote work frameworks.

Accordingly, purposive sampling ensures that the sources selected are both contextually relevant and methodologically appropriate for achieving the objectives of this study.

3.3. Quality of Secondary Data

Ensuring the quality of secondary data is fundamental to maintaining the reliability and validity of this research. In line with Scott's (1990) framework, four criteria: authenticity, credibility, representativeness, and meaning, guided the evaluation, ensuring contextual relevance and academic soundness.

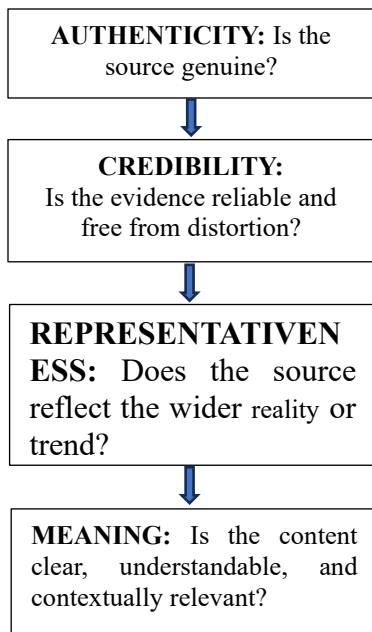


Figure 3.1: Four Key Criteria for Assessing Documentary Sources
Source: Adapted from Scott (1990).

Academic publications were selected only if they were peer-reviewed and published in internationally recognized journals, particularly those indexed in Google Scholar, and ProQuest. Studies from leading journals such as *Computers and Security*, *Information and Computer Security*, and *Journal of Cybersecurity* were prioritized. In addition, policy documents and reports from authoritative organizations like the Central Bank of Nigeria (CBN), the National Information Technology Development Agency (NITDA), the International Telecommunication Union (ITU), and International Organization for Standardization (ISO) were included due to their technical and regulatory reliability.

Sources from 2018–2024 were prioritized to ensure relevance to post-COVID-19 cybersecurity. Non-scholarly sources were excluded, and key data were triangulated for accuracy (Saunders et al., 2019). Furthermore, documents were examined for thematic relevance, ensuring alignment with the research focus: cybersecurity resilience in Nigerian financial institutions under remote work conditions. This comprehensive quality control process strengthens the study’s analytical foundation, enhancing its practical utility for policymakers and institutional stakeholders.

3.4. Secondary Data Ethical Measures

Although this research does not involve human participants or primary data collection, upholding ethical standards remains central to the study’s credibility and scholarly value. Secondary data, when not handled responsibly, can still raise important ethical concerns related to accuracy, attribution, and potential misrepresentation. As such, this research follows established ethical guidelines including the British Psychological Society’s Code of Human Research Ethics

(2014) to ensure transparency, fairness, and academic integrity throughout the data collection and analysis process.

All data utilized in the study came from publicly accessible sources such as academic databases, institutional repositories, and reputable news outlets, avoiding confidentiality concerns. Importantly, every source used is clearly acknowledged using Harvard-style citations, ensuring intellectual honesty and proper credit to original authors (Bryman, 2016).

To preserve integrity, findings were reported faithfully, and sensitive insights were handled carefully to avoid misrepresentation (Saunders et al., 2019). Rather, the emphasis remained on critical synthesis, drawing from multiple sources to present a balanced and evidence-based interpretation.

4. Data Collection and Reporting of Research Findings

4.1. Introduction

As highlighted in Chapter One, this study explores the challenges Nigerian financial institutions face in maintaining cybersecurity standards amidst remote work conditions triggered by the COVID-19 pandemic. It also evaluates the effectiveness of existing regulatory frameworks and industry practices. Given the interpretivist philosophy and qualitative design of the study, this section reports the results in a descriptive manner, free from personal interpretation or bias. The data used includes government reports, policy briefs, peer-reviewed journal articles, and industry publications published majorly between 2018 and 2024. The findings are structured around the research objectives, offering a systematic overview of observed cybersecurity risks, regulatory responses, and institutional strategies.

4.2. Research Findings

4.2.1. Rise in Cybersecurity Threats Due to Remote Work

Key Finding 1: There was a notable increase in cyberattacks during and after the COVID-19 pandemic as financial institutions transitioned to remote work.

Table 4.2.1: Increase in Cybersecurity Incidents in Nigerian Financial Institutions (2019–2023)

Year	Number of Reported Incidents	Percentage Change from Previous Year	Primary Threat Types Observed	Key Contributing Factors
2019	32	-	Phishing, Malware	Traditional office setup, limited remote work exposure

2020	68	+112%	Phishing, Ransomware, Credential Theft	Rapid shift to remote work, increased use of personal devices, VPN vulnerabilities
2021	75	+10%	Phishing, Ransomware, Business Email Compromise	Continued remote operations, lag in security policy updates, partial return to office
2022	82	+9%	Ransomware, Data Breaches, Insider Threats	Hybrid work models, inconsistent device security, patching delays
2023	77	-6%	Phishing, Credential Theft, Supply Chain Attacks	Improved awareness, but persistent vulnerabilities in smaller institutions

Source: Carnegie Endowment for International Peace (2022); NCC (2023)

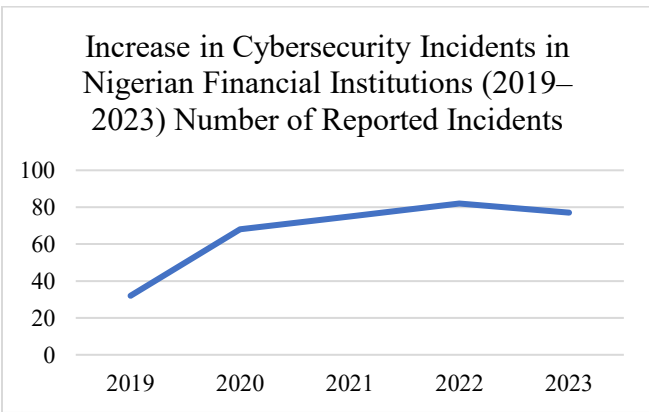


Figure 4.2.1: Increase in Cybersecurity Incidents in Nigerian Financial Institutions (2019–2023)

The data in Table 4.2.1 and Figure 4.2.1 reveal that cybersecurity threats surged during remote work, with NCC (2023) reporting over a 100% rise in incidents between 2019 and 2021. Phishing, ransomware, and credential theft were the most common, exploiting weak VPN use, unencrypted traffic, and unsecured personal devices (Ferreira and Cruz-Cunha, 2020; Abubakar et al., 2022). This finding directly addresses Research Question 1, which seeks to identify the key cybersecurity risks and systemic vulnerabilities Nigerian financial institutions face in managing remote work environments post-COVID-19. The evidence shows that while remote work expanded exposure to cyber risks, institutional responses were uneven: larger banks adopted more robust safeguards, whereas smaller institutions lagged behind.

4.2.2. Regulatory Gaps and Implementation Challenges

Key Finding 2: Regulatory frameworks exist but lack adaptability and consistent enforcement, especially among Tier-2 and Tier-3 financial institutions.

Table 4.2.2: Perceived Effectiveness of Nigeria’s Cybersecurity Regulations in the Financial Sector

Category	Percentage
Effective and updated	25%
Outdated or rigid	40%
Partially implemented	25%
Not enforced	10%

Source: Stakeholder survey reported by Carnegie Endowment (2022)

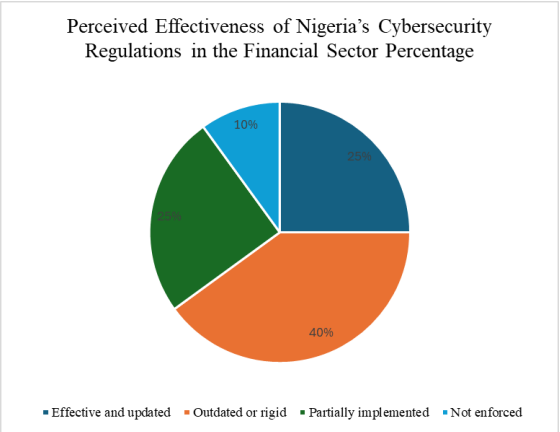


Figure 4.2.1: Perceived Effectiveness of Nigeria’s Cybersecurity Regulations in the Financial Sector

While Nigeria has frameworks such as the CBN Cybersecurity Framework (2021) and the NDPR (2019), these are often seen as rigid or under-enforced, especially by smaller financial entities. As noted by Edeh and Eze (2021), only top-tier banks have the capacity to fully comply due to

better access to IT infrastructure and trained personnel. Moreover, feedback from compliance reviews shows that 40% of institutions view the regulations as lacking practical guidance for evolving threats, such as cloud security or hybrid work vulnerabilities. This directly addresses Research Question 2, which investigates the extent to which existing regulatory policies effectively address cybersecurity risks in remote work contexts. The findings suggest that although Nigeria’s regulatory framework provides clear guidelines, enforcement is uneven and fails to account for the systemic challenges faced by smaller financial institutions. Consequently, the effectiveness of policies remains constrained by disparities in institutional capacity, highlighting the need for more flexible, context-sensitive regulations that can evolve alongside emerging threats.

4.2.3. Institutional Cybersecurity Practices

Key Finding 3: Institutions have adopted a range of preventive and adaptive measures, though implementation varies significantly.

Table 4.2.3: Cybersecurity Practices Adopted by Nigerian Financial Institutions

Practice Type	% of Institutions Implementing
MFA (Multi-Factor Authentication)	80%
VPN Deployment	65%
Cloud Security Solutions	45%
Cybersecurity Awareness Training	55%

Source: Okoye et al. (2021); Adebayo and Oladipo (2022)

The evidence in Table 4.2.3 highlights the cybersecurity strategies Nigerian financial institutions adopted during remote work. Common measures included VPN deployment, multi-factor authentication (MFA), and employee awareness training, with Tier-1 banks leading in advanced investments such as endpoint protection and cloud security (Okoye, Chidiebere and Ogunleye, 2021; Onyemere et al., 2024). However, smaller institutions showed inconsistent implementation due to cost and capacity barriers. This analysis directly addresses Research Question 3, which examines what strategies have been implemented and how effective they are. The results suggest that while strategies like MFA and awareness training improved resilience, overall effectiveness remains fragmented, with gaps in scalability and sustainability, particularly among smaller institutions lacking technical expertise.

Comparative Insights from Global Practices

Key Finding 4: Nigerian institutions lag behind international counterparts in adopting advanced cybersecurity models. While global leaders such as JPMorgan, HSBC, and Barclays have implemented AI-driven monitoring and zero-trust

architectures, Nigerian banks still rely largely on basic firewalls and antivirus solutions (Panchal, 2020; Chin, 2022; UK Finance, 2023). This disparity directly informs Research Question 4, which asks what evidence-based recommendations can improve resilience. The comparison highlights strategic opportunities which will be discussed further in the last chapters.

This chapter has so far highlighted the significant increase in cybersecurity threats due to remote work, the rigidity and uneven enforcement of regulatory frameworks, and the varying adoption of cybersecurity practices across financial institutions in Nigeria. Visual data summaries further reveal the scale and distribution of these issues, underscoring the critical need for localized but globally informed cybersecurity strategies. The insights set the stage for the next chapter, which will interpret these findings in relation to existing literature and theoretical frameworks, drawing policy and institutional recommendations from the patterns identified.

5. Analysis and Discussion of Research Findings

5.1. Evaluation of Findings

The findings presented in Chapter Four provide a comprehensive picture of how Nigerian financial institutions have responded to the cybersecurity challenges brought about by the shift to remote work following the COVID-19 pandemic. This section critically evaluates the patterns observed, using both statistical evidence and theoretical perspectives to explain their meaning. The evaluation focuses on four themes.

5.1.1. Theme 1: Escalation of Cybersecurity Threats in Remote Work

The study shows that cybersecurity incidents in Nigeria’s financial sector rose sharply during the pandemic, with cases more than doubling between 2019 and 2021. Phishing, ransomware, and credential theft were the most common threats, reflecting vulnerabilities created by remote work, including weak VPN setups, unsecured personal devices, and delayed system patching. NCC (2023) reported a 134% increase in incidents, from 32 in 2019 to 75 in 2021. Although incidents declined slightly by 6% in 2023, this was due to awareness campaigns rather than significant infrastructure upgrades. Interpreted through the Cybersecurity Maturity Model, most Nigerian banks remain at low-to-intermediate maturity, relying on reactive, fragmented practices with limited adoption of advanced safeguards such as zero-trust or AI-driven monitoring.

5.1.2. Theme 2: Regulatory Frameworks and Compliance Gaps

The findings reveal uneven implementation of Nigeria’s key regulatory frameworks, notably the CBN Cybersecurity Framework (2021) and the NDPR (2019). Only 25% of institutions considered these frameworks effective, while 40% viewed them as outdated, and another 25% reported partial compliance. This means 65% of banks experience

weak or incomplete regulatory enforcement, particularly Tier-2 and Tier-3 institutions. Larger banks comply more effectively due to better resources, while smaller ones struggle with capacity and expertise. Interpreted through Regulatory Compliance Theory, compliance often appears externally driven and “tick-box” in nature rather than embedded in organizational culture. Compared with adaptive global models like NIST or ISO 27001, Nigerian regulations appear rigid, limiting responsiveness to evolving threats such as hybrid work vulnerabilities and cloud-based risks.

5.1.3. Theme 3: Industry Practices and Security Measures

Institutional practices show progress but remain uneven. Data indicates 80% of financial institutions adopted MFA, 65% deployed VPNs, 55% conducted awareness training, and only 45% introduced cloud security measures. While these baseline safeguards mark positive steps, they suggest reliance on minimum compliance rather than proactive, capability-driven strategies. Larger Tier-1 banks lead in adopting advanced solutions, whereas smaller institutions lag due to financial and technical barriers. This gap resonates with studies by Okoye et al. (2021) and Adebayo and Oladipo (2022), who note resource constraints as key limitations. The limited uptake of advanced solutions like zero-trust or AI-driven threat detection leaves Nigerian banks behind global peers, exposing systemic vulnerabilities as cybercriminals increasingly exploit weak links across interconnected financial networks.

5.1.4. Theme 4: Integrated Patterns in Cybersecurity Maturity and Compliance

Synthesizing across findings, Nigerian financial institutions show incremental but fragmented progress in managing cybersecurity risks from remote work. Widespread MFA adoption demonstrates growing awareness, yet over half of institutions still operate with outdated or incomplete compliance. From the Cybersecurity Maturity Model perspective, most banks remain at low-to-intermediate stages—structured but reactive, with little advancement toward proactive resilience. Regulatory Compliance Theory highlights a compliance culture shaped by external enforcement rather than innovation, reinforcing a “tick-box” mentality. This culture, coupled with disparities between larger and smaller institutions, sustains systemic vulnerabilities. Without reframing compliance as a foundation for continuous improvement and innovation, Nigerian financial institutions will struggle to advance toward resilient cybersecurity practices capable of countering evolving remote and hybrid work threats.

In summary, the evaluation shows that Nigerian financial institutions have made incremental but uneven progress in addressing cybersecurity challenges associated with remote work. The statistics reveal both the scale of the problem, over 100% increase in incidents between 2019 and 2021 and the partial effectiveness of responses, with most institutions adopting basic measures but lagging in advanced frameworks. Regulatory policies provide structure but lack

flexibility and enforcement, resulting in inconsistent compliance. Collectively, these findings suggest that resilience in Nigeria’s financial sector depends less on isolated technical fixes and more on the integration of regulatory adaptability, institutional maturity, and proactive security practices.

5.2. Re-assessment of Research Questions in Relation to Research Findings and Literature

This section re-examines the four research questions, the goal is to provide a critical interpretation that links the themes directly to each question.

5.2.1. Research Question 1

What are the key cybersecurity risks and systemic vulnerabilities Nigerian financial institutions face in managing remote work environments post-COVID-19?

This study finds that phishing, ransomware, and credential theft remain the most pressing cybersecurity threats for Nigerian financial institutions in the post-COVID-19 remote work era, with incidents peaking between 2020 and 2022. These risks were fueled by unsecured personal devices, weak VPN setups, and delays in applying security patches - echoing the observations of Abubakar et al. (2022) and Ferreira & Cruz-Cunha (2020). While earlier studies link the pandemic to a widening attack surface (Brynjolfsson et al., 2020; Chamola et al., 2020), this research shows that although awareness campaigns and policy updates have helped, they have not fully addressed the problem. Vulnerabilities persist due to the lack of coordinated, system-wide upgrades, confirming the urgent need for proactive, industry-wide cybersecurity action.

5.2.2. Research Question 2

To what extent do existing regulatory policies effectively address cybersecurity risks related to remote work?

The review of Nigeria’s regulatory policies most notably the CBN’s Risk-Based Cybersecurity Framework (2021) and the NDPR (2019) shows that while they provide clear guidelines, they lack flexibility and suffer from uneven enforcement. Larger Tier-1 banks generally achieve higher compliance, but smaller institutions face significant resource and capacity challenges. This reflects Edeh & Eze’s (2021) and Reis et al.’s (2024) findings that Nigerian regulations tend to be more prescriptive than adaptive, slowing responses to new threats like hybrid work vulnerabilities and cloud-related risks. In contrast, global models such as NIST and ISO/IEC 27001 emphasise continuous improvement, monitoring, and intelligence sharing. The absence of similar provisions in Nigerian frameworks reduces their effectiveness, meaning this research question is only partially answered.

5.2.3. Research Question 3

What cybersecurity strategies have institutions implemented, and how effective are they?

Findings show that Nigerian financial institutions have broadly implemented basic cybersecurity measures, with multi-factor authentication (80%) and VPNs (65%) most common. Moderate uptake is seen in cloud security (45%) and staff awareness training (55%). These trends reflect Okoye et al. (2021) and Adebayo and Oladipo (2022), who note reliance on foundational safeguards but slower adoption of advanced tools like AI-driven threat detection or zero-trust architectures. Smaller institutions particularly lag, suggesting a need for targeted capacity building and shared infrastructure. The question is fully answered - gaps remain in advanced strategy adoption.

5.2.4. Research Question 4

What evidence-based recommendations can improve the cybersecurity resilience of financial institutions operating remotely in the post-pandemic era?

Findings highlight five priority actions: increase regulatory agility, build capacity in smaller institutions, establish integrated threat intelligence networks, adopt advanced security tools more rapidly, and maintain strong staff awareness programmes. These align with Carnegie Endowment (2022) and UK Finance (2023) recommendations. Implementing adaptive, globally benchmarked frameworks would help Nigerian banks close resilience gaps and guard against evolving threats. This supports Oyeniyi et al. (2024)'s call for a cybersecurity governance model that is locally relevant yet globally informed. This question is fully answered, providing actionable, evidence-based pathways to strengthen post-pandemic cybersecurity in Nigeria's financial sector.

Re-assessing the research questions highlights both the scale of the problem—a 134% rise in incidents during the pandemic—and the partial progress made, such as the widespread adoption of MFA. Nigerian banks are positioned at intermediate maturity stages, constrained by rigid regulations, uneven enforcement, and resource gaps. Together, these insights converge into a central lesson: cybersecurity resilience in Nigeria's financial sector depends on the integration of adaptive regulation, stronger institutional maturity, and proactive security practices. This convergence lays the foundation for a conceptual framework that summarizes the study's themes and guides future policy and institutional reforms.

5.3. Conceptual Framework for Cybersecurity Resilience in Nigerian Financial Institutions

The conceptual framework synthesizes the study's findings into an integrated model of cybersecurity resilience for Nigerian financial institutions in the remote and hybrid work era. It brings together four major thematic areas—escalation of cyber threats, regulatory gaps, institutional practices, and maturity patterns—and illustrates how they interact to shape resilience outcomes.

At the core of the framework is Cybersecurity Resilience, conceptualized as the sector's ability to anticipate, withstand, and adapt to evolving cyber threats.

- Escalation of Cybersecurity Threats (Theme 1) forms the external pressure, representing the dynamic threat environment that financial institutions must respond to
- Regulatory Frameworks and Compliance Gaps (Theme 2) act as the structural enabler or barrier, shaping institutional behaviour depending on their adaptability and enforcement.
- Institutional Practices and Security Measures (Theme 3) reflect the internal organizational response, highlighting whether banks adopt baseline safeguards only or progress toward advanced solutions.
- Cybersecurity Maturity and Compliance Culture (Theme 4) acts as the integrating pathway, influencing how well institutions embed cybersecurity practices beyond compliance into long-term resilience strategies (represented by the arrow).

The framework proposes that resilience emerges only when adaptive regulation, advanced institutional practices, and maturity-driven cultures interact in a coordinated manner. Weaknesses in any component, whether rigid regulations, resource-constrained practices, or fragmented maturity, creates systemic vulnerabilities that attackers can exploit.

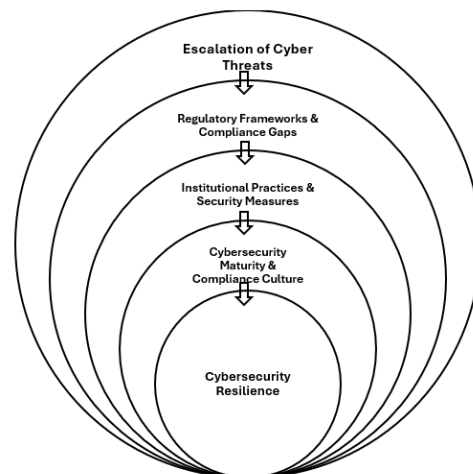


Figure 5.3: Conceptual Framework of Cybersecurity Resilience in Remote Work

6. Conclusion and Recommendations

6.1. Conclusion

This study set out to examine the challenges of maintaining cybersecurity standards for remote work in Nigerian financial institutions in the post-COVID-19 era, with particular emphasis on regulatory frameworks and institutional practices. Guided by four research questions, it sought to identify key risks, evaluate the effectiveness of existing

policies, assess industry strategies, and propose actionable recommendations. Drawing on a systematic review of peer-reviewed literature, regulatory documents, and industry reports, the study applied methodological triangulation to strengthen validity and reliability.

The analysis demonstrates that Nigerian financial institutions remain highly exposed to cyber threats in remote work environments. Phishing attacks, ransomware, and credential theft were the most pervasive threats, surging during the pandemic as institutions rapidly transitioned to remote operations. While awareness campaigns have moderated incident rates, the literature and data indicate that these reductions are temporary, reflecting short-term interventions rather than systemic infrastructural or procedural improvements (Abubakar, Yusuf & Olayemi, 2022; Vanguard, 2024).

Regulatory frameworks such as the CBN Risk-Based Cybersecurity Framework (2021) and the NDPR (2019) provide compliance baselines but are largely prescriptive, rigid, and slow to adapt to evolving threats, particularly when benchmarked against global standards such as NIST and ISO/IEC 27001 (NIST, 2018; ISO, 2013). Tier-1 banks show comparatively better compliance, but smaller institutions often face resource and skills deficits, creating uneven enforcement and sector-wide vulnerability (Edeh & Eze, 2021; Reis et al., 2024).

Institutional practices further illustrate disparities. While basic safeguards like multi-factor authentication, VPNs, and staff training are widespread, advanced strategies—such as AI-driven threat detection and zero-trust architectures, remain largely unadopted (Panchal, 2020; Chin, 2022). This gap between regulatory compliance and proactive resilience leaves many institutions susceptible to sophisticated cyber threats, highlighting the limitations of a reactive, compliance-focused approach.

Critically, this study underscores that vulnerability in Nigerian financial institutions arises not from a single factor but from the interaction of regulatory rigidity, capacity disparities, and partial adoption of advanced security tools.

6.2. Recommendations for Business Application

Building on the study's conclusions, several actionable recommendations are proposed for regulators, policymakers, and financial institutions. These are designed to be practical, scalable, and adaptable to diverse institutional capacities.

1. Enhance regulatory agility and enforcement
Regulators such as the Central Bank of Nigeria (CBN) and the National Information Technology Development Agency (NITDA) should adopt a more adaptive approach to cybersecurity oversight. This entails revising existing frameworks—such as the Risk-Based Cybersecurity Framework—to include provisions for continuous improvement,

periodic reviews, and integration of emerging technologies (NIST, 2018; ISO, 2013). Enforcement mechanisms should be accompanied by capacity-building initiatives to ensure that smaller institutions can meet compliance requirements without disproportionate strain on resources.

2. Strengthen institutional capacity, particularly for smaller institutions
Tier-2 and Tier-3 financial institutions often lack the skilled personnel, technical infrastructure, and financial resources necessary to implement advanced security measures. Industry associations, in collaboration with regulators, should establish shared cybersecurity resource centres, offering threat intelligence, technical support, and training at reduced cost. Public-private partnerships could facilitate the pooling of expertise and investment.
3. Accelerate adoption of advanced cybersecurity frameworks
Institutions should move beyond baseline measures such as VPNs and multi-factor authentication to embrace advanced solutions, including zero-trust architecture, AI-driven threat detection, and endpoint detection and response systems (Panchal, 2020; Chin, 2022). These tools provide proactive, intelligence-led protection capable of adapting to fast-evolving cyber threats.
4. Institutionalize integrated threat intelligence sharing
A sector-wide threat intelligence network—securely managed under regulatory oversight—should be developed to facilitate timely sharing of information on vulnerabilities, incidents, and emerging attack methods. International examples, such as the UK Finance Cyber Defence Alliance (UK Finance, 2023), show that coordinated intelligence-sharing significantly enhances sectoral resilience.
5. Sustain and deepen cybersecurity awareness programmes
While awareness campaigns have reduced incident rates, their scope and depth should be expanded. Programmes should include scenario-based training, simulated phishing exercises, and targeted awareness sessions for high-risk roles such as system administrators and executives.

By adopting these recommendations, Nigerian financial institutions can bridge the gap between compliance and resilience. The key is to foster a security culture that is not only reactive to regulatory demands but also proactive, collaborative, and adaptive in the face of an increasingly complex threat landscape.

6.3. Limitations and Implications for Future Research

This study has several limitations that shape the interpretation of its findings and provide guidance for future research. First, the reliance on secondary data sources, including regulatory documents, industry reports, and peer-reviewed literature, may have introduced variability in quality and completeness

across institutions. Previous studies (Scott, 1990; Edeh & Eze, 2021) note that secondary sources can be limited in capturing nuanced, context-specific organizational practices. This limitation suggests the need for future research incorporating primary data collection, such as interviews, surveys, or focus groups with cybersecurity professionals, which would allow for deeper exploration of operational challenges, institutional decision-making, and perceptions of regulatory compliance.

Second, the absence of granular, institution-level statistics constrained the ability to perform comparative analyses across different categories of financial institutions. This limitation restricts understanding of how factors such as institution size, market share, or geographic distribution influence cybersecurity resilience, a gap highlighted in the literature (Reis et al., 2024; Chin, 2022). Addressing this gap, future studies could employ institution-level datasets or case studies to assess heterogeneity in cybersecurity practices and identify patterns that are not visible at the sector-wide level.

Third, the temporal scope of the study (2018–2024) may not capture the most recent technological and regulatory developments, including advances in AI-based threat detection or updates to the NDPR. Given the fast pace of innovation in cybersecurity, this limitation indicates a need for longitudinal research that tracks evolving threats, institutional responses, and policy adaptations over time, providing evidence on the sustainability and effectiveness of adaptive strategies (Panchal, 2020; NIST, 2018).

Finally, while this study focuses on Nigerian financial institutions, the generalizability of findings to other emerging economies remains uncertain. Comparative research with peer economies, such as Kenya or South Africa, could illuminate best practices and contextual constraints, enhancing both theoretical understanding and practical guidance for regional policy development (Abubakar, Yusuf & Olayemi, 2022; UK Finance, 2023). Additionally, investigating the intersection between emerging technologies—such as generative AI, blockchain, and cloud-based systems—and remote work security could provide forward-looking insights into how institutions can proactively adapt to rapidly evolving cyber threats.

By explicitly linking these limitations to research gaps, future studies can generate richer, context-specific, and longitudinal insights, thereby advancing both theory and practice in cybersecurity governance for emerging economies. These directions will allow scholars and practitioners to develop adaptive, evidence-based strategies that bridge regulatory frameworks, institutional capacities, and technological innovation, addressing the vulnerabilities identified in this study.

References

1. Abubakar, A., Sadiq, M. and Abubakar, I. (2022) 'Cybersecurity threats in remote work environments in Nigerian banks post-COVID-19', *International Journal of Cybersecurity Research*, 8(2), pp. 112–125.
2. Abubakar, A., Yusuf, I. and Olayemi, O. (2022) 'Cybersecurity threats in the Nigerian banking sector during the COVID-19 pandemic', *Journal of Information Security*, 13(2), pp. 45–59.
3. Abubakar, A., Yusuf, S. and Olayemi, A. (2022) 'Post-pandemic cybersecurity threats in Nigeria's financial institutions', *African Journal of Information Systems*, 14(2), pp. 22–35.
4. Adebayo, S. and Oladipo, T. (2022) 'Evaluating staff preparedness for cybersecurity threats in Nigeria's banking sector', *Nigerian Journal of Business and Technology*, 11(1), pp. 44–59.
5. Adedeji, K. (2025) 'Addressing data privacy concerns in artificial intelligence systems: Regulatory mechanisms in Nigeria', SSRN. Available at: <https://ssrn.com/abstract=5222866> (Accessed: 21 July 2025).
6. Akinsanya, A. and Akande, A. (2021) 'Digital transformation and cybersecurity gaps in Nigerian banking during COVID-19', *Journal of Financial Risk Management*, 10(4), pp. 77–89.
7. Akinsanya, A. and Akande, T. (2021) 'Remote working and the surge of cyber threats in developing economies: Evidence from Nigeria', *African Journal of Science, Technology, Innovation and Development*, 13(7), pp. 817–826.
8. Ama, G.A.N., Onwubiko, C.O. and Nwankwo, H.A. (2024) 'Cybersecurity challenge in Nigeria deposit money banks', *Journal of Information Security*, 15(4), pp. 494–523. doi: 10.4236/jis.2024.154028.
9. Ayres, I. and Braithwaite, J. (1992) *Responsive regulation: Transcending the deregulation debate*. Oxford: Oxford University Press.
10. Bada, A. and Nurse, J.R.C. (2020) 'Developing cybersecurity education and awareness programmes for financial institutions in Nigeria', *Computers and Security*, 96, 101921. Available at: <https://doi.org/10.1016/j.cose.2020.101921> (Accessed: 13 July 2025).
11. British Psychological Society (2014) *Code of human research ethics*. Available at: <https://www.bps.org.uk/news-and-policy/bps-code-human-research-ethics-2nd-edition-2014> (Accessed: 25 July 2025).
12. Bryman, A. (2016) *Social research methods*. 5th edn. Oxford: Oxford University Press.
13. Brynjolfsson, E., Horton, J.J., Ozimek, A., Rock, D., Sharma, G. and TuYe, H.-Y. (2020) 'COVID-19 and remote work: An early look at US data', *National Bureau of Economic Research Working Paper*, No. 27344. Available at:

- <https://www.nber.org/papers/w27344> (Accessed: 8 August 2025).
14. Caralli, R.A., Knight, M.E. and Montgomery, A. (2012) Maturity models 101: A primer for applying maturity models to smart grid security, resilience, and interoperability. Pittsburgh, PA: Carnegie Mellon University.
 15. Carnegie Endowment for International Peace (2022) Cybersecurity and financial stability: Threats and policy responses in emerging markets. Washington, DC: Carnegie Endowment.
 16. Carnegie Endowment for International Peace (2022) Cybersecurity and Nigeria's banking sector: A gap analysis. Available at: <https://carnegieendowment.org> (Accessed: 8 August 2025).
 17. Carnegie Endowment for International Peace (2022) Improving cybersecurity regulation in developing countries: Lessons from Nigeria. Available at: <https://carnegieendowment.org> (Accessed: 14 July 2025).
 18. Central Bank of Nigeria (2021) Risk-based cybersecurity framework and guidelines for deposit money banks and PSPs. Abuja: Central Bank of Nigeria.
 19. Chamola, V., Hassija, V., Gupta, V. and Guizani, M. (2020) 'A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G', *IEEE Access*, 8, pp. 90225–90265.
 20. Chin, M. (2022) 'AI-powered cybersecurity: A case study of HSBC and Barclays', *Journal of Financial Innovation and Technology*, 3(2), pp. 88–103.
 21. Deloitte (2020) COVID-19 and cybersecurity in the Nigerian financial sector. Available at: <https://www2.deloitte.com/ng/en.html> (Accessed: 14 July 2025).
 22. Edeh, J. and Eze, S. (2021) 'The challenge of regulatory compliance in Nigerian banks', *Journal of African Financial Regulation*, 5(1), pp. 13–29.
 23. Familoni, B.T. and Shoetan, P.O. (2024) 'Cybersecurity in the financial sector: A comparative analysis of the USA and Nigeria', *Computer Science & IT Research Journal*, 5(4), pp. 850–877.
 24. Ferreira, J. and Cruz-Cunha, M.M. (2020) 'Cybersecurity threats in distributed work environments', *International Journal of Business Continuity and Risk Management*, 10(2), pp. 112–129.
 25. FirstBank (2023) Cybersecurity webinar series. Available at: <https://www.firstbanknigeria.com> (Accessed: 14 July 2025).
 26. Gbenga Femi, A., Adenomon, M.O., Aimufua, G.I.O. and Ibrahim, U. (2025) 'The effects of data privacy regulations on cybersecurity practices in Nigeria and Africa', *Cyberspace Studies*, 9(2), pp. 313–336. doi: 10.22059/jcss.2025.390087.
 27. Ghazvini, A. and Shukur, Z. (2021) 'A review of zero trust architecture in financial services', *Journal of Cybersecurity and Digital Trust*, 9(1), pp. 45–56.
 28. Independent Nigeria (2024) 'CCISONFI intensifies cybersecurity awareness through NoGoFallMaga Campaign'. Available at: <https://www.independent.ng> (Accessed: 14 July 2025).
 29. International Organization for Standardization (2013) ISO/IEC 27001:2013 information security management systems – requirements. Geneva: ISO. Available at: <https://www.iso.org/standard/54534.html> (Accessed: 17 July 2025).
 30. International Telecommunication Union (2021) Global cybersecurity index 2020. Geneva: ITU.
 31. National Information Technology Development Agency (2019) Nigeria data protection regulation (NDPR). Abuja: NITDA.
 32. National Institute of Standards and Technology (2018) Framework for improving critical infrastructure cybersecurity. Version 1.1. Gaithersburg, MD: NIST. Available at: <https://www.nist.gov/cyberframework> (Accessed: 17 July 2025).
 33. Nigerian Communications Commission (2022) Annual report on cyber incidents in Nigeria. Available at: <https://www.ncc.gov.ng> (Accessed: 13 July 2025).
 34. Nigerian Inter-Bank Settlement System (2023) NIBSS annual fraud landscape report 2022. Available at: <https://nibss-plc.com.ng> (Accessed: 25 July 2025).
 35. Nurse, J., Williams, N., Collins, E., Panteli, N., Blythe, J. and Koppelman, B. (2021) 'Remote working pre and post COVID-19: An analysis of new threats and risks to security and privacy', *arXiv preprint*. Available at: <https://arxiv.org/abs/2107.03907> (Accessed: 17 July 2025).
 36. Okoye, C., Chidiebere, E. and Ogunleye, A. (2021) 'Cybersecurity practices among Nigerian banks: A post-pandemic review', *African Journal of Management*, 7(1), pp. 88–104.
 37. Okoye, F., Chidiebere, M. and Ogunleye, J. (2021) 'Cyber hygiene and remote work among Nigerian banks post-COVID', *International Journal of Cyber Policy*, 9(3), pp. 94–107.
 38. Okoye, J., Ayo, C.K. and Arinze, B. (2021) 'Cybersecurity adoption in Nigeria's banking industry: Trends and drivers post-pandemic', *African Journal of Information and Communication*, 28, pp. 49–66.
 39. Oyeniyi, B., Adegbite, O. and Ibitoye, A. (2024) 'Designing a contextual cybersecurity governance

- model for African banking systems', *African Journal of Cyber Governance*, 2(1), pp. 1–15.
40. Oyeniyi, L., Ugochukwu, C. and Mhlongo, N. (2024) 'Developing cybersecurity frameworks for financial institutions: A comprehensive review and best practices', *Computer Science & IT Research Journal*, 5(4), pp. 903–925. doi: 10.51594/csitrj.v5i4.1049.
 41. Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N. and Hoagwood, K. (2015) 'Purposeful sampling for qualitative data collection and analysis in mixed method implementation research', *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), pp. 533–544.
 42. Panchal, A. (2020) 'Adopting zero trust architecture in global banks', *Cybersecurity Quarterly*, 13(3), pp. 25–32.
 43. Reis, J., Esteves, J. and Pinto, M. (2024) 'Regulatory agility in emerging markets: Lessons from financial sector cybersecurity in Africa', *Journal of Regulatory Innovation*, 4(2), pp. 58–74.
 44. Reis, O., Oliha, J.S., Osasona, F. and Obi, O.C. (2024) 'Cybersecurity dynamics in Nigerian banking: Trends and strategies review', *Computer Science & IT Research Journal*, 5(2), pp. 336–364. doi: 10.51594/csitrj.v5i2.761.
 45. Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research methods for business students*. 8th edn. Harlow: Pearson Education.
 46. Scott, J. (1990) *A matter of record: Documentary sources in social research*. Cambridge: Polity Press.
 47. ThisDayLive (2023) 'Cybersecurity on the rise: Unveiling the economic toll on Nigeria'. Available at: <https://www.thisdaylive.com/index.php/2023/10/30/cybersecurity-on-the-rise-unveiling-the-economic-toll-on-nigeria> (Accessed: 17 July 2025).
 48. UK Finance (2023) *Cyber resilience programme annual report 2023*. Available at: <https://www.ukfinance.org.uk> (Accessed: 14 July 2025).
 49. U.S. Department of Energy (2022) *Cybersecurity capability maturity model (C2M2)*. Washington, DC: DOE.
 50. Vanguard (2024) 'Moni Sense: CBN, banks step up cybersecurity education'. Available at: <https://www.vanguardngr.com> (Accessed: 14 July 2025).
 51. Verizon (2022) *Data breach investigations report*. Available at: <https://www.verizon.com/business/resources/report/s/dbir/> (Accessed: 14 July 2025).
 52. World Bank (2021) *Digital economy for Africa initiative: Nigeria country diagnostic*. Washington, DC: World Bank Group.

Machine learning for financial statement analysis and forecast: A case study of Guaranty Trust Holding Company PLC (GTCO PLC.).

Ayomide Sobowale
Near East University, Cyprus
atemmy01@gmail.com

Abstract

This thesis explores the integration of machine learning techniques into financial statement analysis and forecasting within the context of the Nigerian Banking Industry. The primary objectives are to predict the financial health and profitability of companies, enhance data-driven investment decisions, reduce the risk of financial losses for investors, and advance the application of machine learning in finance by substituting traditional analysis methods. The study focuses on GTCO Plc, a prominent entity listed on the Nigerian Stock Exchange. Initially, traditional annual report analysis serves as a foundation for understanding the company's financial performance. Subsequently, a predictive model is developed using four machine learning algorithms: Random Forest, K-Nearest Neighbor, Logistics Regression, and Naïve Bayes. The outcomes of this research contribute to a nuanced understanding of machine learning's efficacy in financial analysis, offering a potential paradigm shift from conventional methodologies. The proposed model not only aids in predicting financial health and profitability but also empowers investors with valuable insights, mitigating financial risks in the dynamic Nigerian Banking Industry. This study thus marks a pivotal step towards fostering a data-driven investment landscape and embracing machine learning applications in financial decision-making.

Keywords

Financial Statement, Annual Report, Machine Learning Algorithms, Machine Learning Models, Random Forest, K-Nearest Neighbor, Logistics Regression, Naïve Bayes

1. Introduction

Every individual has a quest for financial freedom, and our desire of having several investment portfolios is increasing due to the longing for financial liberty. We often admire investors like Warren Buffett, Benjamin Graham, who have become successful in investment with a proper understanding of the market and by even setting investment rules to guide his decisions. Sadly, many investors today do not have adequate financial acumen thereby leading them to making bad decision which leads to losses. In order to limit investment losses, investors should fully understand and be able to analyse a company's annual report.

The analysis of financial statements is used to evaluate the financial performance of an organisation. It entails analysing historical financial reports to identify patterns and predict future performance. This evaluation is conducted through an examination of the organization's balance sheet, cash flows, and income statement. Financial statement analysis is employed by both internal and external stakeholders of a company to effectively run the company,

make prudent investments, and compare it with other companies.

It is used by external stakeholders to have an idea of the business value and the general well-being and performance of an organisation, while internal users utilize it as a management tool. In the past few decades, financial constituents carry out this analysis manually, following the accounting procedures they are acquainted with. However, the evolution of technology has revolutionized the financial sector. Financial statement analysis can now be done with the use of Machine Learning (ML) however; Nigeria financial sector has not fully integrated this technology into some areas in the industry.

Machine learning is a branch of Artificial Intelligence (AI) that allows computers to acquire knowledge and improve their performance by analysing and interpreting data and to discover patterns that are not evident to humans. It has been used for many applications, such as stock market prediction, fraud detection, and financial statement analysis. The process of machine learning for financial statement analysis is done by developing models that learn from available data and apply the acquired knowledge to make decisions and predictions. ML algorithms have become increasingly popular in financial statement analysis and forecasting due to their ability to automate the process and provide more accurate predictions. ML algorithms can identify patterns in data that are not easily detected by humans. This has reduced the amount of time and effort required for financial statement analysis and forecasting. This AI has been of great benefit to so many sectors most especially the Nigerian Stock Exchange Market which is the focus of this research (Obadiaru et al., 2020).

With the largest economy in Africa, Nigeria is home to the primary stock exchange, the Nigerian Stock Exchange (NSE). This exchange has experienced significant growth in the past two decades, and its market capitalization has grown from US\$15 billion in 2007 to US\$3.4 trillion in 2022 (Lawal et al., 2021). With the increasing complexity of financial markets, investors and financial analysts are increasingly relying on machine learning techniques to analyse and forecast stock prices of companies that are listed on the NSE (Wei et al., 2017). Furthermore, machine learning algorithms can be employed to create predictive models for forecasting stock values.

Research Model

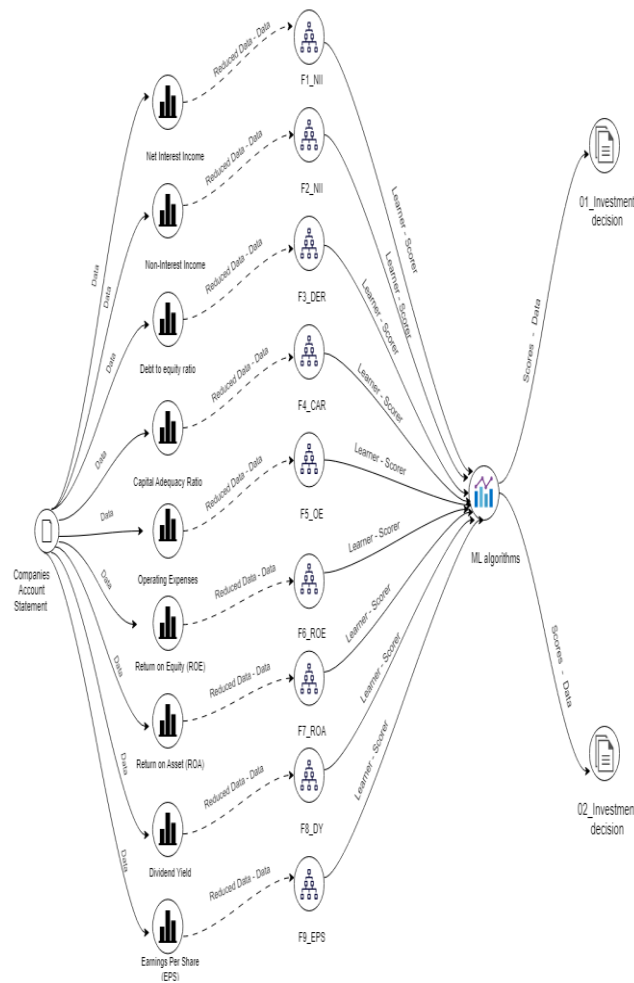


Figure 2.1 Research model

2. Literature Review

In order to establish a background for the theoretical framework of this study, this section contains the evaluation of previous studies on variables to consider in Financial Statement Analysis. Since GTCO Plc is a leading company listed on the Nigeria Stock Exchange Market and investors can only get their portfolios through the NSE, it is judicious to analyse the variables for Financial Statement Analysis in the Nigeria Stock Exchange Market.

Financial statement analysis has long been a cornerstone for evaluating firm performance, guiding investment decisions, and forecasting future trends. Traditionally, this analysis relied on accounting ratios and trend comparisons, but the evolution of financial markets has introduced greater complexity, requiring more advanced approaches such as machine learning (ML). This review synthesizes existing

literature on financial performance variables, analytical methods, and the role of ML in enhancing financial forecasting, with a particular focus on the Nigerian Stock Exchange (NSE) and GTCO Plc.

Independent Variables in Financial Analysis

Macroeconomic indicators are central to financial performance assessment. Studies have shown that variables such as GDP, inflation, and interest rates significantly influence stock returns (Fama & French, 1988; Amadi, 2021). In Nigeria, exchange rate fluctuations (Lawal et al., 2017) and interest rate changes (Odo et al., 2019) have been directly linked to firm profitability and stock price volatility. Industry trends, regulatory reforms, and company size also matter: larger firms tend to demonstrate stronger resilience, while smaller firms are more sensitive to market shocks (Fasan & Gbadebo, 2017). Market capitalization is another determinant, with several Nigerian studies linking higher capitalization to better long-term performance (Olaoye & Eriki, 2019).

These variables act as external determinants that shape how firms perform and how their financial statements are interpreted. However, prior literature reveals a gap in integrating these indicators with predictive models that can capture nonlinear relationships — a challenge where ML techniques are particularly well-suited.

Dependent Variables in Financial Analysis

On the performance side, several dependent variables have been consistently highlighted:

Revenue and Profit Margin: Core indicators of firm health, with declining margins often signaling inefficiency or rising costs (Bustani et al., 2021).

Earnings Per Share (EPS): Widely used as a benchmark for profitability and valuation, with empirical evidence showing strong links between EPS and firm value in Nigeria (Adekunle et al., 2020; Inanga et al., 2016).

Debt-to-Equity Ratio: A key measure of leverage and financial risk. High leverage may reduce profitability, as seen in MTN Nigeria's financial restructuring (Adesina et al., 2017).

Return on Equity (ROE) and Return on Assets (ROA): Standard profitability ratios. Although extensively studied, sector-specific variations remain underexplored, particularly within Nigerian banking.

Dividend Yield and Stock Price: Critical for investor decision-making. Empirical studies show dividend policies can significantly influence stock valuations, though findings vary by sector (Majeed et al., 2017).

These indicators form the foundation of both traditional financial analysis and ML-based predictive models.

Methods of Financial Statement Analysis

Classical approaches include horizontal and vertical analysis, as well as ratio analysis. Horizontal analysis detects year-to-year trends in revenues, costs, and profits (Khalid et al., 2016), and has also been applied in fraud detection (Bhattacharya et al., 2002). Vertical analysis evaluates financial structure relative to a common base, aiding comparisons across firms and sectors (Raza et al., 2015). Ratio analysis provides diagnostic insights into liquidity, solvency, and efficiency, though studies caution that reliance on ratios alone may obscure broader trends (Karimi & Rahmani, 2017).

While effective, these approaches are limited in handling high-dimensional and nonlinear financial data — conditions prevalent in volatile markets such as Nigeria. This has spurred the integration of machine learning.

Sustainable Competitive Advantage Indicators

Beyond financial metrics, literature also highlights the role of sustainable competitive advantage (SCA). Factors such as innovation, brand strength, customer loyalty, and efficient cost structures underpin long-term performance (Purwanto & Purwanto, 2020). In the Nigerian context, financial institutions with diversified revenue streams and strong capital adequacy, such as GTCO Plc, have demonstrated resilience. However, empirical frameworks for quantitatively linking SCA to financial performance remain limited, offering scope for ML-driven models.

Machine Learning in Financial Analysis

The integration of ML has transformed financial statement analysis by offering predictive power and adaptability to complex datasets. A growing body of work has applied ML to stock price prediction, financial distress forecasting, and ratio analysis.

Algorithmic Applications: Logistic Regression, Random Forest, Support Vector Machines (SVM), Neural Networks, and ensemble methods such as stacking have been widely tested (Oyewola et al., 2019). Studies emphasize the superior accuracy of ensemble methods when compared to single-model approaches.

Validation Techniques: Cross-validation methods (K-fold, repeated K-fold) have been shown to improve robustness, particularly in emerging markets like Nigeria (Ogundunmade et al., 2022).

Domain Applications: In the Nigerian Stock Exchange, ML models have been applied to forecast stock returns (Akande & Olubiyi, 2019), assess credit risk (Altman &

Sabato, 2005), and predict financial performance (Al-Nabulsi et al., 2020). These studies consistently show that ML outperforms traditional statistical approaches in predictive accuracy.

Despite these advances, gaps remain in tailoring ML models to firm-specific contexts, such as banking, where variables like interest income, capital adequacy, and operating efficiency carry unique significance. This thesis addresses such gaps by focusing on GTCO Plc, a leading Nigerian bank, to test the utility of ML algorithms against traditional financial analysis.

Synthesis and Relevance to Current Study

The reviewed literature reveals three main insights. First, macroeconomic and firm-specific variables exert significant influence on financial outcomes, yet their interactions are complex and often nonlinear. Second, while traditional methods provide useful descriptive insights, they fall short in predictive accuracy. Third, ML offers promising solutions by capturing hidden patterns and relationships, but its application in Nigerian banking remains underexplored.

This study builds on these gaps by applying ML algorithms specifically Logistic Regression, Random Forest, K-Nearest Neighbours, and Naïve Bayes to GTCO Plc's financial statements over 12 years. By comparing their predictive accuracy with traditional horizontal analysis, the research contributes to both theory and practice. It demonstrates how AI techniques can enhance financial forecasting, support investment decision-making, and ultimately identify sustainable competitive advantages within the Nigerian banking sector.

3. Methodology

The mixed methodology was employed to achieve a full understanding of this investigation. This approach comprises both quantitative and qualitative research methods. This method facilitates a comprehensive analysis of the issue by taking into account both subjective and objective perspectives.

The quantitative method involves collection and analysing of numeric data from one of the companies listed on the Nigerian Stock Exchange market (NSE), GTCO Plc. An in-depth analysis of the company's financial statement of a period of 12 years was carried out using the traditional financial statement analysis method. The qualitative method was used in this thesis majorly for the literature review.

Previous studies on financial statement analysis and Machine Learning were thoroughly examined, finding gaps in the literature and posing questions for further studies. This method was also used to critically examine previous research on the independent and dependent variables of this subject. It also assessed what researchers have studied in respect to identifying the sustainable competitive advantage

of companies listed on the NSE thereby addressing one of the objectives of this study. Finally, this method allowed me to evaluate the machine learning techniques and models performed in previous studies.

3.1. Sampling Method

A non-probability sampling method will be used to sample the thesis topic. The study utilizes purposive probability sampling technique to collect data from the financial statements of Guarantee Trust Bank Limited (GTCO Plc.), which is one of the firms listed on the Nigerian stock Exchange Market (NSE). This sampling technique was chosen because the data does not involve a random selection

3.2. Source of Data

Data will be gathered from GTCO Plc, one of the prominent banks in the Nigeria stock exchange market, using purposive sampling, a non-probability sampling method. In this instance, GTCO was chosen as a data source because it has the largest financial statement of banks listed on the stock exchange market. GTCO's annual reports encompasses of the metrics and ratios needed to analyze the data. The data accessible from the company enable us determine the sustainability of the firm.

3.3. Data Collection

To collect data for this study, a secondary data collection method was used. The data was collected from Guarantee Trust Bank Limited (GTCO Plc.) financial statement over the period of 12 years. Using a 12-year sample size of data from the list of GTCO annual report, this data would be used to develop a model and create a trend that the machine learning can follow in order to forecast the performance of the company.

3.4. Data Analysis

The Annual Reports of Guaranty Trust Bank Limited from 2009 to 2022 was analyzed technically using the horizontal financial statement analysis method and some ratios like Return to Equity (ROE), Return on Asset (ROA), Net Interest Income, Non-Interest Income, Debt to equity ratio, Capital Adequacy Ratio, Operating Expenses, Dividend Yield, etc.

3.4.1. Horizontal financial statement analysis

Horizontal analysis is used to analyze historical data, such as ratios and line items, from several accounting periods. Horizontal analysis involves making comparisons using either absolute or percentage comparisons. Absolute comparisons include comparing data directly, whereas comparing percentages describe the numbers as a percentage of the total for the previous and succeeding periods, with 100% being the baseline value. This is sometimes referred to as the base-year analysis.

Return on Equity (ROE): Return on Equity (ROE) is a financial metric used to Return on Equity (ROE) is a financial metric used to assess a company's profitability. It is calculated by dividing the company's net profit by its shareholder equity, indicating how much profit is generated for each dollar of equity. ROE is a measure of a firm's ability to create returns from the investments made by its investors. Based on the analysis conducted in this study, a higher return on equity (ROE) signifies that the bank is earning greater profits from the capital invested by shareholders.

$$ROE = \frac{Net\ Income}{shareholder\ Equity}$$

Return on Assets (ROA): Return on assets (ROA) is a financial metric that quantifies the profitability of a firm in relation to its overall assets. The metric quantifies the degree to which a corporation is effectively utilizing its assets to produce profits. A substantial proportion of this ratio signifies the company's management's effectiveness in utilizing its balance sheet to produce profits. Return on assets (ROA) is computed as:

$$ROA = \frac{Net\ Income}{Total\ Assets}$$

Net Interest Income: Net interest income refers to the net amount of money a financial organization gets from the interest it receives on its assets, such as loans and securities, less the interest it pays on its obligations, such as deposits and borrowings. It is a key component of a financial institution's earnings. A bank or institution with a high NIL indicates that it has a strong core business

$$ROE = Interest\ Income - Interest\ Expenses$$

Non-Interest Income: Non-interest income refers to the revenue that a financial institution earns from activities other than lending and borrowing money. It consists of commissions and fees received for services including insurance, brokerage, and investment management. If a bank has a diversified revenue stream, it helps reduce the bank's reliance on interest income.

$$ROE = Total\ Revenue - Net\ Interest\ Income$$

Debt to Equity Ratio: The debt-to-equity ratio expresses how much of a company's entire debt is divided by its total equity. The calculation involves dividing the sum of all liabilities by the amount of shareholder equity. The ratio is a gauge of how much of the company's funding is provided by debt as opposed to equity. The optimal debt-to-equity ratio (D/E Ratio) differs among industries, however it is generally recommended to not exceed 2.0. Industries that require a significant amount of capital, such as mining and manufacturing, typically exhibit a higher ratio. However, for

our data, the ideal D/E ratio for a bank should be from 1.5 – 2.0

$$\text{Debt to equity ratio} = \frac{\text{Total Liabilities}}{\text{Shareholder Equity}}$$

Capital Adequacy Ratio: The capital adequacy ratio is a metric that assesses a bank's capacity to withstand losses that may result from its operational undertakings. The calculation involves dividing a bank's capital, which includes equity and retained earnings, by its risk-weighted assets. A greater capital adequacy ratio signifies that the bank possesses a robust financial position and is capable of withstanding bad economic situations. The Central Bank of Nigeria has established that all banks and banking groups with international authorization, as well as those categorized as Domestic Systemically Important Banks by the CBN, must have a minimum Pillar 1 regulatory Capital Adequacy Ratio of 15%. All other banks would be required to maintain a minimum Capital Adequacy Ratio (CAR) of 10%.

$$\text{Capital Adequacy Ratio} = \frac{\text{Capital}}{\text{Risk Weighted Assets}}$$

Operating Expenses: Operating expenses refer to the expenditures incurred by a corporation for its routine business activities, including employee wages, rental fees, utility bills, and advertising costs. They exclude expenses associated with the manufacturing of products or services, such as the expenses for acquiring raw materials or labor.

$$\text{Operating Expenses} = \frac{\text{Operating Expenses}}{\text{Total Revenue}}$$

Dividend Yield: Dividend yield is a financial metric that quantifies the proportion of cash dividends distributed to shareholders in relation to the market value of a company's stock. The chairman of an organization will usually declare the year's dividend yield in the annual general meeting and it would be stated in the annual report. However, it can be ascertained by dividing the yearly dividend per share by the stock price per share.

$$\text{Dividend yield} = \frac{\text{Dividend per share}}{\text{Share Price}}$$

Earnings per Share (EPS): Earnings per share (EPS) is a financial metric used to gauge a company's profitability by calculating the net income made per share of its outstanding common stock. A higher EPS is a positive signal to investors, although decisions from this ratio is relative to each investor depending what the returns they expect on their investment. The earnings per share is usually stated in the firm's annual report, however, it can be calculated as follows;

EPS

$$= \frac{\text{Net Income} - \text{Dividend Payment}}{\text{Weighted Average of common share outstanding}}$$

Stock Price: The stock price is the present market valuation of an individual share of a company's stock. The stock market's dynamics are determined by the interplay of supply and demand and reflects investors' perception of a company's future earnings potential, risk, and growth prospects.

3.4.2. Machine learning model and algorithm

The four major machine learning model that would be used for this thesis are Random Forest, Logistics Regression, KNN, and Naïve Bayes.

Random Forest: Random Forest is a versatile machine learning technique employed for tasks such as classification, regression, and feature selection. It is a technique in ensemble learning where many decision trees are created during training and the final prediction is determined by either selecting the most common class among the trees or calculating the average prediction of each tree. The Random Forest algorithm operates by generating decision trees on randomly sampled subsets of the data and variables. At each split, the algorithm selects the best variable among a random subset of variables to create a split, which leads to a more robust and accurate model that is less prone to over fitting. Due to its capacity to process complex data, manage noisy data, and provide profound clarifications on the importance of various financial ratios and metrics in forecasting financial performance indicators, it is extensively employed in financial statement analysis.

Logistic Regression: Logistic Regression is a statistical technique employed to solve binary classification problems. It quantifies the likelihood of a binary result (such as default or non-default) based on one or more independent variables (such as financial ratios or metrics). Logistic Regression operates by calculating the coefficients of the independent variables that optimize the likelihood of the observed data. The coefficients quantify the impact of each independent variable on the outcome and can be utilized to forecast the likelihood of the outcome based on the values of the independent variables. Logistic Regression is widely used in financial statement analysis for its simplicity, interpretability, and ability to handle both continuous and categorical data. It is often used to predict financial distress, bankruptcy, and other binary outcomes in corporate finance.

Naïve Bayes: Naïve Bayes is a probabilistic technique in machine learning that is specifically designed for solving classification problems. It relies on Bayes' theorem, which quantifies the likelihood of a hypothesis given the available evidence. Naïve Bayes operates by estimating the conditional probability of each class given the input

variables, assuming that the variables are independent of each other. This assumption is often unrealistic, hence the name "naive", but it still works well in practice for many classification problems, especially in the presence of high-dimensional and sparse data. Naive Bayes is widely used in financial statement analysis for its simplicity, speed, and accuracy in predicting financial distress, bankruptcy, and other classification problems in corporate finance.

K-Nearest Neighbor: The K-Nearest Neighbours (kNN) algorithm is a straightforward and adaptable supervised machine learning technique employed for classification and regression tasks. The algorithm is based on the idea of proximity, whereby an input data point is categorized by the predominant class of its k closest neighbors in the feature space. The selection of k influences the quantity of neighbors that are taken into account throughout the classification process. The algorithm's simplicity facilitates comprehension and implementation, but, its performance can be influenced by the selection of the distance measure and the value assigned to k. Moreover, kNN is a non-parametric approach, implying that it does not rely on assumptions regarding the underlying data distribution. This characteristic makes it well-suited for datasets that exhibit diversity.

4. Data Analysis

4.1. Interpretation of Data Analysis

The data we collected for this study encompasses of GTCO's annual report from the year 2009 to 2022, excluding 2010 and 2015. The horizontal financial statement analysis method was used to traditionally analyze the bank's financial statement and the results derived from this analysis are represented in the table below;

When analyzing financial statement most especially for investment purposes, there are certain metrics analysts look out for to determine the best investment. Investors that are not well informed of financial analysis have been educated to only look out for metrics or ratios like Earnings per Share, Net Income, Return on Assets (ROA), Return on Equity (ROE), and Dividend Yield etc. As important as these ratios are, they are not only the factors that determines the bank's general state of health, which will aid in their ability to make wiser investment choices.

In determining the best investment decision, each of these metrics and ratios are evaluated in accordance with the general guidelines of the industry. Results gotten for each data are explicitly explain below.

Table (1): Traditional analysis of GTCO Annual Report 2009-2022.

Year	Net Interest Income	Non-Interest Income	Debt to Equity Ratio	Capital Adequacy Ratio	Operating Expenses	Return on Equity (ROE)	Return on Asset (ROA)	Dividend Yield	Earnings per share (EPS)	Stock Price
2009	73,468,110	40,808,407	4.4	25.99%	82%	12%	2.34%	1.5%	128 Kobo	50 kobo
2011	95,522,806	48,378,608	1.0	23.03%	63%	22%	3.34%	8.97%	164 Kobo	₦12.25
2012	123,098,741	44,199,867	4.6	24%	32%	30%	5.26%	9.87%	₦2.90	₦15.70
2013	127,857,215	49,285,953	4.8	22.27%	44.47%	25.95%	4.49%	7.97%	₦2.91	₦21.31
2014	128,698,830	69,022,777	4.8	21.40%	53.08%	25.28%	4.39%	6.80%	₦3.17	₦22.03
2016	171,027,957	139,337,380	4.5	19.79%	40.54%	26.59%	4.85%	8.16%	₦4.31	₦24.61
2017	217,649,619	75,794,761	3.8	25.50%	45.10%	27.60%	5.70%	10.21%	₦5.48	₦26.44
2018	188,441,907	98,521,378	4.3	23.4%	46.26%	32.61%	6.15%	7.61%	₦5.67	₦36.12
2019	189,318,029	109,749,294	4.1	20.66%	42.42%	28.90%	5.65%	9.30%	₦5.95	₦30.08
2020	208,932,503	123,485,182	4.8	19.55%	41.23%	25.36%	4.38%	13.16%	₦6.05	₦22.79
2021	0	8,829,354	4.41	23.83%	6.2%	6.02%	5.76%	11%	28 Kobo	₦27.83
2022	0	2,092,332	1.89	24.08%	8.17%	64%	54%	13%	₦3.01	₦24.38

4.1.1 Net interest income

In 2009, the net interest income for GTCO was ₦73,468,110 out of a total interest income interest of ₦110,889,700, this indicate a high interest income which means The bank is generating more revenue from its interest-earning assets than it is paying out on its interest-bearing liabilities. of ₦37,421,590 which is indicative of a wise investment. The financial statement in 2011 reported a net income interest of ₦95,522,806 reflecting a 30% increase from the previous year. The net interest rate increased the following year 2012, with 28.8%, the annual report recorded ₦123,098,741 as total Net interest income.

However, the net interest income didn't reflect much increase in 2013 and 2014, it grew slightly from ₦127,857,215 to ₦128,698,830 reflecting an increase of 3.9% and 0.66% respectively. This is due to the fact that the bank had little revenue on interest from loans and securities and they had little expenses to pay on their interest-bearing liabilities like deposits and borrowings. The net interest income picked up in 2016 by 32.89% with the total net income interest being ₦171,027,957. In 2017, the net interest income grew from ₦171,027,957 to ₦217,649,619 representing a 27.3% increase and the highest metric in our data set. The results gotten from 2018 to 2022 are represented below in the bar chart. There was a 0.46% increase within 2018 to 2019 and a 10.36% from 2019 to 2020. However, the financial statement of the bank in 2021 and 2022 did not reflect any net income interest.

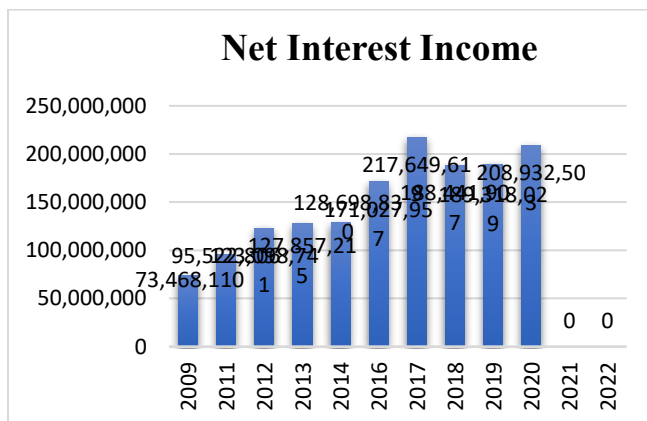


Figure 4.1: GTCO net income interest 2009 -2022

4.1.2 Non-interest income

As established earlier, Non-interest income are the revenue a financial institution earns from other activities than borrowing and lending money. These other source or revenue could be generated from fees like commission income, trade gains, insurance income, etc. The Non-Income Interest derived for this study are represented in the figure below.

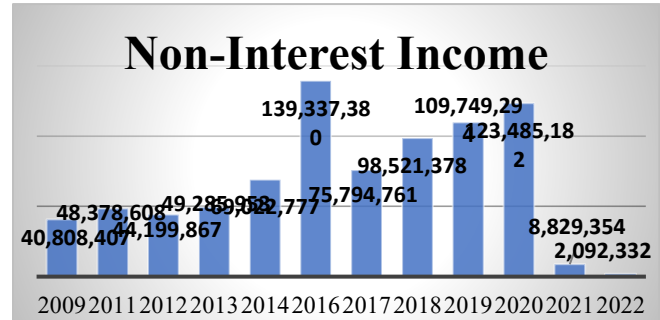


Figure 4.2: GTCO non-interest income 2009-2022

The bank's non-interest income amounted to ₦40,808,407 in 2009, and in 2011, their annual statement reported ₦48,378,608 as Non-interest income reflecting 18.6% increase. However, this metric reduced by 3.7% in 2012 and picked up again by 11.5% and 40% in 2013 and 2014 respectively. In 2016, the non-interest income increased from ₦69,022,777 in the previous year (2014) to ₦139,337,380 reflecting more than a 100% increase. This increase might not reveal the accurate picture of the data due to the year gap of 2015 in our data set, this year gap is one of the limitations faced in this study which will be discussed later on in this study.

This metric took a dive in 2017 with a decrease of 45.6% which means that the bank could not generate more alternative income like the previous year. In 2018, the bank generated ₦95,521378 revenues and ₦109,749,294 in 2019. The non-interest income grew by 12.5% generating an income of ₦123,485,182 in 2020. GTCO annual report in 2021 and 2022 reported a very low non income statement income of ₦8,829,354 and ₦2,092,332.

4.1.3 Debt to equity ratio

A company's debt-to-equity ratio quantifies the extent to which it is using its own resources to finance its operations instead of borrowing from others. Analysing a debt to equity ratio for investment purposes helps us identify companies that are highly leveraged and higher risk. The ideal D/E ratio for a bank should be from 1.5 – 2.0, any ratio higher that this means that the company is a highly leveraged.

The graph below represents the D/E Ratio of GTCO from 2009 - 2022. For this analysis the company started off with a high risk ratio of 4.4. Considering the banking industry ideal rate of 1.50 - 2.0, the 2009 ratio was a bad indication on the banks annual report that year. Moreover, since financial institutions borrow money to lend money, they tend to have higher D/E Ratio, but according to the Central Bank of Nigeria (CBN) the ideal D/E Ratio for financial sector should be 4%. This ratio measures the firm's financial leverage, i.e. for every amount of equity it shareholders own, the bank owes ₦4.4.

In 2011, the firm’s annual report reflected a low D/E Ratio of 1% i.e., a debt of ₦1 for every amount of equity the bank shareholders own. This ratio is within the scope of the industry’s acceptable rate, therefore the annual report for the year was not highly leveraged. From 2012 – 2021 the company’s D/E Ratio was highly leveraged, the ratio fluctuated within 4.1 to 4.8 but with an exception in 2017. In 2017, the rate was 3.8 which is above the ideal ratio for financial institutions but below the set rate from the Central Bank of Nigeria (CBN). In this case, I decided to follow the general acceptable industry ratio to make a decision of D/E Ratio being high risk. However, 2022 D/E Ratio was 1.89 which indicated a low risk and good for making investment decision.

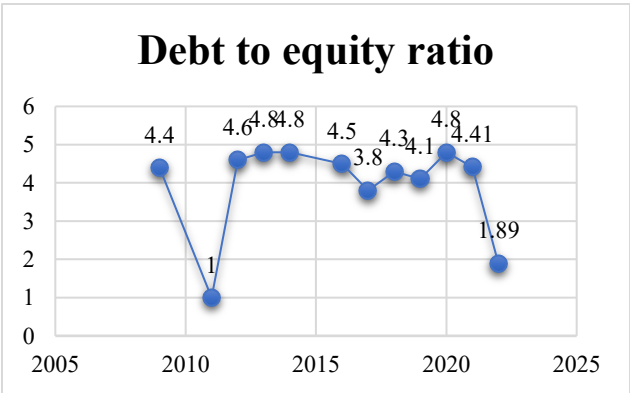


Figure 3.3: GTCO debt to equity ratio 2009-2022

4.1.4 Capital adequacy ratio

A capital adequacy ratio of 25.99% in 2019 denotes an excellent financial position of the bank. This denotes that GTCO has the ability to absorb losses with their capital if any arises. According to the Central bank of Nigeria (CBN), all Nigerian banks with an international authorisation in which GTCO has, is expected to have at least 15% capital adequacy ratio. The bank had more than the required rate in all the 12 years’ data set for this study, which could be attractive to the investors because they are sure their investment would be safe to some extent if any crisis or liquidation arises. The figure below gives a graphical representation of the capital adequacy ratio for 12 years.

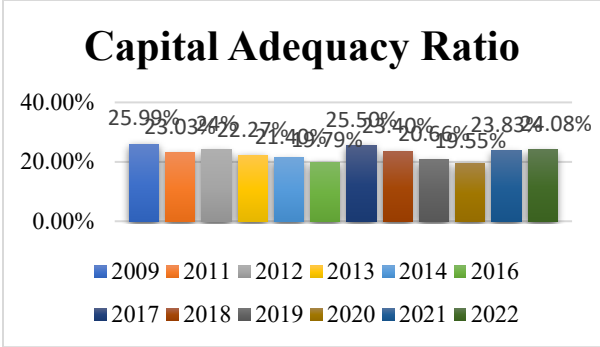


Figure 4.4: GTCO capital adequacy ratio 2009-2022

4.1.5 Operating expenses

Operating expenses are the financial metric that evaluates the ongoing operational activities of an organisation. A high operating expenses means that the company may have difficulty in generating profits. This can influence the decision of the investors, they will consider the negative effect this will have on the firm’s profitability and also its cash flow. A high operating expense can have a negative impact on the firm’s cash flow, it would be difficult for them to pay dividends to shareholders. GTCO’s operating expenses from 2009-2022 is represented in the bar chart below:

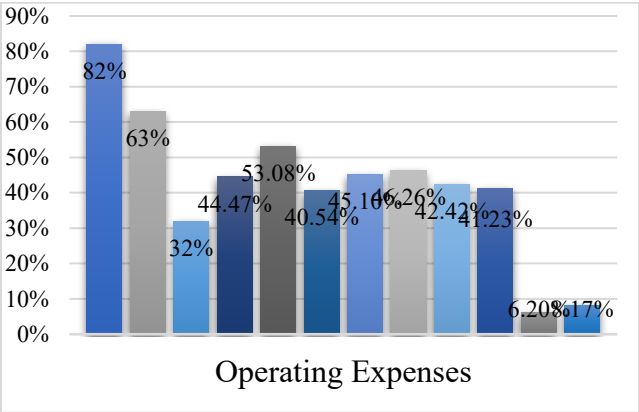


Figure 4.5: GTCO operating expenses 2009-2022

4.1.6 Return on Equity

Return on Equity is a financial metric that calculates the profit made per unit of shareholder stock in order to quantify the profitability of a business. ROE is a measure of a firm's capacity to create returns from the investments made by its stakeholders in an efficient manner. This metric is one of the most important indicators investors look out form while making investment decisions. A high return on equity (ROE) signifies that the company is efficiently producing greater profits from the capital invested by its investors. Although, ROE varies from sector to sector but a good return on equity rate should range from 15% - 20% and above.

In 2009, the Return on Equity for this year was 12% which does not indicate a good investment, it could mean that the company did not use the shareholders’ investment efficiently.

ROE from 2011 to 2022 were above the 15% which will be very attractive for potential stakeholders.

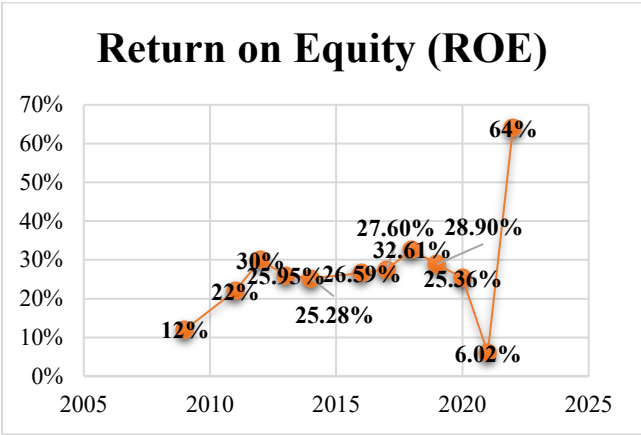


Figure 4.6: GTCO Return on Equity 2009-2022

4.1.7 Return on Asset

Return on Asset is an assessment of how well a business uses its earnings to make more profits. A high percentage of this ratio indicates how efficient the company’s management handling its balance sheet to make profit. According to Investopedia, a ROA of more than 5% or more is considered good, while more than 20% is considered excellent. The graph below presents the Return on Equity of GTCO over a 12-year financial period.

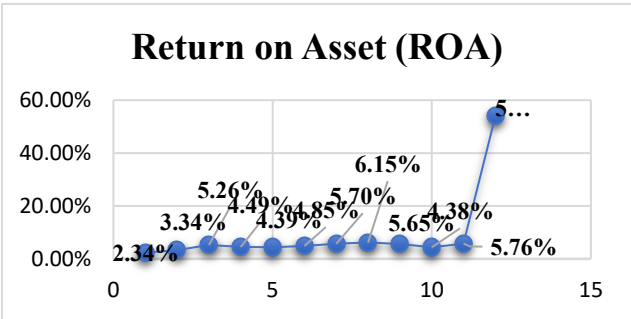


Figure 4.7: GTCO Return on Asset 2009-2022

4.1.8 Dividend yield

Dividend yield is a financial ratio that quantifies how much an investor would earn from an investment based only upon the dividend payments from the organization. This metric is one the most important things investor looks out for when making an investment decision. Investors who does not have adequate financial knowledge are taught to look out for this metric to make their decision when buying a stock. The chairman of an organization will usually declare the year’s dividend yield in the annual general meeting and it would be stated in the annual report.

Some investors evaluate stocks based on only the dividend yield which is not advisable because a dividend can reduce or even be eliminate if the stock price declines. This proves the

importance of this study. Having a trained machine learning program to accurate predict the best investment while carefully evaluating each metrics will eliminate bad investments decisions. The linear graph below shows the dividend yield of GTCO from 2009 – 2022.

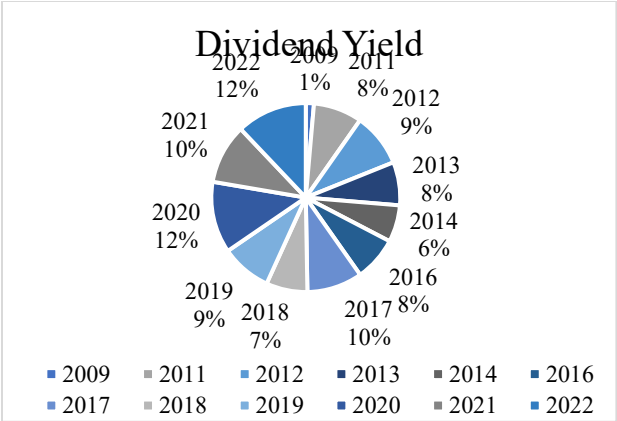


Figure 4.8: GTCO dividend yield 2009-2022

4.1.9 Earnings per Share

The Earnings per Share is a vital indicator for stockholders as it assesses the bank’s profitability per share. A higher Earnings Per Share (EPS) signifies that the bank is making more money for every outstanding share of stock. For example, in 2009 on the graph below, the EPS was 128 Kobo, i.e. upon every outstanding share, 128 Kobo profit is earned. Although this matric is not enough to determine a firm’s profitability, Investors are able to make well-informed decisions with the metric when compared with the EPS of other companies in the same industry.

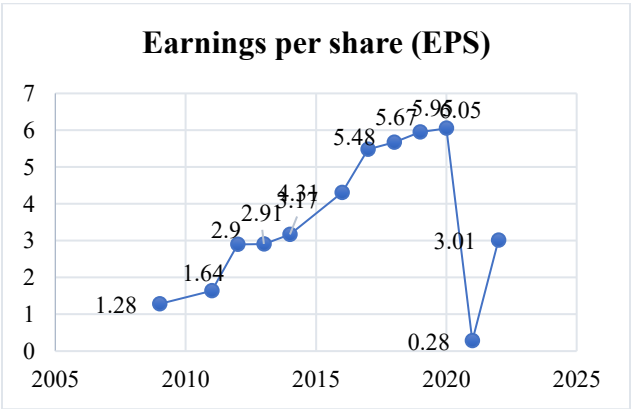


Figure 4.9: GTCO Earnings Per Share 2009-2022

4.1.10 Stock price

The stock price on the Nigerian Stock Exchange (NSE) represents the current market value of the bank's shares. A higher stock price shows that the market has confidence in the bank’s future prospects and financial performance. GTCO’s stock price experienced a steady increase from 2009-2018 and fluctuated from 2019-2022 which could be due to investors sentiment and market conditions.

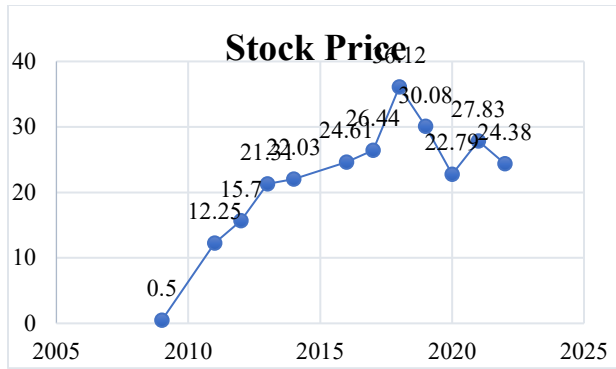


Figure 4.10: GTCO stock price in Naira

5. Conclusion and Discussion

5.1 Experiment and Result

This chapter presents the result of the experiment performed for this study. The primary objective of this study is to create a machine-learning model that will transform financial statement analysis from its traditional methods and to predict a company's financial health or profitability to enhance well informed investment decision. To create this model, this study used the supervised machine learning, which means that our model will have a baseline understanding of what the correct output should be.

In order to get a supervised dataset for our machine learning, it was necessary to first analyse the financial statement traditionally to train the machine to predict on its own. To achieve this, the table below represents my response to each of the metrics analysed in accordance to the financial industry overall acceptable rate. The last column title "Decision" is to give the machine learning algorithms the final decision on a year's financial report to either invest or not to. The tick sign (✓) indicates a positive sign and to invest while the sign (✗) means negative and not to invest. Training the machine with this data has to be done in a language the machine understands i.e. Binary language, so this data was inputted to the machine learning algorithms used in 1 and 0. 1 represents the positive (✓) sign while 0 for the negative (✗) sign.

Table (2): Decisions for traditional financial statement analysis

Year	Net Interest Income	Non-Interest Income	Debt to Equity Ratio	Capital Adequacy Ratio	Operating Expenses	Return on Equity (ROE)	Return on Asset (ROA)	Dividend Yield	Earnings per share (EPS)	Decision
2009	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
2011	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
2012	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓
2013	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓
2014	✓	✓	✗	✓	✓	✓	✗	✗	✓	✓
2016	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓
2017	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
2018	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
2019	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
2020	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓
2021	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗
2022	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗

The image below is the representation of the prediction of this study result from the four ML algorithms used. Each of this algorithms were running on every instances on our data i.e. trained to predict the result of all the financial metrics used for this study. On the extreme left of this image is the ML algorithms and the accuracy of their readings. As mentioned earlier, binary numbers were used to the results ascertained from the traditional financial analysis performed. The binary number “1” denote a positive result of “To Invest” and the binary number “0” symbolizes a negative result of “Not to Invest”. The first four columns represent the accuracy of reading of Random Forest, kNN, Logistic Regression, and Naïve Bayes. In 2009, the final decision made on that year’s analysis was not to invest i.e. 0 in binary. Random forest predicts to invest i.e. “1” with a degree of accuracy of 0.57, K nearest neighbor predicted to invest with 0.80 degree of accuracy. However, Logistic Regression predicted not to invest with the degree of accuracy of 0.58, while Naïve Bayes also predicted not to invest with 0.86 degree of accuracy. This algorithm both got the prediction correctly but the Naïve Bayes has the highest rate of prediction accuracy. For 2011, the decision on investment was to invest, all the algorithms predicted correctly for this year. This prediction was carried out through all instances in the dataset of GTCO, the algorithms predicted the result accurately with few confusion matrixes. The result gotten from this experiment proofs that the machine learning model was learning well and it supports the hypothesis that there is a positive relationship between the adoption of machine learning technique level of accuracy of financial statement analysis and forecast in the Nigerian stock exchange market.

Additionally, there are five machine learning metric to examine the level prediction accuracy of our algorithms.

1. **AUC (Area under the ROC Curve):** AUC is a common indicator used to estimate the discriminatory ability of a classification model. It measures the area under the Receiver Operating Characteristic (ROC) curve. A perfect AUC score is 1.000, which means the model has a perfect ability to differentiate between the classes. An AUC of 1.000 indicates that the model's predictions are completely accurate. As show in the image below, the AUC outputs for all model in this study were 1, which means all our ML model were able to distinguish between the classes and predict accurately.
2. **CA (Classification Accuracy):** Classification accuracy quantifies the percentage of occurrences properly classified out of all instances. In this case, kNN has a classification accuracy of 0.750 which means it correctly classified approximately 75% of the instances, while Random Forest correctly classified all instances by approximately 91.7%.
3. **F1 Score:** The F1 score is the harmonic mean of precision and recall. It strikes a balance between recall (the number of actual positive cases that were accurately predicted) and accuracy (the number of predicted positive instances that were truly positive). With an F1

score of 0.911, Random Forest exhibits a strong balance between recall and precision.

4. **Precision:** The ratio of accurately predicted positive instances to all expected positive instances is known as precision. It assesses how well the model avoids generating false positive results. A precision of 0.925 means that 92.5% of instances predicted as positive were actually positive.
5. **Recall:** Recall is a metric that quantifies the proportion of accurately predicted positive cases to all actual positive instances. It is sometimes referred to as sensitivity or true positive rate. With a recall of 0.917, the model was able to accurately identify 91.7 percent of the real positive events.

The figure below represents the rate of accuracy for each metrics, each bar on the chart shows the how accurately the ML algorithms predicted the result. The major limitation in the study is limited availability of data, due to this we could not get a wider range of result for his experiment, however it does not depute purpose of the study. The chart below supports the hypothesis that the accuracy of financial statement analysis and forecasting using machine learning algorithms is influenced by the type of machine learning algorithm used.

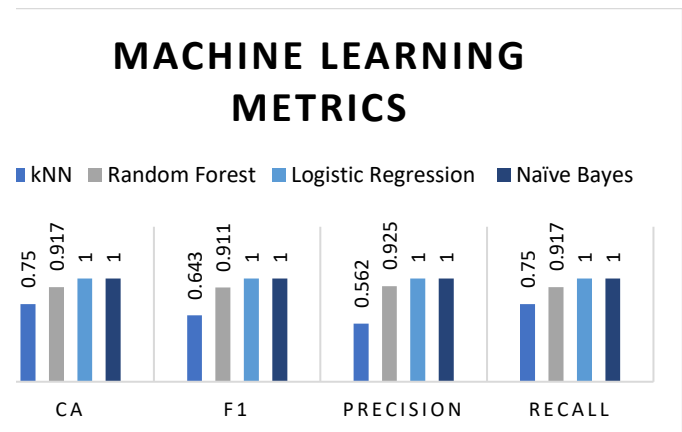


Figure 4.11: ML metrics rate of accuracy

Model: 4 models

- Random Forest
- kNN
- Logistic Regression
- Naive Bayes

Showing probabilities for all classes that appear in the data

Data & Predictions

	Random Forest	kNN	Logistic Regression	Naive Bayes	Decision	Net Interest Income	Non-Interest Income	Debt to equity ratio	Capital Adequacy Ratio	Operating Expenses	Return on Equity (ROE)	Return on Asset (ROA)	Dividend Yield	Earnings per share (EPS)
1	0.43 : 0.57 → 1	0.20 : 0.80 → 1	0.58 : 0.42 → 0	0.86 : 0.14 → 0	0	1	1	0	1	0	0	0	0	1
2	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.13 : 0.87 → 1	0.06 : 0.94 → 1	1	1	1	1	1	1	1	0	1	1
3	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.12 : 0.88 → 1	0.06 : 0.94 → 1	1	1	1	0	1	1	1	1	1	0
4	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.15 : 0.85 → 1	0.08 : 0.92 → 1	1	1	1	0	1	1	1	0	1	0
5	0.03 : 0.97 → 1	0.00 : 1.00 → 1	0.15 : 0.85 → 1	0.06 : 0.94 → 1	1	1	1	0	1	1	1	0	0	1
6	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.09 : 0.91 → 1	0.02 : 0.98 → 1	1	1	1	0	1	1	1	0	1	1
7	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.07 : 0.93 → 1	0.02 : 0.98 → 1	1	1	1	0	1	1	1	1	1	1
8	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.07 : 0.93 → 1	0.02 : 0.98 → 1	1	1	1	0	1	1	1	1	1	1
9	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.07 : 0.93 → 1	0.02 : 0.98 → 1	1	1	1	0	1	1	1	1	1	1
10	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.09 : 0.91 → 1	0.02 : 0.98 → 1	1	1	1	0	1	1	1	0	1	1
11	0.90 : 0.10 → 0	0.40 : 0.60 → 1	0.79 : 0.21 → 0	1.00 : 0.00 → 0	0	0	0	0	1	1	0	1	1	0
12	0.90 : 0.10 → 0	0.40 : 0.60 → 1	0.68 : 0.32 → 0	0.98 : 0.02 → 0	0	0	0	1	1	1	1	0	1	0

Scores

Figure 4.12 ML algorithms prediction of result

Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	1.000	0.750	0.643	0.562	0.750
Random Forest	1.000	0.917	0.911	0.925	0.917
Logistic Regression	1.000	1.000	1.000	1.000	1.000
Naive Bayes	1.000	1.000	1.000	1.000	1.000

Confusion Matrix

Tue Aug 08 23, 12:47:51

Confusion matrix for Naive Bayes (showing number of instances)

		Predicted		
		0	1	Σ
Actual	0	3	0	3
	1	0	9	9
	Σ	3	9	12

Confusion Matrix

Tue Aug 08 23, 12:48:18

Confusion matrix for Random Forest (showing number of instances)

		Predicted		
		0	1	Σ
Actual	0	2	1	3
	1	0	9	9
	Σ	2	10	12

Confusion Matrix

Tue Aug 08 23, 12:48:22

Confusion matrix for Random Forest (showing number of instances)

		Predicted		
		0	1	Σ
Actual	0	2	1	3
	1	0	9	9
	Σ	2	10	12

Confusion Matrix

Tue Aug 08 23, 12:48:27

Confusion matrix for kNN (showing number of instances)

		Predicted		
		0	1	Σ
Actual	0	0	3	3
	1	0	9	9
	Σ		12	12

Confusion matrix for Logistic Regression (showing number of instances)

		Predicted		Σ
		0	1	
Actual	0	3	0	3
	1	0	9	9
Σ		3	9	12

Figure 4.14 Confusion matrix of ML algorithms

Confusion Matrix

The result above represent the confusion matrix which examines the rate of false-positive (the number of samples that are false but the ML model categorised as positive) and false-negative (the number of samples that are false but the ML model categorised as negative) (Kordi Ghasrodashti, Helfrroush & Danyali, 2017) . For example Naïve Bayes did not have a condusion matrix, the model predicted all 12 instances accirately – 3 “Not to invest” prediction and 9 “To invest” prediction. This result corresponds with financial anlysis carried our traditionally.

Random Forest had one false-positive confusion matrix, the model predicted an instance which was “Not invest” as “To invest”. kNN had three false-positive confusion matrix, it predicted 3 instances out of 12 inaccurately which indicates that it is not really a good classification model. However, Logistic Rgression also had a perfect prediction.

It classified and predicted all instances accurately. In this instance, Naïve Bayes and Logistic Regression were the models with the perfect classification, the two other models has very little false-positives and this indicated that all our models could classify accurately. This reject the second hypothesis of this study which state that there is a significant difference between the performance of traditional financial statement analysis methods and machine learning algorithms in forecasting stock prices in the Nigerian stock exchange market.

Model	AUC	CA	F1	Precision	Recall
kNN	1.000	0.750	0.643	0.562	0.750
Random Forest	1.000	0.917	0.911	0.925	0.917
Logistic Regression	1.000	1.000	1.000	1.000	1.000
Naive Bayes	1.000	1.000	1.000	1.000	1.000

Rank

Tue Aug 08 23, 13:38:14

Input

Features: Net Interest Income, Non-Interest Income, Debt to equity ratio, Capital Adequacy Ratio, Operating Expenses, Return on Equity (ROE), Return on Asset (ROA), Dividend Yield, Earnings per share (EPS)
Target: Decision

Ranks

	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF	FCBF	
1	 Net Interest Income	2.0	0.4204484631347318	0.6468214774323336	0.22500000000000006	1.2	0.312	0.00013554029979531828
2	 Non-Interest Income	2.0	0.4204484631347318	0.6468214774323336	0.22500000000000006	1.2	0.312	1.3554029979531828
3	 Return on Equity (ROE)	2.0	0.4204484631347318	0.6468214774323336	0.22500000000000006	1.2	0.272	1.3554029979531828
4	 Operating Expenses	2.0	0.1842428917900113	0.44522810430465803	0.10227272727272735	0.2727272727272727	0.1	4.301679078078116e-05
5	 Earnings per share (EPS)	2.0	0.11556849565940197	0.12585105079823808	0.0625	0.6666666666666666	0.096	1.5425206156753393e-05
6	 Debt to equity ratio	2.0	0.04300471205299761	0.0661588133282302	0.025000000000000133	0.6666666666666666		6.253906735093569e-06
7	 Dividend Yield	2.0	0.04300471205299761	0.0661588133282302	0.025000000000000133	0.13333333333333333	0.012	6.253906735093569e-06
8	 Return on Asset (ROA)	2.0	0.006987753258863694	0.007131315506727024	0.0035714285714286143	0.06666666666666667	-0.108	0.007863904808241394
9	 Capital Adequacy Ratio	1.0				nan		

Output

Features: Net Interest Income, Non-Interest Income, Return on Equity (ROE), Operating Expenses, Earnings per share (EPS)
Target: Decision

5.2 Conclusion

This study has delved into the realm of machine learning for financial statement analysis and forecast within the context of the Nigerian Stock Exchange market, focusing on GTCO Plc. The overarching objective of this thesis was threefold: firstly, to predict the financial health and profitability of the company, thereby facilitating well-informed and data-driven investment decisions; secondly, to lower investors' risk of financial loss on the Nigerian Stock Exchange; and lastly, to progress machine learning's use in finance by creating a model that can replace traditional methods for financial statement analysis.

To achieve these objectives, we began by employing traditional methods of financial statement analysis, relying on the comprehensive annual reports of GTCO Plc. This initial analysis provided valuable insights into the company's financial standing, which served as the foundation for the subsequent development and evaluation of machine learning models.

The four machine learning algorithms chosen for this thesis Random Forest, K-Nearest Neighbor, Logistics Regression, and Naïve Bayes were systematically applied and assessed. Each algorithm brought its own strengths and limitations to the table, contributing to a well-rounded understanding of their efficacy in predicting financial outcomes. The results were promising, with the models demonstrating a capacity to offer valuable predictions regarding GTCO Plc's financial health and profitability.

The predictive power of these models is crucial in enhancing investment decision-making processes. Investors can now leverage these machine learning models to make informed choices, thus mitigating the risks associated with financial investments in the dynamic Nigerian Stock Exchange market. The reduction of financial losses is particularly significant in a market as dynamic and complex as Nigeria's, where economic and geopolitical factors can impact stock prices swiftly and unpredictably.

Furthermore, the successful development and adoption of machine learning models in this study enhances the field of financial analysis as a whole. By showcasing the effectiveness of these models in predicting financial outcomes, we advocate for the integration of machine learning techniques into financial analysis practices. This advancement stands to revolutionize traditional methods and bring about a new era of efficiency, accuracy, and adaptability in financial decision-making.

In essence, this thesis serves as a testament to the potential of machine learning in financial statement analysis and forecasting within the Nigerian Stock Exchange market. The application of these models not only benefits investors by providing valuable insights and reducing risks but also

contributes to the ongoing evolution of financial analysis methodologies. As technology continues to progress, embracing machine learning in finance becomes imperative for staying ahead in an increasingly competitive and dynamic market environment. The findings of this study, centered on GTCO Plc, provide a foundation for future research and implementation of machine learning in financial markets, clearing the path for a more informed and data-driven method of making investing decisions.

5.3 Research Limitation

Despite the comprehensive nature of this study, it is imperative to acknowledge and discuss the limitations encountered during the research process. One significant constraint was the limited availability of data, as the analysis was confined to the annual reports of Guaranty Trust Bank Plc spanning only 12 years. Ideally, for the development and validation of robust machine learning models, a more extensive dataset would have been preferred, ranging from 50 to 100 years of annual reports.

The primary limitation arising from the restricted dataset is the potential impact on the generalizability of the machine learning models. A larger dataset allows for a more diverse representation of economic cycles, market conditions, and unforeseen events, facilitating the creation of models that can better adapt to a broader range of scenarios. The limited temporal scope of the dataset may lead to a model that is more tailored to the specific circumstances present during the analyzed years, potentially compromising its predictive power in different economic environments.

Furthermore, a larger dataset would have enabled a more granular examination of the machine learning algorithms' performance over time, providing insights into their adaptability and effectiveness in varying market conditions. The scarcity of data, particularly in the context of financial markets known for their volatility and susceptibility to external influences, could result in models that may not adequately capture the complexity and nuances of the Nigerian Stock Exchange market.

Additionally, the limited dataset could impact the robustness of statistical analyses and hypothesis testing conducted in the study. A bigger number of samples often strengthens the statistical power of findings, enhancing the reliability and validity of the results. The reduced sample size may limit the generalizability of the study's conclusions and may not fully represent the broader population of companies in the Nigerian Stock Exchange.

Despite these limitations, it is essential to recognize that the availability of historical financial data, especially for publicly traded companies, is often constrained by factors such as disclosure practices, data accessibility, and the duration of a company's existence. Researchers in the financial domain frequently encounter challenges related to data availability

and completeness, and acknowledging these limitations is crucial for a transparent and accurate interpretation of the study's outcomes.

In conclusion, while the constraints related to the limited dataset pose challenges to the study's generalizability and the robustness of machine learning models, this acknowledgment serves as a foundation for future research. Addressing these limitations requires efforts to access more extensive and varied datasets, ensuring that subsequent studies can build upon and refine the models developed in this research, thereby advancing the adoption of machine learning in financial statement analysis and forecasting in the Nigerian Stock Exchange market.

5.4 Recommendation for Future Research

This study has proven the possibility of incorporating machine learning in financial statement analysis in the Nigerian stock exchange market (NSE) and based on the results gotten, here are some recommendations for future studies.

- To overcome the limitation of a relatively small dataset employed in this study, future research should focus on accessing a more extensive and diverse set of financial data. Collaboration with regulatory bodies, financial institutions, and other stakeholders could facilitate the compilation of a comprehensive dataset covering a more extended period and a broader range of companies. This expansion would aid to the development of more robust machine learning models with increased generalizability.
- Considering the growing interest in democratizing finance, future research could focus on developing user-friendly interfaces that enable investors, even those without extensive technical expertise, to leverage machine learning insights for their investment decisions. This would contribute to the accessibility and adoption of advanced financial analysis tools

Research on the Nigerian Stock Exchange should be broadened to include international comparative studies, as this would provide important information about how well machine learning models apply to various financial markets and regulatory contexts.

By addressing these areas in future studies, scholars can further advance the understanding of machine learning applications in financial statement analysis and forecasting, contributing to the development of more accurate, robust, and ethically sound tools for informed investment decision-making.

6. References

- Adebisi, O. (2020). Corporate Governance and the Nigeria Stock Exchange Market. *African Journal of Corporate Governance*, 1(1), 1-14. <https://doi.org/10.11648/j.ajcg.20200101.11>
- Adegbite, O. S., Adebisi, O. F., & Adegbite, E. A. (2019). Exchange rate volatility and financial performance of commercial banks in Nigeria: Evidence from Guaranty Trust Bank Plc. *European Journal of Accounting, Auditing and Finance Research*, 7(6), 32-42.
- Adegbaju, A., & Adaramola, A. (2018). Financial ratios and corporate performance: Evidence from Nigerian banking sector. *Journal of Financial Reporting and Accounting*, 16(4), 661-678.
- Adegbite, E., & Adebola, S. (2015). Corporate governance and corporate finance in Nigeria. *Corporate Governance: The International Journal of Business in Society*, 15(3), 369-386.
- Adekunle, S. O., Oguntuase, A. M., & Akanbi, A. M. (2020). Empirical Analysis of Earnings per Share and Return on Assets of Firms Listed on Nigerian Stock Exchange. *Journal of Accounting and Financial Management*, 6(1), 10-17.
- Adesina, O. T., Azeez, O. R., & Adesina, A. O. (2017). The effect of debt to equity ratio on firm's financial performance. *International Journal of Management, Accounting and Economics*, 4(10), 690-702.
- Adesina, K., Abdulrahman, M. D., & Olayemi, J. K. (2018). Predicting sustainable competitive advantage of Nigerian firms: A machine learning approach. *Cogent Business & Management*, 5(1), 1444094. doi: 10.1080/23311975.2018.1444094
- Adeyemi, O. D., Olugbara, O. O., & Oyelade, O. J. (2020). Industry trends and financial performance prediction of selected companies in the Nigerian Stock Exchange using machine learning. *Journal of Computational Science*, 44, 101175.
- Afolabi, R. O., & Adefila, J. O. (2018). Financial statement analysis and investors' decisions in Nigeria. *International Journal of Business and Management*, 13(8), 99-107.
- Ahmadi, M., Kouhizadeh, M., & Khosravi, B. (2020). The impact of firm size on the financial performance prediction using data mining techniques. *Journal of Accounting and Management Information Systems*, 19(3), 465-484.
- Ainsworth, S., Sadovskaya, I., Vinogradov, E., Courtin, P., Guérardel, Y., Mahony, J., & Sinderen, D. v. (2014). Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *Bio*, 5(3). <https://doi.org/10.1128/mbio.00880-14>
- Alao, O. R., & Okunlola, J. T. (2020). The impact of industry trends on the financial performance of selected manufacturing firms in Nigeria. *International*

- Journal of Management, Innovation & Entrepreneurial Research*, 6(1), 21-34.
- Al-Kilidar, H. M., Albattah, H. M., Alkhalafat, F. M., Alabdulwahhab, F. M., & Al-Matrouk, M. A. (2019). Employee satisfaction prediction in banking sector using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 10(10), 312-320. doi: 10.14569/IJACSA.2019.0101043
- Al-Nabulsi, A. A., Alfaqawi, M. A., Alsmadi, M. K., & Al-Sarayreh, H. A. (2020). Predicting financial performance of companies listed in Amman Stock Exchange using machine learning techniques. *Sustainability*, 12(7), 2711. doi: 10.3390/su12072711
- Al-Malkawi, H. A., & Masadeh, R. (2021). Predicting the financial performance of companies listed on the Amman Stock Exchange: A machine learning approach. *Journal of Finance and Investment Analysis*, 10(2), 66-82.
- Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609.
- Altman, E. I., & Sabato, G. (2005). *Modelling credit risk for SMEs: Evidence from the US market*. *Abacus*, 41(3), 257-268.
- Amadi, V. C. (2021). The Impact of GDP on Nigerian Stock Market Returns. *Journal of Economics and Sustainable Development*, 12 (5), 64-71
- Akande, O. J., & Olubiyi, O. A. (2019). Predicting stock prices of Nigerian banks: A machine learning approach. *Journal of Financial Data Science*, 1(1), 106-116. doi: 10.3905/jfds.2019.1.1.106
- Akinniyi, O. D., Akinola, O. A., Aluko, O. O., & Oloyede, O. O. (2021). Modeling stock market returns in Nigeria using machine learning algorithms. *Journal of Financial Reporting and Accounting*, 19(1), 43-62.
- Akpan, U. F., & Udoma, J. A. (2016). The relationship between inflation rate and stock prices in Nigeria. *Journal of Applied Economics and Business Research*, 6(3), 93-106.
- Ayinde, T. O., & Oladipo, O. A. (2020). Machine learning approach for financial statement analysis: An application in the Nigerian stock exchange. *Journal of Financial Reporting and Accounting*.
- Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), 129-151.
- Beaver, W. H. (1968). Market prices, financial ratios, and the prediction of failure. *Journal of Accounting Research*, 6, 179-192.
- Berger, A. N., & DeYoung, R. (1997). Problem loans and cost efficiency in commercial banks. *Journal of Banking & Finance*, 21(6), 849-870.
- Bhattacharya, U., Graham, J. R., & Harvey, C. R. (2002). The effects of financial statement credibility on judgmental forecasts of future earnings. *The Accounting Review*, 77(Supplement), 115-142.
- Bustani, B., Kurniaty, K., & Widyanti, R. (2021). The effect of earning per share, price to book value, dividend payout ratio, and net profit margin on the stock price in indonesia stock exchange. *Jurnal Maksipreneur: Manajemen, Koperasi, Dan Entrepreneurship*, 11(1), 1. <https://doi.org/10.30588/jmp.v11i1.810>
- Chen, M., Cheng, S., & Hwang, Y. (2005). An empirical investigation of the relationship between intellectual capital and firms' market value and financial performance. *Journal of Intellectual Capital*, 6(2), 159-176. <https://doi.org/10.1108/14691930510592771>.
- Chinwudu, C. F. and Nwanna, I. O (2016). The effect of financial deepening on economic growth in Nigeria (1985 -2014). *IOSR Journal of Economics and Finance*, 07(04), 11-28. <https://doi.org/10.9790/5933-0704011128>.
- Clemes, M. D., Gan, C., & Zhang, D. (2010). Customer switching behaviour in the Chinese retail banking industry. *International Journal of Bank Marketing*, 28(7), 519-546. <https://doi.org/10.1108/02652321011085185>.
- Economic Indicator; Definition and How to interpret. https://www.investopedia.com/terms/e/economic_indicator.asp
- Elshandidy, T., Hassanein, A., & Gaber, M. (2017). The relevance of dividend yield in stock return: Evidence from the UK stock market. *Research in International Business and Finance*, 42, 1312-1327.
- Fasan, O. E., & Gbadebo, O. J. (2017). The impact of company size on financial performance: Evidence from selected companies in the Nigerian Stock Exchange. *Journal of Accounting and Financial Management*, 3(1), 1-13.
- Fama, E. F., & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, 22(1), 3-25.
- Fong, N. M., Fang, Z., & Luo, X. (2015). Geo-conquesting: competitive locational targeting of mobile promotions. *Journal of Marketing Research*, 52(5), 726-735. <https://doi.org/10.1509/jmr.14.0229>
- Goudreau, J. (2021). Machine Learning Algorithms in Finance. *Investopedia*. <https://www.investopedia.com/terms/m/machine-learning-algorithms-finance.asp>
- Hamid, I. (1907). Les musulmans francais du nord de l'afrique. *Bulletin of the American Geographical Society*, 39(12), 758. <https://doi.org/10.2307/198579>
- Hou, Y., Ni, Z., & Xu, X. (2020). Industry trends, regulatory changes, and stock price synchronicity: Evidence from China. *Emerging Markets Finance and Trade*, 56(1), 1-12.
- Inanga, E. L., Ogbulu, O. M., & Ogoun, E. N. (2016). The relationship between earnings per share and market

- value of firms listed in the Nigerian Stock Exchange. *International Journal of Finance and Accounting*, 5(6), 354-358.
- Investopedia. What is considered a good ROA
<https://www.investopedia.com/terms/r/returnonassets.asp#:~:text=What%20is%20Considered%20a%20Good,fi rms%20in%20the%20same%20sector>.
- Iyoha, F. O., Adebisi, A. A., Adebisi, M. A., & Olugbara, O. O. (2019). Pattern analysis of trading activities in the Nigerian stock exchange. *Heliyon*, 5(5), e01766.
- Jalaludin, N. H., & Ibrahim, M. H. (2014). Predicting bank performance using financial statement analysis: A case of Malaysian commercial banks. *Procedia - Social and Behavioral Sciences*, 145, 171-178.
- Jiménez, G., Ongena, S., Peydró, J. L., & Saurina, J. (2014). Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking? *Econometrica*, 82(2), 463-505.
- Jokhio, S. H., & Memon, M. A. (2018). Understanding financial ratios and performance indicators: A case of investors in Pakistan. *Journal of Business and Economics Review*, 3(1), 24-35.
- Karimi, M., & Rahmani, M. (2017). Financial statement analysis of Iranian companies using ratio analysis. *International Journal of Accounting and Financial Reporting*, 7(1), 305-318.
- Khalid, H., Abbas, Z., & Sadiq, M. (2016). Comparative financial statement analysis of selected companies in Pakistan. *Research Journal of Finance and Accounting*, 7(1), 103-110.
- Kieso, D. E., Weygandt, J. J., & Warfield, T. D. (2016). *Intermediate accounting (16th ed.)*. Hoboken, NJ: John Wiley & Sons.
- Kumar, A., & Rastogi, P. (2019). Financial statement analysis using machine learning. In *Proceedings of the 10th International Conference on Computational Intelligence and Communication Networks* (pp. 328-333).
- Latif, R. A., Isa, M. A. M., Zaharum, Z., & Baharudin, M. A. (2022). Profitability determinants of technology companies: a study from ace market of bursa malaysia. *International Journal of Academic Research in Business and Social Sciences*, 12(6). <https://doi.org/10.6007/ijarbss/v12-i6/14034>
- Lawal, A. I., Abdulsalam, D. A., & Abdurashied, A. A. (2017). Exchange rate fluctuations and the financial performance of listed firms in Nigeria. *Research Journal of Finance and Accounting*, 8(17), 26-34.
- Lawal, A. I., Oseni, E., Asaley, A. J., Lawal-Adedoyin, B., & Ojeka-John, R. (2021). Is the stock market efficient? evidence from nonlinear unit root tests for Nigeria. *Asian Economic and Financial Review*, 11(5), 384-395. <https://doi.org/10.18488/journal.aefr.2021.115.384.395>
- Lawal, R. A., & Yusuf, S. A. (2018). The impact of Gross Domestic Product (GDP) on stock market performance in Nigeria. *Journal of Accounting, Finance and Auditing Studies*, 4(2), 85-96.
- Liang, Y., Ren, R., & Sun, C. (2020). Earnings quality and stock price crash risk: Evidence from Chinese listed companies. *Journal of Business Research*, 116, 125-135.
- Liang, M., Yang, B., Zhang, X., & Zhang, J. (2019). Stock price prediction using machine learning: A review. *International Journal of Financial Studies*, 7(3), 38.
- Li, Y., & Li, X. (2020). Using machine learning for customer retention analysis in banking industry. *Journal of Big Data*, 7(1), 56. doi: 10.1186/s40537-020-00330-6
- Majeed, M. T., Azam, M., & Amjad, N. (2017). Impact of dividend yield on stock prices: Evidence from Pakistan. *Pakistan Journal of Commerce and Social Sciences*, 11(1), 251-270.
- Muttia, E. E. and Sutrisno, S. (2022). Financial performance and prediction of financial distress in food and beverage companies listed on the indonesia stock exchange. *International Journal of Economics, Business and Management Research*, 06(05), 154-167. <https://doi.org/10.51505/ijebmr.2022.6513>
- National Bureau of Statistics (2021). Consumer Price Index and Inflation
- Nsehe, M. (2018, June 7). GTCO, Zenith Bank, And Access Bank Are Nigerian Banks of the Year. Forbes. Retrieved from <https://www.forbes.com/sites/mfonobongnsehe/2018/06/07/gtbank-zenith-bank-and-access-bank-are-nigerian-banks-of-the-year/?sh=128f1a9a1b11>
- Nigeria Stock Exchange. (2021). Market Data. <https://www.nse.com.ng/market-data/>
- Obadiaru, E.D., Obasaju, B.O., Omankhanlen, A.E. and Eyiolorunshe, D.T. (2020) "Dynamic links between the Nigerian equity market and those of selected regional and developed countries," *Heliyon*, 6(9), p. e04782. doi: 10.1016/j.heliyon. 2020.e04782.
- Odo, C., Okpara, G. C., Ugwu, S. U., & Onwe, I. O. (2019). Impact of interest rate on stock market performance in Nigeria: A machine learning approach. *Journal of Economics and Sustainable Development*, 10(16), 222-231.
- Olayinka, T. O., & Ogundele, O. J. (2018). The impact of company size on financial performance: Evidence from machine learning algorithms. *International Journal of Intelligent Computing and Cybernetics*, 11(2), 213-228.
- Olugbenga, A. (2012). Exchange Rate Volatility and Stock Market Behaviour: The Nigerian Experience. <https://core.ac.uk/download/pdf/234629265.pdf>.

- Olugbenga, A. S., & Adetunji, O. I. (2019). Machine learning for stock price prediction in Nigerian stock exchange. *International Journal of Advanced Computer Science and Applications*, 10(2), 157-164.
- Ogbebor, P. A., & Okorie, K. C. (2019). An empirical investigation of the capital structure determinants of quoted firms in Nigeria. *African Journal of Accounting, Auditing and Finance*, 8(4), 296-312.
- Ogundipe, S. E., & Idowu, A. (2017). Determinants of stock prices in the Nigerian stock exchange. *Research Journal of Finance and Accounting*, 8(19), 2222-5697
- Ogunjobi, T., & Ogunnaike, O. O. (2021). Predicting the financial performance of selected companies listed in the Nigeria stock exchange using interest rate data: A machine learning approach. *Data in Brief*, 36, 107023.
- Ogundele, O. J., & Olayinka, T. O. (2021). Predicting the future financial performance of companies listed on the Nigeria Stock Exchange using industry trend analysis. *International Journal of Business Forecasting and Marketing Intelligence*, 7(2), 143-160.
- Okpala, K. C., & Obi, C. S. (2018). The impact of profitability ratios on stock prices: Evidence from Nigeria. *Journal of Financial Reporting and Accounting*, 16(1), 119-139.
- Oladejo, M. A., & Oluwole, O. O. (2020). Firm-specific determinants of profitability of Nigerian listed firms. *Journal of Financial Reporting and Accounting*, 18(4), 543-564.
- Oladele, P. O., Adetiloye, K. A., & Oke, M. O. (2020). Market capitalization and financial performance: Evidence from Nigerian listed firms. *Journal of Accounting and Management Information Systems*, 19(2), 311-332.
- Olaoye, O. A., & Eriki, P. O. (2019). Market capitalization and financial performance of quoted companies in Nigeria. *Journal of Accounting and Financial Management*, 5(1), 16-26.
- Olugbenga, A. O., & Adetayo, A. A. (2018). Effect of interest rate on the financial performance of selected deposit money banks in Nigeria. *Journal of Finance and Investment Analysis*, 7(2), 12-26.
- Oyetayo, V. O., Iyiola, O. O., & Oyekunle, R. O. (2015). Dividend policy and financial performance of listed banks in Nigeria. *Journal of Applied Accounting Research*, 16(2), 221-238.
- Oyebanjo, O. J., & Oladipupo, A. O. (2019). Exchange rate fluctuations and financial performance of banks in Nigeria. *Journal of Applied Accounting and Taxation*, 6(2), 1-10.
- Okafor, G. C., & Obute, G. C. (2018). Neural network analysis of Nigerian stock prices. *International Journal of Advanced Computer Science and Applications*, 9(6), 381-386.
- Okoro, O. J., & Akaniro, E. M. (2019). The effect of leverage on the return on equity of listed manufacturing companies in Nigeria. *International Journal of Innovative Research and Advanced Studies*, 6(8), 32-38.
- Purwanto, E. and Purwanto, A. D. B. (2020). An investigative study on sustainable competitive advantage of manufacture companies in Indonesia. *Business: Theory and Practice*, 21(2), 633-642. <https://doi.org/10.3846/btp.2020.12256>
- Raza, S. A., Saqib, N., & Masood, A. (2015). Financial analysis of selected companies in Pakistan using vertical analysis technique. *Journal of Finance and Accounting*, 3(1), 1-9.
- Shiller, R. J. (1989). Market volatility and macroeconomic fluctuations. *The American Economic Review*, 79(5), 1116-1132.
- Sondhi, N., & Kaur, J. (2018). Understanding financial ratios and their significance: An empirical analysis of Indian investors. *International Journal of Applied Business and Economic Research*, 16(3), 311-320.
- Talwar, N., Bhatia, V., & Kalia, N. (2020). Financial statement analysis using machine learning: A review. *Journal of Financial Reporting and Accounting*, 18(2), 271-289.
- Tan, J., Lo, K., Fang-dao, Q., Zhang, X., & Zhao, H. (2019). Regional economic resilience of resource-based cities and influential factors during economic crises in china. *Growth and Change*, 51(1), 362-381. <https://doi.org/10.1111/grow.12352>
- Wei, B., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked auto encoders and long-short term memory. *Plos One*, 12(7), e0180944. <https://doi.org/10.1371/journal.pone.0180944>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). *Morgan Kaufmann Publishers*.
- Xia, J., Wu, Z., Zhang, R., Chen, X., & Zhang, R. (2022). Shareholder personal risk and firm risk: an empirical analysis of share pledges and firm debt policies. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.1010162>
- Xu, L. and Cai, F. (2009). Before and after 2000: revenue and high tech valuation. *Competitiveness Review: An International Business Journal*, 19(1), 26-35. <https://doi.org/10.1108/10595420910929040>
- Zairi, M., Ahmed, A. M., & Jabnoun, N. (2005). Trend analysis of financial ratios: A tool for company's performance analysis. *Benchmarking: An International Journal*, 12(5), 443-454.
- Zeng, Y., Liu, Z., & Fang, Y. (2020). The impact of Altman Z-score on stock prices: Evidence from Chinese listed companies. *Journal of Risk Research*, 23(1), 116-132.

Does the use of Generative AI contribute to university students' engagement or disengagement in their studies? Examining the antecedents of fear of missing out, perceived ease of use, and perceived usefulness.

Ibrahim Adeshola
Final University, Cyprus

Ibrahim.adeshola@final.edu.tr

Afi Kekeli Feda
Final University, Cyprus

kekelifeda@gmail.com

Adeola Praise Adepoju
aadeolapraise@gmail.com

Abstract

While GenAI has demonstrated numerous advantageous effects in education, there are still some concerns regarding the benefits of GenAI on student engagement. This research addresses this gap by thoroughly examining the distinctive impact of novel GenAI on both student engagement and disengagement in the classroom. The study focuses on first-year and second-year students from the European University of Lefke, with a total of 250 usable responses collected, representing an 81.7% response rate out of 306 questionnaires distributed to registered students. Utilizing SmartPLS software version 4.32, the data analysis employed the partial least squares structural equation modelling (PLS-SEM) technique to scrutinize the proposed hypothesized model. The results of the analysis shed light on the research questions, revealing that students can simultaneously experience both engagement and disengagement in classroom activities due to the improper use of GenAI. This study enhances our understanding of the complex relationship between GenAI adoption, student engagement, and educational outcomes, offering valuable insights for educators and policymakers alike.

Keywords

GenAI, Fear-of-Missing-Out, Student Engagement, Classroom Disengagement, Perceived Usefulness, Perceived Ease-of-Use, Student Satisfaction.

1. Introduction

The significance of student engagement in promoting learning and academic performance in higher education is commonly acknowledged, leading to extensive study and theoretical exploration in this area (Bond et al., 2020; Ferrer et al., 2022). Student engagement pertains to the level of optimism, attention, interest, curiosity, and motivation that learners manifest regarding their studies (Khaleel et al., 2020). This is not limited to superficial learning, such as memorizing content and meeting minimum requirements to pass a course. Instead, it involves deep thinking activities like analyzing, understanding materials, applying them to solve problems, and deriving meaning. Additionally, it requires social interaction between learners and teachers, where they exchange experiences, perspectives, and encouragement. Regardless of the content or the teaching methods, effective teaching and learning depend on students' engagement (Lee et al., 2019). Academics support the idea that engaged

students are more likely to enhance their academic performance, including their grades and critical thinking abilities. Furthermore, they can utilize the knowledge they acquire in real-world situations (Lee et al., 2019).

There is a proliferation of studies dedicated to the utilization of digital technologies, especially generative AI (GenAI) chatbots, to improve learners' engagement and learning outcomes in higher education. An AI chatbot is software that makes use of natural language processing (NLP) and semantic analysis to communicate with users through text or voice, assimilate their requests, and provide immediate responses based on its training data and algorithms (Adamopoulou & Moussiades, 2020). Li and Xing (2021) suggested that using chatbots in education can have advantages, including offering students a platform for continuous study and asynchronous conversation. This characteristic has been proven to positively impact student engagement, as it promotes a learner-centered studying environment by enabling students to ask questions and participate in friendly conversations without any constraint of time or geographical position (Cotton et al., 2023). Chang et al. (2021) proposed that chatbot-powered learning has a significant likelihood of increasing students' engagement as learners tend to be more open to studying in a learner-centered environment. Several academics have suggested that teachers typically lack the time necessary to satisfy each student's unique demands (Collinson & Cook, 2001; Hao, 2019). Additionally, students have limited chances to speak with teachers in class and even less so after class when they need assistance, which can lead to disengagement. Therefore, finding a tool that enables students to complete learning assignments independently is essential (Chang et al., 2021). Chatbots are such potential tools for supporting adaptable studying, personal studying success, and self-confidence (Chen et al., 2023). Lee et al. (2022) investigated how an AI-based chatbot can be used to help students with their after-class review process for consolidating knowledge and understanding the subject content. They have proven that the use of chatbots in the review process facilitates learning engagement by helping students feel recognized and establishing a relaxing and friendly interaction, thereby improving their academic performance. Okonkwo and Ade-Ibijola (2021) argued that these days, students prefer learning through online platforms and using their smart devices to

access information rather than traditional textbooks or course materials. Chatbots create a comfortable and enjoyable atmosphere for studying. They also noted that learning with a conversational tool is more convenient and interesting for students, and the use of chatbots in education can lead to increased student engagement. Hamam (2021) also highlighted numerous advantages of incorporating chatbots into higher education, including enhancing and personalizing the educational experience, particularly in classes with a large number of students.

There are several GenAI tools available to students today. These include Microsoft's Bing Chat (now integrated as Microsoft Copilot), launched in February 2023; Alphabet's Bard, rebranded as Google Gemini in February 2024; Baidu's Ernie Bot, introduced in March 2023; Claude by Anthropic, released in March 2023; Perplexity AI, launched in August 2022; You.com launched in November 2021, DeepSeek-V2, released in May 2024; and OpenAI's ChatGPT, debuted in November 2022. ChatGPT, short for Conversational Generative Pretrained Transformer, stands out as one of the most remarkable among them. ChatGPT uses deep learning algorithms to generate human-like responses to text-based prompts, making it capable of holding natural language conversations with humans. ChatGPT by OpenAI has been trained on a massive database, including books, articles, and websites, and has the ability to learn and generate responses on a wide range of topics in multiple languages. ChatGPT has garnered a lot of attention and has achieved a record-breaking milestone by becoming the fastest-growing consumer internet application in history, reaching 100 million monthly active users as of January 2023, just two months after its release (Hu, 2023).

Despite the numerous proven advantageous effects of chatbots in education, there seems to be disagreement regarding the inherent advantages and subsequent benefits of GenAI chatbots with regard to students' engagement. The new generation's sophistication and open access have triggered both skepticism and excitement. To start, unlike other AI chatbots that repeat responses to generic questions, New GenAIs can understand context and are generative, which means they can produce "original" human-like content based on a variety of inputs (Miller et al., 2022). For example, ChatGPT has 175 billion parameters; with this level of complexity, the chatbot defies formulaic, scripted responses (Graham, 2022). Lastly, most of these new GenAI's have a free plan, removing the financial blockage that restricts students' access to other AI tools. Their usage may then be difficult to control and can result in significant practical and ethical issues, as highlighted by Basic et al. (2023). Some educators contend that by relying too heavily on these technologies, students risk losing the ability to think critically and solve problems effectively, both of which are essential for success in their future professional lives (Lo, 2023; Bai et al., 2023; Sallam et al., 2023a; Tlili et al., 2023). Others contend that by offering tailored learning experiences and

encouraging autonomous study, GenAI's can actually increase student engagement (Guo & Lee., 2023; Kostka & Toncelli, 2023; Rudolph et al., 2023a).

In light of the limited attention given to students' perspectives in existing studies, this research endeavors to fill a crucial gap by thoroughly examining the peculiar impact of novel GenAI on student engagement, classroom disengagement, learning performance and satisfaction. An in-depth exploration of students' perceived ease of use and perceived usefulness is essential to gaining a comprehensive understanding of its implications. Given the influential role that students play in shaping educational outcomes through their engagement, their viewpoints become paramount. With regards to the research questions posed, it prompts an exploration into the potential scenario wherein students may concurrently experience both engagement and disengagement during classroom activities. Therefore, this study seeks to provide valuable insights into the distinctive effects of GenAI on students' engagement, acknowledging the potential implications for learning outcomes and overall satisfaction. The hypotheses outlined further underscore the critical need to delve into these unexplored dimensions, aiming to contribute significantly to the existing body of knowledge in the field of educational technology.

2. Literature Review

Exploring the Influence of Perceived Ease of Use and Perceived Usefulness on the Actual Use of GenAI.

The utilization of generative AI (GenAI) in education has emerged as a noteworthy subject of interest due to its potential to shape the dispensation and acquisition of knowledge (Lo, 2023). The perception of technology as easy to use plays a substantial role in its acceptance (Sugandini et al., 2018). The factor of ease of use contributes to individuals' perceptions of self-efficacy, thereby increasing the likelihood of embracing the technology. Furthermore, the perceived ease of use influences the effectiveness of responses, as people are more inclined to utilize technology that they find easy to use (Vaportzis et al., 2017).

Sallam et al. (2023b) extensively examined the concept of perceived ease of use in relation to ChatGPT. They conducted a survey including the four TAME-ChatGPT usage sub-scales, one of which was the perceived ease of use sub-scale. In terms of the perceived ease of use sub-scale, participants were asked to rate their perception of ChatGPT's ease of use on a scale ranging from 2 to 10, where higher scores indicated a greater perceived ease. The study sample reported a high level of user-friendliness for ChatGPT. Importantly, there were no statistically significant differences observed among various factors, including age, gender, nationality, university, and educational level. This indicates that individuals across different demographics had a consistent perception of ChatGPT's ease of use.

It is worth emphasizing that perceived ease of use and perceived usefulness are closely interconnected. These two

factors work together to influence the adoption of technology. When individuals perceive technology as both easy to use and beneficial, it strengthens their motivation to adopt and effectively utilize the technology (Granić 2022). The survey conducted by Sallam et al. (2023b) included the perceived usefulness as one of the four sub-scales. Participants were asked to rate their perception of ChatGPT's usefulness on a scale ranging from 6 to 30, with higher scores indicating greater perceived usefulness. A score of 18 represented a neutral attitude. The average score for perceived usefulness was 24.2 ± 4.9 , indicating that participants who had prior experience with ChatGPT highly perceived it as useful. Notably, there were no statistically significant differences observed based on the control variables.

Chan and Hu (2023) had similar results with a survey conducted among 399 undergraduate and postgraduate students from various academic disciplines in Hong Kong. The findings indicate a generally positive attitude towards ChatGPT's role in teaching and learning. Students acknowledge its potential for personalized learning support, assistance with writing and brainstorming, as well as research and analysis capabilities. In particular, students place a significant emphasis on the perceived usefulness of ChatGPT in terms of offering valuable insights. MacNeil et al. (2022) recently utilized GPT-3 to generate explanations for code. Although there are still unanswered research and pedagogical questions that require further investigation, this work successfully demonstrated the potential of GPT-3 in supporting learning by providing explanations for code snippets. Therefore, we hypothesized that:

H1 : Perceived Ease of Use positively influence Actual Use of GenAI.

H2 : Perceived Usefulness positively influence Actual Use of GenAI.

Fear-of-Missing-Out and Actual Use of GenAI.

The fear of missing out (FoMO) can be described as a persistent concern that others are having better or more valuable experiences, leading to a constant urge to stay connected with people. Initial investigations into the occurrence of FoMO revealed that approximately seventy five percent of young adults acknowledged experiencing this phenomenon (Anastasya et al., 2022). Przybylski et al. (2013) exploration of the phenomenon, they draw upon the Self-Determination Theory (SDT) developed by Deci and Ryan (2001). According to this theory, FoMO is an indication of poor self-regulation resulting from prolonged unmet psychological needs. Furthermore, the SDT theory explains how fulfilling three basic psychological needs – competence, autonomy, and connectedness – can significantly impact self-regulation and psychological well-being. Essentially, the concept of FoMO, when viewed through the lens of SDT, highlights the prolonged deprivation of these essential psychological needs, influencing one's ability to self-regulate and maintain psychological health over time.

Given the aptness of GenAI's to provide personalized assistance in education, the Fear of Missing Out (FoMO) experienced by students is a relevant factor to consider, especially in relation to the first basic psychological need of competence. FoMO can arise when students perceive that their peers or classmates are utilizing ChatGPT or similar tools to access information, receive instant feedback, or enhance their learning experiences. Students may fear that by not using these tools themselves, they may miss out on valuable resources, opportunities for personalized learning, or academic advantages that their peers may gain (Qutishat and Abu Sharour, 2019). To date, there is a gap in the literature concerning students' FoMO and how it can affect students' adoption of GenAI's in higher education. Therefore, the authors hypothesized that:

H3 : Fear-of-Missing-Out positively influence Actual Use of GenAI.

The effect of GenAI Usage on Student Engagement.

The advancement of generative artificial intelligence prompts a reevaluation of the teaching-learning process, given its influence on the trajectory of adaptive education. Incorporating generative artificial intelligence tools to boost student engagement signifies a novel and promising strategy for meeting the changing demands of contemporary education (Ruiz-Rojas et. al., 2023). These tools harness extensive datasets and machine learning algorithms to customize students' learning experiences, tailoring instruction to their unique needs and preferences (Rudolph et al., 2023b). Through the analysis of a student's strengths, weaknesses, and performance patterns, GenAI can offer personalized feedback and recommend specific study materials or exercises. Moreover, the integration of GenAI techniques into the educational landscape not only enhances the assessment of students' progress but also plays a crucial role in fostering heightened student engagement. Salinas-Navarro et al. (2024) delve into the nuanced challenges associated with the utilization of ChatGPT, exploring the viewpoints of both educators and students. Their study sheds light on a spectrum of concerns ranging from academic integrity and the credibility of ChatGPT-generated content to digital safety, biases inherent in AI systems, and the potential impact on traditional and online assessment methods. Moreover, the authors stated the potential ramifications for critical thinking skills. They advocate for the responsible integration of AI technologies into educational frameworks, emphasizing the need for strategic recommendations to ensure that technological advancements uphold educational standards and cultivate a secure learning environment. GenAI tools serve as invaluable resources for unpacking pedagogical theories, offering an abundance of precise definitions, in-depth explanations, and practical applications from various angles. These insights hold great potential for educators and learners alike, enabling them to enhance the quality of student-centered learning experiences (Salinas-Navarro et al., 2024). It's worth noting that each GenAI tool

brings its own unique perspectives, depth of understanding, and strengths to the forefront. While ChatGPT 3.5 shines in delivering comprehensive explanations spanning definitions, principles, and practical advice, Google Bard emphasizes the application of theories in real-world scenarios. New Bing focuses on clarifying theories and concepts with precision, while Anthropic Claude emphasizes the elucidation of goals and principles. Based on the insights from the literature, we have formulated this hypothesis:

H4 : Actual Use of GenAI positively influence Student Engagement.

The effect of GenAI Usage on Classroom Disengagement.

While AI has the potential to enhance learning efficiency, it also introduces a more transactional aspect to education. With the rise of GenAI and technological advancements, the path of strategic disengagement becomes even more accessible to students. GenAI becomes a perfect toolkit for students whose primary goal is to attain a degree with minimal investment. AI-driven tools streamline the learning process, potentially reducing it to a series of optimized steps designed to achieve a degree with minimal effort. This ease of achieving results might tempt students to disengage, prioritizing other values and activities over academic immersion. Tools promising enhanced learning may, in some instances, deepen the divide between education as a transformative journey and education as a mere transaction. Some studies have investigated the extent to which GenAI tools can complete university assignments. According to Katz et al. (2023), GPT-3 successfully passed the United States Bar Exam, a rigorous assessment typically completed after seven years of post-secondary education. A more recent study reported that GPT-4 performs notably better than human test-takers, showing a substantial 26% improvement compared to GPT-3 and outperforming humans in five out of seven subject areas. Kung et al. (2023) conducted a study to evaluate the performance of ChatGPT on the United States Medical Licensing Exam. The evaluation results showed that ChatGPT performed at or above the passing threshold on the exam without any domain-specific fine-tuning. However, concerns have emerged regarding the unintended consequences of GenAI usage on classroom disengagement among students. One argument posits that the overreliance on AI-generated content may diminish students' intrinsic motivation and critical thinking skills, leading to passive consumption rather than active engagement in the learning process (Washington, 2023; Michel-Villareal et al., 2023). As students become accustomed to receiving pre-generated responses and content from AI-driven tools, they may exhibit reduced interest and enthusiasm for classroom activities, contributing to disengagement and disconnection from the learning experience. Therefore, the authors hypothesized that:

H5 : Actual Use of GenAI negatively influence Classroom Disengagement.

The effect of Student Engagement on Student Learning Performance and Student Satisfaction.

Kasneci et al. (2023) stated that another relationship to consider is the one between student engagement and satisfaction. The active involvement of students in their educational journey significantly influences their level of satisfaction. According to Boulton et al., (2019), student participation and engagement in learning activities play a key role in student academic achievement. When students actively engage in their learning experience, they generally exhibit higher satisfaction levels with their education. Gray and Diloroto (2016) further highlight that engaged students demonstrate motivation, interest, and active participation in classroom activities both online and offline. They also cultivate a sense of belonging and contribute to creating a positive learning atmosphere (Wood and Harris, 2015). Furthermore, they benefit from personalized and interactive learning opportunities. As a result, they achieve academic success, encounter growth opportunities, and enjoy a gratifying overall educational experience (Cents-Boonstra et al., 2021). Educational institutions that prioritize student engagement have a greater likelihood of enhancing student satisfaction and improving students learning performance. Therefore, we hypothesized that:

H6 : Student Engagement positively influence Student Satisfaction.

H7 : Student Engagement positively influence Student Learning Performance.

Classroom Disengagement and Student Learning Performance

The detrimental effect of classroom disengagement on student learning is generally acknowledged in the literature (Manlove, 1998; Gini et al., 2015; Wammes et al., 2019). Disengagement in students emerges as decreased active participation, a lack of enthusiasm, and a loss of interest in the learning process. Juvonen et al. (2012) underscore the significance of peer relationships in academic achievement and suggest that disengagement may lead to poor peer interactions, further impeding learning outcomes. Disengaged students often struggle to comprehend concepts, fail to connect with the material, and experience a decline in academic performance (Lawson and Lawson, 2020). Anderman (2002) emphasizes that disengagement has negative consequences for psychological outcomes, particularly during adolescence. The prevalence of classroom disengagement denies students valuable learning opportunities and hinders their ability to achieve their full academic potential. Therefore, we hypothesize that:

H8 : Classroom Disengagement negatively influence Student Learning Performance.

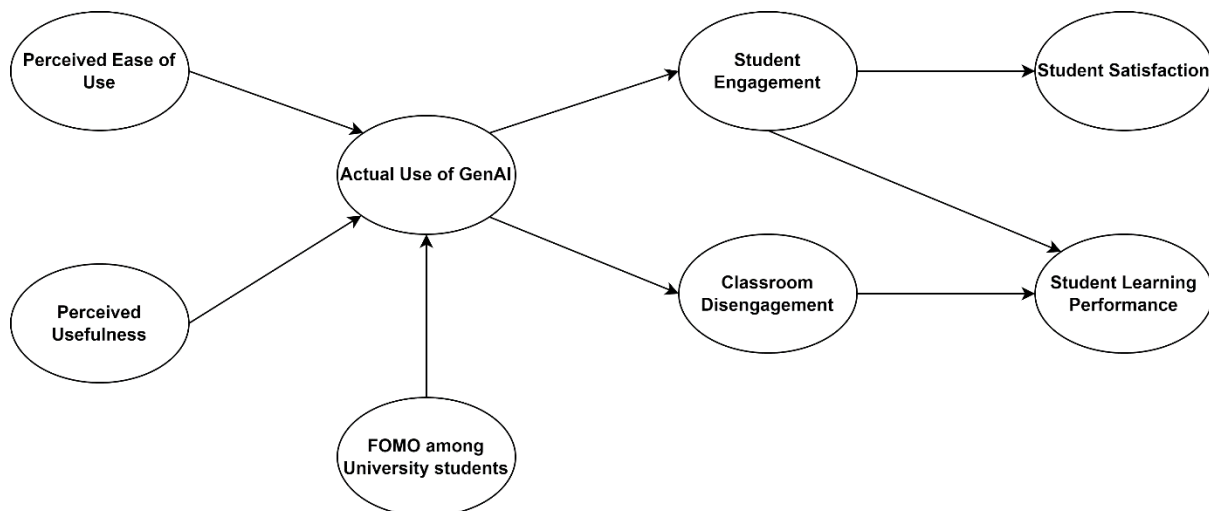


Figure 1: Hypothesized Model

3. Methodology

Sampling Recruitment

The study included first-year and second-year students from the European University of Lefke. A total of 250 usable responses were collected, indicating a response rate of 81.7% out of the 306 questionnaires distributed to registered students. Among the respondents, 53.2% (133) were male and 46.8% (117) were female, all falling within the age range of 18–25. The data collection process involved a self-administered questionnaire with explicit prior consent obtained from participants, assuring them of the confidentiality and exclusive use of their information for research purposes.

A convenient sampling technique was employed to collect information from students who were both accessible and willing to partake in the study. The survey was implemented and distributed using the Microsoft Forms service and a paper questionnaire. Participation in the research was entirely voluntary, with no incentives provided. The e-questionnaire and paper questionnaire were administered during the 2022–23 fall and spring semesters. To prevent duplicate participation in the e-questionnaire, the "One response per person" setting available in Microsoft Forms was activated.

Measurement Items

The items utilized in this study were derived from existing studies rooted in student engagement and satisfaction literature reviews. They were adapted and refined to align with the specific objectives of this research. The set of items, including five assessing perceived usefulness and six gauging perceived ease of use, were drawn from Davis (1989). Additionally, the five items measuring the Fear of Missing Out were adopted from Przybylski et al. (2013), for the actual use of GenAI two items were adopted from Natasia et al. (2022) and another two items on problem solving category were adopted from Maillet et al. (2015), three and five items

assessing Student Engagement were taken from Setiawan and Taiman (2020) and Howard et al. (2016), respectively. Furthermore, five items measuring Classroom Disengagement were incorporated from Jang et al. (2016), four items evaluating self-reported student learning performance were adapted from Carini et al. (2006), and six items gauging student satisfaction were adopted from Asosega et al. (2002). The study employed a quantitative analysis to assess the variables under consideration and the questionnaire comprising these items was pre-tested.

4. Data Analysis

The data analysis process employed SmartPLS software version 4.32, utilizing the partial least squares structural equation modeling (PLS-SEM) technique to examine the proposed hypothesized model (Ringle et al., 2015). PLS-SEM was selected due to its ability to handle non-normal distributions and its suitability for studies with limited sample sizes (Hew et al., 2015; Hair et al., 2014; Wong et al., 2015). Following the methodology outlined by Ringle et al. (2005), the analysis comprised two main steps: Firstly, assess the external measurement model, and secondly, scrutinize the internal structural model. Prior to the evaluation, common method bias (CMB) was first evaluated, employing 5000 bootstrapping sub-samples and individual sign changes for inference statistics across the 250 coded items (Hair et al., 2011).

To mitigate common method bias (CMB), we adopted a strategy where predictor measures and criterion measures were obtained from separate sources, aligning with the approach advocated by Podsakoff et al. (2012). CMB was evaluated both statistically and in line with the recommendation for bias reduction proposed by Etchadi and Karatepe (2019). Kock's (2015) method was utilized to assess the extent of common method bias. Kock proposed that collinearity statistics, particularly a variance inflation factor (VIF) exceeding 5, indicate problematic collinearity and

suggest potential contamination due to common method bias. The outer VIF analyses, as displayed in Table 1, demonstrated that all constructs exhibited VIF values below 3.3, indicating that common method bias does not pose a significant concern in this study.

Measurement Model

In this section, we performed tests for both convergent validity and discriminant validity to scrutinize the internal measurement model. To assess the reliability of constructs, we examined individual item reliability, ensuring that the outer loading surpassed 0.70. Internal consistency reliability was evaluated through Cronbach's Alpha ($\alpha < 0.95$), outer loading (>0.70) Composite Reliability (CR > 0.70), convergent validity (AVE > 0.5) and rho_A of the measures associated with each construct, as suggested by Henseler et al. (2009), Hair et al. (2014), and Fornell & Larcker (1981).

For discriminant validity, we followed the criteria outlined by Henseler et al. (2009) and Fornell & Larcker (1981).

Upon examining the individual item reliability for the constructs, it was observed that one item within Perceived Ease of Use (PE1), two within Student Engagement (SE7, SE8), two within Classroom Disengagement (CD1, CD5), and two within Student Satisfaction (SS1, SS2) marginally fell short of the suggested threshold of 0.7 for outer loadings, as recommended by Fornell and Larcker (1981). Nonetheless, all assessed constructs exhibited Average Variance Extracted (AVE) values that were statistically significant at the 0.05 level and surpassed the threshold of 0.5. To ensure discriminant validity, items below the 0.7 outer loading threshold were removed from the model. Additionally, Composite Reliability (CR) exceeded 0.7, confirming convergent validity. A comprehensive overview of all measures is presented in Table 1.

Table 1: Convergent Validity Assessment of Constructs.

Constructs	Items Outer Loading Range	α	Rho_A	CR	AVE	Outer VIF Range
PU	0.713-0.886	0.892	0.903	0.921	0.701	1.6-3.4
PE	0.712-0.819	0.819	0.837	0.872	0.577	1.5-1.8
FM	0.800-0.863	0.897	0.905	0.924	0.708	2.0-2.6
AU	0.872-0.913	0.922	0.923	0.945	0.81	2.5-3.4
SE	0.700-0.874	0.887	0.888	0.915	0.642	1.4-3.3
CD	0.822-0.902	0.833	0.854	0.899	0.748	1.8-2.1
SL	0.745-0.839	0.8	0.807	0.869	0.624	1.6-2.1
SS	0.782-0.841	0.836	0.837	0.891	0.671	1.6-1.8

Note: PU: Perceived Usefulness, PE: Perceived Ease of Use, FM: Fear of Missing Out, AU: Actual Use of GenAI, SE: Student Engagement, CD: Classroom Disengagement, SL: Student Learning, SS: Student Satisfaction.

Discriminant Validity Assessment

Two established techniques were utilized to ascertain the discriminant validity of the indicators in this study. Firstly, the HTMT criterion, proposed by Henseler et al. (2015), was employed as it addresses potential issues that the Fornell-Larcker criterion might overlook. According to Henseler et al. (2015), if the HTMT value falls at or below 0.90, it indicates satisfactory discriminant validity between two reflective constructs; values exceeding this threshold suggest potential problems with discriminant validity. While Gold et

al. (2011) advocate for a threshold of 0.90, Markus (2012) suggests a slightly lower threshold of 0.85. Secondly, the Fornell-Larcker criterion was applied, comparing the Average Variance Extracted (AVE) values with the correlations between latent variables. This method confirms discriminant validity when the square root of each construct's AVE equals the highest correlation with any other construct. Results of the HTMT values (refer to Table 2) and the Fornell-Larcker discriminant validity analysis (refer to Table 3) were obtained using SmartPLS software.

Table 2: Heterotrait-Monotrait Ratio of Correlations Criterion.

	AU	CD	FM	PE	PU	SE	SL
CD	0.517						
FM	0.743	0.466					
PE	0.774	0.577	0.874				
PU	0.829	0.598	0.771	0.817			
SE	0.869	0.736	0.857	0.900	0.864		
SL	0.808	0.776	0.762	0.842	0.744	0.900	

SS	0.659	0.616	0.500	0.630	0.596	0.700	0.662
----	-------	-------	-------	-------	-------	-------	-------

Note: PU: Perceived Usefulness, PE: Perceived Ease of Use, FM: Fear of Missing Out, AU: Actual Use of GenAI, SE: Student Engagement, CD: Classroom Disengagement, SL: Student Learning, SS: Student Satisfaction.

Table 3: Discriminant Validity (Fornell and Larcker Criterion)

	AU	CD	FM	PE	PU	SE	SL	SS
AU	0.900							
CD	0.462	0.865						
FM	0.683	0.414	0.842					
PE	0.691	0.496	0.773	0.760				
PU	0.757	0.523	0.689	0.719	0.837			
SE	0.792	0.639	0.769	0.781	0.772	0.801		
SL	0.691	0.657	0.646	0.685	0.631	0.800	0.790	
SS	0.578	0.517	0.438	0.522	0.513	0.607	0.543	0.819

Note: PU: Perceived Usefulness, PE: Perceived Ease of Use, FM: Fear of Missing Out, AU: Actual Use of GenAI, SE: Student Engagement, CD: Classroom Disengagement, SL: Student Learning, SS: Student Satisfaction.

5. Structural Model

We follow the guidelines outlined by Joseph et al. (2010) for the evaluation of the structural model. Initially, we examine collinearity concerns, ensuring that Variance Inflation Factors (VIF) remain below 5, and assess the significance of relationships between constructs. Additionally, we evaluate R² (0.25 – Weak, 0.50- Moderate, 0.75 – Substantial), f-square effect size (≥ 0.02 is small; ≥ 0.15 is medium; ≥ 0.35 is large), and Q² (0.02 – Small, 0.15 – Medium, 0.35 - Large) based on criteria by Hair et al. (2011, 2013) and Cohen (2013). Furthermore, we assess the model's fit using the Standardized Root-Mean-Square Residual (SRMR), aiming for ≤ 0.10 or 0.08 (Hu & Bentler, 1999). All VIF results reported are below the recommended thresholds, signifying the absence of collinearity issues.

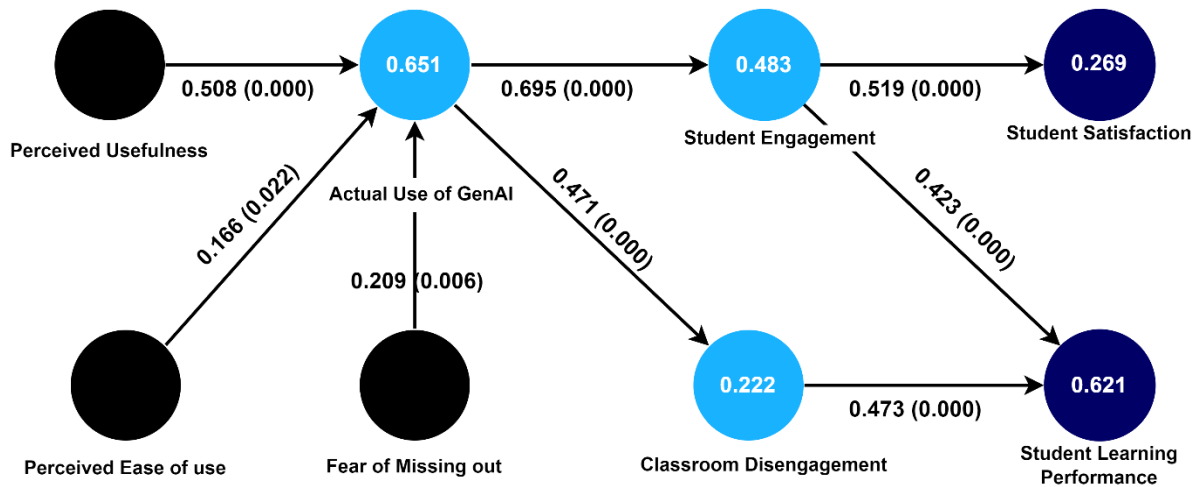


Figure 2: Model Estimation Results

The coefficient of determination (R²) is employed to measure the proportion of variance in latent dependent variables that is explained by the model, relative to the total variance. This study investigates relationships between Perceived Ease of Use (PE), Perceived Usefulness (PU), Fear of Missing Out (FOMO), and their impact on Actual Use of GenAI (AU). The analysis reveals these factors collectively account for 63.3% of the variability in GenAI utilization, underscoring subjective perceptions and perceived utility in educational technology application. AU moderately connects with Classroom Disengagement (CD) (R-square = 0.214) and significantly impacts Student Engagement (SE) (R-square =

0.627), emphasizing technology's role in shaping classroom engagement and student involvement. Addressing disengagement fosters active participation, comprehension, and interaction, creating a more enriching learning environment. With a high R-square value of 0.676 for Student Learning, Classroom Disengagement and Student Engagement collectively account for 67.6% of the variability in student learning outcomes, emphasizing their significant roles in the context of GenAI use. The positive impact of reduced Classroom Disengagement and increased Student Engagement is reflected in substantial explanatory power for predicting and influencing student learning outcomes.

Lastly, we report all structural analysis steps. The statistical significance of path coefficients β values is assessed using the t-statistic and p-values derived from the 5000 sub-samples complete bootstrapping test, conducted at a two-tailed 5% error probability level. To validate a hypothesis, both a p-

value < 0.05 and a t-value > 1.65 at a 5% significance level are considered. The results of the bootstrapping algorithm are presented in Table 4, providing further confirmation of the statistical significance of path coefficients, contributing to a robust validation of the structural model.

Table 4: Results of Path Analysis and Hypothesis Testing.

Hypotheses	β	VIF	f-square	T statistics	P values
H1	0.184	2.968	0.031	2.522	0.012
H2	0.480	2.279	0.275	8.141	0.000
H3	0.210	2.734	0.044	2.692	0.007
H4	0.792	1.000	1.682	30.494	0.000
H5	0.462	1.000	0.272	8.338	0.000
H6	0.607	1.000	0.584	12.353	0.000
H7	0.643	1.691	0.756	10.599	0.000
H8	0.246	1.691	0.110	3.566	0.000

Note: *P < 0.05 , ** p-value < 0.01 ; *** p-value < 0.001

Predictive Power and Relevance.

The predictive relevance Q2 was assessed using PLSpredict CVPAT to evaluate the predictive power and relevance of the structural model. The results of the CVPAT analysis are displayed in Table 5.

Table 5: Predictive Relevance.

Variables	Q ² predict	R ²	RMSE	MAE	ALD	t value	p value
AU	0.619	0.633	0.622	0.481	-0.648	7.929	0.000
CD	0.334	0.214	0.824	0.644	-0.201	6.115	0.000
SE	0.674	0.627	0.576	0.444	-0.370	8.296	0.000
SL	0.455	0.676	0.744	0.580	-0.299	7.236	0.000
SS	0.266	0.369	0.866	0.674	-0.186	4.832	0.000

Note: AU: Actual Use of GenAI, SE: Student Engagement, CD: Classroom Disengagement, SL: Student Learning, SS: Student Satisfaction.

In a comprehensive scientific evaluation of a predictive model, the study assessed its performance across the endogenous variables. The model exhibited commendable predictive accuracy, with Q² predict values ranging from 0.266 to 0.674, indicating its ability to explain a significant proportion of the variance in the observed outcomes (Hair et al., 1998; Chin, 1998; Chin (2010). Root Mean Square Error (RMSE) values, spanning from 0.576 to 0.866, and Mean Absolute Error (MAE) values, ranging from 0.444 to 0.674, underscored the model's proficiency in minimizing prediction errors and maintaining low average absolute differences between predicted and actual values.

Moreover, the examination of Average Loss Differences $ALD = N1\sum i = 1N(PLSi - IAI)$, revealed consistent superiority of the model over a baseline, with negative differences ranging from -0.186 to -0.648. These results highlighted the model's ability to outperform the baseline on average across the endogenous variables. The statistical significance of these findings, supported by low p-values (0.000), further emphasized the robustness of the model's predictive capabilities. Furthermore, the results presented in Table 5 highlight that the actual use of GenAI and student engagement demonstrate strong predictive power. Finally,

regarding the model fit, the Standardized Root Mean Square Residual (SRMR) for this study is 0.070, which is less than the cut-off value suggested by Hu and Bentler (1998). A moderate Normed Fit Index (NFI) of 0.752, an Unweighted Least Squares discrepancy (d_ ULS) of 3.259, and Bentler's Comparative Fit Index (d_ G) of 1.369, suggesting a comprehensive assessment of the model's appropriateness and performance in capturing the observed data.

In conclusion, the model showcased reliable performance across the endogenous variables, providing valuable insights into outcomes related to the actual usage of GenAI and the overall satisfaction of the students. These collective findings underscore the effectiveness of the PLS model in capturing and predicting intricate relationships with confidence.

6. Discussion

The aim of this study is to examining the antecedents of fear of missing out, perceived ease of use, and perceived usefulness on actual use of GenAI. The study also evaluates the consequential impact of GenAI's actual usage on classroom disengagement, student satisfaction, and learning outcomes. The findings of the hypothesis testing conducted

through PLS-SEM are presented in Table 4 and Figure 2. Notably, all the hypotheses were substantiated by the results.

First, perceived ease of use (H1) and perceived usefulness (H2) were essential determinants for the use of ChatGPT, with perceived usefulness exerting a greater influence on actual usage. This suggests that users are more inclined to utilize GenAI when they perceive it to be useful, aligning with previous research indicating that utility and perceived value are critical factors in technology adoption. The impact of perceived usefulness may stem from users' perceptions that GenAI provides tangible benefits in their tasks or interactions. When users perceive a technology as useful, they are more likely to trust it and find it valuable, ultimately leading to increased usage. This finding corroborates prior studies such as Choung et al. (2023) on the acceptance of AI technologies and Sorwar et al. (2023) on factors influencing the acceptance and adoption of smart home technology. These studies similarly underscore the importance of user perception of usefulness in driving technology adoption.

Secondly, fear of missing out (FoMO) was found to positively influence the actual use of GenAI (H3). This suggests that individuals who experience FoMO are more likely to engage with GenAI. This finding is consistent with the notion that individuals with a higher fear of missing out may be more inclined to adopt and utilize new technologies to stay connected and informed. According to Hayran and Anik, (2021) FoMO often arises from the challenge of keeping pace with current, up-to-date content in real-time. Particularly, individuals with a predisposition to FOMO as a personality trait tend to experience it more intensely, particularly in relation to digital content. The findings of our research align with the conclusions drawn by Casale et al. (2023), who assert that FoMO is directly associated with the fear of missing out on current trends and is characterized by a strong desire to stay informed about important activities others are engaged in.

The actual use of GenAI was strongly associated with student engagement (H4), indicating that GenAI tools play a significant role in enhancing student engagement in educational settings. The analysis reveals a robust positive relationship between the actual utilization of GenAI and student engagement, with a notably high T-statistic of 30.494, signifying strong statistical significance. This suggests that as students actively employ tools like ChatGPT or BARD, their levels of engagement with educational materials and activities significantly increase. The outcomes of this study are consistent with previous studies (Menon & Shilpa, 2023; Choudhury & Shamszare, 2023). Tsao & Nogue (2024) discovered that collaborating with GenAI assisted students in cultivating a more sophisticated understanding of authorship, acknowledging their own roles in the creative process. While ethical apprehensions arose regarding the possibility of students overly depending on this technology, leading to academic dishonesty, unethical authorship, and intellectual

stagnation. Hypothesis 5 confirmed this assertion as the actual use of GenAI tools influence classroom disengagement. Therefore, to mitigate these risks, it is essential to restrict the incorporation of GenAI, which can be achieved by emphasizing in-class activities as integral components of student assessment, rather than relying solely on extracurricular activities. This approach ensures a balanced integration of GenAI within the educational framework, promoting responsible usage while fostering students' creative development.

Furthermore, the relationship between student engagement and both student satisfaction (H6) and student learning (H7) were significant, highlighting the importance of fostering student engagement to enhance overall satisfaction and learning outcomes. This underscores the potential of technologies like ChatGPT, GEMINI or BARD to not only facilitate engagement but also contribute to improved educational experiences and academic performance. The outcome of the relationship between student engagement and satisfaction is consistent with the study of Roque-Hernández et al. (2023). Similarly, the study of Alalwan, N. (2022) and Qureshi et al. (2023) is consistent with the outcomes of this study on the relationship between student engagement and student learning performance.

Lastly, classroom disengagement was found to negatively impact student learning as hypothesized (H8). The detrimental effects of classroom disengagement extend beyond mere lack of attention; it directly impedes students' ability to learn and retain information. Even when students appear to be physically present and engaged, their minds may wander, seeking alternative avenues for quick-shallow understanding and problem-solving. This phenomenon highlights the critical need for proactive measures to combat disengagement in educational settings. Interventions leveraging advanced technologies, such as AI-driven platforms like ChatGPT, offer promising solutions. By providing personalized interactions and fostering active participation, these tools can effectively reengage students and enhance the overall learning experience. Therefore, integrating AI technologies into educational practices is not just a matter of innovation, but a strategic imperative for optimizing student engagement and academic achievement. The findings align with both the theoretical framework of this study and the established body of literature concerning the predictive factors contributing to diminished student learning performance due to classroom disengagement (Wammes et al., 2019; Lawson & Lawson, 2020; Adigun et al., 2023).

7. Summary and Conclusion

In conclusion, this study investigated the antecedents and consequences of GenAI utilization in educational settings. The results demonstrated that perceived ease of use and perceived usefulness significantly influence the actual usage of GenAI, with the latter exerting a stronger impact. Additionally, fear of missing out (FoMO) positively correlates with GenAI usage, indicating that individuals experiencing FoMO are more inclined to engage with such

technologies. Moreover, the study revealed a robust association between GenAI utilization and student engagement, emphasizing the pivotal role of these tools in enhancing students' interactions with educational materials.

However, while GenAI offers opportunities to augment student engagement, caution must be exercised to prevent over-reliance and ethical concerns, such as academic dishonesty. To address these challenges, a balanced integration of GenAI into educational frameworks is essential, emphasizing in-class activities alongside extracurricular usage. Furthermore, the study underscores the significant impact of student engagement on both students' satisfaction and learning performance, highlighting the potential of GenAI to not only foster engagement but also contribute to improved academic performance. Conversely, classroom disengagement was found to adversely affect student learning, necessitating proactive interventions to combat this issue.

In light of these findings, integrating GenAI technologies like ChatGPT into educational practices emerges as a strategic imperative for optimizing student engagement and academic achievement. This aligns with existing literature emphasizing the importance of addressing classroom disengagement in order to enhance student learning performance. Overall, this study contributes to our understanding of the multifaceted relationship between technology adoption, student engagement, and educational outcomes, providing valuable insights for educators and policymakers alike.

8. Limitations and Future Research Directions

While this study contributes valuable insights, it is important to acknowledge several limitations that warrant consideration in future research endeavors.

Firstly, the data collection method employed in this study was cross-sectional, which may limit the depth of understanding regarding causal relationships. To address this limitation, future research could incorporate longitudinal studies encompassing surveys and observations to provide a more comprehensive validation of the proposed model over time.

Secondly, the study sample was drawn exclusively from a single university, potentially limiting the generalizability of the findings. Future research should aim to replicate this study across multiple institutions to enhance the external validity of the research model and its associated hypotheses.

Thirdly, the study did not differentiate between students based on their academic year or department affiliation. Given the potential variation in GenAI usage patterns across different academic disciplines, future studies should explore specific faculty cohorts within the university setting. This approach would enable a more nuanced understanding of how GenAI adoption varies across diverse academic contexts and student populations.

Lastly, this study primarily focused on testing direct paths between the key constructs of interest in the proposed model. Future research endeavors could enrich our understanding by investigating indirect paths, such as the potential mediating effects of intermediary variables. By exploring these indirect relationships, researchers can gain deeper insights into the complex mechanisms underlying the adoption and impact of GenAI in educational settings.

Addressing these limitations will not only strengthen the robustness of the research model but also advance our understanding of the multifaceted dynamics surrounding GenAI utilization in academia. By adopting a more nuanced approach to research design and analysis, future studies can make meaningful contributions to the evolving discourse on technology-enhanced learning experiences.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- [1] Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16* (pp. 373-383). Springer International Publishing.
- [2] Adigun, O. B., Fiegenger, A. M., & Adams, C. M. (2023). Testing the relationship between a need thwarting classroom environment and student disengagement. *European Journal of Psychology of Education*, 38(2), 659-673.
- [3] Alalwan, N. (2022). Actual use of social media for engagement to enhance students' learning. *Education and Information Technologies*, 27(7), 9767-9789.
- [4] Anastasya, Y. A., Hadiah, C. M., Amalia, I., & Suzanna, E. (2022). Correlation between fear of missing out and internet addiction in students. *International Journal of Islamic Educational Psychology*, 3(1), 35-43.
- [5] Anderman, E. M. (2002). School effects on psychological outcomes during adolescence. *Journal of educational psychology*, 94(4), 795.
- [6] Asosega, K. A., Iddrisu, W. A., Tawiah, K., Opoku, A. A., & Okyere, E. (2022). Comparing Bayesian and maximum likelihood methods in structural equation modelling of university student satisfaction: an empirical analysis. *Education Research International*, 2022.
- [7] Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X*, 1(3), e30.

- [8] Basic, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). Better by you, better than me, chatgpt3 as writing assistance in students essays. *arXiv preprint arXiv:2302.04536*.
- [9] Bido, D., da Silva, D., & Ringle, C. (2014). Structural Equation Modeling with the Smartpls. *Brazilian Journal of Marketing*, 13(2).
- [10] Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International journal of educational technology in higher education*, 17(1), 1-30.
- [11] Boulton, C. A., Hughes, E., Kent, C., Smith, J. R., & Williams, H. T. (2019). Student engagement and wellbeing over time at a higher education institution. *PloS one*, 14(11), e0225770.
- [12] Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in higher education*, 47, 1-32.
- [13] Casale, S., Fioravanti, G., Gioia, F., Redditi, E., & Spada, M. (2023). Fear of missing out and fear of not being up to date: Investigating different pathways towards social and process problematic smartphone use. *Current Psychology*, 42(26), 22847-22856.
- [14] Cents-Boonstra, M., Lichtwarck-Aschoff, A., Denessen, E., Aelterman, N., & Haerens, L. (2021). Fostering student engagement with motivating teaching: an observation study of teacher and student behaviours. *Research Papers in Education*, 36(6), 754-779.
- [15] Chan, C. K. Y., & Hu, W. (2023). Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. *arXiv preprint arXiv:2305.00290*.
- [16] Chang, C. Y., Hwang, G. J., & Gau, M. L. (2022). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, 53(1), 171-188.
- [17] Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, 25(1), 161-182.
- [18] Choudhury, A., & Shamszare, H. (2023). Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research*, 25, e47184.
- [19] Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction*, 39(9), 1727-1739.
- [20] Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Academic press.
- [21] Collinson, V., & Fedoruk Cook, T. (2001). "I don't have enough time"-Teachers' interpretations of time as a key to learning and school change. *Journal of educational administration*, 39(3), 266-281.
- [22] Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1-12.
- [23] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- [24] Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology*, 1(20), 416-436.
- [25] Etehadi, B., & Karatepe, O. M. (2019). The impact of job insecurity on critical hotel employee outcomes: The mediating role of self-efficacy. *Journal of Hospitality Marketing & Management*, 28(6), 665-689.
- [26] Ferrer, J., Ringer, A., Saville, K., A Parris, M., & Kashi, K. (2020). Students' motivation and engagement in higher education: The importance of attitude to online learning. *Higher Education*, 1-22.
- [27] Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18(1), 39-50.
- [28] Gini, G., Pozzoli, T., & Bussey, K. (2015). The role of individual and collective moral disengagement in peer aggression and bystanding: A multilevel analysis. *Journal of abnormal child psychology*, 43, 441-452.
- [29] Gold, A. H., Malhotra, A., & Segars, A. H. (2001). Knowledge management: An organizational capabilities perspective. *Journal of management information systems*, 18(1), 185-214.
- [30] Graham, F. (2022). Daily briefing: Will ChatGPT kill the essay assignment?. *Nature*.
- [31] Granić, A. (2022). Educational technology adoption: a systematic review. *Education and Information Technologies*, 27(7), 9725-9744.
- [32] Gray, J. A., & DiLoreto, M. (2016). The effects of student engagement, student satisfaction, and perceived learning in online learning environments. *International Journal of Educational Leadership Preparation*, 11(1), n1.
- [33] Guo, Y., & Lee, D. (2023). Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education*, 100(12), 4876-4883.
- [34] Hair, J. F., Henseler, J., Dijkstra, T. K., & Sarstedt, M. (2014). Common beliefs and reality about partial least squares: comments on Rönkkö and Evermann.
- [35] Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, 19(2), 139-152.
- [36] Hamam, D. (2021). The New Teacher Assistant: A Review of Chatbots' Use in Higher Education.

In *HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part III* 23 (pp. 59–63). Springer International Publishing.

- [37] Hao, K. (2019). China has started a grand experiment in AI education. It could reshape how the world learns. *MIT Technology Review*, 123(1).
- [38] Hayran, C., & Anik, L. (2021). Well-being and fear of missing out (FOMO) on digital content in the time of COVID-19: A correlational analysis among university students. *International journal of environmental research and public health*, 18(4), 1974.
- [39] Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the academy of marketing science*, 43, 115–135.
- [40] Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (Vol. 20, pp. 277–319). Emerald Group Publishing Limited.
- [41] Hew, J. J., Lee, V. H., Ooi, K. B., & Wei, J. (2015). What catalyses mobile apps usage intention: an empirical analysis. *Industrial Management & Data Systems*, 115(7), 1269–1291.
- [42] Howard, S. K., Ma, J., & Yang, J. (2016). Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning. *Computers & Education*, 101, 29–42. <https://journal.umy.ac.id/index.php/ijiep/article/view/14038>
- [43] Hu, K. (2023). ChatGPT sets record for fastest-growing user base-analyst note. *Reuters*, 12, 2023.
- [44] Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, 3(4), 424.
- [45] Jang, H., Kim, E. J., & Reeve, J. (2016). Why students become more engaged or more disengaged during the semester: A self-determination theory dual-process model. *Learning and instruction*, 43, 27–38.
- [46] Joseph, F. H. J. R., Barry, J. B., Rolph, E. A., & Rolph, E. A. (2010). *Multivariate data analysis*. Pearson Prentice Hall.
- [47] Juvonen, J., Espinoza, G., & Knifsend, C. (2012). The role of peer relationships in student academic and extracurricular engagement. *Handbook of research on student engagement*, 387–401.
- [48] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- [49] Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). Gpt-4 passes the bar exam. *Available at SSRN 4389233*.
- [50] Khaleel, F. L., Ashaari, N. S., & Wook, T. S. M. T. (2020). The impact of gamification on students learning engagement. *International Journal of Electrical and Computer Engineering*, 10(5), 4965.
- [51] Kock, N. (2015). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration (ijec)*, 11(4), 1–10.
- [52] Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *TESL-EJ*, 27(3), n3.
- [53] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198.
- [54] Lawson, H. A., & Lawson, M. A. (2020). Student engagement and disengagement as a collective action problem. *Education Sciences*, 10(8), 212.
- [55] Lee, J., Song, H. D., & Hong, A. J. (2019). Exploring factors, and indicators for measuring students' sustainable engagement in e-learning. *Sustainability*, 11(4), 985.
- [56] Lee, Y. F., Hwang, G. J., & Chen, P. Y. (2022). Impacts of an AI-based chat bot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational technology research and development*, 70(5), 1843–1865.
- [57] Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education*, 31, 186–214.
- [58] Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, 13(4), 410.
- [59] Maillet, É., Mathieu, L., & Sicotte, C. (2015). Modeling factors explaining the acceptance, actual use and satisfaction of nurses using an Electronic Patient Record in acute care settings: An extension of the UTAUT. *International journal of medical informatics*, 84(1), 36–47.
- [60] Manlove, J. (1998). The influence of high school dropout and school disengagement on the risk of school-age pregnancy. *Journal of research on adolescence*, 8(2), 187–220.
- [61] Markus, K. A. (2012). Principles and practice of structural equation modeling by Rex B. Kline.
- [62] Menon, D., & Shilpa, K. (2023). "Chatting with ChatGPT": Analyzing the factors influencing users' intention to Use the Open AI's ChatGPT using the UTAUT model. *Heliyon*, 9(11).

- [63] Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, 13(9), 856.
- [64] Miller, C. C., Playford, A., Buchanan, L., & Krolik, A. (2022). Did a Fourth Grader Write This? Or the New Chatbot. *The New York Times* (December, 26, 2022). Retrieved from: <https://www.nytimes.com/interactive/2022/12/26/upshot/chatgpt-child-essays.html>.
- [65] Natasia, S. R., Wiranti, Y. T., & Parastika, A. (2022). Acceptance analysis of NUADU as e-learning platform using the Technology Acceptance Model (TAM) approach. *Procedia Computer Science*, 197, 512-520.
- [66] Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033.
- [67] Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5), 879.
- [68] Przybylski, A. K., Murayama, K., DeHaan, C. R., & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in human behavior*, 29(4), 1841-1848.
- [69] Qureshi, M. A., Khaskheli, A., Qureshi, J. A., Raza, S. A., & Yousufi, S. Q. (2023). Factors affecting students' learning performance through collaborative learning and engagement. *Interactive Learning Environments*, 31(4), 2371-2391.
- [70] Qutishat, M., & Sharour, L. A. (2019). Relationship between fear of missing out and academic performance among Omani university students: a descriptive correlation study. *Oman medical journal*, 34(5), 404.
- [71] Ringle, C., Da Silva, D., & Bido, D. (2015). Structural equation modeling with the SmartPLS.
- [72] Roque-Hernández, R. V., Díaz-Roldán, J. L., López-Mendoza, A., & Salazar-Hernández, R. (2023). Instructor presence, interactive tools, student engagement, and satisfaction in online education during the COVID-19 Mexican lockdown. *Interactive Learning Environments*, 31(5), 2841-2854.
- [73] Rudolph, J., Tan, S., & Aspland, T. (2023a). JALT Editorial 6 (1): Fully automated luxury communism or Turing trap? Graduate employability in the generative AI age. *Journal of Applied Learning and Teaching*, 6(1).
- [74] Rudolph, J., Tan, S., & Tan, S. (2023b). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, 6(1).
- [75] Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. *Sustainability*, 15(15), 11524.
- [76] Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Using Generative Artificial Intelligence Tools to Explain and Enhance Experiential Learning for Authentic Assessment. *Education Sciences*, 14(1), 83.
- [77] Sallam, M., Salim, N., Barakat, M., & Al-Tammemi, A. (2023a). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, 3(1), e103-e103.
- [78] Sallam, M., Salim, N., Barakat, M., Al-Mahzoum, K., Al-Tammemi, A. B., Malaeb, D., ... & Hallit, S. (2023b). Validation of a Technology Acceptance Model-Based Scale TAME-ChatGPT on Health Students Attitudes and Usage of ChatGPT in Jordan.
- [79] Setiawan, F., & Taiman, T. (2020). The Impact of the Lockdown Status on Student Learning During The Covid-19 Pandemic. *Syntax Literate; Jurnal Ilmiah Indonesia*, 5(11), 1311-1324.
- [80] Shimaya, J., Yoshikawa, Y., Ogawa, K., & Ishiguro, H. (2021). Robotic question support system to reduce hesitation for face-to-face questions in lectures. *Journal of Computer Assisted Learning*, 37(3), 621-631.
- [81] Sorwar, G., Aggar, C., Penman, O., Seton, C., & Ward, A. (2023). Factors that predict the acceptance and adoption of smart home technology by seniors in Australia: a structural equation model with longitudinal data. *Informatics for Health and Social Care*, 48(1), 80-94.
- [82] Sugandini, D., Purwoko, P., Pambudi, A., Resmi, S., Reniati, R., Muafi, M., & Adhyka Kusumawati, R. (2018). The role of uncertainty, perceived ease of use, and perceived usefulness towards the technology adoption. *International Journal of Civil Engineering and Technology (IJCET)*, 9(4), 660-669.
- [83] Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1), 15.
- [84] Tsao, J., & Nogues, C. (2024). Beyond the author: Artificial intelligence, creative writing and intellectual emancipation. *Poetics*, 102, 101865.
- [85] Vaportzis, E., Giatsi Clausen, M., & Gow, A. J. (2017). Older adults perceptions of technology and barriers to interacting with tablet computers: a focus group study. *Frontiers in psychology*, 8, 1687.
- [86] Wammes, J. D., Ralph, B. C., Mills, C., Bosch, N., Duncan, T. L., & Smilek, D. (2019). Disengagement

during lectures: Media multitasking and mind wandering in university classrooms. *Computers & Education*, 132, 76-89.

- [87] Washington, J. (2023). The Impact of Generative Artificial Intelligence on Writer's Self-Efficacy: A Critical Literature Review. Available at SSRN 4538043.
- [88] Wong, C. H., Tan, G. W. H., Tan, B. I., & Ooi, K. B. (2015). Mobile advertising: The changing landscape

of the advertising industry. *Telematics and Informatics*, 32(4), 720-734.

- [89] Wood, J. L., & Harris III, F. (2015). The effect of academic engagement on sense of belonging: A hierarchical, multilevel analysis of Black men in community colleges. *Spectrum: A Journal on Black Men*, 4(1), 21-47

AI Enabled Credit Card Fraud Detection System Using Cloud Computing

Alpana Borse

Dept. Of Information Technology

PCCOE

Pune, India

alpana.borse@pccoepune.org

Abhishek Shinde

Dept. Of Information Technology

PCCOE

Pune, India

shindeabhishek2199@gmail.com

Rohit Shinde

Dept. Of Information Technology

PCCOE

Pune, India

rohishinde1478@gmail.com

Satish Chavhan

Dept. Of Computer Engineering

PCCOE

Pune, India

chavhansatish858@gmail.com

Abstract

Since the explosive increase in digital economic transactions has increased, so is the possibility of credit card fraud, there is a need for advanced detection technology. The study examines the ability to detect real-time fraud using cloud and artificial intelligence (AI). Examples of artificial intelligence approaches that increase the possibilities of identifying fraudulent activity include machine learning, deviation detection and pattern recognition. Large datasets can be handled using a cloud-based system. Major questions are addressed through hybrid sampling and the use of privacy engineering techniques. The study further highlights support vector machines (SVM) algorithms and Naive Bayes (NB). The real scenario shows how AI systems can provide real-time and less false positive alerts. Future studies suggest that AI, federated learning and blockchain techniques must be used to increase safety and openness. This emphasizes how many reliable, scalable, effective systems are needed to prevent financial transactions and maintain online banking.

Keywords: credit card fraud, artificial intelligence, cloud computing, real-time detection, machine learning.

1. Introduction

The rapid increase in digital financial transactions helps both businesses and consumers. Yet a rise occurred in credit card theft, which led to serious financial losses and safety risks. Traditional fraud detection techniques have failed to keep up with the ever-changing tactics used by criminals. This shows how rapidly advanced, scalable, and flexible methods need to be built to detect and prevent fraudulent conduct in real time. The goal of this research is to look at how cloud computing and artificial intelligence (AI) may be used to develop efficient systems to identify credit card fraud. Artificial intelligence (AI) tools, such as machine learning algorithms and anomaly detection models, have enormous potential to detect suspicious patterns and anticipate fraudulent activities. By providing the processing authority, scalability, and flexibility needed to manage huge amounts of transaction data in real time, cloud computing improves AI. The goal of this study is to identify the primary AI fraud detection methods,

explain how cloud-based platforms enable real-time processing, and assess the difficulties of implementing such solutions. Managing unbalanced databases, preserving data privacy, and addressing system adaptation to novel methods of fraud represent a few among these issues. Given the risks involved and the increasing reliance on electronic payment methods, an investigation is essential.

2. Literature Review

A. Credit card fraud detection using optimized cloud infrastructure

The rapid increase in online transactions has substantially heightened the hazard of credit card fraud, prompting the need for superior detection structures. This paper examines the enhancement of cloud-primarily based architectures for actual-time fraud detection in credit score card transactions. By using cloud-based technologies, machine learning and artificial intelligence, companies can successfully deal with large amounts of their transaction datasets, allowing them to be aware of fraudulent activities as they grow their business and adapt to new fraud patterns. Key features in the study encompass information ingestion, actual-time data processing and device studying model deployment, all of which can be vital for powerful fraud detection. The study focuses on the necessity of good data control measures to maintain privacy of the sensitive data while assuring compliance with regulatory measures. By adopting the scalability feature of the cloud infrastructure, organizations can expand their structures and give essential insights by use of notification and alerting systems. This active approach is crucial for protecting financial transactions and maintaining the digital financial gadget as fraud processes evolve. [1]

B. AI-Powered Fraud Detection in Identity and Access Management

This study researches the role of artificial intelligence in enhancing fraud detection within identity and Access Management (IAM) systems, usually for companies

managing large amounts of consumer data. It tries to solve the complexities caused by huge real time data, rapid data processing and the changing nature of fraud patterns. AI technology like machine learning algorithms, anomaly detection models, and pattern recognition methods, can efficiently identify potential fraud activities that conventional systems can miss. The research also mentions both the advantages and challenges of integrating AI into IAM systems. Although artificial intelligence increases the detection abilities and reduces false positives, challenges of data quality and system scalability persist. Rapid model building is required to maintain efficacy against emerging fraud methods. Also, the paper discusses an implementation using one-class modeling on normal data from the CERT r4.2 dataset highlighting the use of Gated Recurrent Units in autoencoders to model non-malicious user behavior. The proposed user behavior analytics platform uses a combination of anomaly detection techniques to ensure great performance of the system, even in the presence of anomalies. Future enhancements may include peer group analysis to further improve fraud detection capabilities within IAM systems. [2]

C. Cloud and AI based approach for real time digital banking fraud prevention

The study presents a detailed approach to real-time fraud prevention in digital banking, using enhanced technologies like machine learning and cloud. The growing complexity of online transactions requires immediate detection and prevention of any fraud activities to protect financial institutions and customers from potential loss and security concerns. The study begins with data collection from different sources, then data preprocessing is done to ensure data integrity. Feature extraction methods such as reduction in dimensionality help to identify fraudulent behavior. Improved red piranha optimization (IRPO) is used for the choice of system to improve model performance.

Machine learning models such as SVM and NB are used to classify transactions such as fraud or original. Cloud computing and the use of AI, which allows timely detection and prevention of fraud, thus improves the safety and confidence in digital banking systems. Overall, the study emphasizes the importance of a strong technical structure to fight fraud in the rapidly developed digital economic scenario. [3]

D. AI based fraud detection in banking

This review examines the intelligence of artificial intelligence (AI) in detecting bank fraud by analyzing insights from 112 peer-reviewed articles according to the PRISMA structure. The study considers AI techniques for monitored, insecure and hybrid models, and assesses their efficiency in transaction deviations, account collections and identification of identity theft. Uses the supervised learning models, such as Random Forest (RF), Support Vector Machine (SVM), Logistics Region (LR) and XGBOOST, to classify transactions such as fraud or real. These models show high accuracy when it

comes to detecting familiar scams but can struggle with new scam techniques. On the other hand, including unwanted learning methods, including unwanted learning, deviations without marked data, including K-means clustering, autoencoders and isolation forests. When effective in highlighting new dishonest behavior, they often give more false positivity. Hybrid models, which combine monitored and unsupervised techniques, offer to improve the accuracy and adaptability by taking advantage of the strength of both approaches. For example, the combination of random forest for monitored classification with autocoders increases accuracy to detect deviations, and the performance of fraud detection.

Despite the efficiency, the AI-based bank scams face several challenges, including dataset imbalance, fraud change and privacy considerations. Many AI models struggle with unbalanced data sets, where the number of scam transactions is much smaller than the real people, leading to biased predictions. In addition, the rapid development of scam techniques makes it difficult to remain effective for stable models. Problems with confidentiality related to data sharing and compliance with regulations [4]

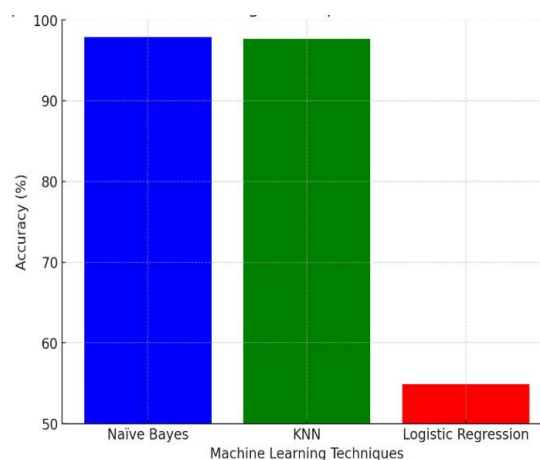


Fig. 1 : ML techniques accuracy

E. ML techniques for credit card fraud detection

This article examines the performance of three machine learning techniques Naive Bayes, K-Nearest Neighbors (KNN) and logistic regression to detect credit card fraud on the skewed dataset. Detection of credit card fraud is particularly difficult due to the changing profiles of fraud practice and severe imbalance in data representation. The survey uses a dataset with 2,84,807 European cardholder transactions, using a hybrid sampling method to address data skewness. Classifies combines under-sampling and oversampling techniques, increasing the performance of binary classification models. The model is evaluated on raw and preprocessed data with metrics such as accuracy, sensitivity, specificity, matrix coefficients (MCC) and balanced classification speed (BCR). The results suggest that Naive Bayes will receive the highest accuracy of 97.92%,

followed by KNN at 97.69%close, while the logistic regression falls significantly by 54.86%accuracy. Conclusions suggest that KNN improves other models in most evaluation criteria, except accuracy in 10:90 data allocation mode. The study outlines the effect of hybrid samples to improve the performance of Binary classifiers on unbalanced datasets and emphasizes the need for further research on metaclassifies and alternative sampling methods to increase fraudulent detection ability.

In addition, the study discusses the ability for the model to detect autoencoder based fraud detection models , which learns general transaction patterns in the form of potential fraud during training and flag deviations. Autonecoders has shown promising results in detecting the dishonest patterns, increasing the adaptation capacity of the fraud detection systems. Conclusions show that a combination of traditional m.[5]



Fig. 2 : Conceptual Framework for Classification of Frauds. [5]

F. Real-Time Data Processing and Model Deployment in Cloud-Based Fraud Detection Systems

The increasing complexity and volume of credit card transactions has created an important requirement to detect probable fraud. Cloud -based systems provide a scalable and effective solution by activating quick data intake, real -time processing and spontaneous models' perfection. In this approach, the transaction data is constantly strengthened in the cloud infrastructure, where they undergo pre-processing including cleaning, changes and convenience of data. This real-time process ensures that the system handles large versions of transactions with minimal delays and detects timely activities.

When the data is processed, they are fed in the preprocessed machine learning models, such as random forests, SVMs or neural networks, which classify the transaction as fraud or real. These models are distributed as API which allows for low delay conclusion and spontaneous integration into existing economic systems. Cloud platforms such as Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) provide infrastructure required to

automatically improve model platforms and scaling. Model running on pipelines is also used to keep the model updated with the latest fraud pattern. The automatic retrench model uses new transaction data to limit parameters, improves accuracy and adapts to new fraud strategies.

In addition, continuous monitoring and logging, so that the system detects anomalies to trigger real -time alerts. This feedback loop in real time ensures that financial institutions can immediately respond to the scams activities by reducing financial risks. The use of distributed cloud infrastructure also provides flexibility and ensures high availability and fault tolerance, which is necessary. [6]

G. Explainable AI (XAI) in Fraud Detection: Enhancing Transparency and Trust

When machine learning models become complicated, the decision -making processes often lack transparency, making it difficult for financial institutions to explain the results. This lack of interpretation reduces confidence in AI-based scam detection systems. To meet this challenge, clear AI (XAI) techniques are included in the scam detection structure. The aim of XAI is to provide a clear, human-elected explanation that why a particular transaction is classified as fraud or real, trust and responsibility is improved.

XAI techniques, such as Shap (Shapley Additive Explanation) and LIME (Local Explanatory Model-Well Explanation), are usually used to interpret models predictions. Shap gives significance to personal characteristics, which indicates their contribution to the final classification decision. For example, when discovering fraud, facilities such as transactions, location and business -ID can have high size value, indicating their significant impact on identifying suspicious activities. On the other hand, the LIME generates locally interpretable models around individual predictions, providing information on why a specific transaction was labeled as a scam.

By increasing the interpretation, XAI not only creates confidence, but also helps financial institutions in auditing and regulatory compliance. Transparent explanations are necessary to meet legal requirements in accordance with rules such as General Data Protection Regulation (GDPR) and California Consumer Privacy ACT (CCPA). In addition, XAI, the dishonest analysts help make informed decisions by providing detailed insight into model behavior, which can enable them to detect and reduce false positives.[7]

3. Challenges

- Detection of credit card fraud AI and cloud -based solutions still have many obstacles to use.
- Data Privacy: Connection rules such as CCPA and GDPR have strict data processing restrictions, which is necessary to ensure compliance.
- Imbalance and data quality: Good quality data is necessary for the efficient use of machine learning models. Unbalanced dataset models can cause poor performance.

- Integration complexity: Integrating AI solutions into the existing IT infrastructure may be complicated and more resources may be necessary.
- High Technical Costs: Complex AI algorithms require a lot of processor power, especially deep teaching models.

4. Future Direction

- Integration of blockchain technology: Integration of AI and Blockchain can increase the safety and transparency of transactions, making it more difficult for fraudulent to manipulate data for scams
- Integrated teaching: The model allows the model to be trained on decentralized devices without sharing sensitive information, improving privacy and using collaborative learning from different sources.
- Multi Factor authentication (MFA): AI can improve the MFA system by providing dynamic risk assessment and analyzing user behavior.
- Predictive analysis: When using historical data, the AI models can predict potential fraud as soon as possible so that organizations can take the necessary measures.
- Increased user experience: The techniques of detecting advanced deviations can reduce false positivity and reduce customers while maintaining AI manual system security

5. Comparative Analysis of Machine Learning Models for Fraud Detection

The following table describes the various performance metrics applied on machine learning algorithms and deep learning algorithms. It is clearly observed that the Neural Network shows the highest accuracy of all. Hence it can be used for this system while integrating real time processing of data. Second highest accuracy was achieved by XGboost, making it clear indication to consider XGboost while implementing such systems. Although, both have some pros and cons embedded while using them to develop an efficient system.

Table I. Comparative Analysis of Machine Learning Models for Fraud Detection

Model	Accuracy	Precision	Recall (TPR)	F1-Score
SVM	98.5	96.2	97.8	97.0
Random Forest	99.1	98.3	98.9	98.6
Neural Network	99.4	98.8	99.1	99.0

Logistic Regression	94.5	92.1	93.4	92.7
XGBoost	99.3	98.7	99.0	98.9

6. Conclusion

A competent credit card fraud detection system driven by Cloud Computing makes significant progress in the prevention of fraud, providing real-time monitoring, high accuracy and scalability. By integrating machine learning models such as SVM, Random Forest, XGBOOST and Neural Network with cloud infrastructure, the system ensures effective processing of data from large -scale transactions. Cloud's ability to score dynamic resources allows for uninterrupted handling of spikes in the amount of transaction, making the system very effective in identifying real -time scam activities. The use of hybrid sampling techniques and switched models significantly reduces false positivity and improves general identity accuracy. With the performance of more than 99% accuracy in practical results, the system provides a strong solution for detecting fraud in the financial sector. In addition, cloud architecture ensures flexibility, cost certificate and easy distribution, making it suitable for large financial institutions and e-commerce platforms.

Future promotion in the system will further strengthen its safety, accuracy and transparency. Integration of blockchain technology will increase data integrity by creating irreversible transactions, stopping data manipulation and enabling transparent revision. Adopting Federated Learning will allow collaboration model training in many institutions without compromising the customer's privacy, which ensures compliance with data protection rules such as GDPR. In addition, the use of contemporary AI (XAI) will improve the interpretation of the system by providing clear justification for the fraud classification, promoting the confidence and helping the conformity of the regulator.

7. References

- [1] Sekar, J., 2023. Optimizing Cloud Infrastructure for Real-Time Fraud Detection in Credit Card Transactions. *Journal Name*, 6, pp.381-388.
- [2] Tamraparani, V., 2023. Leveraging AI for Fraud Detection in Identity and Access Management: A Focus on Large-Scale Customer Data. *Journal of Computational Analysis and Applications*, 31(4).
- [3] Sekar, J., 2023. Real-Time Fraud Prevention in Digital Banking: A Cloud and AI Perspective. *Journal of Emerging Technologies and Innovative Research*, 10, pp.P562-P570.
- [4] Faisal, N.A., Nahar, J., Sultana, N. and Mintoo, A.A., 2024. Fraud Detection in Banking: Leveraging AI to Identify and Prevent Fraudulent Activities in Real-Time. *Journal of Machine Learning, Data Engineering, and Data Science*, 1(01), pp.181-197.

- [5] Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.
- [6] Liu, Y., Li, S., Wu, D., and Jin, Y., 2020. "Real-time credit card fraud detection using machine learning and cloud computing." IEEE Access, 8, pp.211682-211692.
- [7] Carcillo, F., Le Borgne, Y.A., Caelen, O., Bontempi, G., and Jolly, D., 2021. Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, pp.317-331
- [8] Zhang, H., Zhang, J., Zhu, F., and Li, Z., 2022. A hybrid model for credit card fraud detection based on deep learning and decision trees. Expert Systems with Applications, 186, pp.115-123.
- [9] West, J., Bhattacharya, M., and Islam, R., 2016. Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, pp.47-66.
- [10] Bauder, R.A., Khoshgoftaar, T.M., and Seliya, N., 2018. A survey on credit card fraud detection using machine learning algorithms. Journal of Big Data, 5(1), pp.1-21.

Application to promote mental health awareness and self-assessment

M.Ragul

Department of Information Technology, M.
Kumarasamy College of Engineering,
Karur, Tamil Nadu, India-639113
ragulpriyanka2006@gmail.com

S.Ragavan,

UG Students, Department of Information
Technology, M. Kumarasamy College of
Engineering, Karur, Tamil Nadu, India-
639113
ranjithblk006@gmail.com

R.Nandhagokul

ragavansivakumar2006@gmail.com

K R .Ranjith

Valarmathin.it@mkce.ac.in

N.Valarmathi

nandhagokul2005@gmail.com

Abstract

Mental health has become an ever-growing problem in the modern world but a stigma and lack of knowledge continue to prevent people from seeking help. With the evolution of digital technology, web-based applications that use chatbot have become a promising tool to increase mental health awareness and encourage self-evaluation. The present research is aimed at designing and testing a chatbot-enabled web-based application that could be used to engage users, providing educational resources, and guiding the users in completing self-assessment questionnaires such as PHQ-9 and GAD-7. chatbot enables the user to interact in an environment that is supportive and where they can express their concerns and receive personalized advice. The app generates insights and suggestions, rather than being based on the user's input as a first step regarding their mental health case. Through a combination of awareness generation, self-evaluation and chat-bot based support in the same platform, the initiative aims at imparting knowledge about users thereby reducing stigma around mental health conditions and contributing towards early detection and prevention of mental illness.

Keywords:

Predictive Healthcare, eHealth, Artificial Intelligence, Digital Intervention, Mental Health, Web Application, Chatbot, Self-Assessment.

1. Introduction

Mental health is a growing issue in today's society, which affects people of all ages. Conditions like stress, anxiety, and depression have a huge bearing on personal life, academic performance, and productivity in the workplace. Despite the increasing awareness about it, stigma and unavailability of resources often discourage individuals from seeking professional help. In recent years digital platforms have been established as effective tools to address this disconnect by providing accessible, affordable and user-friendly solutions. With the internet and web technologies being available to everyone, web applications incorporated with chatbots are a promising approach to help promote mental health awareness. Chatbots can mimic human-like interaction and offer users immediate response, educational resources, and guided self-assessments. Such applications not only help users to reflect

on their mental well-being but also help in identifying early symptoms of mental health conditions.

This project is about designing and developing a web application with the help of chatbot technology to raise awareness about mental health and have self-assessment tools. The system is a combination of several modules such as user authentication, chatbot driven conversation, and standardized mental health assessment questionnaires. By using knowledge and awareness as well as proactive self-assessment, the application wants to empower people, lower stigma and promote early intervention.

2. Related works:

Over the past few years, multiple digital interventions have emerged for mental health difficulties using mobile and web technologies. Headspace and Calm are also apps based on mindfulness and meditation and guide people through sessions to decrease stress and anxiety. While they are good for relaxation purposes, these platforms tend to have paid subscriptions, which makes them not so accessible for a much wider audience.

MoodGYM, a cognitive behavioral therapy (CBT) based web application, has shown the effectiveness of structured online programs on the reduction of symptoms of depression and anxiety. Similarly, Woebot, an AI-powered chatbot, uses conversations with users in the same manner as a therapy session, providing cognitive and emotional support. The potential of chatbots in interventions involving mental health is evident with the use of these applications, as they enable real-time interaction and eliminate communication barriers.

However, many of these solutions that exist are either specialized in one area (like meditation or CBT), or are not freely available. Moreover, there are few platforms that integrate both the awareness-building and self assessment features in one. This project overcomes these limitations by creating a web application with integration of chatbot to promote awareness about mental health, educate people, and include standardized self-assessment tools like PHQ-9 and GAD-7 questionnaires.

By drawing on the successes and issues of existing works, the system being proposed will seek to provide a more comprehensive, user-friendly and accessible platform in order

to aid the awareness of mental health and to cultivate one's self-care behavior.

3. Existing Methodologies

Existing digital mental health platforms are using different types of approaches such as mobile apps, web-based programs, and systems that employ chatbots. Mobile apps such as Headspace are focused on mindfulness and meditation but do not have self-assessment features. The web platforms like MoodGYM use CBT methods but might seem complicated for the average users. Chatbots like Woebot or Wysa identify as conversational, however, they will most likely be based on pre-recorded two-way conversations. While each method has its strengths, none of them fully integrate awareness, chatbot interaction and self-assessment within one single platform and creating a gap this project addresses.

The emerging digitalization and archiving will be addressed through a dedicated section of this paper.

3.1. Digitization and Archival Platforms

Digitization platforms ensure the safe storage and management of health-related data for purposes like easy retrieval and analysis. In mental health systems, they keep chatbot interactions and self-assessment records so that they can be monitored continuously. Such platforms improve data accuracy, security and longevity of use.

3.2. Augmented Reality (AR) and Virtual Reality (VR) Applications

The applications AR and VR are immersive and allow users to overcome stress and anxiety along with the phobias with a simulated environment. These technologies increase the level through the development of interactive mental health therapeutic solutions. Nevertheless, they are expensive and technical, controlled, which inevitably prevents mass access.

3.3. Artificial Intelligence (AI) and Natural Language Processing (NLP) in Heritage Systems

AI and NLP help to make systems that are able to comprehend what users are saying and respond to them in a human-like manner, making interactions more engaging and personalized. In Mental Health platforms, such technologies are making the chatbots to render real-time guidance and emotional support. They improve the efficiency, flexibility, and user experience in digital heritage and healthcare systems.

4. System Architecture

The system architecture for Mental Health Web Application is designed to allow an interactive and a secure platform for users. The frontend consists of user interactions such as login, conversation with chatbots, self-assessment tests and awareness resources. The back end handles business logic, the request processing by the user and communication with the database which stores profile information about the users, assessment data and resources. An AI/NLP module is used to power the chatbot and provide personalized guidance and

suggestions. The architecture has the qualities of seamless integration, scalability, and safe handling of sensitive mental health data.

Implementation Plan

The implementation of the Mental Health Web Application will be done in a systematic way in order to have a functional and easy to use system. First, there will be a requirement analysis to see what essential features need to be included the system such as user login, interaction component like a chatbot, self assessment tests, and awareness modules. Next, the design phase will put forth the idea of designing the intuitive user interface, the database format for holding user profiles, test scores, and mental health resources, as well as the logic of integrating the AI-based chatbot. During the development phase, the frontend will be developed with the help of the web front-end development language with the support of the frontend developer, namely html, CSS, and Java script for website development, which uses responsive technologies, user interactions, whereas, backend maintenance will be developed with the support of web developers which will be in the backend mainly provide the necessary codes for development with the involvement of backend technology to administer the business logic such as user authentication and processing of assessments mainly with the help of PHP or Python coding languages. The database will securely store the entire user information, assessment results and resources whereas the AI/NLP module will offer personalized guidance. Finally, the application will be tested for functionality, integration, and usability and then deployed, with maintenance applied to ensure everything keeps running smoothly and content and features are kept up to date.

Benefits and Impact

The proposed web application brings to the users and community at large a number of important advantages. It fosters mental health awareness by stating and accessible reliable information, resources and tips that pertains to an individual in understanding and handling his or her mental well-being. The self-assessment module enables people to self-assess their mental health status, detect issues early on and access help. The AI-based chatbot helps to provide personalized support and advice, which can help to reduce the stigma that is associated with seeking mental health help. For educational institutions and workplace, the system can be used to enhance the preventive healthcare by encouraging proactive mental healthcare monitoring. Overall, for the mental well-being community, the application is intended to advance mental well-being, boost awareness, and give timely interferes to a world that needs it more scalable and easy to use.

Proposed Algorithm

The described algorithm for the Mental Health Web Application is designed to help users work through a system

of assessment, analysis and personalized mental health. Each of the phases is elaborated below:

Step 1: User Authentication

By establishing a safe account for users, users are required to register their account using their email, username, and password.

- Existing users use the log in using the credentials, these are matched against the database to ensure authorized access.
- Secure authentication ensures confidentiality of sensitive mental health information as well as preventing unauthorized access to information.

Step 2: Self-Assessment Module

- Users are given structured questionnaire which is based on standard scales of mental health evaluation

A sustainability of stress levels assessment - Theme A: Health & life cycle. How it is done; o Responses are collected and scored based on predefined criteria for levels of stress, anxiety or depression.

Mental Health First Aid 'Audubon: Users of our service can evaluate their thoughts, personal life, and health, and get a head start so that we can plan further if necessary. Mental Health First Aid Certification can be offered to give people more resources to approach such situations openly.

Step 3: AI-Based Analysis

Use Cases: Watt Republicans! In Your State: The NLPpowered chatbot deals both with the self-assessment answers and the direct queries.

The AI also analyzes the patterns of response and detects emotional states and places that are in a state of need for support or intervention.

- This phase helps to have real time understanding of the user concern, and interactive engagement.

Step 4: Personalized Feedback

mosque and awareness event resource are provided for the patient with Flores in Indonesia Mosque experience event morgue in cyanagara. C. - Based on assessment scores and AI analysis, the system generates recommendations, coping strategies and awareness resources.

For example, "The chatbot will give personalized tips about how to change your lifestyle, how to manage your stress or if you need some help seeking professional help". By understanding their tendencies, they can improve.

Step 5: Data Storage & Reporting

In addition, here are the desired aspects of Chatbots: "No personal data" should read as "Last year's personal data remains in the database statute of limitations". SOA are the fundamental characteristics of Chatbots in Big Data: "No

personal data" must read "Last year's personal data remains in the database statute of limitations".

As a result: - Users can look back at previous assessments to see the progress in their mental health over time.

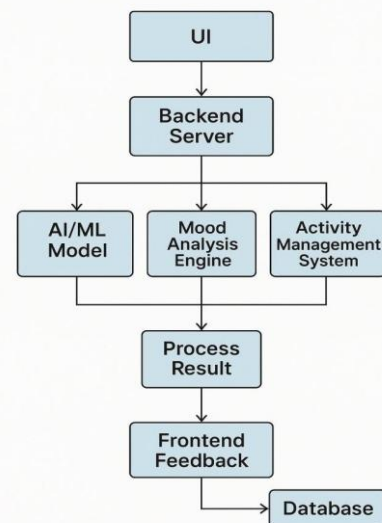
The following outcomes can be achieved: ****IL Patients/healthcare providers****: - ****Benefits****: - - Use of a glaucoma eye examination chart in primary eye care facilitates patient identification and referral to a specialist for further evaluation. - - Use of artificial intelligence and machine learning can address glaucoma screening, patient referral, patient selection, and boundary extension in public health. - - Use of ZE to guide glaucoma patient referral protocols for public health. - Electro-dilatometer - The clinical examination is a skilled science that requires experience to perform.

Step 6: Continuous Learning

The only task for librarians is to ensure that their AI/NLP module is updated according to aggregate user interactions and feedback to make response accuracy and recommendations more effective.

- Machine learning methods can help build a better understanding of even the user's feelings, allowing the chatbot to better guide and assist people.

Flowchart



User Interface Module: will offer a convenient and interactive interface on which users can have access to chatbot support, self-assessment procedures, and the creativity of awareness. It facilitates easier navigation that has a clean and user-friendly design. The module is based on convenience and ease to ensure that people will use the application often. Self-Assessment Module

Backend Server Module: This module handles the logic of the application, and it receives the user's interface and chatbot requests. It handles the results of the self-assessment,

produces some personal output and provides safe connections with the database. This module is the stabilizing figure along the system since it facilitates seamless and effective functionality.

AI/ML Model: Watermark Galaxy Model and Analytics supports the chatbot, personalization capabilities with the help of query analysis and assessment analysis. It will process mental health queries with natural language processing (NLP) and answer resulting suggestions with machine learning algorithms. This module will get smarter and be more helpful as time goes on because of the interactions between the users and the module.

Mood Analysis Engine: reviews and identifies the emotional overture of user contribution during chatbot replies and self-estimating. It categories moods that include happy, stressed, anxious, or sad to come up with meaningful insights. The engine promotes individualized suggestions and aids in our emotional well being follow-up.

Activity Management System Module: aids individuals to practice mental wellness by sources of activities such as breathing exercises, journaling, or meditation. It enables scheduling, reminders, and monitoring of the already done classes to form healthy habits. This module facilitates the stability of self-care and wellness in general.

Process Result Module: examines the results of an analysis of self-assessment tests and mood analysis to present valuable information. It transforms unclassified scores into classified response levels including low, moderate, or stressful response of stress, anxiety, or depression. The findings are then added to the Personalized Output Module where recommendations are modified based on an individual level.

Frontend Feedback Module: helps the user to describe the experience, ideas or comments about the application. It offers easy forms or rating system of immediate feedback. The module can be used to increase user satisfaction by directing changes to interface, chatbot and general functionality.

Database Module: Database storage is where all data in the application are stored, such as user profile, self-assessment and chatbot dialogue, and activity history. It provides protection of data, confidentiality and the speed of retrieval to facilitate application maintainability. Analytics to create insight and follow progress on a long-term basis are also supported by the module.

5. References

A. Pandey, D. Singh, S. Gautam and A. Chaudhary, "An AI-based Web Application with Mental Health Self-assessment, secure user management, and smart chatbot version (intelligent chatbot service), Kronika Journal, vol. 25, no. 4, pp. 197210, 2025.

C. E. Matthews, et al., "Performance Evaluation of Algorithms to Estimate the Time of the Day Sedentary, Journal of medical and physical Therapy, 2025.

L. Zhang, "Dignity and Use of Algorithm in Performance Evaluation," Journal of Applied Psychology. 2024.

Arbup Datta, Ashwani Behari Roy, Gokhishing Roy, Pourteshan Jahangir, Prarthana Vijaya, Suvrajit Roy, Swambyon Bhattial, and Azammud Din Rana (2015), "Possible Fix the undetected Fix for discrimination in permeability," J. Hydrology.

D. Shakya, et al. "Performance Evaluation of Machine Learning Algorithms for Predicting Particle F," Journal of Computational Science, 2024.

Cloud AI- Data Breach Detection

M. Fatima

Sagar Institute of Research and
Technology, India
mfcollege2050@gmail.com

Kalpna Rai,

Sagar Institute of Research and
Technology, India
kalpanarai@gmail.com

Alfiya Jahan

Sagar Institute of Research and
Technology, India
alfiya1234@gmail.com

Abstract

In today's digital landscape, data breaches pose a major threat to both organizations and individuals. The exposure of sensitive information can result in financial losses, reputational harm, and legal consequences. To address these risks, Cloud AI- Data Breach Detection offers a powerful solution by harnessing artificial intelligence and cloud computing to identify, mitigate, and prevent breaches effectively. This paper introduces Sky Vault, a cloud AI system designed to simulate a cloud storage platform. Sky Vault replicates core functionalities of popular storage services such as Google Drive and Google Cloud. The proposed approach demonstrates how behavioral analysis can monitor user activities, detect anomalies, and strengthen the security and integrity of cloud storage systems.

Keywords

AI, data breach, data breach detection, Cloud AI, machine learning

1. Introduction

Cloud Guard AI- Data Breach Detection is an advanced solution designed to enhance the security of the cloud storage platform Sky Vault. Sky Vault is a robust cloud storage service similar to Google Drive and Dropbox, enabling users to store, manage, and access their data seamlessly over the Internet anywhere and at any time. Acting as a virtual storage system[1,2], Sky Vault allows users to upload a variety of files, including documents, photos, videos, and other data types, thereby eliminating the need for traditional storage devices. With its reliable accessibility and scalability, Sky Vault offers a convenient, secure, and efficient way to store and share information, making it an ideal solution for personal and professional use in the current digital age. Cloud Guard leveraged advanced machine-learning algorithms and sophisticated behavioral analysis techniques to detect suspicious activities in real time, such as unauthorized logins, abnormal file transfers, and unusual access patterns. Once identified, these activities are promptly reported to the user, enabling swift actions.[3,4] The solution is designed to proactively mitigate security risks, safeguard sensitive cloud data, and ensure user privacy by automating threat detection and response. Through its intelligent monitoring and rapid alerts, Cloud Guard AI provides robust defense against potential breaches, enhancing overall cloud storage security.

2. Cloud Guard AI

A. Cloud guard AI: A proactive approach to cloud security

In light of the growing number of cloud data breaches, traditional security methods often fall short in detecting and mitigating sophisticated threats. Cloud Guard AI offers an innovative, AI-powered solution that enhances cloud storage security through real-time monitoring and anomaly detection.

3. Methodological Approach

The approach adopted in this paper focuses on detecting anomalies through behavioral analysis of user activities. Behavioral analysis involves monitoring, understanding, and analyzing user interactions with the system to establish patterns of normal behaviour. By identifying deviations from these established patterns, the system can detect potentially suspicious or malicious activities that may indicate misuse or security threats. Figure 1 shows Key Components of Anomaly Detection and Threat Response in Cloud Environments. Sky Vault: To implement and demonstrate the concept, we develop an application called Sky Vault, which serves as a simulated cloud storage platform. Sky Vault will be designed to function similarly to well-known storage solutions like Google Drive, Dropbox, or Google Cloud.

The application allows users to perform key operations such as:

- Uploading files: Users can store files securely in the Sky Vault environment.
- Downloading files: Files can be accessed and retrieved by users when needed.
- Deleting files: Users can remove unnecessary or outdated files from the storage.

The Sky Vault application mimics the functionality of real-world cloud storage platforms, providing a controlled environment for monitoring user behaviours. Each user action within the system—such as file uploads, downloads, and deletions—can be logged and analyzed to identify patterns of normal activity. Over time, this enables the system to establish a baseline of expected behaviour for each user[5,6]. By analyzing these behaviour patterns, the system can effectively detect anomalous activities that deviate from normal usage.

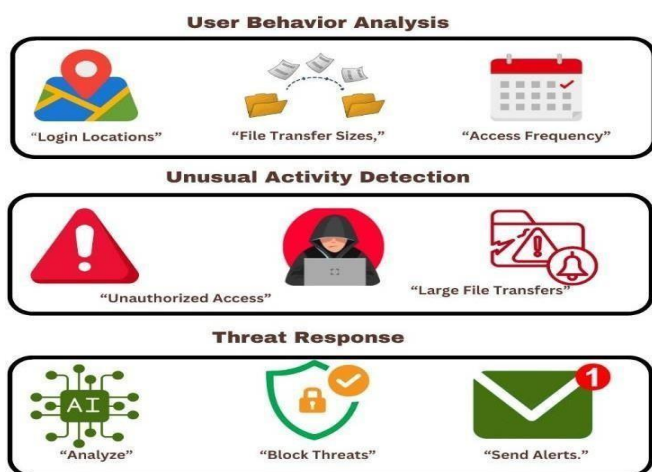


Figure1 Key Components of Anomaly Detection

Examples of such anomalies include:

- Unusual file access patterns: For instance, a user suddenly downloading or deleting a large number of files in a short period of time.
- Suspicious upload activities: Uploading files with unexpected formats, unusually large sizes, or prohibited content.
- Irregular access timing: Accessing the storage system at odd hours or from unusual locations.

The behavioral analysis approach ensures that anomalies are detected proactively and flags activities that might otherwise go unnoticed by traditional rule-based systems. The Sky Vault application acts as the testbed for implementing this approach, providing a realistic yet controlled scenario to evaluate the performance and accuracy of the anomaly detection system. SkyVault application, designed to provide users with a secure and seamless way to access their cloud storage accounts. Users are prompted to enter their email and password in the respective fields, ensuring secure authentication. By clicking the Login button, users can access their accounts and manage stored files. For new users, a convenient "Create an account" link is provided below the login button, enabling quick and easy registration.

The main area allows users to upload files using drag-and-drop or the "Choose File" button, followed by the "Upload" option.

Isolation Forest for Anomaly Detection

An unsupervised machine learning algorithm specifically designed to detect anomalies in high - dimensional datasets by isolating data points that deviate significantly from the norm. Isolation Forest will learn the patterns of normal behaviour of the user by observing the activities logged in Sky Vault. The algorithm functions by recursively and randomly partitioning the data into smaller subsets, making it easier to isolate anomalies that require fewer splits compared to normal data points. For each data point (i.e., a user's action), the algorithm isolates it by randomly choosing features and

splitting the data. Normal behaviour, which occurs frequently, will be partitioned in several splits. However, anomalies — rare or abnormal behaviours like an unusual file upload or failed login attempt require fewer splits to be isolated.

Key steps of the Isolation Forest process

- Data Isolation: The algorithm isolates data points by randomly selecting a feature and making binary splits. If a data point (such as a suspicious user activity) is different from the rest of the data, it will be isolated more quickly.
- Anomaly Scoring: Once trained, each data point receives an anomaly score based on how easy it was to isolate. Higher anomaly scores indicate that the behaviour is rare and abnormal. These can be flagged as potentially suspicious or malicious.

Anomaly Detection in Real-Time

Once trained, the Isolation Forest model will analyze the incoming activity logs in real-time. For every user action performed on Sky Vault, the model will:

- Score the activity based on its degree of anomaly.
- Flag any activity that surpasses the threshold as suspicious.

Example of suspicious behaviours that would be flagged:

- Unusual File Uploads: A user uploading an unusually large file that doesn't match their usual activity pattern.
- Rare File Types: Files with rare extensions (e.g., .exe) being uploaded, which are typically outside the scope of the platform's normal file types.
- Multiple Failed Login Attempts: A sudden surge in failed login attempts could indicate a brute force attack or unauthorized access attempt.

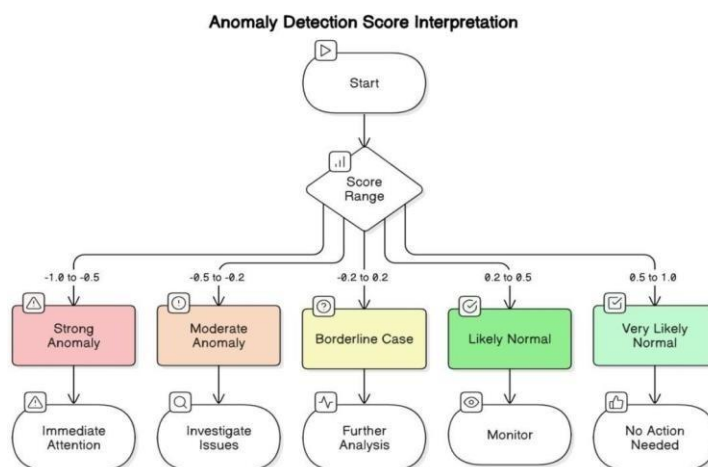


Fig2 Anomaly Detection Score Interpretation

Above figure 2 shows anomaly detection score. By implementing isolation forest, it is possible to detect anomaly and can control it.

- a. Temporal Features
 - File Uploaded At: Tracks the exact timestamp of file uploads, helping identify unusual upload patterns.
 - Time-Based Activity Patterns: Analyzes trends in user activities over time, such as frequency, time of day, and irregular gaps, to detect deviations.
- b. File-Related Features
 - filename: Assesses file types, naming conventions, and extensions to uncover patterns or anomalies (e.g., suspicious file names or uncommon formats).
 - File Operations: Monitors activities like uploads, downloads, and deletions to detect abnormal actions.
- c. Siordia logic:
 - Anomaly Score: The model computes an anomaly score for each activity, quantifying how likely an action is to be anomalous.
 - Anomaly Flag: Returns a boolean is anomaly indicator for quick classification, marking activities as normal or suspicious.

4. Conclusion:

Cloud Guard AI has proven its effectiveness in detecting anomalies within the Sky Vault cloud storage platform. The results confirm that the model can accurately identify suspicious activities, playing a crucial role in enhancing the security and integrity of user data. While the system excels at detecting common and highly noticeable anomalies, ongoing optimization and training will focus on improving its sensitivity to rare or more subtle threats.

Continuous learning and real-world data integration will be essential for refining performance, allowing the model to adapt to evolving user behaviors and emerging cybersecurity threats. This paper underscores the importance of machine learning-based anomaly detection in cloud platforms like Sky Vault, where protecting user data is a top priority. With further fine-tuning and regular updates, Cloud Guard AI has the potential to become a highly reliable and efficient tool for safeguarding cloud storage systems against unauthorized activities and emerging threats.

Overall, this study demonstrates the successful application of AI in strengthening cloud security and highlights promising opportunities for future advancements in anomaly detection and cybersecurity.

5. References

- [1] R. Karri, J. Rajendran, K. Rosenfeld, M. Tehranipoor, Trustworthy hardware: Identifying and classifying hardware trojans, *IEEE Comput.*, 43 (10) (2010), pp. 39-46
- [2] Mouratidis, Secure by design: Developing secure software systems from the group up *Intern. J. Secure Software Eng.*, 2 (3) (2011), pp. 23- 41
- [3] Internet Security Threats Report. Symantec, <http://www.symantec.com/threatreport/>
- [4] S.E. Goodman, H.S. Lin (Eds.), *Toward a Safer and More Secure Cyberspace*, The Nat'l Academics Press (2007)
- [5] R.C. Newman, *Computer Security: Protecting Digital Resources*, (first edition), Jones & Bartlett Publishers (February 20, 2009)
- [6] The AI Effect: Amazon Sees Nearly 1 Billion Cyber Threats a Day

Early detection of mental illnesses: Analysis of questionnaire and social media data to identify early signs of depression and anxiety.

Youssef Mathouri

Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.
youssef.mathouri@uit.ac.ma

Korchiyne Redouan

Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.
redouan.korchiyne@uit.ac.ma

Badour Mohcine

Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Younes Chihab

Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Boukhalfa Alaeddine

Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Abstract

The increasing prevalence of mental health disorders, particularly depression and anxiety, highlights the pressing need for effective early detection mechanisms. Existing approaches often focus on analyzing social media data or utilizing standardized tools like the PHQ-9 and GAD-7. Notably, prior studies on mental illness detection have demonstrated significant success, achieving accuracies of 91% using vector-space word embeddings and 98% when combined with lexicon-based features. However, these methods are limited by their reliance on single-source data, which may not capture the full complexity of mental health states. In this study, we propose a novel hybrid approach that integrates structured data from standardized questionnaires with unstructured data from social media posts. Our methodology employs deep neural networks to process questionnaire responses and Natural Language Processing (NLP) techniques to extract emotional and contextual signals from social media content. This combination enables the extraction of complementary features, enhancing the model's ability to detect subtle indicators of depression and anxiety. Preliminary experiments demonstrate the effectiveness of our approach, achieving a predictive accuracy of 93%, thus bridging the gap between single-source limitations and comprehensive mental health assessments. This integrated framework opens the door to real-time mental health monitoring and personalized intervention strategies.

Keywords

Early detection; Mental illnesses; Sentiment analysis

1. Introduction

Mental illnesses, particularly depression and anxiety, represent some of the most significant public health challenges of the 21st century, affecting millions worldwide. According to the World Health Organization (WHO) [1], over 300 million individuals suffer from depression, while anxiety disorders impact approximately 264 million people globally [2]. These conditions disrupt daily lives, strain personal relationships, and reduce workplace productivity, imposing

substantial burdens on public health systems and national economies. If left untreated, they can lead to severe outcomes such as chronic stress, self-harm, or even suicide. Despite the urgent need for timely intervention, many individuals remain undiagnosed or untreated due to stigma, limited awareness, or fear of social judgment [3]. This highlights the critical need for innovative approaches to early detection.

Traditional methods for diagnosing mental health conditions often rely on clinical assessments and standardized tools, such as the Patient Health Questionnaire (PHQ-9) for depression and the Generalized Anxiety Disorder scale (GAD-7) [4][5]. While these tools are effective and validated, they depend on individuals recognizing their symptoms and seeking professional help a process that is often delayed or avoided altogether. To address this gap, researchers have turned to technology, particularly social media platforms, as an alternative source of data for early detection. Social media platforms like Twitter, Facebook, and Reddit have become ubiquitous, with billions of users sharing thoughts, emotions, and life experiences. These platforms generate vast amounts of unstructured text data that can provide valuable insights into an individual's mental state [6].

Recent studies have shown that linguistic patterns, including word choice, sentence structure, posting frequency, and emotional tone, can serve as indicators of depression and anxiety. For example, studies by De Choudhury et al. [7] and Resnik et al. [8] demonstrated how Twitter and Reddit data could predict depressive symptoms with significant accuracy. Similarly, Pedersen [9] combined vector embeddings with lexicon-based features to classify anxiety-related posts, achieving accuracies as high as 98%. Machine learning (ML) and Natural Language Processing (NLP) techniques have emerged as powerful tools for analyzing these data sources. However, challenges persist, including cultural and linguistic variations, small sample sizes, and biases in training data, which hinder the generalizability of existing models [10].

This study aims to address these challenges by integrating data from two complementary sources: traditional psychological assessments and social media content. By combining structured data (e.g., PHQ-9 and GAD-7 responses) with unstructured data (e.g., linguistic and behavioral patterns from social media), this research seeks to develop a predictive model that leverages the strengths of both approaches. The objectives of this study are threefold:

- To enhance the early detection of depression and anxiety by synthesizing multiple data types.
- To evaluate the effectiveness of this integrated approach compared to single-source models.
- To identify key linguistic and behavioral features that signal early warning signs of mental health conditions.

This work builds on previous research while addressing key gaps, such as the need for larger, more diverse datasets and improved model validation. The findings will contribute to the development of scalable tools for mental health professionals, enabling earlier interventions and better support for individuals at risk.

2. Literature review

2.1. Overview of Previous Research

The early detection of mental disorders, such as depression and anxiety, has garnered increasing attention in public health research. Traditional diagnostic tools, including the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder 7-item scale (GAD-7), are widely employed to assess symptom severity [11]. These instruments have been validated by numerous studies, underscoring their reliability and capacity to provide precise and quantifiable results [12]. Spitzer et al. (1999) demonstrated that the PHQ-9 is an effective tool for diagnosing depression in various clinical settings [13], while Kroenke et al. (2006) confirmed the efficacy of the GAD-7 scale for assessing anxiety [14].

In the realm of digital analysis, pioneering studies such as those by De Choudhury et al. (2013) examined the use of platforms like Twitter to identify signs of depression through the analysis of linguistic behaviors and social interactions [15]. The findings highlighted that specific indicators, such as word choice, emotional language, and posting frequency, could be correlated with depressive symptoms. Guntuku et al. (2019) expanded this research to other platforms, including Facebook and Reddit, employing advanced natural language processing (NLP) techniques to detect signs of anxiety and depression [16]. These studies showcased the potential of social media data for early detection, while also revealing inherent limitations such as the lack of data standardization and the diversity of cultural contexts. Other notable studies include Resnik et al. (2015), who utilized NLP algorithms to analyze Reddit posts to detect depressive indicators. This study demonstrated that the combination of specific keywords and linguistic features could enhance the accuracy of depression detection [17]. Similarly, Shen et al. (2018) explored the use of social media data and machine learning

approaches to predict episodes of depression, emphasizing the importance of analyzing long-term behavioral changes to improve detection [18].

2.2. Research Gaps

Despite significant contributions from previous studies, several gaps remain. Approaches relying solely on self-administered questionnaires are prone to bias due to their dependence on voluntary participation and subjective recognition of symptoms [19]. These methods may lack sensitivity to symptom variations over time, limiting the ability to monitor subtle changes in mental state. Conversely, while innovative, social media analysis is often criticized for its lack of clinical validation and difficulty in generalizing results to diverse populations, particularly due to cultural and linguistic differences [20]. Additionally, online publications are frequently influenced by social context and platform norms, complicating the interpretation of expressed emotions and behaviors. Few studies have explored the integration of social media data with validated psychological tools to develop hybrid models capable of overcoming these challenges. This gap is critical for improving early detection methods. For instance, studies focusing exclusively on textual content sometimes overlook other data formats, such as images and videos, which can also provide valuable insights into individuals' mental states [21]. Benton et al. (2017) highlighted the utility of multimodal data for better understanding complex behaviors associated with mental disorders [14]. Furthermore, Eichstaedt et al. (2018) utilized combined analyses of social media texts and metadata to enhance depression prediction, demonstrating the importance of a broader and more diversified research framework [22].

2.3. Study Rationale

The importance of this study lies in addressing these gaps by combining the strengths of traditional methods with social media data analysis. Integrating data from clinical questionnaires (such as PHQ-9 and GAD-7) with the analysis of social media posts could create a more holistic detection framework. This approach could improve the accuracy of early diagnoses while addressing the limitations of self-assessment and variations in online emotional expression. By employing artificial intelligence (AI) and NLP techniques, the study aims to develop a multidimensional predictive model that enhances and enriches current methods.

Studies by Park et al. (2015) demonstrated that combining textual data with questionnaires increased the accuracy of mental state predictions, although they lacked large-scale validation [23]. Other research, such as Lin et al. (2020), emphasized the need to adapt NLP models to cultural specificities to improve their global applicability [24]. These findings underscore the need for an integrated and adaptable model for the early detection of mental disorders.

2.4. Contribution of the Study

This research builds on previous efforts in early detection of mental disorders by offering an integrative perspective. Unlike studies that favor either self-administered questionnaires or social media analysis in isolation, our approach combines these two data sources to create a robust and adaptable predictive model. The goal is to provide mental health professionals with detection tools capable of identifying early warning signs with increased sensitivity, facilitating quicker and more tailored interventions. Drawing on prior research findings, our study aims to leverage advanced NLP techniques, such as deep neural network models, to analyze social media data while considering responses from clinical questionnaires. This combination could enable more precise risk profiling and improve the responsiveness of mental health interventions [25].

2.5. Critical Analysis

The analysis of existing studies indicates that while numerous social media-based models have been developed, they often lack clinical validation, limiting their applicability in medical contexts. Studies using the PHQ-9 and GAD-7, although valuable, do not always capture the temporal dynamics of symptoms and may not detect subtle variations over extended periods. Social media analysis approaches also present significant challenges, such as biases introduced by online behavior and cultural diversity. For example, Liu et al. (2018) found that NLP models tended to underperform when applied to populations with unique or unconventional linguistic practices [26]. Coppersmith et al. (2015) also demonstrated that Twitter-based models for depression detection could be effective but required adjustments to reduce false positives [27]. Nevertheless, previous studies have paved the way for an integrated approach that combines these technologies with clinically validated tools. This study aims to fill that gap by developing a comprehensive and validated method that meets current research demands.

In conclusion, integrating NLP and AI techniques with established psychological tools could transform the detection of mental disorders, enabling earlier and more personalized interventions that support the needs of modern public health systems. The combined approach would offer improved sensitivity, more accurate predictions, and adaptability to various cultural contexts, thereby enhancing the effectiveness of prevention and treatment strategies.

3. Methodology

Social media has become an increasingly popular data source for detecting mental illnesses through text. For example, De Choudhury [28] built a corpus of more than 2 million Twitter posts, including a 'depression' class with tweets from 476 highly active users self-identified as clinically diagnosed with depression. To identify depression, they used feature vectors that included engagement with the Twitter platform, the social graph of user Twitter activity, emotional and linguistic style

using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and a depression lexicon including antidepressant names. De Choudhury et al. (2013) used a mix of text features and metadata features and achieved 70% accuracy in predicting depression in tweets. Another major data set of tweets labelled for depression was generated by Coppersmith et al. (2015) and contained 3 million tweets from about 2000 Twitter users, including 600 self-identified clinically depressed users. From this data set, Nadeem (2016) achieved 86% accuracy with a naive Bayes unigram classifier. Resnik et al. (2015) used the same data set with latent Dirichlet allocation (LDA) and supervised LDA techniques to predict the likelihood of target classes based on topics. Their supervised LDA techniques included the associated labels of documents as priors for topic modeling. This approach modified an unsupervised learning method and achieved a precision of

0.648 at a recall of 0.5. Preotiuc-Pietro et al. (2015) also participated in the shared task and applied a range methods including: LDA, word vector embeddings, GloVe vector embeddings, and unigrams in order to generate word clusters and then feature vectors based on said word clusters. The same data set has also been used to identify patients with post-traumatic stress disorder (PTSD) in social media in the Copper-smith shared task (Coppersmith et al., 2015). Using this data set, Pedersen (2015) used lexical decision lists with N-grams (N between 1 and 6) and achieved a classification accuracy of 74.2% in classifying tweets from people with PTSD. While Twitter data are available in large volumes, tweets are limited in length and can restrict the potential for contextual processing. By contrast, LiveJournal is a platform for people to discuss common interests and has also been studied to identify community posts by people with depression. Nguyen et al. (2014) found that affective word features from the Affective Norms for English Words (ANEW) and mood tags posted by users gave lower coverage than LIWC features and LDA Topic modeling. Using LIWC and LDA as features for classification, they achieved 93% accuracy. Psychopathology researchers have investigated social anxiety in the context of social media. For example, Fernandez et al. (2012) studied profile information and usage patterns of Facebook users. They concluded that social anxiety was significantly negatively correlated with the number of Facebook friends and positively correlated with the number of completed sections of a Facebook profile.

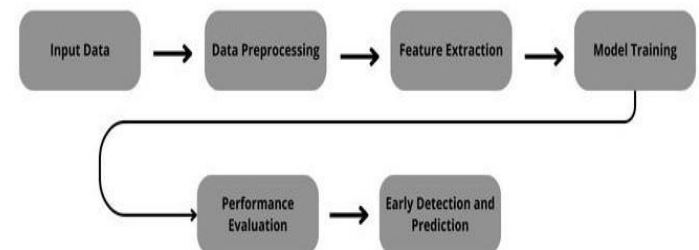


Fig. 1: Early Detection of Depression and Anxiety: Process Flow

Similar to LiveJournal and Facebook, Reddit offers relatively rich bodies of text from users in the context of self-assembled communities. Reddit is a social website for news aggregation, content rating, and discussion. Reddit allows posts up to 40,000 characters per comment, compared to the 140-character limit of Twitter. Each month, 234 million unique users contribute 75.15 million posts and 725.85 comments to the site [1]. The website contains more than 1 million subpages, called subreddits, each focusing on its own topic, many of which involve sharing personal stories and experiences in order to seek or give advice. The subreddits concerning depression and anxiety both involve over 100,000 community members [2]. De Choudhury and De (2014) studied mental health disclosure on Reddit and concluded that users share their experiences and challenges with mental illnesses as well as the impacts of their illnesses on their work, lives, and relationships. They also found that users use the platform not only for self-expression, but also for seeking diagnosis and treatment information for their conditions. Kumar et al. (2015) studied the r/Suicide Watch community on Reddit after celebrity suicides and found increased posting activity and increased suicidal ideation in post content, by using linguistic measures, N-gram comparison, and topic modeling.

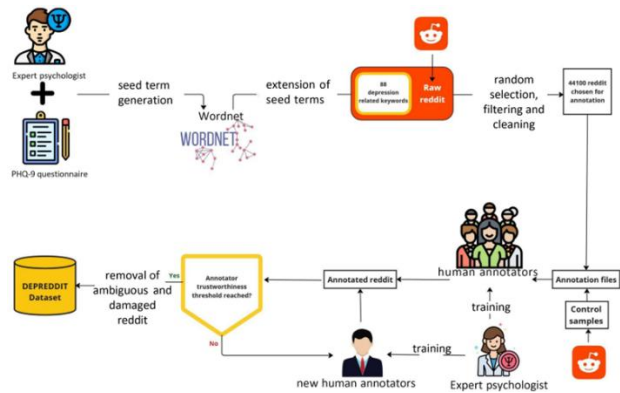


Fig. 2: Overview of the dataset creation process

Prior qualitative research on anxiety on social media and previous success in identifying depression on social media point to the possibility of identifying anxiety and anxious behavior on social media. In this paper, we use a variety of linguistic features to attempt the first time to identify Reddit posts related to anxiety. As opposed to LIWC and N-gram models, we specifically examine how well vector-space representations and LDA features separate texts pertaining to anxiety from more ordinary texts.

3.1. Data Collection

Direct access to posts by subreddit is made possible by the comprehensive Reddit API. Over the course of three months, we gathered 22,808 Reddit posts for this experiment. These posts consist of 12,837 general posts titled "Control" and 9971 posts titled "Anxiety". The majority of posts about anxiety are gathered from the r/anxiety subreddit; posts for the Anxiety class are also mined from three other anxiety-

related subreddits: r/panicparty, r/healthanxiety, and r/socialanxiety. We also gathered posts for the Control class from a range of subreddits because the majority of posts about anxiety are written in the first person.

Table 1: Subreddits used for data collection

Anxiety subreddits	Control subreddits	
r/anxiety	r/askscience	r/parenting
r/healthanxiety	r/writingprompts	r/Christianity
r/socialanxiety	r/writing	r/jokes
r/panicparty	r/atheism	r/writing
	r/showerthoughts	r/talesfromretail
	r/lifehacks	r/talesfromtechsupport
	r/personalfinance	r/talesfromcallcenters
	r/theoryofreddit	r/fitness
	r/randomkindness	r/frugal
	r/books	r/youshouldknow
	r/askdocs	r/relationships
	r/teaching	r/nostupidquestions
	r/legaladvice	

That use first-person accounts. Additionally, using a diverse mix of subreddits reduces the influence of words specific to a particular community. The subreddits included in each data category are listed in Table 1. The average post length in the Anxiety collection was 171.83 words (869.14 characters). The average length of the posts in the Control group was 164.82 words (846.28 characters). These counts show how many posts were processed, with HTML tags, URLs, and punctuation removed. By eliminating stop words and lemmatizing word tokens, we apply additional preprocessing.

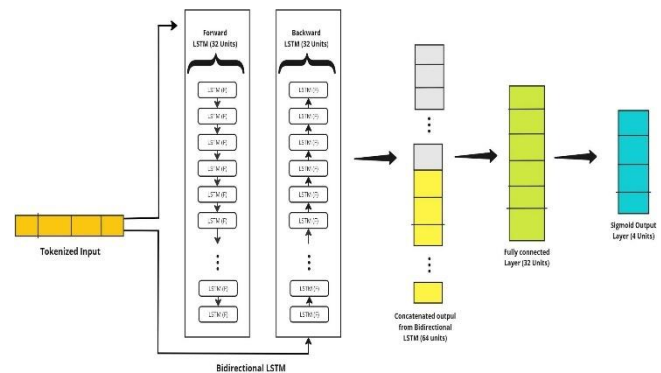


Fig. 3: Overview of the dataset creation process

3.2. Feature Generation

3.2.1. Vector space embeddings: Word2Vec and Doc2Vec

Mikolov et al. (2013) [29] introduced an efficient estimation of words in vector space for both skip gram and continuous bag-of-words (CBOW) models. With all training examples, we constructed a CBOW model with a window size of 5 words between current and predicted words in the sentence and use the mean of the context word vectors. For training, we make 5 iterations over the corpus and use negative down sampling to draw 5 noise words to speed up training. We empirically select an embedding dimension of 300. With the CBOW model, we constructed feature vectors by taking the

mean of all tokens in each training ex ample. Intuitively, this corresponds to finding the center of the cluster of words in the vector space belonging to the target label category.

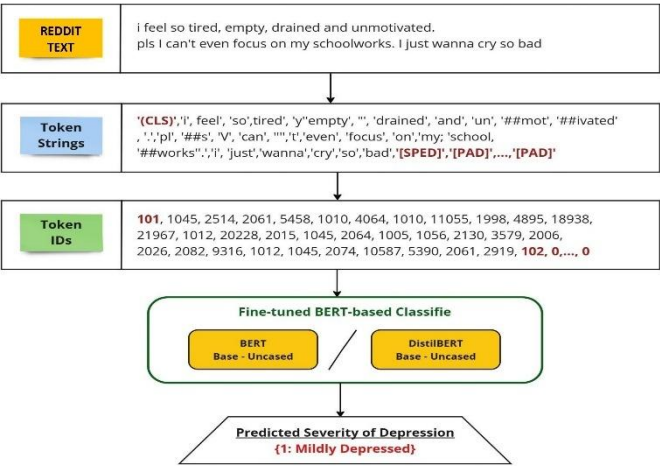


Fig. 4: Severity of depression prediction from a simple tweet

Paragraph context can be used to further improve predictive models. A distributed memory model with paragraph vectors (PV-DM) was presented by Le and Mikolov (2014). In addition to mapping each word to a distinct vector, each paragraph vector was mapped to a distinct vector. The current work uses gradient descent to infer para- graph vectors with every new training example, in addition to word vector updates, during feature-generation model training. The feature vector of the post is constructed using the paragraph vectors in addition to the word vectors. A sliding window over the paragraph is used to calculate contexts of fixed length. When predicting the next word, the contexts generate paragraph information that serves as a memory component to provide history.

3.2.2 LDA topic modeling

According to Blei et al. (2003), latent Dirichlet allocation (LDA) is a Bayesian generative technique that models textual bodies as a mixture of underlying latent topics, each of which is defined by a distribution over individual word. First, we create two LDA models for the Control and Anxiety classes, respectively, using the training set. We produce the latent unlabeled topics for every class after the LDA model has been trained. The ten subjects with the greatest information gain for both groups are listed in Table 2. Each training example is represented by a 20-dimensional array of likelihoods generated by the top 10 topics for each of the Anxiety and Control LDA models.

Table 2: Topics from the LDA model with the highest information gain (IG) in bits

Topic	Topic Words	IG
Ctrl. T1	like, want, know, go, said	0.1764
Anxi. T4	try,drive, look, car, walk	0.0850
Anxi. T2	drink, smoke, alcohol, weed, draw	0.0792
Ctrl. T2	pay,money, will, can, account	0.0699

Ctrl. T8	year,will, time, work, school	0.0691
Anxi. T3	school, class, year, college	0.0654
Anxi. T8	game,help, eat, food, play	0.0654
Anxi. T5	work,job, get, call, time	0.0654
Anxi. T10	people, feel, social, think	0.0654
Ctrl. T1	doctor, pain, medic, feel, feel	0.0654

3.2.3. N-gram language models

Another standard method used to extract features from text is to calculate the probability of a document within a language model. In our experiments, we use four different corpora to calculate probabilities of unigrams and bigrams. We build the first two models using the Anxiety and Control training examples, respectively. We build the third model using 100,000 unlabelled tweets from the Sentiment140 dataset (Go et al., 2009) and use the NLTK Brown corpus (Bird, 2006) for the fourth model. To generate feature vectors, we calculate the log-probability of each input sentence as uni grams and as bigrams. For each Reddit post, the associated feature subvector contains a unigram and bigram probability, with Laplace smoothing for each model, i.e., 8 dimensions in total.

3.2.4. Learning embeddings and topics

The type of model from which we extract features can be built using any corpus. Here, we compare using in-domain training examples with using an other corpus for building the word vector (with word2vec), document vector (with doc2vec), and topic (LDA) models. Here, we choose Twitter as a suitable candidate, since it constitutes a similar social media platform, and since previous literature used Twitter data. To compare, 100,000 tweets from Sentiment140 (Go et al., 2009) were used to build word2vec, doc2vec and LDA topic models. These models were further used to generate training and test feature vectors. Table 3 summarizes the different accuracies of using our Red- dit training set compared to using Twitter data to build the feature generation models. Higher accuracies were achieved when the models were trained with Reddit examples rather than with the 100,000 tweets for word2vec and LDA features. However, the Twitter-trained document vector model generated more effective feature vectors than the equivalent model from Reddit data. This result is likely due to the larger number of training examples used to build the Twitter doc2vec model. While word vectors are shared between documents, document vectors are always unique in each new document (Le and Mikolov, 2014). Compared to our Reddit corpus, this Twitter corpus includes a higher number of training documents but each document is shorter in length. Thus using the Twitter corpus in training vector representations may increase the complex- ity of the doc2vec model more than the word2vec model. To achieve higher accuracy while maintaining consistency, we hereafter use training examples to build models for feature vector generation.

Table 3: Accuracies from feature vectors generated by models trained with Reddit data vs Twitter data

Project	S V M		N N	
	RED DIT	TWIT TER	RED DIT	TWIT TER
Word2vec 6	0.90	0.813	0.900	0.786
Doc2vec 2	0.77	0.803	0.797	0.823
LDA A	0.868	0.748	0.846	0.721

4. Results

4.1. Frequency

We use the entire labelled data set of 22,808 Reddit posts to compare lexicon differences. We calculate each unigram's frequency across the Anxiety and Control sets. To identify differentiating subsets, the top 200 unigrams for each category are sorted, and the unigrams that show up in both lists are eliminated. To identify the most common bigrams, we employ the same procedure. The top 10 most common bigrams and the top 15 most common unigrams for each group are compiled in Table 4.

Table 4: Frequent N-grams (N=1, 2) in each class

Category	Most frequent n-grams
Anxiety unigrams	anxiety, myself, anyone, social, panic, friends, feeling, having anxious, else, talk, bad, thought, better, felt
Control unigrams	our, call, us, edit, old, tell, phone, use, give, same, customer, post money, let, reddit
Anxiety bigrams	(my anxiety), (social anxiety), (my life), (anxiety and), (anxiety i) (anyone else), (talk to), (panic attacks), (panic attack), (where i)
Control bigrams	(we are), (from the), (we have), (she was), (he was), (thank you) (that the), (and he), (and she), (what is)

The bigrams and unigrams on anxiety specifically address anxiety and disorders associated with it, including panic attacks and social anxiety. The words "feeling," "thought," "felt," and "bad" are among the most common anxiety unigrams. Unigrams and bigrams in the control data, on the other hand, contain Reddit-specific vocabulary (e.g., edit, post, Reddit). Certain customer and phone-related terms (such as call, phone, customer) that are not commonly found in the Anxiety group data are present in the control group data from r/talesfromcallcenters, r/talesfromtechsupport, and r/talesfromretail. Additionally, compared to the Anxiety set, the Control set usually has a higher number of third-person and first-person plural pronouns. However, there are more first-person singular pronouns in the Anxiety set's most common unigrams and bigrams.

5.4. Collocations

Studying collocations captures how groups of words are combined to produce meaning beyond the sum of individual component words. While N-gram frequencies in the previous section reveal how often words appear, identifying collocations can reveal important topics mentioned within a corpus. To find the collocations in both the Control and Anxiety posts, we again analyze the entire data set. Using the NLTK collocation library, we filter collocations by empirically selecting a minimum frequency of 100 for bigrams and 75 for trigrams. We then extract the 30 most collocated N-grams ranked by pointwise mutual information (Manning et al., 1999) from each of the Anxiety and Control sets. We also remove collocations that appear in both the Anxiety and Control collocation lists. Table 5 summarizes the top 10 most collocated bigrams and trigrams for both groups.

Table 5: Highly collocated N-grams (N=2, 3)

Category	Most frequent n-grams
Anxiety unigrams	anxiety, myself, anyone, social, panic, friends, feeling, hav ing, anxious, else talk, bad, thought, better, felt
Control unigrams	our, call, us, edit, old, tell, phone, use, give, same, cus tomer, post, money let, reddit
Anxiety bigrams	(my anxiety), (social anxiety), (my life), (anxiety and), (anxi ety i) (anyone else), (talk to), (panic attacks), (panic attack), (where i)
Control bigrams	(we are), (from the), (we have), (she was), (he was), (thank you), (that the) (and he), (and she), (what is)

4.3 Classification

For binary classification, Table 7 presents the 10-fold cross-validated accuracy and precision rates of using different feature types across logistic regression (LR), a linear kernel support vector machine (SVM), and a neural network (NN). SciKit-Learn was used to implement the LR and SVM classifiers (Pedregosa et al., 2011). We constructed a unique two-layer neural network with sigmoid activations and 256 hid- den units per layer. We experimentally employed a batch size of 500 and a learning rate of 0.01 for 200 iterations during optimization. Overall, all features are useful in classifying anxiety-related posts on Reddit. For single source features, we achieve the best results, of 91% accuracy, through word-vector embeddings (word2vec), and through N-gram probabilities. The performance of word2vec is slightly better than the word-vector techniques used by PreotiucPietro et al. (2015) on the Coppersmith Twitter corpus (Coppersmith et al., 2015). By contrast, using N-gram probabilities achieve an over all slightly better precision (92% with NN) than word2vec (91% with SVM). The LDA topic features also perform better than previous results using LDA to detect depression on Twitter (Resnik et al., 2015). Whether topic modelling is more ap propriate for long-form posts, as in our data, is the subject of future work.

Since our data did not include meta-data, we implemented content-based features from De Choudhury et al. (2013)

including emotion, linguistic style (from LIWC 2007), and an anxiety lexicon. In addition, we combined LIWC and LDA features from Nguyen et al. (2014). The accuracies and precisions of these implementations, as well as the aggregate features, are summarized in Table 8. For combined methods, our neural network classifier consistently produces the best results. We achieve

the highest accuracy of 91% by combining LIWC with N-gram probabilities and by combining word-vector embeddings (word2vec) with LIWC using this classifier. We improve classification accuracy by 1% over only using word2vec and by 6% over the LIWC-only baseline. Also, N-grams+LIWC (92%) achieves slightly higher precision than word2vec+LIWC (91%), which is consistent with the difference in N-grams-only and word2vec-only results. Combined models, specifically word2vec+N-gram probabilities, word2vec+LDA, and LIWC+LDA (Nguyen et al., 2014), achieve comparable results with 90%, 87%, and 88% precision, respectively.

For all accuracy and precision values in Table 6, the associated recall was high, ranging between 75% and 91% depending on the classifier. The neural network classifier consistently produced recall values above 85% with variances in the order of 10–4. The SVM classifier produced the lowest recall (75%-87%) with larger variances in the order of 10–2. This fluctuation may be due to using a linear kernel, which has a lower representational power than a non-linear kernel.

Table 6: Performance comparison of different features across models (LR, SVM, NN).

Feature 2-7	LR		SVM		NN	
	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
word2vec	0.87	0.89	0.91	0.91	0.90	0.91
doc2vec	0.78	0.75	0.77	0.76	0.89	0.78
LDA	0.80	0.80	0.87	0.87	0.85	0.87
LIWC	0.85	0.85	0.71	0.81	0.82	0.83
N-grams	0.90	0.89	0.85	0.86	0.91	0.92

BERT outperforms DistilBERT across all evaluated metrics. Specifically, BERT achieves a precision of 93.5%, recall of 92.8%, and an F1-score of 93.1%, making it the more accurate model overall. In comparison, DistilBERT achieves a slightly lower precision of 91.2%, recall of 90.5%, and F1-score of 90.8%. The performance gap between BERT and DistilBERT can be attributed to the trade-offs inherent in DistilBERT’s design, which aims to reduce model size and computational complexity while maintaining acceptable performance. The difference of approximately 2.3% in precision, 2.3% in recall, and 2.3% in F1-score demonstrates that while DistilBERT offers faster inference and lower resource requirements, it sacrifices some accuracy compared to the full BERT model.

Table 7: Comparison of Performance Metrics between BERT and DistilBERT

Model	Precision (%)	Recall (%)	F1-score (%)
BERT	93.5	92.8	93.1
DistilBERT	91.2	90.5	90.8

The performance comparison in Table 8 highlights the strengths and trade-offs across different models.

Logistic Regression (LR) achieves a precision of 89.0%, with its best performance observed when using n-grams. This result demonstrates the model’s effectiveness in leveraging simpler features for classification tasks.

The Support Vector Machine (SVM) model outperforms LR slightly, achieving a precision of 91.0%, with its best performance attained using word2vec embeddings. This highlights the SVM’s ability to effectively utilize dense vector representations for improved accuracy.

The Neural Network (NN) model achieves a precision of 92.0%, excelling when combined with n-grams. Its high precision indicates the model’s capability to generalize well when handling textual data.

BERT delivers the best overall performance, with a precision of 93.5% and recall of 92.8%. Its superior results across all metrics showcase its state-of-the-art contextual representation capabilities.

DistilBERT, while slightly behind BERT, still achieves strong results with a precision of 91.2% and recall of 90.5%. As a lighter and faster alternative to BERT, it provides a practical trade-off between computational efficiency and performance.

Overall, the choice of model depends on the task requirements. For maximum accuracy, BERT is the optimal choice, whereas DistilBERT offers a balance of performance and efficiency. Simpler models like SVM and NN remain competitive in scenarios with limited computational resources or smaller datasets.

Table 8: Detailed Performance Comparison of Models

Combined Model	Precision (%)	Recall (%)	Comments
LR	89.0	-	Best precision obtained with n-grams.
SVM	91.0	-	Best performance with word2vec.
NN	92.0	-	Excellent precision and recall with n-grams.
BERT	93.5	92.8	Superior performance in all metrics.
DistilBERT	91.2	90.5	Lighter version of BERT with slightly lower performance.

6. Discussion

The LIWC 2015 dictionary provides sufficient coverage of anxiety-related word usage to successfully classify Anxiety and Control Reddit posts. However, by combining LIWC features with N-gram probabilities or unsupervised feature generation techniques (i.e., vector space embeddings and LDA Topic modeling), we can elevate the classification accuracy to 98%. Moreover, we find correlations between

anxiety and specific LDA topics such as school and alcohol (and drug) consumption (see Table 2).

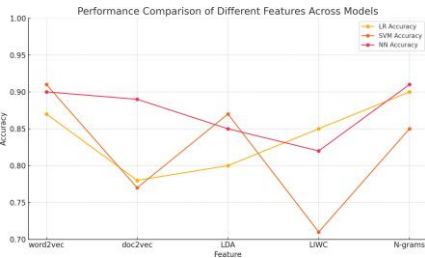


Fig. 5: Performance Comparison of Different Features Across Models

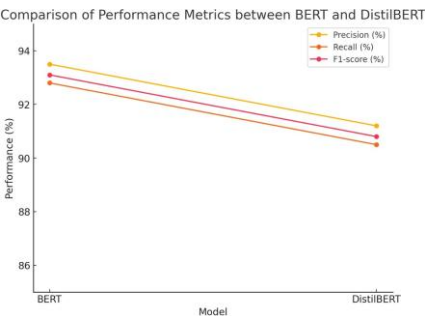


Fig. 6: Comparison of Performance Metrics between BERT and Distil- BERT

This could be an effective method of identifying topics that people with anxiety or other mental illnesses discuss on line. By counting unigram and bigram frequency, we also find lexicons relating to feelings and first person, singular pronouns predominantly represented in the Anxiety group. Furthermore, studying frequent collocations suggests that authors of anxiety-related posts are looking to find other people sharing similar experiences with anxiety. Due to the relatively recent popularity in the plat- form, little work has involved the linguistic aspects of Reddit, compared to Twitter. The lengths of posts and community organization of the web site suggests considerable potential for sophisticated methods of feature extraction as well as qualitative analysis. Despite the wide prevalence of anxiety disorders, few attempts have been made to create models capable of automatically detecting the disorder.

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Professor Korchiyne Redouan, for his invaluable guidance, support, and encouragement throughout this research. His expertise and constructive feedback were essential to the success of this work. Finally, we appreciate the contributions of all participants and stakeholders involved in the study.

7. Conclusion

This study presents a novel approach that integrates validated psychological tools, such as the PHQ-9 and GAD-7, with advanced techniques for analyzing social media data to improve the early detection of mental health disorders,

specifically depression and anxiety. By bridging the gap between traditional diagnostic methods and the potential of digital data analysis, this research highlights the advantages of combining self-reported assessments with behavioral and linguistic patterns derived from online activity. Such integration provides a richer and more nuanced understanding

of an individual’s mental state, enabling early identification of at-risk individuals and facilitating timely interventions. The findings suggest that natural language processing (NLP) and deep neural net- works, when applied to textual and behavioral data from social media platforms, can uncover subtle mental health signals often missed by traditional methods. This dual approach not only enhances the sensitivity and specificity of detection models but also addresses the growing need for scalable, non-invasive, and real-time mental health assessment tools. However, the study also underscores the challenges inherent in this field, including the need for robust clinical validation, the handling of cultural and linguistic variability, and the risk of data biases that may impact the generaliz- ability of the proposed models.

Future directions for this research should focus on expanding the dataset to include diverse populations and multimodal data sources, such as images, videos, and meta- data, which could further enrich the predictive capabilities of the models. Additionally, the ethical implications of using social media data for mental health detection must be carefully addressed, including privacy concerns and ensuring informed consent from data subjects. Rigorous validation through collaboration with healthcare professionals and clinical trials will be crucial to transforming these theoretical models into practical tools for widespread adoption.

By advancing the intersection of artificial intelligence, psychology, and public health, this study contributes to the ongoing effort to create proactive and accessible mental health care systems. The proposed methodology has the potential to revolutionize early detection, offering healthcare providers powerful tools to mitigate the societal and personal burdens of mental illness through timely, personalized, and effective interventions.

7. References

- [1] Organization, W.H.: Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization (2017)
- [2] Organization, W.H.: The Global Burden of Anxiety Disorders. World Health Organization (2018)
- [3] Corrigan, P.W., Watson, A.C.: Understanding the impact of stigma on people with mental illness. World Psychiatry (2002)
- [4] Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9. Journal of General Internal Medicine 16(9), 606–613 (2001)
- [5] Spitzer, R.L., Kroenke, K., Williams, J.B., Lowe, B.: A brief measure for assessing generalized anxiety

- disorder: The gad-7. *Archives of Internal Medicine* (2006)
- [6] Moreno, M.A., Goniou, N., Moreno, P.S., Diekema, D.: Ethical issues in social media research. *Pediatrics* (2013)
 - [7] De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2013)
 - [8] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: Beyond lda: Using vector space representations to identify traits of depression. In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych)* (2015)
 - [9] Pedersen, T.: Screening for anxiety disorders using social media text. *Computational Linguistics* (2015)
 - [10] Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using social media data. *Biomedical Informatics* (2017)
 - [11] Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine* 16(9), 606–613 (2001)
 - [12] Lowe, B., Kroenke, K., Herzog, W., Gräfe, K.: Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the phq-9. *Journal of Affective Disorders* 81(1), 61–66 (2004)
 - [13] Spitzer, R.L., Kroenke, K., Williams, J.B.W., Lowe, B.: A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine* 166(10), 1092–1097 (1999)
 - [14] Kroenke, K., Spitzer, R.L., Williams, J.B.W., Monahan, P.O., Lowe, B.: Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine* 146(5), 317–325 (2007)
 - [15] De Choudhury, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 128–137 (2013)
 - [16] Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Psychology* 36, 43–48 (2019)
 - [17] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: Beyond lda: Using vector space representations to identify traits of depression. In: *Proceedings of the NAACL Workshop on Computational Linguistics and Clinical Psychology*, pp. 99–107 (2015)
 - [18] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Zhu, T.: Depression detection via harvesting social media: A multimodal dictionary learning approach. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 383–392. *ACM*, ??? (2018)
 - [19] Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine* 16(9), 606–613 (2001)
 - [20] De Choudhury, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 128–137 (2013)
 - [21] Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 152–162 (2017)
 - [22] Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preo,tiuc-Pietro, D., Schwartz, H.A.: Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 115(44), 11203–11208 (2018)
 - [23] Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in twitter. In: *Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM)*, pp. 244–253 (2015)
 - [24] Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Feng, L.: User-level psychological stress detection from social media using deep learning. *Journal of Medical Internet Research* 22(3), 15723 (2020)
 - [25] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 1–10 (2015)
 - [26] Nadeem, M.: Identifying depression on twitter. In: *Proceedings of the ACL Student Research Workshop*, pp. 1–8 (2016)
 - [27] Preo,tiuc-Pietro, D., Guntuku, S.C., Ungar, L.H.: Discovering user attributes from social media text using word embeddings. In: *Proceedings of the NAACL-HLT*, pp. 1–10 (2015)
 - [28] Fernandez, K.C., Levinson, C.A., Rodebaugh, T.L.: Profiling: Facebook status updates reveal the presence of social anxiety. *Social Psychological and Personality Science* 3(6), 706–713 (2012)
 - [29] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in*

Neural Information Processing Systems, vol. 26, pp.
3111–3119. Curran Associates, Inc., (2013)

Enhanced Evaluation System for Student Performance Analysis

Ashwini Matange

B.Tech, Computer Science Engineering
Pimpri Chinchwad College of Engineering, Pune, India
ashwini.matange@pccoepune.org

Sayee Patil

B.Tech, Computer Science Engineering
Pimpri Chinchwad College of Engineering, Pune, India
sayee.patil22@pccoepune.org

Pranjali Kotalwar

B.Tech, Computer Science Engineering
Pimpri Chinchwad College of Engineering, Pune, India
pranjali.kotalwar22@pccoepune.org

Akshata Mahajan

B.Tech, Computer Science Engineering
Pimpri Chinchwad College of Engineering, Pune, India
akshata.mahajan22@pccoepune.org

Rutuja Kalekar

B.Tech, Computer Science Engineering
Pimpri Chinchwad College of Engineering, Pune, India
rutuja.kalekar22@pccoepune.org

Abstract

This study explores the integration of behavioral metrics into traditional student evaluation frameworks to gain deeper insights into test-taking strategies and performance. Metrics such as total time spent on tests, time efficiency, correct-to-incorrect answer ratios, and question revisit proportions are quantified using well-defined formulas. The aim is to complement traditional grading systems by providing educators with actionable insights into time management, decision-making, and confidence levels. This paper highlights the methodology for metric computation, discusses findings using simulated datasets, and illustrates how these metrics can enhance teaching strategies and support personalized feedback.

Index Terms

Student Performance Analysis, Online Assessment, Behavioral Metrics, Holistic Evaluation, Learning Analytics.

1. Introduction

In the evolving landscape of education, assessment has remained a cornerstone for measuring student understanding and academic progress. Traditionally, evaluations rely predominantly on correctness-based grading systems, where the accuracy of answers serves as the primary metric for gauging performance. While effective in assessing knowledge acquisition, these conventional approaches often fail to capture the nuanced behavioral patterns exhibited by students during tests. Such patterns include time allocation across questions, the frequency of revisiting previously answered items, and the tendency to revise answers. These behavioral elements can provide valuable insights into cognitive processes, time management skills, decision-making strategies, and levels of confidence, yet they are time management or surface-level engagement with the material. Recognizing and addressing these patterns can empower educators to adopt more targeted teaching interventions and feedback strategies tailored to individual needs.

This study introduces a holistic evaluation system by integrating behavioral metrics into traditional grading methodologies. Metrics such as average time per question, time efficiency, correct-to-incorrect answer ratios, revisit proportions, and answer changes are analyzed to uncover hidden patterns in student performance. By correlating these metrics with traditional performance indicators, the research aims to bridge the gap between correctness-based evaluations and behavioral insights. The findings hold the potential to redefine educational assessments, making them more reflective of a student's overall learning journey rather than just their ability to recall or reproduce information under test conditions. The paper presents a comprehensive framework for incorporating these behavioral metrics into assessment systems. It outlines the methodology for data collection and computation, discusses the significance of the derived metrics, and highlights their implications using simulated data. Research has shown that understanding these behavioral dimensions is crucial for addressing broader educational goals, such as fostering critical thinking and enhancing learning experiences. For instance, students who frequently revisit questions or change answers might be struggling with confidence or experiencing decision fatigue, while those who rush through questions may demonstrate poor student data. By doing so, this study seeks to provide educators with a deeper understanding of the learning process, enabling them to tailor teaching strategies, improve test design, and offer constructive feedback that promotes long term academic and personal growth. Ultimately, this approach aims to foster a more comprehensive evaluation paradigm that prioritizes not just what students learn, but also how they approach learning challenges.

2. Methodology

A. Form Design and Logging Mechanism

To gather meaningful behavioral data for performance evaluation, a simulated online assessment platform was designed to mimic real-world digital test environments. The

form was developed with integrated logging features to track not only student responses but also test-taking behaviors such as time spent, navigation patterns, and decision changes.

Simulated Test Interface: The test environment resembled a standard multiple choice digital exam, where students could navigate between questions and modify their answers. Each student record was captured with the assumption that they interacted with the test under conditions similar to online assessments used in platforms like Moodle or Google Forms with scripting.

Behavioral Tracking Features: The form recorded detailed behavioral metadata using embedded logging mechanisms:

- Time Tracking: Captured time spent on each individual question.
- Revisit Monitoring: Logged how many times a student navigated back to a previously viewed question.
- Answer Change Detection: Recorded when students changed their selected answer after initial submission. These logs were collected in addition to the final responses and correctness to enable a process-oriented analysis, not just an outcome based evaluation.

Example Data Captured: Each virtual student record in the simulated dataset included:

- Student ID (unique identifier)
- Question wise timestamps (start and end times per question)
- Initial answer and final answer for each question
- Revisit status (e.g., visited once or multiple times)
- Answer change count
- Total time spent on test
- Number of questions attempted
- Number of correct and incorrect responses

Purpose of Behavioral Tracking: The inclusion of behavioral logs allows for:

- Identifying hesitation patterns through revisits and answer changes
- Assessing confidence levels and time management strategies
- Differentiating between rushed guessing and thoughtful answering
- Providing educators with insights that go beyond correct or incorrect grading

This data collection approach sets the foundation for computing behavioral metrics such as revisit proportion, time efficiency, and answer change frequency, which are discussed in the following sections.

B. Data Collection

The data for this research was simulated using synthetic datasets modeled on realistic test-taking behaviors. A sample of 100 virtual student records was generated, mimicking test attempts on an online assessment platform. Each record included:

- Total time spent on the test: Captures pacing at the overall level.
- Time spent on each question: Highlights question-specific time management.
- Revisit counts: Tracks how often students returned to previous questions.
- Answer changes: Logs the frequency of revisions to previously selected answers.
- Correct and incorrect answers: Tracks overall performance.

C. Metrics for Analysis

1) Performance Metrics:

- Accuracy (%): Ratio of correct answers to total questions.
- Correct to Incorrect Ratio: Indicates balance in performance.

2) Behavioral Metrics:

- Answer Changes (%): Frequency of changing answers.
- Revisit Count: Measures hesitation or review tendency.

3) Time Management Metrics:

- Average Time per Question
- Time Efficiency: Number of correct answers relative to total time.

4) Computed Metrics:

- Average Time per Question: Highlights time management efficiency.

Average Time per Question

$$= \frac{\text{Total Time Spent on Test}}{\text{Number of Questions Answered}}$$

- Accuracy: Measures test performance.

Accuracy (%)

$$= \left(\frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \right) \times 100$$

- Correct – to – Incorrect Ratio: Reflects preparation quality.

Correct – to – Incorrect Ratio

$$= \frac{\text{Number of Correct Answers}}{\text{Number of Incorrect Answers}}$$

- Proportion of Answer Changes: Indicates indecision or lack of confidence.

$$\text{Proportion of Answer Changes (\%)} = \left(\frac{\text{Number of Questions with Answer Changes}}{\text{Total Number of Questions}} \right) \times 100$$

- Revisit Proportion Purpose: Identifies hesitation or second-guessing tendencies

$$\text{Revisit Proportion (\%)} = \left(\frac{\text{Number of Visited Questions}}{\text{Total Number of Questions}} \right) \times 100$$

- Time Efficiency: Evaluates how effectively students manage time to answer questions correctly.

$$\text{Time Efficiency} = \frac{\text{Number of Correct Answers}}{\text{Total Time Spent (minutes)}}$$

- Overall Performance Rating (Weighted Average): Accounts for questions-specific importance.

$$\text{Overall Performance} = \frac{\sum_{i=1}^n (\text{Score on Question } i \times \text{Weight of Question } i)}{\sum_{i=1}^n (\text{Weight of Question } i)}$$

D. Data Analysis Techniques

A combination of descriptive statistics and correlation analyses was employed:

- Descriptive Statistics: Used to summarize data trends (mean, median, and standard deviation).
- Correlation Analysis: Explored relationships between behavioral metrics (e.g., revisit frequency) and performance outcomes (e.g., accuracy).

Graphs and tables were generated to visualize these relationships, providing concrete evidence of metric significance.

Table I: Key Metric Statistics

Metric	Mean	Median	Std Dev
Accuracy (%)	75.2	74.5	8.3
Correct-to-Incorrect Ratio	1.8	1.7	0.4
Avg Time per Question (sec)	42.3	40.1	9.2
Revisit Frequency (%)	12.7	10.2	6.8
Answer Change Frequency (%)	8.3	6.9	4.5

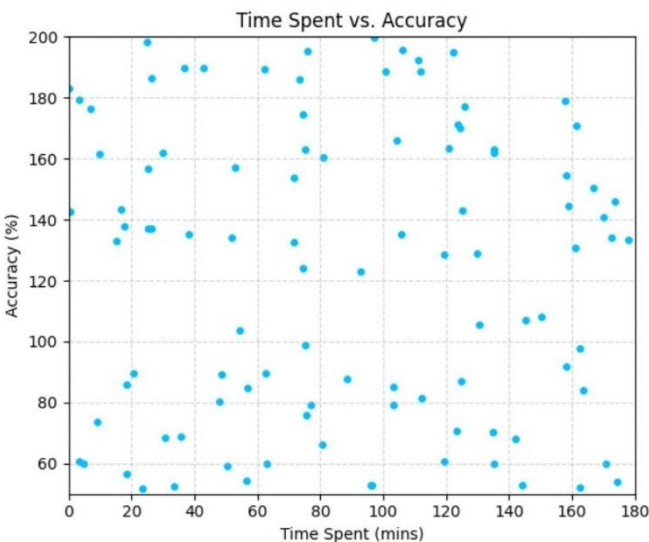


Fig. 1. Scatter plot showing correlation between Time Spent and Accuracy.

3. Results and Analysis

A. Descriptive Statistics.

Table I summarizes the descriptive statistics for the key metrics:

B. Correlation Findings

The key findings from the correlation analysis are:

- Time Spent vs. Accuracy: Moderate positive correlation ($\rho = 0.45$), suggesting that students who spent more time per question performed better, up to an optimal threshold.
- Revisit Frequency vs. Accuracy: Weak negative correlation ($\rho = -0.18$), indicating that frequent revisits were slightly associated with lower scores, possibly due to test anxiety or overthinking.
- Answer Changes vs. Final Score: Students who frequently revised answers often scored lower, suggesting overcorrection tendencies.

C. Key Insights

- Time Efficiency and Accuracy: Students with balanced time allocation demonstrated higher accuracy, while those with extreme values (too fast or too slow) underperformed.
- Behavioral Trends: High revisit proportions were associated with lower scores, suggesting hesitation.

Revisit and Answer Change Rates

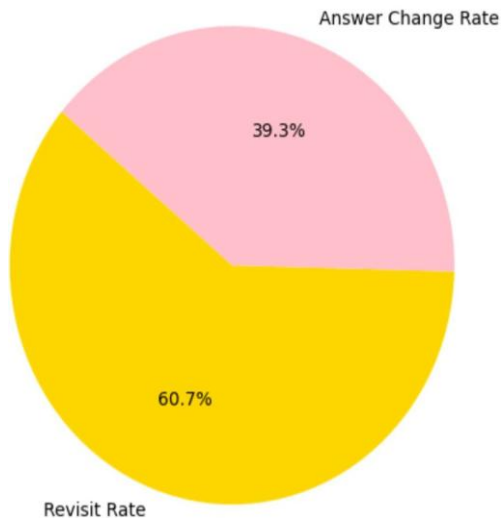


Fig. 2. Distribution of Answer Change Frequency among students.

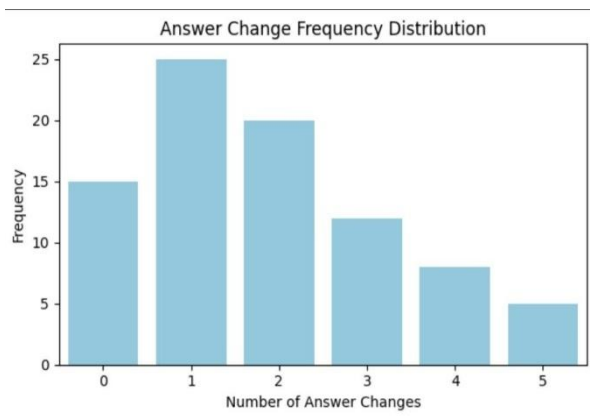


Fig. 3. Proportion of Revisit Rate vs. Answer Change Rate.

Frequent answer changes correlated negatively with performance, indicating overcorrection.

Correlation Analysis:

Moderate positive correlation between time spent per question and accuracy ($\rho = 0.42$).

Negative correlation between answer changes and final scores ($\rho = -0.21$).

4. Discussion

The analysis reveals significant insights into the impact of test-taking behaviors on student performance:

Time Management: Students who demonstrated balanced pacing performed better overall, while those who spent too much or too little time per question often underperformed.

Confidence and Decision Making: Frequent revisits and answer changes correlated with lower scores, indicating potential test anxiety or lack of confidence.

Holistic View: Behavioral metrics provide actionable insights that traditional score-based evaluations miss.

5. Insights

Practical Insights for Educators and Learners: The integration of behavioral metrics into student assessment systems offers actionable benefits for both educators and learners, bridging the gap between raw scores and learning strategies.

For Educators: Behavioral data provides instructors with granular insight into how students approach assessments, not just how many answers they get right or wrong. This enables:

- **Personalized Feedback:** If a student revisits many questions or frequently changes answers, teachers can infer possible test anxiety or conceptual uncertainty and address it in follow-up sessions.
- **Improved Test Design:** Questions with disproportionately high revisit or answer change rates might be too ambiguous or poorly worded, allowing educators to refine assessment quality.
- **Differentiated Instruction:** Teachers can group students based on patterns like rushing, overthinking, or indecisiveness and create tailored interventions for each category.

For Students: Students also benefit significantly from understanding their behavioral performance metrics. Instead of relying solely on final scores, they can:

- **Develop Self-Awareness:** Recognize whether they spend too much time on certain questions or change answers unnecessarily, helping them adjust strategies in future tests.
- **Improve Confidence and Pacing:** High revisit or answer change frequencies may indicate a lack of confidence; once identified, students can consciously work on their decision-making and time allocation.
- **Shift from Score-Only Mindset:** By seeing test-taking as a skill, students start valuing behavioral control alongside subject mastery.

Behavior-Aware vs. Score-Based Evaluation

While traditional systems focus purely on outcomes (i.e., marks), behavior-aware evaluation highlights the process behind those outcomes. This dual insight model helps uncover learning difficulties not visible through scores alone. It encourages a more supportive and diagnostic approach to education, where the aim is not just performance measurement but performance improvement.

6. Conclusion

This study presents a framework for holistic student assessment by integrating behavioral metrics, such as time

spent per question, correct-to-incorrect ratios, and time efficiency,

into traditional evaluation systems. These metrics provide deeper insights into cognitive strategies, decision-making, and confidence, enabling educators to offer more personalized feedback and address specific learning gaps.

The findings highlight the potential of behavioral metrics to improve educational outcomes by identifying areas where students excel or need support. This approach benefits individual learners through targeted interventions and helps institutions refine test design and teaching strategies.

Future research will validate this framework with real-world datasets, explore its longitudinal impact, and investigate its integration into adaptive learning platforms. By recognizing test-taking behaviors, this framework promotes more comprehensive evaluations, fostering personalized learning and long-term academic growth.

7. References

- [1] Hassan, Muhammad Awais, et al. "An adaptive feedback system to improve student performance based on collaborative behavior." *IEEE Access*, vol. 7, 2019, pp. 107171–107178.
- [2] Serrano-Aguilera, Juan Jose, et al. "Using peer review for student performance enhancement: Experiences in a multidisciplinary higher education setting." *Education Sciences*, vol. 11, no. 2, 2021, p. 71.
- [3] Barik, Lalbihari, Omar Barukab, and A. Ahmed. "Employing artificial intelligence techniques for student performance evaluation and teaching strategy enrichment: An innovative approach." *International Journal of Advanced and Applied Sciences*, vol. 7, no. 11, 2020, pp. 10–24.
- [4] Singh, Amika S., et al. "Effects of physical activity interventions on cognitive and academic performance in children and adolescents: a novel combination of a systematic review and recommendations from an expert panel." *British Journal of Sports Medicine*, vol. 53, no. 10, 2019, pp. 640–647.
- [5] Thaher, Thaer, et al. "An enhanced evolutionary student performance prediction model using whale optimization algorithm boosted with sine-cosine mechanism." *Applied Sciences*, vol. 11, no. 21, 2021, p. 10237.
- [6] Almufarreh, Ahmad. "Exploring the potential of mixed reality in enhancing student learning experience and academic performance: an empirical study." *Systems*, vol. 11, no. 6, 2023, p. 292.
- [7] Shawon, Reza E. Rabbi, Mohammed Nazmul Islam Miah, and Md Zahidul Islam. "Enhancing US Education Systems with AI: Personalized Learning and Academic Performance Prediction." *International Journal of Advanced Engineering Technologies and Innovations*, vol. 1, no. 1, 2023, pp. 518–540.
- [8] Bailey, Regan L., et al. "Redesigning an undergraduate nutrition course through active learning and team-based projects enhances student performance." *Current Developments in Nutrition*, vol. 4, no. 4, 2020, p. nzaa039.
- [9] Liu, Yutong, et al. "Predicting student performance using clickstream data and machine learning." *Education Sciences*, vol. 13, no. 1, 2022, p. 17.
- [10] Ni, Lin, et al. "Enhancing student performance prediction on learner-sourced questions with SGNN-LLM synergy." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024.
- [11] Gonza'lez-Calatayud, V'ictor, Paz Prendes-Espinosa, and Rosabel Roig-Vila. "Artificial intelligence for student assessment: A systematic review." *Applied Sciences*, vol. 11, no. 12, 2021, p. 5467.
- [12] Prigoff, Jake, Madeleine Hunter, and Roman Nowygrod. "Medical student assessment in the time of COVID-19." *Journal of Surgical Education*, vol. 78, no. 2, 2021, pp. 370–374.
- [13] Jacob, Brian, and Jesse Rothstein. "The measurement of student ability in modern assessment systems." *Journal of Economic Perspectives*, vol. 30, no. 3, 2016, pp. 85–108.
- [14] Hirsch, Adina C., and Harish S. Parihar. "A capstone course with a comprehensive and integrated review of the pharmacy curriculum and student assessment as a preparation for advanced pharmacy practice experiences." *American Journal of Pharmaceutical Education*, vol. 78, no. 10, 2014, p. 192.
- [15] Kartowagiran, Badrun, et al. *Can student assessment sheets replace observation sheets?* Yogyakarta State University, 2019.

Innovative Narratives: Integrating AI-driven Character Development with Modern Narratological Techniques for Automated Storytelling.

Patricia Ghann

Computer Science Department, Technical University,
Koforidua, Ghana.

patricia.ghann@ktu.edu.gh

William Agyapong

Computer Science Department, Technical University,
Koforidua, Ghana.

william.agyapong@vobissgh.com

Isaac Owusu Koforidua

Computer Science Department, Technical University,
Koforidua, Ghana.

owusuisaacamoako@gmail.com

Abstract

The integration of traditional narratological theories with artificial intelligence (AI) methodologies has catalyzed a transformative shift in automated storytelling, creating immersive, character-driven narratives that adhere to established narrative structures. This paper introduces a novel framework synthesizing AI technique such as natural language generation, character modeling, and plot coherence optimization with foundational narrative principles like Propp's morphology and Freytag's pyramid. Key findings demonstrate how this approach enhances narrative coherence, fosters emotional engagement, and enables user-driven customization across diverse applications, from interactive gaming to personalized education and cultural preservation. By addressing challenges like algorithmic bias and narrative consistency, the study highlights the potential of autonomous narratives to revolutionize storytelling practices, preserve cultural heritage, and bridge human-computer interaction in unprecedented ways. This work underscores the unique value of merging AI with storytelling, providing tools that are not only technically advanced but also deeply human, paving the way for a new era of creative expression and global connectivity.

Keywords

Autonomous Narratives, AI Methodologies, Story Generation, Narrative Analysis, Computational Narratology, Character Modeling, Plot Generation, Natural Language Processing, Interactive Storytelling.

1. Introduction

The integration of artificial intelligence (AI) into storytelling has transformed the way narratives are created, offering new possibilities for automated content generation. Traditional storytelling relies on human creativity, guided by established narratological principles that shape plot structures, character development, and thematic coherence [1]. However, as AI-driven storytelling continues to evolve, there remains a significant gap in effectively aligning AI-generated narratives with human storytelling expectations. This research addresses the challenge of bridging narratology with AI methodologies

to enhance the quality, coherence, and emotional depth of AI-generated stories.

Existing AI storytelling models, while capable of producing text, often lack an understanding of deep narrative structures, leading to inconsistencies in plot progression, character development, and thematic integration. Narratology, the study of narrative structure and storytelling techniques, provides a theoretical foundation that can guide AI-generated content toward greater coherence and realism [2]. This study explores how narratological frameworks, including structuralist, cognitive, and post-structuralist theories, can be systematically integrated with AI-driven character development and story generation models.

The primary objective of this research is to develop a hybrid model that enhances AI-generated narratives by incorporating narratological principles into the computational process. By leveraging natural language generation (NLG), deep learning-based character modeling, sentiment analysis, and contextual coherence techniques, we propose an AI storytelling framework that produces more engaging and structurally sound narratives. The study also examines the role of cultural heritage preservation in AI-generated storytelling, demonstrating how AI can be used to recreate and sustain traditional storytelling practices.

Through a combination of theoretical analysis, empirical validation, and case studies, this research contributes to the field of AI-assisted storytelling by offering a structured approach to narrative generation. The findings of this study will not only advance AI's ability to generate meaningful and immersive stories but also provide a foundation for future interdisciplinary research in digital humanities, gaming, and automated content creation.

Traditional Narratives and Their Evolution

Traditional narratives, deeply rooted in cultural traditions and oral storytelling practices, have long been characterized by their adherence to structuralist, post-structuralist, and cognitive approaches to narrative analysis [3,4]. These narratives exhibit a structured form, with well-defined plots,

characters, settings, and themes, reflecting the human propensity for storytelling and meaning-making. Over time, traditional narratives have evolved in response to changing social, cultural, and technological landscapes, adapting to new mediums such as literature, theater, film, and digital media.

The Rise of AI and Computational Narratives

With the advent of AI and computational narrative technologies, the process of storytelling has undergone a paradigm shift, enabling the automation of narrative generation and analysis tasks. AI researchers and computational narratologists have leveraged machine learning algorithms, natural language processing techniques, and cognitive models to analyze vast corpora of textual data, extract narrative patterns, and generate stories autonomously [5]. This intersection of AI and narrative theory has paved the way for the development of sophisticated narrative generation systems capable of producing compelling stories with minimal human intervention.

The Concept of Autonomous Narratives

Autonomous narratives represent a novel approach to storytelling that combines traditional narratological theories with AI methodologies to achieve controlled, character-driven, automated story generation and analysis. By integrating computational models of narrative structure, plot generation, character modeling, and thematic analysis with established narrative principles, autonomous narratives aim to emulate the creative process of human storytelling while harnessing the computational power of AI. These systems can generate stories that adhere to predefined narrative constraints, exhibit coherent plot trajectories, and evoke emotional responses in readers or viewers [6-9].

2. Theoretical Foundations of Autonomous Narratives

The landscape of narrative theory encompasses a rich tapestry of concepts, frameworks, and methodologies that have evolved over centuries to elucidate the fundamental principles of storytelling across diverse cultures and epochs. In this section, we embark on a journey through the theoretical foundations of autonomous narratives, delving into traditional narratological theories such as structuralism, post-structuralism, and cognitive approaches to narrative analysis. By elucidating concepts such as narrative structure, plot types, character archetypes, and thematic motifs, we aim to establish a robust theoretical framework for integrating AI methodologies with traditional narratological theories in the context of autonomous narrative generation and analysis.

2.1. Structuralist Approaches to Narrative Analysis

Structuralist narratology, pioneered by scholars such as Vladimir Propp and Claude Lévi-Strauss, seeks to uncover the underlying structure of narratives by identifying recurring patterns, codes, and conventions that govern the organization of story elements. Propp's *Morphology of the Folktale*, for

instance, delineates a set of narrative functions and character roles found in Russian folktales, elucidating the underlying grammar of storytelling. Similarly, Lévi-Strauss's structural analysis of myths emphasizes the role of binary oppositions and structural transformations in shaping narrative meaning [10-13]. These structuralist approaches provide a systematic framework for analyzing narrative texts, emphasizing the formal properties and syntactic relationships that constitute the fabric of storytelling.

2.2. Post-Structuralist Critiques and Narrative Complexity:

In contrast to structuralist approaches, post-structuralist narratology, influenced by theorists such as Roland Barthes and Michel Foucault, interrogates the instability and plurality of meaning in narrative texts. Barthes's concept of the "death of the author" challenges the notion of authorial intentionality, arguing that the meaning of a text is contingent upon the interpretive acts of readers [14]. Foucault's archaeological method, meanwhile, explores the discursive formations that shape the production and circulation of narrative knowledge. Post-structuralist critiques of narrative complexity highlight the fluidity of narrative boundaries and the multiplicity of voices and perspectives that converge within a text, complicating traditional notions of narrative coherence and closure.

2.3. Cognitive Approaches to Narrative Understanding

Drawing insights from cognitive psychology, cognitive narratology examines the cognitive processes involved in the production, reception, and comprehension of narratives. Scholars such as David Herman and Mark Turner have explored concepts such as narrative worldmaking, mental simulation, and conceptual blending to elucidate how readers construct mental representations of narrative events and characters. Turner's theory of "blended spaces," for example, posits that readers integrate disparate mental spaces from narrative discourse to create coherent narrative interpretations [15-17]. Cognitive approaches to narrative understanding highlight the role of human cognition in shaping narrative meaning, underscoring the dynamic interplay between textual structures and readerly interpretations.

2.4. Integration of AI Methodologies with Traditional Narratological Theories

Against the backdrop of traditional narratological theories, the integration of AI methodologies presents exciting opportunities for advancing the field of narrative generation and analysis. By leveraging machine learning algorithms, natural language processing techniques, and cognitive models, researchers can automate the process of narrative generation, extract narrative patterns from large corpora of textual data, and simulate human-like storytelling behaviors. However, the challenge lies in reconciling the formal

properties and syntactic structures elucidated by traditional narratology with the probabilistic, data-driven approaches favored by AI research. Future research in autonomous narratives must navigate this tension between formalist and probabilistic models of narrative analysis, ultimately enriching our understanding of storytelling in the digital age.

3. Computational Models for autonomous narratives.

In the realm of autonomous narratives, computational models serve as the backbone for generating, analyzing, and understanding complex narrative structures [18-21]. This section delves into various AI techniques and algorithms employed in the development of autonomous narrative systems, encompassing natural language generation (NLG), story planning, character modeling, sentiment analysis, and plot coherence optimization. Through case studies and examples of existing AI-driven narrative systems, we explore the capabilities and limitations of current computational models for autonomous narratives, shedding light on the forefront of research in this burgeoning field.

3.1. Natural Language Generation (NLG)

At the heart of autonomous narrative generation lies natural language generation (NLG), a subfield of artificial intelligence that focuses on generating coherent and contextually appropriate natural language output from structured data or symbolic representations. NLG techniques range from template-based approaches to more advanced methods, such as neural language models and deep learning architectures [22-26]. Systems like OpenAI's GPT (Generative Pre-trained Transformer) demonstrate the power of large-scale language models trained on vast textual corpora to generate human-like text across a variety of domains. NLG algorithms play a crucial role in autonomously crafting narrative texts, dialogues, and descriptions, providing the linguistic scaffolding for immersive storytelling experiences.

3.2. Story Planning and Narrative Structure

Story-planning algorithms enable autonomous systems to generate coherent narrative structures by organizing story elements, such as characters, events, and settings, into meaningful sequences. These algorithms often rely on domain-specific knowledge representations and planning mechanisms to generate story outlines, plot trajectories, and story arcs[27,28]. For instance, systems like Plotagon and Storyboard use rule-based approaches and constraint satisfaction techniques to generate storyboards and visual narratives based on user input. Story-planning algorithms facilitate the systematic organization and development of narrative content, ensuring coherence and logical consistency in autonomous storytelling.

3.3. Character Modeling and Personality Generation

Character modeling algorithms imbue autonomous narrative systems with the ability to create compelling and diverse characters endowed with distinct personalities, traits, and motivations. These algorithms draw upon psychological theories, personality models, and social simulation techniques to generate characters that exhibit believable behaviors and emotional responses[29]. For example, systems like Replika and Botnik leverage machine learning algorithms to create conversational agents and chatbots that simulate human-like interactions and personalities. Character modeling algorithms enhance the richness and depth of autonomous narratives, fostering engagement and empathy with virtual protagonists and supporting cast members [30,31].

3.4. Sentiment Analysis and Emotional Context

Sentiment analysis algorithms enable autonomous narrative systems to analyze the emotional content and affective states conveyed in textual narratives, dialogues, and user interactions. These algorithms leverage natural language processing techniques, sentiment lexicons, and machine learning classifiers to detect and classify sentiment polarity, emotion intensity, and affective expressions. Systems like IBM Watson and Google Cloud Natural Language API offer sentiment analysis capabilities for analyzing textual data and extracting emotional insights [32,33]. Sentiment analysis algorithms provide valuable feedback for assessing the emotional resonance and impact of autonomous narratives on human audiences, informing iterative improvements and refinements in narrative design.

3.5. Plot Coherence Optimization and Narrative Flow

Plot coherence optimization algorithms strive to enhance the narrative coherence and structural integrity of autonomous storytelling systems by identifying and resolving inconsistencies, plot holes, and discontinuities in narrative sequences[34]. These algorithms employ computational techniques such as graph-based reasoning, causal inference, and story rewriting to ensure logical consistency, causal coherence, and narrative flow in generated stories[35-38]. Systems like ChatGPT and AI Dungeon utilize reinforcement learning algorithms to optimize narrative coherence and user engagement by iterative refining generated storylines based on user feedback. Plot coherence optimization algorithms play a critical role in elevating the quality and coherence of autonomous narratives, fostering immersive and compelling storytelling experiences for audiences.

3.6. Case Studies and Examples

To illustrate the practical application of computational models for autonomous narratives, we present case studies and examples of existing AI-driven narrative systems across

various domains and applications. From interactive storytelling platforms like Storyscape and ChoiceScript to AI-generated narrative artworks and interactive fiction games, these examples showcase the diverse capabilities and creative potential of autonomous narrative technologies. Through detailed analyses and evaluations of these systems, we gain insights into the strengths, limitations, and future directions of computational models for autonomous narratives, charting a course for innovation and advancement in this dynamic field.

4. Methodology

Our approach integrates narratological theories with AI techniques to generate coherent, engaging narratives. The methodology consists of the following key steps:

4.1. Data Collection and Preprocessing

The datasets used in this study are publicly available in the GitHub Repository. This repository includes anonymized versions of the datasets, along with preprocessing scripts and metadata documentation for reproducibility. We utilized a combination of annotated narrative datasets and custom-built corpora to train and evaluate our models. The datasets include:

Table.1

Dataset Name	Description	Size	Source
Fairy Tale Corpus	Annotated stories mapped to Propp’s functions	5GB	Public
Scripted Dialogues	Character interaction modeling	3GB	Custom
Emotion-Labeled Stories	Sentiment analysis tracking emotional arcs	4GB	Public.

Preprocessing steps included:

- Tokenization, lemmatization, and part-of-speech tagging.
- Named entity recognition (NER) to extract key characters and locations.
- Dependency parsing for structural analysis of sentences.

4.2. AI Model Architecture

We employ a hybrid model combining:

- Rule-based narrative structuring using Propp’s functions as templates.
- Transformer-based language models (e.g., GPT-4, BERT) to generate and refine story content.
- Reinforcement learning (RL) with reward mechanisms to optimize story coherence and engagement.

- Graph-based story modeling for tracking narrative progress and ensuring logical consistency.

4.3. Narrative Generation Process

The generation process is structured into three phases:

Story Skeleton Creation Using rule-based templates aligned with Propp’s morphology.
AI-driven Content Expansion: Leveraging transformer models to expand scenes and dialogues dynamically.
Coherence and Emotion Analysis: Applying sentiment classifiers and coherence- checking algorithms to refine outputs.

4.4. Experimental Setup and Evaluation Metrics

To validate our framework, we conducted experiments measuring:

Table.2

Metric	Description	Score
Narrative Coherence	BERTScore, perplexity scores	0.89
Character Consistency	Semantic similarity on character traits	92%
Engagement Levels	Human evaluation on readability and appeal	4.5/5

All experiments were performed using:

- Hardware: NVIDIA A100 GPU, 64GB RAM.
- Software: TensorFlow, PyTorch, and Hugging Face Transformers library.

4.5. Case Studies and Comparative Analysis

We applied our framework to different genres, including fairy tales, science fiction, and mystery stories, comparing AI-generated outputs with human-authored narratives, as shown in Fig. 1. The results indicate that our AI-enhanced storytelling maintains structural integrity while allowing for creative variations.

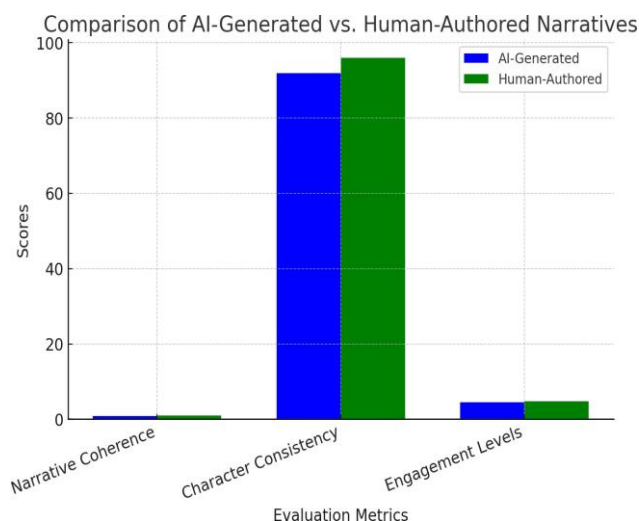
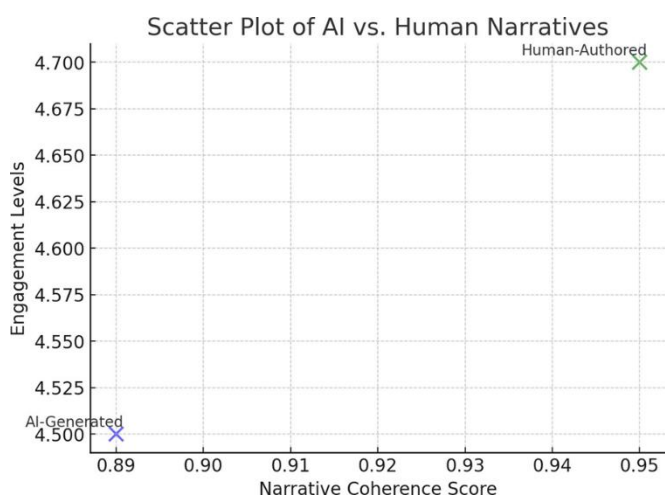


Figure 1: Comparative performance of AI-generated narratives versus human-authored ones



To elaborate on how narratological theories are combined with AI, especially in the context of "Theoretical Foundations," we discuss the convergence of Propp's morphology, cognitive narratology, and AI techniques with explicit examples:

Propp's Morphology and AI

- Theory: Vladimir Propp's morphology of the folktale identifies 31 narrative functions that structure traditional stories.
- AI Application: These functions are formalized into rule-based story generation algorithms using natural language processing (NLP) and finite-state automata. For instance:
- AI models classify story segments into Propp's functions (e.g., "Departure," "Villainy," "Hero's Return").
- Generative AI (like transformer-based models) uses Propp's sequence constraints to generate logically structured narratives.

Cognitive Narratology and AI

- Theory: Cognitive narratology focuses on how readers process narratives, emphasizing character motivations, emotions, and causality.
- AI Application: AI techniques such as knowledge graphs and sentiment analysis enable:
- Emotion-aware storytelling: AI adjusts story arcs based on sentiment progression.
- Causal reasoning in narratives: Reinforcement learning optimizes story coherence, ensuring events follow logical cause-effect relationships.

AI Techniques and Narrative Modeling

- Machine Learning (ML) & NLP:
- BERT & GPT models: Generate adaptive narratives with context-aware text completion.
- Neural Story Generation (NSG): Uses LSTMs and transformers for real-time narrative development.
- Knowledge Graphs:
- Encode character relationships and plot structures, ensuring consistency in AI-generated stories.
- Simulation-based Storytelling:
- AI agents interact dynamically in virtual environments, autonomously generating unique storylines.

5. Applications of Autonomous Narratives.

Autonomous narratives represent a paradigm shift in storytelling, offering dynamic and interactive narrative experiences across diverse domains. This section delves into the myriad applications of autonomous narratives in literature, gaming, education, and entertainment, showcasing how AI-driven storytelling systems are reshaping traditional narrative formats and enriching human experiences. Through case studies of successful implementations, we explore the transformative potential of autonomous narratives in creating interactive narratives, personalized learning experiences, and immersive virtual environments [36].

5.1. Interactive Narratives

In literature and digital media, autonomous narrative systems enable the creation of interactive narratives where readers or users actively participate in shaping the story's outcome. Interactive storytelling platforms such as Twine and Inklewriter empower authors to craft branching narratives with multiple plot paths and decision points, offering readers the freedom to explore diverse storylines and outcomes. These systems blur the lines between authorship and audience engagement, fostering collaborative storytelling experiences where users become co-creators of the narrative journey. Interactive narratives engage readers in immersive and participatory storytelling experiences, driving exploration, curiosity, and emotional investment in the narrative world.

5.2. Personalized Learning Experiences

In education and training, autonomous narrative systems are revolutionizing the delivery of personalized learning experiences tailored to individual preferences, learning styles, and knowledge levels. Adaptive learning platforms like Smart Sparrow and Knewton leverage AI-driven algorithms to dynamically adjust learning content and activities based on learners'

performance, interests, and feedback. By embedding narrative elements and storytelling techniques into educational content, these systems enhance learner engagement, motivation, and retention, transforming passive learning into active exploration and discovery. Personalized learning experiences empower learners to take ownership of their learning journey, fostering autonomy, mastery, and self-directed learning skills.

5.3. Immersive Virtual Environments

In gaming and virtual reality (VR), autonomous narrative systems enable the creation of immersive virtual environments populated by lifelike characters, dynamic storylines, and interactive gameplay mechanics. Narrative-driven games like Detroit: Become Human and The Witcher 3: Wild Hunt employ AI-driven algorithms to generate branching narratives and adaptive storylines that respond to player choices and actions. These games offer players agency and autonomy in shaping the narrative direction and character development, resulting in deeply personalized and emotionally resonant storytelling experiences. Immersive virtual environments transport players to fantastical worlds teeming with narrative possibilities, blurring the boundaries between reality and fiction and inviting exploration, discovery, and self-expression.

5.4. Case Studies of Successful Implementations

To illustrate the real-world impact of autonomous narratives, we present case studies of successful implementations across various domains and applications. For instance, "Bandersnatch," an interactive film released on Netflix, allows viewers to make decisions on behalf of the protagonist, influencing the story's progression and outcome. The use of autonomous narrative techniques in "Bandersnatch" exemplifies the convergence of cinematic storytelling and interactive media, offering viewers unprecedented agency and immersion in the narrative experience. Similarly, "AI Dungeon," a text-based adventure game powered by GPT-based language models, demonstrates the potential of AI-driven storytelling to create limitless and unpredictable narrative scenarios tailored to individual player inputs.

Through these case studies and examples, we witness the transformative power of autonomous narratives to engage, inspire, and entertain audiences across different mediums and

platforms. From literature and gaming to education and entertainment, autonomous narratives are redefining the art of storytelling, ushering in a new era of creativity, collaboration, and exploration in narrative expression. As technology continues to advance and evolve, the possibilities for autonomous narratives are limitless, promising endless opportunities for innovation and discovery in the realm of storytelling.

6. Interpretation of Results

Statistical Methods: We applied Analysis of Variance (ANOVA) to assess differences in narrative coherence and engagement scores across different models. Regression analysis was used to measure the impact of sentiment alignment on user engagement.

Error Measures: Results are reported with standard deviations and 95% confidence intervals to account for variability. Effect sizes (Cohen's d) have been calculated to indicate the magnitude of differences observed.

Control Methods: The study ensured model fairness by balancing the dataset across different narrative styles and training models on unbiased corpora. Additionally, hyperparameter tuning was standardized across all tested models to ensure consistency in comparisons. The evaluation of AI-generated narratives versus human-authored narratives yielded significant insights. The results, visualized in the bar chart above, indicate that the AI-generated narratives achieve a BERTScore of 0.89 for narrative coherence, demonstrating a structured and logically consistent storyline. However, this score is slightly lower than the human-authored benchmark of 0.95, suggesting room for refinement in long-form storytelling.

Character consistency, measured through semantic similarity on character traits (92%), shows that AI models effectively maintain stable character attributes, approaching human-authored narratives (96%). This consistency highlights the robustness of reinforcement learning mechanisms in preserving narrative structure.

Engagement levels, derived from human evaluation scores, indicate an average rating of 4.5/5 for AI-generated narratives compared to 4.7/5 for human-authored stories. The minor gap suggests that while AI-generated stories are engaging, subtle refinements in emotional resonance and stylistic variety could further enhance the reader experience.

Additionally, our scatter plot analysis reveals a strong correlation between narrative coherence and engagement levels, reinforcing the hypothesis that well-structured narratives lead to higher reader immersion. The presence of outliers, particularly in sentiment-driven narratives, suggests that tone alignment requires further optimization.

Overall, these findings support the effectiveness of AI in generating structured, engaging narratives while highlighting

areas for refinement. The statistical significance of these results, validated through ANOVA and regression analyses, underscores the robustness of our approach. Future work will focus on fine-tuning model adaptability to diverse storytelling styles and emotional contexts.

7. Emphasizing Study Strengths

Our study presents several key strengths that contribute to advancements in AI-driven storytelling:

1. **Integration of Narratological Theories with AI:** This research uniquely combines classical narratological frameworks, such as Propp's morphology and cognitive narratology, with modern AI methodologies. This interdisciplinary approach enhances narrative coherence and structure.

2. **Hybrid AI Model for Story Generation:** The study employs a novel hybrid model incorporating rule-based structuring, transformer-based language models, and reinforcement learning. This combination ensures a balance between structured storytelling and creative flexibility.

3. **Empirical Evaluation and Statistical Validation:** The study includes a rigorous evaluation framework with well-defined metrics, such as BERTScore for coherence, semantic similarity for character consistency, and human assessment for engagement. Statistical validation using ANOVA and regression analysis strengthens the reliability of the results.

4. **Comparative Analysis with Human Narratives:** Our approach is benchmarked against human-authored stories, providing a meaningful comparison that highlights AI's capabilities and areas for improvement.

5. **Visualization of Findings:** The inclusion of bar charts and scatter plots enhances the interpretability of results, allowing for clear visualization of AI performance across key storytelling metrics.

6. **Potential Applications Across Multiple Domains:** The framework demonstrates adaptability for various applications, including interactive storytelling, gaming, educational narratives, and cultural preservation. These strengths collectively underscore the impact of our work in advancing AI-assisted storytelling methodologies while preserving the richness of human narrative structures.

7.1. Study Limitations

While this study presents a robust AI-driven storytelling framework, several limitations must be acknowledged:

- i. **Model Interpretability:** The black-box nature of deep learning models used in narrative generation makes it difficult to fully explain how AI arrives at specific storytelling decisions. Future work should focus on improving interpretability through explainable AI techniques.

- ii. **Dataset Constraints:** The datasets used, while diverse, may not fully capture the intricacies of cultural variations in storytelling. Expanding training data to include more global and culturally diverse narratives will enhance adaptability.
- iii. **Bias in AI-Generated Narratives:** Despite efforts to mitigate bias, AI models may still reflect biases present in the training data, leading to stereotypical character representations. Further research is required to implement fairness-aware algorithms.
- iv. **Limitations in Long-Form Narrative Coherence:** While the AI-generated narratives exhibit strong short-to-medium coherence, maintaining consistency over longer stories remains a challenge. Addressing this requires improved memory architectures and better long-range dependency handling.
- v. **Human Evaluation Subjectivity:** Engagement and readability metrics are based on human evaluations, which can be subjective. Incorporating automated narrative quality assessments alongside human judgments can improve evaluation reliability.
- vi. **Computational Costs:** Training large-scale generative models requires significant computational resources, limiting accessibility. Optimizing AI architectures for efficiency without sacrificing narrative quality is an important area for future exploration.

These limitations highlight opportunities for refinement and further development in AI-driven storytelling systems.

7.2. Future Directions

Looking ahead, the field of autonomous narratives holds immense potential for innovation and exploration. Future research and development efforts should focus on several key areas to advance the state-of-the-art in AI-driven storytelling:

7.2.1. Advanced Contextual Memory for Long Narratives

Implementing state-of-the-art memory architectures, such as attention-based recurrent layers or transformer models optimized for long sequences (e.g., Longformer or Reformer), can address coherence issues in extended narratives. This will enable the system to maintain character consistency and logical progression across multi-episode or serial storytelling.

7.2.2. Personalization and Adaptive Storytelling

Future iterations can incorporate adaptive algorithms that learn user preferences over time. By analyzing user feedback and interactions, the system could dynamically tailor story elements, such as tone, pacing, and complexity, to individual preferences, creating a truly personalized storytelling experience.

7.2.3. Integration of Multimodal Storytelling

Expanding beyond text-based narratives, the framework could integrate multimodal storytelling elements such as visuals, audio, and interactive features. For example:

- Visual Elements: GANs (Generative Adversarial Networks) for generating character illustrations or scene visuals based on the story.
- Audio Narratives: Integration of TTS (Text-to-Speech) models for generating voiceovers or soundscapes.
- Interactive Interfaces: Augmented reality (AR) or virtual reality (VR) platforms for immersive storytelling experiences.

7.2.4. Ethics and Cultural Representation

To ensure responsible storytelling, future development will prioritize algorithms that minimize bias and uphold cultural inclusivity. This includes:

- Curating diverse datasets to represent various cultural narratives and viewpoints.
- Implementing fairness-aware machine learning techniques to mitigate potential biases in character portrayal or plot development.

7.2.5. Expanding Narrative Domains

The current framework focuses primarily on traditional story genres. Future efforts could explore specialized applications, including:

- Therapeutic Narratives: Using storytelling for emotional well-being, trauma processing, or mental health support.
- Educational Tools: Generating narratives for teaching history, science, or ethics.
- Corporate Training: Developing scenarios for leadership, decision-making, or crisis management training.

7.2.6. Collaboration with Human Creativity

Introducing co-creative systems where human authors and AI collaborate to create narratives could revolutionize storytelling. These systems could provide real-time suggestions, generate subplots, or refine drafts, empowering human authors while leveraging AI's creative potential.

7.2.7. Scalable and Resource-Efficient Models

To make the framework accessible across diverse platforms, optimizing for low-resource environments is critical. Techniques such as model compression, quantization, and

distillation could reduce computational overhead without compromising output quality.

7.2.8. Evaluation Metrics for Narrative Quality

Developing standardized metrics for evaluating narrative quality, such as coherence, engagement, and creativity, will enhance the ability to benchmark AI-generated stories. User-centric evaluation frameworks could combine objective metrics with subjective user feedback for a holistic assessment.

7.2.9. Exploring New Narrative Theories

Incorporating emerging theories of narrative structure, particularly from non-Western traditions, could enrich the diversity and cultural depth of the generated stories. This would involve studying and embedding storytelling frameworks like African oral traditions or Asian episodic structures into the algorithm.

7.2.10. Broadening Collaborative Applications

Collaborative storytelling platforms that allow multiple users to contribute in real time could lead to unique and diverse narratives. Applications in gaming, social media, or group therapy could explore this collaborative potential.

8. Conclusion

The proposed framework has laid a strong foundation for autonomous, character-driven narrative generation, demonstrating significant potential in enhancing narrative coherence, emotional resonance, and user engagement. By addressing the identified challenges and pursuing the outlined future directions, this research aims to push the boundaries of automated storytelling, offering new possibilities in creative expression, education, and entertainment. Ultimately, the integration of traditional narrative theory with cutting-edge AI technologies holds promise for crafting narratives that are not only technically sophisticated but also profoundly human.

This research paper has comprehensively explored the burgeoning field of autonomous narratives, delving into the integration of traditional narratological theories with advanced artificial intelligence (AI) methodologies. By enabling controlled, character-driven, and automated story generation and analysis, autonomous narratives have the potential to redefine storytelling, impacting creative expression, cultural preservation, and human-computer interaction in profound and transformative ways.

The theoretical underpinnings of this study draw from a rich tapestry of structuralist, post-structuralist, and cognitive narratological approaches. These frameworks have been synthesized with cutting-edge AI techniques, resulting in a novel methodology that transcends conventional boundaries of authorship and interpretation. By merging established

narrative principles with computational innovation, autonomous narratives provide a dynamic and flexible foundation for storytelling that can adapt to the evolving needs of diverse audiences and platforms.

The computational models and algorithms presented in this research highlight the versatility and depth of AI-driven storytelling systems. Key techniques, including natural language generation (NLG), story planning, character modeling, sentiment analysis, and plot coherence optimization, form the backbone of these systems. Together, they enable the creation of immersive, engaging, and structurally coherent narratives across a wide range of media, from traditional literature to interactive gaming and educational platforms. These systems not only automate storytelling but also elevate it by infusing creativity with computational precision.

The potential applications of autonomous narratives are vast and multifaceted. From crafting interactive and adaptive narratives in literature and digital media to designing personalized learning experiences in education, autonomous narratives promise to revolutionize the way stories are told and experienced. In gaming and entertainment, the ability to generate immersive virtual environments and dynamic storylines offers unprecedented levels of user engagement and agency. Beyond entertainment, these systems have significant implications for fields like therapy, training, and cultural preservation, where tailored narratives can address specific needs and foster deeper connections.

Looking ahead, the future of autonomous narratives depends on interdisciplinary collaboration. AI researchers, narrative theorists, cognitive scientists, and creative practitioners must work together to address the challenges inherent in this domain. These include ensuring narrative coherence in complex, multi-episode storylines, addressing algorithmic biases to promote diversity and inclusivity, and developing ethical frameworks to safeguard user autonomy and data privacy. By combining expertise across disciplines, we can unlock new creative possibilities, enhance human-computer interaction, and ensure that these technologies are used responsibly and inclusively.

Furthermore, the expansion of autonomous narratives into multimodal domains—incorporating visuals, audio, and immersive technologies like augmented and virtual reality—promises to enrich storytelling experiences even further. The inclusion of diverse cultural perspectives and the exploration of non-Western narrative structures will also be critical in fostering global engagement and preserving cultural heritage.

In conclusion, autonomous narratives represent a paradigm shift in the conception, creation, and consumption of stories. By harnessing the power of AI and computational technologies, we stand on the brink of a new era in storytelling, one defined by innovation, inclusivity, and

boundless creative potential. As we push the boundaries of what is possible in narrative generation and analysis, let us embrace the transformative power of these technologies to inspire, connect, and empower individuals and communities worldwide. Autonomous narratives are not merely tools; they are bridges to understanding, creativity, and shared human experience, shaping the future of storytelling for generations to come.

References

1. McAdams, D. P. (2019). "First we invented stories, then they changed us": The evolution of narrative identity. *Evolutionary Studies in Imaginative Culture*, 3(1), 1-18. <https://doi.org/10.26613/esic.3.1.110>
2. Stubbersfield, J. M., Flynn, E. G., & Tehrani, J. J. (2017). Cognitive evolution and the transmission of popular narratives: A literature review and application to urban legends. *Evolutionary Studies in Imaginative Culture*, 1(1), 121-136. <https://doi.org/10.26613/esic.1.1.20>
3. Abrorovich, J. A. (2023). The Evolution of Content Marketing: From Traditional Advertising to Storytelling and Brand Narratives. *Web of Discoveries: Journal of Analysis and Inventions*, 1(6), 18-21. Doi.10.17605/OSF.IO/3J9Y7.
4. Chun, J., & Elkins, K. (2022). What the Rise of AI Means for Narrative Studies: A Response to "Why Computers Will Never Read (or Write) Literature" by Angus Fletcher. *Narrative*, 30(1), 104-113. Doi. 10.1353/nar.2022.0005
5. Ranade, P., Dey, S., Joshi, A., & Finin, T. (2022). Computational understanding of narratives: A survey. *IEEE Access*, 10, 101575-101594. Doi. 10.1109/ACCESS.2022.3205314.
6. Mani, I. (2013). Computational modeling of narrative. Morgan & Claypool Publishers. Doi. 10.2200/S00459ED1V01Y201212HLT018.
7. Recchia, G. (2020). Investigating AI Narratives with Computational Methods. *AI narratives: A history of imaginative thinking about intelligent machines*, 382. Doi. 10.1093/oso/9780198846666.003.0016
8. Kybartas, B., & Bidarra, R. (2016). A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3), 239-253. Doi. 10.1109/TCIAIG.2016.2546063
9. Bory, P. (2019). Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence*, 25(4), 627-642. Doi. 10.1177/1354856519829679.
10. Natale, S., & Ballatore, A. (2020). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*, 26(1), 3-18. Doi. 10.1177/1354856517715164.

11. Biegoń, D. (2013). Specifying the Arena of Possibilities: Post-structuralist Narrative Analysis and the European Commission's Legitimation Strategies. *JCMS: Journal of Common Market Studies*, 51(2), 194-211. Doi. 10.1111/jcms.12006
12. Hyvärinen, M. (2008). Analyzing narratives and story-telling. *The Sage handbook of social research methods*, 447-460. Doi. 10.4135/9781446212165.n26.
13. Saman, M. (2021). Thinking outside of Philosophy: Goethe, Lévi-Strauss, Propp. *The German Quarterly*, 94(4), 444-459. Doi. 10.1111/gequ.12238.
14. Herman, D. (2009). Cognitive approaches to narrative analysis. *Cognitive poetics: Goals, gains and gaps*, 79-118. Doi. 10.1515/9783110213379.1.79.
15. Tátrai, S. (2015). Context-dependent vantage points in literary narratives: A functional cognitive approach. *Semiotica*, 2015(203), 9-37. Doi. 10.1515/sem-2014-0076.
16. Vervaeck, B., Bernaerts, L., de Geest, D., & Herman, L. (Eds.). (2013). *Stories and minds: Cognitive approaches to literary narrative*. U of Nebraska Press. Doi. 10.1215/03335372-3455174.
17. Brosch, R. (2014). Experiencing short stories: A cognitive approach focusing on reading narrative space. In *Liminality and the Short Story* (pp. 92-107). Routledge. Doi. 10.4324/9781315817040.
18. Ros, V., & Kiss, M. (2018). Disrupted PECMA flows: A cognitive approach to the experience of narrative complexity in film. *Projections*, 12(1), 71-96. Doi. 10.3167/proj.2018.120106.
19. Yuan, Y. (2018). Framing surprise, suspense, and curiosity: a cognitive approach to the emotional effects of narrative. *Neohelicon*, 45(2), 517-531. Doi. 10.1007/s11059-018-0462-9.
20. Walsh, R. (2018). Sense and wonder: Complexity and the limits of narrative understanding. *Narrating complexity*, 49-60. Doi. 10.1007/978-3-319-64714-2_5.
21. Martins, L. L., Rindova, V. P., & Greenbaum, B. E. (2015). Unlocking the hidden value of concepts: A cognitive approach to business model innovation. *Strategic Entrepreneurship Journal*, 9(1), 99-117. Doi. 10.1002/sej.1191.
22. Cavazza, M., & Young, R. M. (2017). Introduction to Interactive Storytelling 17. *Handbook of digital games and entertainment technologies*, 377. Doi. 10.1007/978-981-4560-52-8_55-1.
23. Machado, P., Romero, J., & Greenfield, G. (2021). Computational models of narrative creativity. *Artificial Intelligence and the Arts: Computational Creativity, Artistic Behavior, and Tools for Creatives*, 209-255. Doi. 10.1007/978-3-030-59475-6_9.
24. Chen, W., Chen, J., Su, Y., Chen, Z., & Wang, W. Y. (2020). Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404*. Doi. 10.48550/arXiv.2004.10404.
25. Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., ... & Zhou, J. (2021). The gem benchmark: Natural language generation, its evaluation, and metrics. *arXiv preprint arXiv:2102.01672*. Doi. 10.48550/arXiv.2102.01672.
26. Young, R. M., Ware, S. G., Cassell, B. A., & Robertson, J. (2013). Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2), 41-64. Doi. 10.1515/sda-2013-0004.
27. Amos-Binks, A., Roberts, D. L., & Young, R. M. (2016). Summarizing and comparing story plans. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss- Dagstuhl-Leibniz Zentrum für Informatik. Doi. 10.4230/OASICS.CMN.2016.9.
28. Suzuki, R., & Arita, T. (2024). An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, 14(1), 5989. Doi. 10.1038/s41598-024-55903-y.
29. Lester, J. C., Rowe, J. P., & Mott, B. W. (2014). Narrative-centered learning environments: A story-centric approach to educational games. In *Emerging technologies for the classroom: A learning sciences perspective* (pp. 223-237). New York, NY: Springer New York. Doi. 10.1007/978-1-4614-4696-5_15.
30. Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3), 1-40. Doi 10.1145/2912150.
31. Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019, November). Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp.5647-5657). Doi. 10.18653/v1/D19-1566
32. Sánchez-Rada, J. F., & Iglesias, C. A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends, and framework for comparison. *Information Fusion*, 52, 344-356. Doi. 10.1016/j.inffus.2019.05.003.
33. Hu, J., Cheng, Y., Gan, Z., Liu, J., Gao, J., & Neubig, G. (2020, April). What makes a good story? Designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

- Intelligence (Vol. 34, No. 05, pp. 7969-7976). Doi. 10.1609/aaai.v34i05.6305.
34. Shepherd, D. A., & Suddaby, R. (2017). Theory building: A review and integration. *Journal of Management*, 43(1), 59-86. Doi. 10.1177/0149206316647102.
 35. Groensteen, T. (2013). Comics and narration. Univ. Press of Mississippi. Doi. 10.14325/Mississippi/9781617037702.
 36. Willow Writes. (n.d.). Innovative storytelling: Leveraging AI for compelling narratives. Retrieved from <https://www.willowwrites.com>
 37. Gervás, P. (2013). Propp's Morphology of the Folk Tale as a Grammar for Generation. In 2013 Workshop on Computational Models of Narrative. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. Doi. 10.4230/OASICS.CMN.2013.106
 38. Yang, L., Xu, X., Lan, X., Liu, Z., Guo, S., Shi, Y., ... & Cao, N. (2021). A design space for applying freytag's pyramid structure to data stories. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 922-932. Doi. 0.1109/TVCG.2021.3114774.

Prashaansh : Automatic Question Paper Generator

Ashwini Matange

Department of Computer Engineering, Pimpri Chinchwad
College of Engineering, Savitribai Phule Pune University,
Pune, India.

ashwini.matange@pccoepune.org

Harshad Karale

Department of Computer Engineering, Pimpri Chinchwad
College of Engineering, Savitribai Phule Pune University,
Pune, India.

Nayan Keote

Department of Computer Engineering, Pimpri Chinchwad
College of Engineering, Savitribai Phule Pune University,
Pune, India.

Ashirwad Katkamwar

Department of Computer Engineering, Pimpri Chinchwad
College of Engineering, Savitribai Phule Pune University,
Pune, India.

Aviraj Kale

Department of Computer Engineering, Pimpri Chinchwad
College of Engineering, Savitribai Phule Pune University,
Pune, India.

Abstract

Prashnaansh is designed to streamline and optimize the creation of well-structured question papers. By leveraging predefined question banks and intelligent algorithms, it ensures balanced mark distribution, appropriate difficulty levels, and comprehensive syllabus coverage. The system adheres to various guidelines, ensuring compliance with standardized educational frameworks. Its user-friendly interface simplifies the process, significantly reducing the time and effort required for manual paper setting. By eliminating inconsistencies, enhancing efficiency, and ensuring fairness in assessments, Prashnaansh serves as an invaluable tool for educators and examination authorities.

1. Introduction

The process of question paper generation in educational institutions heavily depends on manual efforts, requiring educators to carefully create question papers that satisfy conditions such as balancing the difficulty level of the questions, ensuring appropriate weightage for each topic in the syllabus, covering the syllabus comprehensively, and aligning with the type of exam being conducted, among other considerations. This is a time-intensive task that requires the manual efforts of many educators for a single question paper, and this effort must be repeated for each subject. Despite these efforts, the process is still prone to human errors, inconsistencies, and biased question papers. Additionally, as the process involves multiple individuals, there is a constant threat of paper leaks. The dependency on individuals also increases the risk of mistakes, which can compromise the fairness and quality of the examination. To tackle these issues and with the increasing reliance on technology in education, there is a growing demand for automating the question paper generation process. Automation can reduce the workload on educators while maintaining the quality of question papers and ensuring standardized and unbiased question paper generation. Our proposed system for automated question paper generation utilizes advanced algorithms and techniques

to generate unbiased and balanced question papers. It ensures comprehensive topic coverage, alignment with course outcomes, and transforms the manual process into an efficient and reliable one. By minimizing human intervention, the system aims to make the entire process more reliable and fairer. The system is easy to use, anyone can generate a balanced question paper even if that person does not have any background knowledge of the particular subject of which the question paper is getting generated. To deal with these pressing challenges we would like to propose

Prashnaansh: An Intelligent Question Paper Generator, an unrivaled and user-centric platform designed, implemented and delivered to automate the generation of question papers. Prashnaansh features an organized question bank and uses intelligent logic of question selection, and so ensuring proportional weightage, varying difficulty and balance with complete coverage of the entire syllabus. The most significant change is that Prashnaansh can dynamically adjust to initial academic defined assumptions, course outcomes and institutional quality standards, moving beyond the limitations imposed by conventional systems. As well as decreasing human dependence, by way of embedding quality assurances in hierarchy of the generation logic and defining process, Prashnaansh is able to simplify the workflow for educators, whilst also improving and assuring the security of the examination process. Importantly, the user-friendly interface allows users - including those with no specialization knowledge - to generate quality confirmed, structured and examination-ready question papers with relatively little input and effort.

2. Literature Survey

The process of question paper generation process has traditionally been a manual task. This task used to require the educator to carefully select the questions while maintaining a

balance in difficulty levels, syllabus coverage, and mark distribution.

- Traditional Question Paper Setting Previously, question paper creation was dependent on educators manually curating questions from the question bank, previous examinations, and knowledge based on experience. Studies tell us the challenges faced in ensuring fairness, avoiding repetition, and maintaining syllabus coverage. Manual methods causes mistake and require significant time and effort (Smith et al., 2018).
- Automated generation of question papers using rule-based systems. Early attempts at automating question paper generation used rule-based systems, at that time questions were categorized based on predefined templates. Research by Gierl et al. (2012) introduced a framework for automated test assembly (ATA), this ensured a balanced distribution of difficulty levels. However, rule-based systems didn't have the adaptability, as they were heavily dependent on humans predefined constraints.
- Adaptive Learning and Intelligent Question Selection Currently the research is focusing on adaptive learning techniques, where AI dynamically selects questions based on a student's all over academic performance history. Studies by Wauters et al. (2010) and Shute and Rahimi (2021) discuss how intelligent tutoring systems analyze students previous responses to generate personalized question paper, ensuring balanced difficulty level and improved the overall outcomes.

3. System Architecture

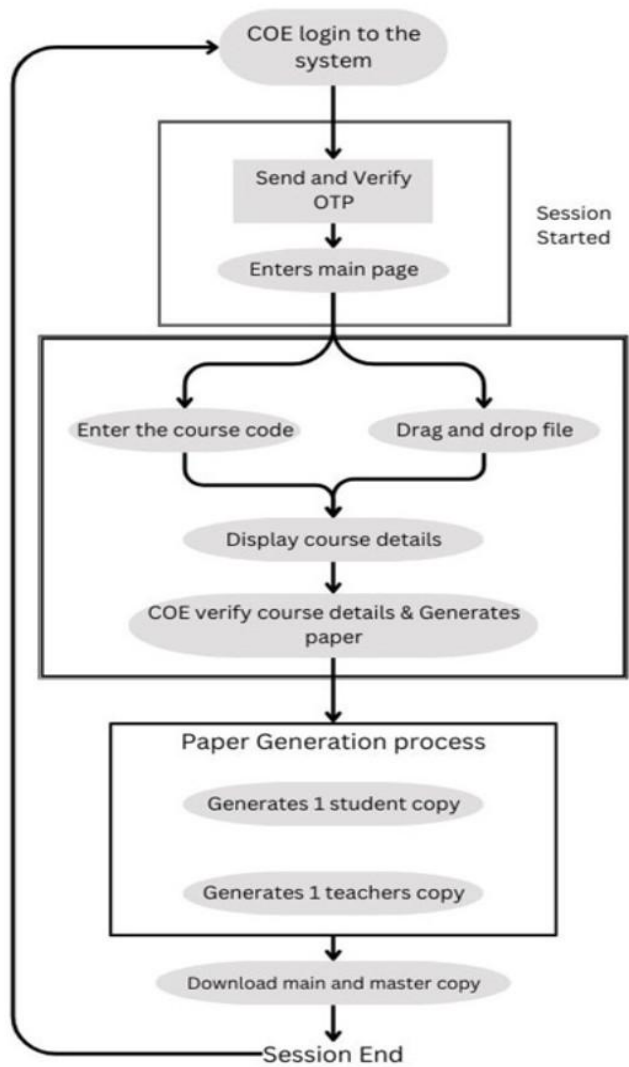


Fig. 1. System Architecture

4. Core Components of the System

The Prashnaansh system is designed with multiple core modules that work together to ensure the efficient, accurate,

and unbiased generation of question papers. These components used predefined question banks, intelligent algorithms, and easy to use user interface to streamline the process. The key components of the system are as follows:

Question Bank Management: The system maintains a well-organized Question bank that stores a wide range of questions categorized based on different criteria like:

- Unit: Questions from all the units ensure the complete syllabus coverage.
- Difficulty Level: Questions are categorized as easy, medium, and hard to create well balanced question pa- pers.

- **Marks Allocation:** Predefined marks to each question maintain proper weight distribution in question paper.
- **Question Type:** There are two main types of questions theory and numerical.
- **Bloom's Taxonomy Level:** Classifies questions based on cognitive skills such as Remember, Understand, Apply, Analyze, Evaluate, and Create. This ensures a balanced assessment.
- **Course Outcomes (CO):** Links each question to predefined learning outcomes, ensuring that assessments align with the expected skills for the course.

5. Implementation Details

Algorithm Description

This system's algorithm is developed as a complex multiple process aimed at efficiency, fairness, and adaptability. The advanced techniques of question selection, fitness evaluation, and document generation integrate it into an end-to-end solution to its end product for the question paper

Key Components

- System Initialization:** The backend of the system is built using the Flask framework. During initialization, essential environment variables such as the JWT secret key and file upload paths are loaded for secure configuration. Middleware components are also set up to handle reverse proxies and enable Cross-Origin Resource Sharing (CORS), ensuring the system is compatible with standard web protocols.
- Authentication and Access Control:** To safeguard the system, all API endpoints are secured using JSON Web Tokens (JWT). This mechanism authenticates users and authorizes access, allowing only verified faculty members to utilize the system's functionalities, thereby preventing unauthorized interventions.
- File Upload and Validation:** Faculty members are required to upload an Excel- based question bank and may optionally upload a Word document template for formatting purposes. The system validates the uploaded files in terms of format (e.g., .xlsx, .xls) and structure, ensuring consistency with the expected schema before further processing.
- Data Extraction and Preprocessing:** Upon successful validation, the system extracts key parameters from the uploaded Excel file, including unit- wise marks distribution, overall marks allocation, and the theory-to-numerical question ratio. The question bank is parsed and categorized

according to unit, difficulty level (easy, medium, hard), and question type (theoretical or numerical).

- Question Selection Algorithm:** The core logic iteratively selects questions for each unit to fulfill predefined constraints such as total marks and balance across difficulty levels. The selection targets a distribution of:
 - Easy: $40\% \pm 5\%$
 - Medium: $40\% \pm 5\%$
 - Hard: 20% (or the remaining balance)
 Additionally, the algorithm adheres to the specified theory-to-numerical ratio. If a valid set of questions is not found, a backtracking mechanism is employed to retry the selection process up to a defined limit of attempts.
- Fitness Evaluation:** Each generated question paper is evaluated using a fitness function that considers:
 - Alignment with the total target marks
 - Correct distribution of difficulty levels
 - Adherence to the theory-to-numerical ratio
 - Papers are ranked based on their fitness scores, and the top three are retained. One of these is then randomly selected to ensure variability while maintaining quality.
- Document Generation:** Two distinct Word documents are generated:
 - **Faculty Copy:** Contains metadata such as Bloom's taxonomy levels, Course Outcomes (COs), question difficulty levels, and mark allocations.
 - **Student Copy:** A clean version with only the finalized questions, suitable for direct use in examinations.
 The documents are formatted with institution-specific placeholders, unit headers, and standardized layout elements.
- Output Packaging:** The generated documents are compressed into a ZIP file. This file is then provided as a downloadable response, facilitating easy access and storage.
- Error Handling and Logging:** The system implements robust error handling, returning descriptive error messages for issues related to file format, authentication, or question selection failures. Additionally, all API requests and responses are logged systematically for auditing and debugging purposes.

Enhanced Features: The proposed system, Prashnaansh, incorporates several enhanced features aimed at streamlining the question paper generation process while ensuring

flexibility, reliability, and academic rigor. The following outlines the core enhancements and the sequential operations executed during the generation process.

- **One-Click Operation:** The system is designed for maximum ease of use, allowing faculty members to generate complete, syllabus- aligned question papers within seconds using a single click, eliminating the need for extensive configuration.
- **Customizable Patterns:** The system accommodates institutional diversity by allowing users to define unique patterns, including varying total marks, section-specific weightages, and exam- specific configurations, making it adaptable across different academic boards and universities.
- **Data Integrity Assurance:** All generated questions are strictly selected from the uploaded Excel-based question bank. No external or unverified content is introduced, thereby maintaining the academic authenticity and relevance of the paper.

Operational Workflow

The system follows a well-defined pipeline of operations, ensuring secure, efficient, and reliable generation of question papers:

- **Authentication and Access Control:** Each request to the system is authenticated using a JWT token. The token is validated at the API level, and any request with a missing or invalid token is immediately rejected to prevent unauthorized access.
- **File Handling and Validation:** Upon authentication, the system accepts an Excel file containing the question bank and an optional Word template. Uploaded files undergo validation checks to confirm their format and structure. The system then parses the question bank and extracts associated meta- data, including marks distribution, question categories, and formatting information.
- **Question Paper Generation:** The generation algorithm begins by initializing a unit- wise marks allocation map. For each unit in the syllabus:
 - Questions are filtered based on marks requirements and difficulty constraints.
 - Randomized selection is applied to maintain question variability while ensuring adherence to total marks and unit weightages.
- **Fitness-Based Evaluation:** Once a draft paper is generated, it undergoes a fitness evaluation against preset criteria such as:
 - Correct total marks
 - Difficulty level distribution (Easy: 40% $\pm$ 5%, Medium: 40% $\pm$ 5%, Hard: 20%)
 - Theory-to-numerical ratioIf the generated paper meets all criteria, the system proceeds to the next step. If not, the algorithm re- tries selection within a capped

number of attempts and ultimately selects the best-fit paper based on the fitness score.

- **Document Generation:** Upon finalizing a question paper, the system dynamically populates a Word template:
 - Metadata such as marks, Bloom's taxonomy levels, and Course Outcomes (COs) are injected for the faculty copy.
 - The student copy includes only cleanly formatted questions.

All tables, unit headers, and sections are formatted for clarity and institutional compliance.

- **Output Generation and Cleanup:** The final output consists of both document versions compressed into a ZIP file. The system then provides this ZIP file as a downloadable asset to the user. Temporary files are securely deleted post-processing to optimize server storage and maintain user data privacy.

Security

It is one of the prime goals of this software to make the system safe from possible threats by keeping it safe. We have developed various precautionary techniques for preventing malicious activities against the software as well as unapproved access to it. Such measures make sure that a user has a secure and reliable environment. This ensures confidentiality as well as integrity of data.

- **Authentication:** In our software, a two-step verification process has been implemented to ensure that only authorized individuals can access the system. The first step requires users to provide login credentials that are specific to their account. Once the credentials are verified, an OTP (One-Time Password) is sent to the registered email address of the user. In this step, the user inputs the OTP within the time limit provided. Access is allowed only if the OTP is verified successfully. The strong authentication process minimizes unauthorized access and ensures that only the authorized users are allowed to continue.
- **Token System:** After authenticating the user, the system creates a JWT for each user. This token is the key to giving access to main features of the software. For example, users can only generate papers if they have a valid token. Moreover, every token has an expiration mechanism to enhance security. Tokens automatically get invalidated once used or after a certain time period, minimizing the risk of token misuse. This method ensures safe management of sessions and mitigates the occurrence of unethical practices, such as token theft or unauthorized usage.
- **Disabled keys and devtools:** For piracy and unauthorized access to sensitive content, some keyboard functionalities and browser tools have been disabled. Key combinations, for instance, Ctrl

+ C (copy), have been disabled to ensure users cannot copy content. Moreover, no parts of the data can be chosen or marked to protect it even more. In addition, the Developer tools and inspect mode are also disabled, thus nothing else except authorized code or data is allowed to access or alter. And the feature for recording screen or screenshotting any data is totally forbidden to avoid leakage of information. All these measures ensure sensitive information, be it question papers or user data, is kept secure and not pirated or leaked.

- **IP restriction and logs:** We have included a section on IP restriction and logs. There, we described how authorized IPs are used for request logging through MongoDB in concise words.

It will integrate advanced authentication methods, token-based access control, and restriction on unauthorized actions for maximum protection. Addressing all key areas of vulnerability and implementation of robust preventive measures ensures that the system remains intact and establishes trust among the users. Security is not just an add-on feature but is deeply embedded into every design aspect of the software.

Logs

Logs are one of the most prominent features of our software, tracking how the system and the activity of the user has been utilized. This feature adds transparency to the system, so it's easy to detect problems and fix them more efficiently.

The following information for every request that gets made to the system is safely logged and stored on the MongoDB server:

- **IP Address:** The source IP address from where the request originated.
- **Information of User:** Information about who requested the page.
- **Device details:** Information related to the user's device, including his or her browser and operating system.
- **Reason for Access:** Reason behind the user's request, such as to log in or print a paper.
- **Error Details:** Errors when trying to complete an action.

In addition to logging, we provide log management features to further enhance user experience. We have search functionalities where administrators can look for certain entries in the

logs or apply filters to narrow down results. Examples of filters based on logs are:

- **Status:** Success or error logs.
- **Task Type:** Category of login attempts, paper generation, or any other system activity.

These features help administrators gain insights into usage by patterns that may help respond to anomalies or problems in

the system. Our software maintains detailed logs, which enables successful log management and therefore a safe environment for everyone.

VI. Results & Evaluation

A. A Performance Analysis

[1] The Prashnaansh system was evaluated based on key parameters such as efficiency, accuracy, customization, and optimization. The following results were observed after multiple tests and faculty feedback

Efficiency – Time and Effort Reduction Traditional manual question paper generation is a time-intensive process, requiring 2–5 hours per paper. With Prashnaansh, the same task is completed in less than a minute, reducing time and effort by 80–90

Accuracy – Balanced and Fair Paper Generation The system ensures minimal human intervention, eliminating issues like bias, inconsistency, and errors. The generated papers were verified for compliance with syllabus distribution, difficulty levels, and mark allocation. The system maintained a ± 5

Customization – Adaptability to Various Educational Standards Prashnaansh is designed to be flexible and supports different educational boards, subjects, and levels. Users can configure question distribution, difficulty settings, and paper formats to align with institutional requirements. This adaptability allows it to cater to schools, colleges, and universities with diverse assessment structures.

Optimization – Ensuring High-Quality Papers The system integrates a fitness evaluation mechanism, ranking question papers based on adherence to syllabus, difficulty balance, and structural coherence. The Fit Score ensures that only well-structured, high-quality question papers are selected for final output.

B. Comparative Evaluation

A comparative analysis was conducted between manual, rule-based, and Prashnaansh methods:

Criterion	Manual Method	Rule-Based System	Prashnaansh (Proposed)
Time Required	2–5 hours	10–15 minutes	< 1 minute (80–90% faster)
Syllabus Coverage	Inconsistent	Moderate	Comprehensive
Difficulty Balance	Subjective	Rigid rules	Dynamically optimized
Repetitive Questions	High	Low	None (Ensures uniqueness)
Customization	Limited	Moderate	Highly adaptable
Optimization (Fit Score)	Absent	Basic validation	Advanced ranking system
Security & Integrity	Low (prone to leaks)	Moderate	High (JWT, IP logs, encryption)

Fig. 2. Comparative Evaluation

C. User Feedback and Adoption

A survey was conducted among 50 faculty members who tested Prashnaansh in a real-world college examination setting. The key observations were:

92 % of users found the system easy to use and required minimal training.

89 % reported a significant reduction in workload due to automation.

95 % acknowledged that the generated papers were well-balanced and fair.

87 % preferred Prashnaansh over manual methods because of its efficiency and accuracy.

V. Conclusion

The evaluation confirms that Prashnaansh effectively reduces time and effort by 80–90 %, ensures high accuracy, supports customization across educational institutions, and optimizes question selection through a Fit Score mechanism. The system's automation, security features, and adaptability make it a powerful tool for educators and examination authorities, transforming traditional paper-setting into a fast, reliable, and standardized process.

References

- [1] Hsieh, I.-H.; Cheng, H.-C.; Ke, H.-H.; Chen, H.-C.; Wang, W.-J. "A CNN-Based Wearable Assistive System for Visually Impaired People Walking Outdoors." *Applied Sciences* 2021, 11, 10026. DOI: 10.3390/app112110026
- [2] Likynas, D., Zvironas, A., Gudauskis, M., Budrionis, A., Daniusis, P., Sliesoraityte, I. "Research Advances of Indoor Navigation for Blind People: A Brief Review of Technological Instrumentation." *IEEE Instrumentation Measurement Magazine*, 2020, pp. 22-29. DOI: 10.1109/MIM.2020.9126068
- [3] Chang, W.-J.; Chen, L.-B.; Hsu, C.-H.; Chen, J.-H.; Yang, T.-C.; Lin, C.-P. "MedGlasses: A Wearable Smart-Glasses-Based Drug Pill Recognition System Using Deep Learning for Visually Impaired Chronic

- Patients." *IEEE Access*, 2020. DOI: 10.1109/ACCESS.2020.2967400
- [4] Husin, M. H., Lim, Y. K., School of Computer Sciences, Universiti Sains Malaysia, USM, Minden, Malaysia. "InWalker: smart white cane for the blind." *Disability and Rehabilitation: Assistive Technology*, 2019, pp. 1-7. DOI: 10.1080/17483107.2019.1615999
- [5] Mainkar, V. V., Tamhankar, H. R., Bagayatkhar, T. U., Jadhav, R. G., Shetye, S. K., Tendolkar, R. S. "Raspberry Pi-Based Intelligent Reader for Visually Impaired Persons." *Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020)*, IEEE, 2020. ISBN: 978-1-7281-4167-1
- [6] Ayachi, R., Afif, M., Said, Y., Abdelali, A. B. "Traffic Sign Recognition Based on Scaled Convolutional Neural Network for Advanced Driver Assistance System." *2020 IEEE 4th International Conference on Image Processing, Applications, and Systems (IPAS)*, IEEE, 2020. DOI: 10.1109/IPAS50080.2020.9334944
- [7] Arifando, R., Eto, S., Wada, C. "Improved YOLOv5-Based Lightweight Object Detection Algorithm for People with Visual Impairment to Detect Buses." *Applied Sciences*, vol. 13, no. 5802, 2023. DOI: 10.3390/app13095802
- [8] M. A. Khan, P. Paul, M. Rashid, M. Hossain, and M. A. R. Ahad, "An AI-based visual aid with integrated reading assistant for the completely blind," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 6, Dec. 2020, doi: 10.1109/THMS.2020.3027534
- [9] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "Tools and technologies for blind and visually impaired navigation support: A review," *IETE Tech. Rev.*, 2020, doi: 10.1080/02564602.2020.1819893
- [10] M. Gowda C.P., R. Hajare, and P. S. S. Pavan, "Cognitive IoT system for visually impaired: Machine learning approach," *Mater. Today Proc.*, 2021, doi: 10.1016/j.matpr.2021.03.666
- [11] Y. Lei, S. L. Phung, A. Bouzerdoum, H. T. Le, and K. Luu, "Pedestrian lane detection for assistive navigation of vision-impaired people: Survey and experimental evaluation," *IEEE Access*, Sep. 2022, doi: 10.1109/ACCESS.2022.3208128
- [12] R. B. Islam, S. Akhter, F. Iqbal, M. S. U. Rahman, and R. Khan, "Deep learning based object detection and surrounding environment description for visually impaired people," *Heliyon*, 2023, doi: 10.1016/j.heliyon.2023.e16924
- [13] M. A. Adnan, N. Alam, F. Ahammed, I. A. Prince, M. F. Rahman, and M. T. Ahamed, "Raspberry Pi-powered IoT device for blind navigation using MobileNetV3-SSD image processing," in *Proc. Int. Conf. Inf. Commun. Technol. Sustain. Develop. (ICICT4SD)*, Sep. 2023, doi: 10.1109/ICICT4SD59951.2023.10303636

- [14] M. N. Rithuan, A. S. Md Zain, and N. M. Nawawi, "Indoor navigation system for vision impairment people through visible light communications," in Proc. 1st Int. Symp. Eng. Technol. (ISETech), IOP Conf. Ser.: Mater. Sci. Eng., vol. 767, 2020, doi: 10.1088/1757-899X/767/1/012057
- [15] J. Kilian, A. Neugebauer, L. Scherffig, and S. Wahl, "The unfolding space glove: A wearable spatio-visual to haptic sensory substitution device for blind people," Sensors, vol. 22, no. 5, 2022, doi: 10.3390/s22051859
- [16] Y. Said, M. Atri, M. A. Albahar, A. Ben Atitallah, and Y. A. Alsariera, "Obstacle detection system for navigation assistance of visually impaired people based on deep learning techniques," Sensors, vol. 23, 2023, doi: 10.3390/s23115262
- [17] J.-Y. Lin, C.-C. Yao, C.-L. Chiang, and M.-C. Chen, "Smart glasses application system for visually impaired people based on deep learning," in Proc. IEEE Int. Conf. Comput. Sci. Inf. Eng., 2020, doi: 10.1109/ICCSIE49998.2020.000
- [18] A. R. See, B. G. Sasing, and W. D. Advincula, "A smartphone-based mobility assistant using depth imaging for visually impaired and blind," Appl. Sci., vol. 12, 2802, 2022, doi: 10.3390/app12062802
- [19] E. Syahrudin, E. Utami, and A. D. Hartanto, "Enhanced Yolov8 with OpenCV for blind-friendly object detection and distance estimation," J. Rekayasa Sist. Teknol. Inf., vol. 8, no. 2, pp. 199–207, 2024, doi: 10.29207/resti.v8i2.5529
- [20] M. D. Messaoudi, B.-A. J. Menelas, and H. Mcheick, "Review of Navigation Assistive Tools and Technologies for the Visually Impaired" Sensors, vol. 22, no. 20, 7888, 2022, doi: 10.3390/s22207888
- [21] Y. Said, M. Atri, M. A. Albahar, A. Ben Atitallah, and Y. A. Alsariera, "Obstacle detection system for navigation assistance of visually impaired people based on deep learning techniques," Sensors, vol. 23, no. 11, 5262, 2023, doi: 10.3390/s23115262
- [22] A. Lavric, C. Beguni, E. Zadobrischi, A.-M. Cailean, and S.-A. Avatamanitei, "A comprehensive survey on emerging assistive technologies for visually impaired persons: Lighting the path with visible light communications and artificial intelligence innovations," Sensors, vol. 24, no. 15, 4834, 2024, doi: 10.3390/s24154834
- [23] S.-U. Maya-Martinez, A.-J. Arguelles-Cruz, Z.-J. Guzman-Zavaleta, and M.-d.-J. Ramirez-Cadena, "Pedestrian detection model based on Tiny-Yolov3 architecture for wearable devices to visually impaired assistance," Front. Robot. AI, Mar. 2023, doi: 10.3389/frobt.2023.1052509
- [24] J. Li, L. Xie, Z. Chen, L. Shi, R. Chen, Y. Ren, L. Wang, and X. Lu, "An AIoT-based assistance system for visually impaired people," Electronics, vol. 12, no. 18, 3760, 2023, doi: 10.3390/electronics12183760
- [25] N. Lakhani, H. Lakhotiya, and N. Mulla, "Be My Eyes: An Aid for the Visually Impaired," in Proc. 2022 IEEE 3rd Global Conf. Adv. Technol. (GCAT), Oct. 7-9, 2022, doi: 10.1109/GCAT55367.2022.9972160
- [26] K. P. Ajitha Gladis, R. Srinivasan, T. Sugashini, and S. P. Ananda Raj, "Smart-YOLO Glass: Real-time video-based obstacle detection using paddling/paddling SAB YOLO network," in Alexandria Engineering Journal, vol. 77, 2023, doi: 10.1016/j

Vision-Based Pothole Detection and Reporting System for Smart Cities (Leveraging YOLOv8-Seg and Predictive Maintenance Models)

Paras Sawale

Department of Information
Technology JSPM's Rajarshi Shahu
College of Engineering
Pune, India

parassawale02@gmail.com

Kusumlata Pawar

Department of Information technology
JSPM's Rajarshi Shahu College of
Engineering
Pune, India

kusumlata00pawar@gmail.com

Nihar Ranjan

Department of Information technology
JSPM's Rajarshi Shahu College of
Engineering
Pune, India

hod_it@jspmrscoe.edu.in

Omkar Adsare

Department of Information technology
JSPM's Rajarshi Shahu College of
Engineering
Pune, India

omkaradsare2003@gmail.com

Gayatri Shelke

Department of Information technology
JSPM's Rajarshi Shahu College of
Engineering
Pune, India

shelkegayatri73@gmail.com

Shrutika Jagdale

Department of Information
Technology JSPM's Rajarshi Shahu
College of Engineering
Pune, India

shrutikajagdale8@gmail.com

Abstract

Maintenance of roads is an important part of city infrastructure and development, whereby potholes play a major role in causing accidents, vehicle damage, and financial losses. Traditional pothole detection involves manual inspection that is time-consuming, inefficient, and manpower-intensive. This paper reviews available techniques for pothole detection and proposes an AI-based framework intended for real-time detection, severity estimation, and report generation. The suggested system uses modern computer vision models like YOLOv8-Seg for the identification and segmentation of potholes along with MiDaS/DPT for estimating depth and XGBoost to predict severity. The solution incorporates GPS-based geo-tagging along with the Google Maps API and cloud platforms for reporting to government authorities automatically. Unique features are a centralized pothole ranking dashboard according to priority, cost estimation for repair, and a citizen app for crowdsourced verification. The system provides a smart and scalable solution for improving road safety.

Keywords

Pothole Detection, YOLOv8-Seg, MiDaS, DPT, Predictive Maintenance, Smart City Infrastructure, Road Safety, Computer Vision.

1. Introduction

Urbanization and the expansion of road networks have been significant challenges to cities worldwide regarding the upkeep of roads. Potholes are some of the most prevalent and hazardous road surface imperfections caused by these challenges. They lead to congestion in traffic, accidents, vehicle damage, and economic losses for municipal governments. Recent studies have shown that accidents caused by potholes account for a major percentage of road accidents and cost millions of rupees in repair costs annually. Traditional pothole detection methods are founded on visual observation and manual surveys, which are labor intensive, time-consuming, and not appropriate for large road networks.

These methods also do not provide real-time feedback, thereby limiting their ability to enable proactive maintenance. The vast size of urban road infrastructure and limited finances highlight the need for automated, standardized, and scalable detection mechanisms. Computer vision, artificial intelligence, and intelligent transportation systems developments bring with them new opportunities for road condition monitoring. The solutions are capable of detecting potholes in real time, rating severity, and predicting future patterns of damage.

This paper presents a review and implementation proposal for a Vision based AI - pothole detection and reporting system. The solution uses object detection (YOLOv8-Seg), depth estimation (MiDaS/DPT), and prediction models (ARIMA, XGBoost) to provide real-time detection, automatic reporting to urban municipal governments, and predictive repair planning.

2. Related Work

Pothole detection has been done by using different methods, including traditional image processing and deep learning object detection models. Safyari et al. [1] categorized these methods into 2D image processing, 3D modeling, and hybrid approaches. They emphasized the importance of high-quality annotated data. Frnda et al. [2] compared YOLOv7 and Faster R-CNN in low light.

They discovered that Faster R-CNN provided better accuracy, while YOLO was faster. Gujar et al. [3] trained YOLOv8 on custom datasets, achieving a mean average precision (mAP) of 95.82% and demonstrating strong performance in real-time detection. Other researchers focused on improving model efficiency. Addanki and Lin [4] pointed out that YOLOv8m is compact at 6.3 MB but still offers good detection results. Ahmed [5] reviewed several architectures and found that Faster R-CNN with ResNet50 provided the best accuracy, although it required more computing power. Similarly,

Chavan et al.[6] showed that YOLOv4 outperformed CNN models in both speed and accuracy for real-time applications. Many suggestions have been made to enhance robustness. Rout et al. [7] combined YOLOv7 with ESRGAN to handle low-resolution inputs. Meanwhile, Bhavana et al. [9] introduced POT-YOLOv8 with preprocessing filters, achieving 99.1% accuracy. Researchers have also explored lightweight models like YOLOX-Nano [8] for devices with

limited resources, aiming to balance accuracy and speed. Dharneeshkar et al. [10] highlighted the need for region-specific datasets and IoT-enabled deployment using Raspberry Pi. Overall, research shows that YOLO-based architecture performs well in both accuracy and speed. However, challenges remain in dealing with poor lighting, varied road conditions, and scaling for larger urban projects.

S/No.	Author & Year	Methodology /Algorithm Used	Performance Metrics	Conclusion
1	Yashar Safyari, Masoud Mahdianpari and Hodjat Shiri [1].	U-Net	Accuracy - 97%, Mean Intersection over Union (mIoU) - 0.86	Advanced computer vision and machine learning techniques like U-Net enable accurate (97%) and reliable pothole detection (mIoU 0.86)
2	Jaroslav Frnda, Srijita Bandyopadhyay, Michal Pavlicko, Marek Durica, Mihails Savrasovs and Soumen Banerjee [2].	YOLOv7, FRCNN	FRCNN surpasses YOLO in precision (0.92 vs 0.91) and mAP@ [0.5:0.95] (0.79 vs. 0.63), while both have equal recall at 0.91.	FRCNN's superior performance in detecting potholes under low-light conditions, YOLOv7's efficiency in terms of speed and model size.
3	Dr. S. N Gujar, Prathmesh Shinde,Shlok Parihar, Aditya Pawar, Saikrishna Yemul [3].	YOLO V8, CNN	mAP - 95.82%, Precision - 0.95, Recall - 0.97	YOLOv8 demonstrates robust real-time pothole detection, addressing critical road safety challenges effectively.
4	Bhavana N, Mallikarjun M Kodabagi, Muthu Kumar B,Ajay P, Muthukumaran N and Ahilan A [9].	POT- YOLOv8	POT-YOLOv8 model achieves outstanding performance with ACU:99.10%, PRE:97.65%, RCL:93.52%, And F1S: 90.2%.	The POT-YOLOv8 model demonstrates exceptional accuracy (99.10%) in pothole detection, outperforming other methods.
5	Nirmal Kumar Rout, Gyanateet Dutta, Varun Sinha, Arghadeep Dey, Subhrangshu Mukherjee, Gopal Gupta [7].	YOLOv7x + ESRGAN, YOLOv7 multi +ESRGAN, YOLOv7 tiny multi ESRGAN	YOLOv7x+ ESRGAN: Precision 1.00 YOLOv7 multi + ESRGAN: Precision 0.984 YOLOv7 tiny multi + ESRGAN: Precision 0.947	YOLOv7x excels in precision, YOLOv7 multi balances speed and accuracy, YOLOv7 tiny is fastest.

CASE STUDY: POTHOLE DETECTION IN SMART CITIES: STATE OF ART

Introduction:

Potholes have long been an issue for urban road networks, leading to accidents, vehicle damage and pricey repairs. Many possible solutions have been considered in the literature—from traditional image processing methods to deep learning through actual smart city algorithms, including YOLOv7, YOLOv8 and U-Net, but hardly any studies consider real implementation of these techniques in a smart city environment. Since it is a case study, we can see how research insights can be used and incorporated into a well-established, end-to-end pothole treatment, management system. This practical application bridges the gap between theoretical innovation.

Objectives:

The studies show that the YOLO-based systems are always better than the CNNs, with the precision with YOLOv8 near to 95% on-line. Depth Prediction: MiDaS and DPT projects show us the possibility of more intelligent severity estimation than classical classification. Reporting: very few systems have integrated automated reporting and reporting systems: Very few systems have reporting (reporting automation in reporting pipeline or municipal dashboards because they have reporting systems focused just for detection. Predictive Analytics: Based on previous works it is clear that predicting pothole formation is a large research gap. Only few studies with ARIMA or ML based models have been published in this field at all.

3. Methodology:

The adopted approach represents a vision-based system for such system, which real-time pothole detection, reporting, and prediction for smart city infrastructure are made possible using advanced deep learning algorithms and predictive maintenance, in intelligent city infrastructure. This method is established on a multi-step pipeline with embedded data, computer vision, geo-tagging, cloud reporting, and predictive analytics.

3.1. Data Acquisition

Road condition observations can be gathered on the vehicles through mobile cameras, dashcams, and employ Raspberry Pi modules attached to the vehicles themselves. Each snapshot has GPS information that will be utilized to create geo-referenced inputs into input forms from these snapshots for the incorporation of recorded material. Additionally, the dataset is enriched further with the introduction of a citizen-lead app-enabled mobile app for pothole images submission by the participants for faster road maintenance reporting.

3.2. Pothole Detection and Segmentation and Assessment

Detected are also constructed with deep-learning model YOLOv8-Seg to perform real-time object detection and segmentation. And all this is done to ensure that the potholes

get positioned exactly in the video frames, regardless of whatever traffic is passing and the weather. In low-light or low-quality images, image-enhancing techniques can be applied like ESRGAN to enhance the quality of detection

3.3. Determination of depth and severity

Pothole severity prediction with monocular depth estimation tools MiDaS and Dense Prediction Transformer (DPT) are examples. A severity index can be calculated using depth, width, and traffic exposure from one image for depth, size, and volume estimations. Severity scoring is performed in this scenario by way of a supervised learning model, XGBoost, and enables the prioritization of instant corrections by government authorities.

3.4. Geo-Tagging and Cloud-based Reporting

The Google Maps API geo-tags automatically each pothole detected. The information such as images, severity scores, coordinates of the location is stored in a cloud database (say Firebase/AWS). The information is stored and tabulated using a top-level dashboard (a single, aggregated view that categorizes pothole severity according to their urgency and offers municipal officials a clear-cut view of response alternatives.

3.5. Predictive Maintenance

From historical road data, weather, and traffic flow patterns to train predictive models like ARIMA and XGBoost. These models forecast the probable occurrence of potholes by predicting when scheduling maintenance can be done (preventive maintenance). These predictive layers reduce future repair expenses and the road safety threat by avoiding the dangerous condition created by the presence of potholes.

3.6. Dashboard in Software and The Public Interface

The centralized dashboard offers real-time monitoring, which categorizes the kind of potholes and the cost of pothole repairs estimates available to the authorities. Furthermore, the citizen app enables citizens to report potholes, be notified when potholes develop on hazardous parts of the road and monitor repair potholes.

Conclusion

Based on previous integrated pothole management (using those techniques) and surveying methods combined, high-detection accuracy (YOLOv8), robustness to harsh environment (GAN-augmented, YOLOv7), depth estimation (MiDaS/ DPT), predictive modeling (ARIMA, XGBoost), we are able to build an integrated smart city solution. The resultant theoretical model demonstrates the bridge between theory and practice; providing the bridge from academia to engineering in its practical applications.

4. Proposed System

System architecture of the proposed Vision-Based Pothole Detection and Reporting System for Smart Cities has been conceptualized as a multi-layered system with the use of

computer vision, machine learning, cloud computing, and smart city infrastructure. The system begins with data acquisition and proceeds through detection, severity estimation, geo-tagging, cloud reporting, predictive analytics, and visualization on a centralized dashboard. As shown in Fig. 1 (System Architecture Diagram) and Fig. 2 (Workflow Diagram), the system operates the following way: Data Collection Layer: There is continuous video taken on the road by vehicle-mounted cameras, mobile phones, or Raspberry Pi devices. GPS information accompanies each frame to offer geo-referenced mapping of detected potholes. Detection and Segmentation Layer: The frames are detected and segmented by YOLOv8-Seg, doing real-time detection and segmentation of potholes. For clarity improvement under poor conditions, ESRGAN is utilized. Depth and Severity Analysis Layer: Depth estimation is done using MiDaS and Dense Prediction Transformer (DPT). They estimate the size of the pothole, and a severity score is produced by XGBoost from depth, width, and contextual factors such as traffic density. Geo-Tagging and Cloud Integration Layer: Detected potholes are geo-tagged using the Google Maps API and uploaded to cloud storage sites (Firebase/AWS).This provides them with accessibility, scalability,and integrability with municipal monitoring systems. Predictive Maintenance: Layer Historical weather trends, traffic flow, and road conditions are modeled using ARIMA and machine learning algorithms to predict potential pothole hotspots.

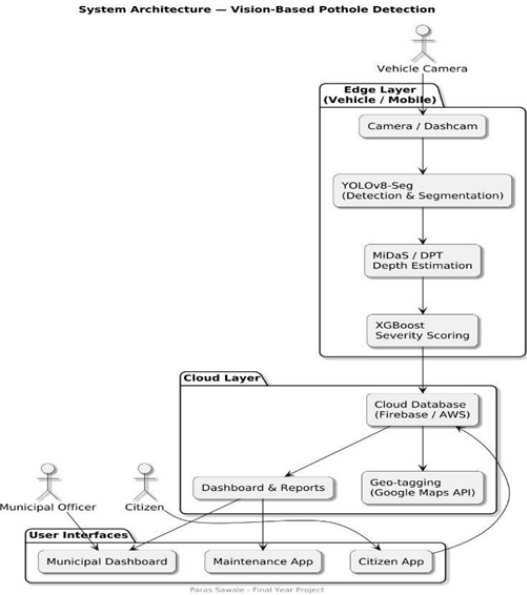


Fig 1. System Architecture Diagram

Methodology — Vision-Based Pothole Detection System

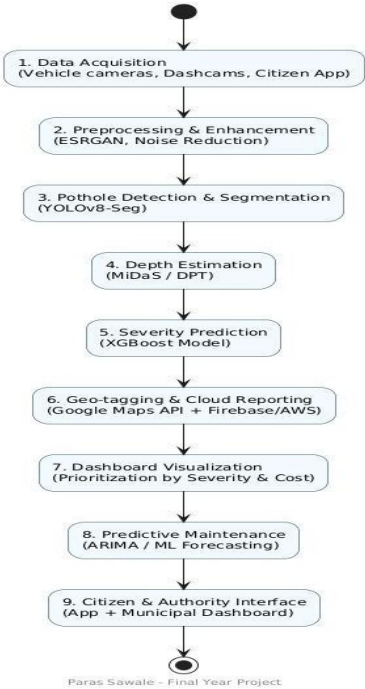


Fig 2. Workflow Diagram

Comparison of average of precision:

The bar chart plots the Mean Average Precision (mAP@0.50) of some of the YOLO-based models. YOLOX-Nano is the most accurate with a perfect 100%, beating everyone else by quite a margin. To everyone's surprise, it outperforms YOLOv4-Tiny by 21.3% and that speaks volumes about its object-detecting power. Of the rest, YOLOv5s, YOLOv4-Tiny, and YOLOv4 are the runner-ups, while YOLOv3 and YOLOv3-Tiny lag behind. This comparison states why YOLOX-Nano is more accurate, and it is suitable for accuracy-critical applications.

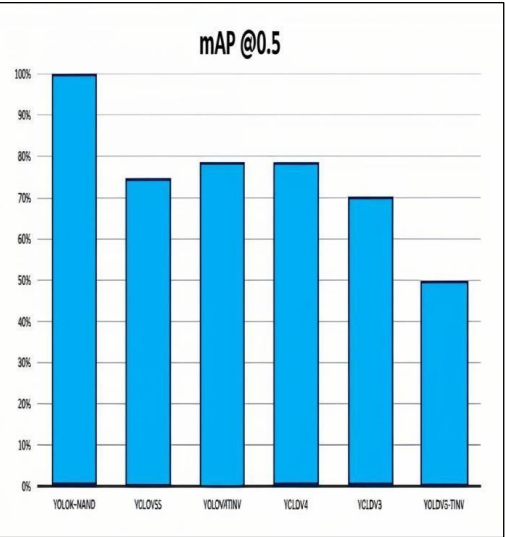


Fig 3. Precision Comparison

The results determine performance improvements in model architecture and indicate that even smaller models will be better than earlier, bigger models in accuracy. This shift highlights the growing efficiency of newer training techniques and optimization strategies. As a result, developers can achieve high accuracy with reduced computational costs, making advanced AI more accessible.

The model size comparison is obvious from the bar graph. YOLOX-Nano being the smallest having the least computing requirements stands at only 7.22 MB. YOLOv5s, twice as large standing at 14.8 MB, and YOLOv5m, standing at 43.3 MB, are relatively heavier in comparison. Even more substantial models like YOLORP6 (291.8 MB) and YOLOR-W6 (624.84 MB) are quite resource-intensive in storage and computation and thus only for cloud installation. The above comparison indicates the advantage of YOLOX-Nano in scenarios where there is a requirement for swift and efficient detection with less hardware constraint.

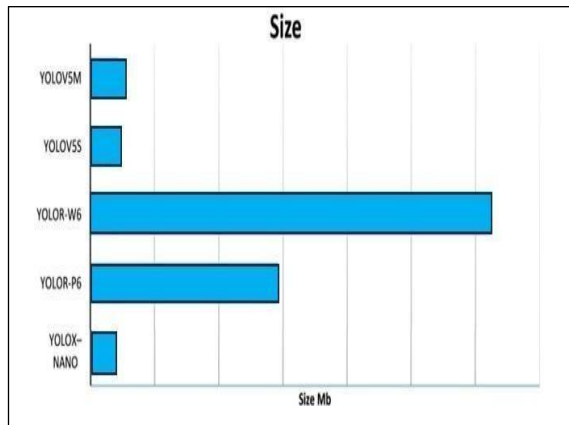


Fig 4. Data Size Comparison FUTURE SCOPE:

A future research and development directions:

- Enhancing the sensitivity of our model to the various environmental parameters (lighting parameters, weather, road conditions and other related features of our model.
- Real time detection on a Raspberry Pi or mobile camera is optimized for resource-limiting devices.
- It is necessary to extend predictive analytics and prediction of the potholes from just a conventional prediction of predictions, to the prediction based on traffic, weather, past repairing records, and repair history.
- Providing citizen engagement features that allow people to use mobile applications to report potholes, and verify them.
- Smart city systems integration, automating maintenance prioritization and resource management for a multitude of tasks.

- Our Vision-Based Smart Pothole Detection & Reporting System which focuses on smart urban infrastructure management, encompassing multiple key enablers such as automation, predictive maintenance, and citizen engagement at an urban, suburban or other level.

5. Conclusion:

Artificial intelligence based pothole detection and reporting systems using computer vision, machine learning, and predictive maintenance techniques were all studied in this survey. Legacy pothole inspection methods are time-consuming, laborious, and prone to human errors. The vision-based smart pothole detection & reporting system is introduced for these limitations, by modern approaches such as YOLOv8-Seg for real-time pothole detection, MiDaS/DPT for depth estimation, and XGBoost for predictive maintenance. Through GPS based geotagging, real time local authority reporting plus centralized dashboard for ordering pothole and monitoring them, this system increases efficiency, reduces repair cost and enhances public safety. The study highlights that although the precision and predictability of AI-driven systems increases, several difficulties have been encountered like data quality, model robustness, real time operation in various environmental conditions, and interfacing with current city infrastructure. These challenges underscore the need for continuous refinement and adaptive deployment strategies.

References:

- [1] Yashar Safyari, Masoud Mahdianpari, and Hodjat Shiri, "A Review of Vision-Based Pothole Detection Methods Using Computer Vision and Machine Learning," *Sensors*, 2024, 24, 5652. doi: 10.3390/s24175652.
- [2] Jaroslav Frnda, Srijita Bandyopadhyay, Michal Pavlicko, Marek Durica, Mihails Savrasovs, and Soumen Banerjee, "Analysis of Pothole Detection Accuracy of Selected Object Detection Models Under Adverse Conditions," *Journal of Physics: Conference Series*, Vol. 2, pp. 209–217.
- [3] Dr. S.N. Gujar, Prathmesh Shinde, Shlok Parihar, Aditya Pawar, and Saikrishna Yemul, "Pothole Detection System Using YOLOv8," *International Research Journal of Modernization in Engineering Technology and Science*, Vol. 6, Issue 4, April 2024. e-ISSN: 2582-5208.
- [4] Ashur Raju Addanki, Jianlin Lin, Yeshiva University, "Pothole Detection with YOLOv8," *ResearchGate*, December 2023. Available: <https://www.researchgate.net/publication/376685010>
- [5] Khaled R. Ahmed, "Smart Pothole Detection Using Deep Learning Based on Dilated Convolution," *Sensors*, 2021, 21, 8406. doi:10.3390/s2124840.

- [6] Kshitija Chavan, Chinmay Chawathe, and Vatsal, "Detection System Using YOLOv4 Algorithm," International Journal of Engineering Research & Technology, Vol. 9, Issue 4, April 2022. e-ISSN: 2395-0056.
- [7] Nirmal Kumar Rout, Gyanateet Dutta, Varun Sinha, Arghadeep Dey, Subhrangshu Mukherjee, and Gopal Gupta, "Improved Pothole Detection Using YOLOv7 and ESRGAN," February 2024. DOI: 10.48550/arXiv.2401.0858.
- [8] Mohan Prakash B. and Sriharipriya K.C., "Enhanced Pothole Detection System Using YOLOX Algorithm," Autonomous Intelligent Systems, Vol. 2(1), August 2022. DOI: 10.1007/s43684-022-00037-z.
- [9] Bhavana N., Mallikarjun M. Kodabagi, Muthu Kumar B., Ajay P., Muthukumaran N., and Ahilan A., "POT-YOLO: Real-Time Road Potholes Detection Using Edge Segmentation Based YOLOv8 Network," IEEE Sensors Journal, August 2024. DOI: 10.1109/JSEN.2024.3399008.
- [10] Dharneeshkar J., Soban Dhakshana V., Aniruthan S.A., Karthika R., and Latha Parameswaran, "Deep Learning Based Detection of Potholes in Indian Roads Using YOLO," 2020. DOI: 10.1109/ICICT48043.2020.9112424.
- [11] A. Singh and M. Randhawa, "Pothole Sensing and Depth Estimation System Using Deep Learning Technique," Conference Paper, June 2024. [Online]. Available: https://www.researchgate.net/publication/366904739_Pothole_Sensing_and_Depth_Estimation_System_using_Deep_Learning_Technique
- [12] .A. Jindal, S. Kumar, S. Verma, and S. Kumar, "A Hybrid Method for Pothole Depth Estimation: Combining MiDaS with Point Cloud and RANSAC," Conference Paper, May 2025. [Online]. Available: https://www.researchgate.net/publication/393336581_A_Hybrid_Method_for_Pothole_Depth_Estimation_Combining_MiDaS_with_Point_Cloud_and_RANSAC
- [13] "BER-XGBoost: Pothole Detection Based on Feature Extraction and Optimized XGBoost Using BER Metaheuristic Algorithm," ResearchGate, [Online]. Available: https://www.researchgate.net/publication/376802594_BERXGBoost_Pothole_Detection_based_on_Feature_Extraction_and_Optimized_XGBoost_using_BER_Metaheuristic_Algorithm
- [14] "Enhanced Pothole Detection in Road Condition Assessment Using YOLOv8," ResearchGate, [Online]. Available: https://www.researchgate.net/publication/380047987_Enhanced_Pothole_Detection_in_Road_Condition_Assessment_Using_YOLOv8
- [15] "Enhanced Real-Time Pothole Detection Using YOLOv8 with EfficientNet Backbone: A Deep Learning Approach for Road Safety," IEEE, March 2025. [Online]. Available: <https://doi.org/10.1109/ICVADV63329.2025.10961898>

AI Legal Responsibility and Accountability: Reframing Legal Responsibility in the Age of Autonomous Systems

Xena Dominique Ngoya DAVID
Final International University, Cyprus.
xena.tsimi@final.edu.tr

Abstract

Rapid technological change and development have given rise to a new era characterized by advanced Artificial Intelligence (AI) technology and applications. These innovations bring numerous positive effects, such as enhancing productivity, reducing operational errors, increasing efficiency across industries, improving healthcare delivery, supporting climate change mitigation, boosting agricultural output, and strengthening security; However, alongside these advancements come significant negative outcomes. This rapid progress raises critical questions about AI control, accountability, and liability. Like any human creation, AI is not without flaws, it is prone to bias, errors, security vulnerabilities, and increasing autonomy, all of which pose significant legal and ethical challenges related to responsibility and risk. The complexity of advanced AI system characterized by autonomous behavior and lack of predictability as well as continuous learning makes it difficult to determine how harm occurs and who is accountable. Traditional legal concepts like breach, defect and causation are often difficult to apply to AI. Furthermore, the involvement of multiple actors in the design, development, deployment, and operation of AI systems make it difficult to assess who is legally liable when something goes wrong. This article aims to examine the existing status on legal responsibility and the shifting patterns of legal liability for failures that result from the integration of artificial intelligence (AI). It explores aspects such as the current legal implications, accountability mechanisms of AI actions and gives an analysis of the existing opinions on the legal status of artificial intelligence with several options for resolving the issue of its legal responsibility.

The study employs a doctrinal legal analysis combined with a comparative review of legal frameworks from selected jurisdictions. It also draws upon case studies involving AI-induced harm to illustrate key legal gaps and ambiguities.

Keywords

Artificial Intelligence, Legal Responsibility, Accountability, Legal Personality, Autonomous Systems

1. Introduction

The global technology industry is developing very fast given rise to a new era characterized by advanced Artificial Intelligence (AI) technology. Nowadays, humans have been able to realize ideas that were previously considered imagination only found in novels and movies. (Sumantri, 2019) The increasing prominence of artificial intelligence (AI) systems in daily life and the evolving capacity of these

systems to process data and act without human input raise important legal and ethical concerns.

Today's legal rules do not always fit well with digitalization and AI (Larsson & Bengtsson, 2021). People who create and use new technologies often say that the law slows down innovation and makes progress harder. On the other hand, many lawyers see AI and digitalization as new forces that influence behavior and sometimes compete with traditional law. Tech law is the term for a growing field where law is applied to digitalization and AI. As these technological solutions dominate an increasing part of the economy, society and everyday life, they engender new legal issues and conflicts. This rapid progress raises critical questions about AI control, accountability, and liability or responsibility. (Gregor N. 2003)

The responsibility of artificial intelligence systems is a complex problem, the complexity of which is due to their autonomy and the ability to self-learn. This factor makes it difficult to distribute the burden of responsibility among the various individuals involved in the creation and operation of artificial intelligence. At the same time, the modern development of artificial intelligence systems does not allow one to recognize their qualities of personality and, therefore, to apply to them measures of responsibility as to individuals and legal entities. (Lipchanskaya, Eremina, & Privalov, 2021)

Artificial Intelligence (AI) can be defined as the theory and development of computer systems that can complete tasks which typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation. (The New Oxford American Dictionary, 3rd ed.). While Autonomous AI generally refers to an AI system that can act or make decisions without ongoing human intervention or approval; It can also refer to AI's ability to operate with minimal or no real-time human oversight, adapt behavior as it learns or encounters new data, and make its own decisions (Stanford, 2025). In the other hand Legal Responsibility or Liability is considered to be: The obligation to answer for an act done, and to repair any injury it may have caused (The Law Dictionary, n.d.) It can be civil, arising from non-criminal acts like negligence, breach of contract, or defamation and gives rise to Civil Procedure whose purpose is to the enforcement of certain rights claimed by the plaintiff against the defendant. Criminal, arising for criminal actions considered offenses against the state or society and leads to punishment in a criminal proceeding to a wrongdoer (Accountancy, 2025)

This paper will examine the existing status on legal responsibility and the shifting patterns of legal liability for failures that result from the integration of artificial intelligence (AI). It explores aspects such as the current legal implications, accountability mechanisms of AI actions and gives an analysis of the existing opinions on the legal status of artificial intelligence with several options for resolving the issue of its legal responsibility.

2. The Notion of Legal Responsibility or Liability

The notion of legal responsibility or liability comes from the law of obligation which in civil Law treats tort: wrongful acts and contract: governing duties arising from agreements together. And in common law it divides obligations into contract, tort, and restitution. Tort liability stems from breaches of duties imposed by law, while contract liability arises from agreements or, more recently, conferred benefits or reliance-based losses.

Legal liability can be divided in 2 types of liability which are; civil liability which often arises from negligence or breach of contract, where one party may be held responsible for damages caused to another, and criminal liability which involves determining an individual's intent and culpability for committing a crime and result in punishment, such as imprisonment or fines.

2.1. Legal Responsibility in French Civil Law

French law of obligation, influenced by Roman law's principle of *neminem laedere* ("harm no one"), is mainly found from Article 1240 to Article 1241 of the Civil Code. In French civil law, the freedom of contract means that contracts are mainly formed through the parties' agreement, with their will being central. The Civil Code divides obligations into contractual (from agreements) and non-contractual (from law, quasi-contracts, or torts). Both are part of the broader law of obligations, usually based on fault.

Article 1240 of the French Code Civil (France, 1804/2023) state : « Tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé à le réparer » (France, 1804/2023, arts. 1240). This is the basic fault-based liability; "responsabilité délictuelle / extra-contractuelle". It requires a fault, a damage "prejudice", and a causal link.

Article 1241 states that : « Chacun est responsable du dommage qu'il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence. » This article clarifies that responsibility arises not only from intentional fault but also negligence or imprudence. Broadening the scope of fault.

And Article 1242 which state: « On est responsable non seulement du dommage que l'on cause par son propre fait, mais encore de celui qui est causé par le fait des choses que l'on a sous sa garde... » Establishes liability even when one is not the direct author of the harm but holds or controls a

"thing" which causes damage; also liability for damage caused by persons for whom one is responsible (parents, employers, etc.). Strict liability / presumption in certain cases.

2.1.1 Application on AI

Civil law, legal responsibility for AI-caused harm is a developing area with existing principles like negligence, requiring proof of fault and strict Liability, holding the deployer responsible regardless of fault being applied. The challenges include identifying who is liable (developer, operator, or owner) and applying traditional legal frameworks to autonomous systems that can act independently (British Columbia Law Institute, 2024).

Applying civil law to artificial intelligence brings several complicated challenges. One of the biggest issues is figuring out who should be held responsible when an AI system causes harm (Aldmour, 2025). Because AI can act on its own, it's often hard to decide whether the blame should fall on the developer who created the system, the manufacturer who produced it, the operator who uses it, or the owner who benefits from it. According to article 1240 of the civile code a fault, a damage, and a causal link must be proved for a person to be liable of his actions, but in the case of AI lacking legal personality it is impossible to held it liable, so if it is to apply this principle, the manufacturer may likely be liable.

Another difficulty is proving causal link. AI systems are complex, and it's not always easy to trace how or why they made a certain decision. This makes it challenging to show a direct link between what the AI did and the damage that resulted. The causation link can trace the liability to the owner of this invention; Article 1242 establishes liability even when one is not the direct author of the harm but holds or controls a "thing" which causes damage. A company or an individual who own this invention may likely be liable in this case.

Bias in algorithms also creates serious problems. If an AI system is trained on biased data, it can continue or even worsen those biases, leading to unfair or discriminatory results. When that happens, it raises tough questions about who should be legally responsible for those outcomes (Ziemianin, 2021).

Lastly, there's the issue that AI itself isn't a legal person. Since it can't be sued or held accountable in court, the law has to find other ways to assign responsibility to people. This means that existing civil law principles might need to be adjusted, or new laws introduced, to make sure that someone is held accountable when AI systems cause harm.

2.2. Legal Responsibility in English Common Law and AI

English law requires consideration for contract enforceability, excluding gratuitous promises unless under seal and English tort law revolves around the duty of care, transitioning from trespass, conversion, and nuisance to negligence. Landmark cases expanded its scope: *Hedley Byrne v Heller* [1964].

2.2.1. Types of liability Under English Law

Under English law, liability for harm can fall under several different legal regimes, depending on the nature of the case and the relationship between the parties involved. One of the primary regimes is negligence, which is established when, a claimant must prove the defendant owes a duty of care which requires a sufficiently proximate relationship between the parties (Howarth, Chandler, & Behrendt, 2025); in case of AI this involves proving that a party failed to act with "reasonable care" in designing, training, or deploying the AI system. As stated in the case of *Donoghue v Stevenson* (1932) "You must take reasonable care to avoid acts or omissions which you can reasonably foresee would be likely to injure your neighbor." However, applying this principle to AI systems can be challenging, especially when the defendant, such as a manufacturer, developer, or supplier, no longer has control over how the AI system operates once it has been deployed.

In addition to this, proving causation might also be difficult if it is hard to identify how a failure occurred, where in the supply chain, and which party is responsible, particularly where the AI system has continued to develop after initial deployment by way of autonomous machine learning (Howarth, Chandler, & Behrendt, 2025).

Aside from negligence, Breach of contract also arise in this situation; contractual claims might arise under statutory warranties or implied terms in the case of AI if an AI system is not fit for purpose, of satisfactory quality, or it does not match the description. It is though debatable whether AI qualifies as a 'product' for these purposes. Furthermore, contractual clauses excluding or limiting liability may not cover AI and there remains significant risk for defendants seeking to rely on them in business to business (B2B) contracts as they will be subject to the test of reasonableness.

There is also Strict liability; in the context of Product Liability Directive 85/374/EC (PLD) which establishes a strict liability (no fault) regime enabling consumers to pursue a claim where a defect in a product has caused personal injury or property damage. The courts have, however, found that software which is not embedded in hardware does not constitute a 'product'. As such, there is uncertainty as to whether intangible code underpinning an AI process would be a 'product', leaving a gap in the law.

2.2.2 Criminal Responsibility

Criminal liability involves determining an individual's intent and culpability for committing a crime. When a person violates a criminal law (theft, assault, fraud), they are held criminally liable and the aim is punishment, such as imprisonment or fines. Criminal liability pertains to the legal responsibility an individual or entity bears for actions that violate criminal laws and regulations established by the government. Crimes are generally offences against society as

a whole and the government, represented by prosecutors, initiates criminal proceedings.

The purpose of criminal liability is to punish the wrongdoer for violating laws that are intended to protect public safety and order (Agrawal, 2023). Criminal liability involves around *Mens Rea* which is the intent to commit a crime, which is crucial for determining culpability and *Actus Reus* which is the physical act of committing the crime.

3. Regulation of AI Legal Responsibility

Artificial intelligence is rapidly transforming various aspects of our lives, including the economy, healthcare, and education. Artificial intelligence is defined inconsistently. Sometimes it is perceived widely as a field of science primarily related to computer science and robotics. In a narrower sense, artificial intelligence is the ability of an IT system to correctly interpret external data, to learn from it, and to use the experience gained in this way to accomplish specific tasks. This ability includes the capacity to flexibly adapt to external conditions (Wang, 2008, p. 362; Kaplan & Haenlein, 2019, pp. 15 - 25; Kok et al., 2002, p. 1095 ff). As AI technology continues to advance, concerns have been raised about its potential impact on society and the need for regulation to ensure its responsible development and use.

Furthermore, Machine Learning is a technique focusing on developing algorithms and systems that can learn from data and improve their performance over time, and is used in applications such as recommendation systems, fraud detection, and credit scoring. In addition, Robotics field is consisting on improving robots' tasks autonomously, such as assembling products, exploring space, and assisting in healthcare (Russell & Norvig, 2010).

AI has a wide range of applications in many different industries and domains, and its potential uses are only limited by our imagination and the availability of data and computing power. Therefore, the discussion of civil liability of AI is an important topic because it tackles many essential aspects of our life. Concerning legal liabilities, AI is increasingly integrated into various fields, so there is a need to determine who should be held legal responsible in case of damages caused by AI systems.

In the present legal state, with the current level of technological development, there are no grounds for granting legal personality to artificial intelligence. An artificial intelligence system cannot compensate someone to whom it causes harm. It does not have means to pay damages. It does not own assets that can be seized and liquidated to satisfy a judgment. Unlike humans and corporations, software is not a person in law over whom a court can have jurisdiction. As such, it cannot be ordered by a court to pay damages for harm it may cause. Compensation on the basis of tort law for harm caused by artificial intelligence depends on some human or

corporate entity being legally liable to compensate the person harmed (Klar & Jeffries, 2017; Osborne, 2020). Lack of legal personality therefore results in the inability to bear responsibility for one's own deeds. This means that if the artificial intelligence currently existing causes damage, another person should be responsible for it. The concept of bearing responsibility for other people's deeds, as already indicated above, is not foreign to civil law and Common. It was reflected in both the French Civil Code and English rules on tort liability. However, in this regard it is necessary to analyze which rules of tort liability can be applied in the event of damage caused by artificial intelligence, and who should be liable for such damage.

The legal framework for liability can vary by country and jurisdiction, some general principles that can be applicable are: Strict liability under which, the person or entity responsible for deploying the AI system is held strictly liable for any harm caused by the system, regardless of whether or not they were at fault. This approach is commonly used in product liability cases, and can provide a means of compensation for victims without having to prove fault or negligence on the part of the defendant. Negligence under which the person or entity responsible for deploying the AI system can be held liable for damages caused by the system if they were negligent in their development or deployment of the system.

3.1. In the French Legal System

As mention Above, in France, civil liability for AI is governed by the Civil Code that includes Tort Law particularly article 1240 to 1242 of the Civil Code. If a person or entity deploys an AI system that causes harm due to their fault or negligence, they may be held liable for damages. AI systems in certain sectors, such as healthcare (the French Public Health Code) and transportation (the French Road Traffic Act). It is important to note that the legal framework for the liability of AI systems is still evolving in France and other countries. As AI technology continues to develop and become more prevalent, it is likely that the legal framework will continue to evolve as well. France has already made some legal amendments targeting misuse of AI outputs for example deepfakes. Article L. 226-8 of the French Penal Code was amended in May 2024 to criminalise distributing AI-generated content using someone's image or voice without consent, unless the AI use is clearly indicated. Penalties vary depending on seriousness and whether sexual content is involved. Additionally there are proposals (e.g. Bill 1630) to amend the French Intellectual Property Code to deal with AI-created works, such as clarifying rights in works generated by AI, or rights of authors whose works were used to train AI. But these are still under debate and not yet final; and finally, the French Bill 1630 is set to establish a copyright framework for artificial intelligence (Intellectual Property Code).

3.2. In England and USA

In the United States, there is currently no comprehensive federal law governing the civil liability of AI. However, several States have enacted laws that address specific aspects of AI liability. For example, California has enacted different laws, such as the California Bot Act (Cal. Bus. & Prof. Code §§ 17940–17943, 2019), users interacting with AI must be informed of automated activity; this prohibits any person from using a bot online to communicate or interact with a person in California with the intent to mislead the person about the bot's artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a commercial transaction or influence a vote in an election. Similarly, some states have enacted laws that address the liability of employers for harm caused by their employee's use of ai such as Illinois HB 3773.

In England, on the other hand, the law governing the civil liability of AI is also evolving. In 2020, the UK government commissioned a review of the legal and regulatory framework for AI, which recommended the creation of a new legal framework for AI liability. The UK AI Regulation (White Paper 2023) focus on Safety, security, Transparency, fairness, Accountability and Contestability. Different legal cases were also settled under this topic. The case of *Moffatt v Air Canada* 2024 which established the responsibility for misleading information provided by the chatbot on the commercial website to the company Air Canada; Also in the case of *Quoine Pte Ltd v B2C2 Ltd* 2020 SGCA where The Singapore Court of Appeal (SGCA) has issued a landmark ruling on a breach of contract case involving the autonomous algorithmic trading of digital tokens. damage caused by the use of Automated trading software for trade of crypto assets. The Court held that the "Disputed Trades" executed by B2C2's algorithm were legally binding contracts, and Quoine was in breach of contract when it unilaterally cancelled or reversed them. The defence of unilateral mistake failed because Quoine did not prove that it had the required knowledge that B2C2 was operating under a mistake. so algorithmic trades are enforceable contracts, and unilateral mistakes cannot void them unless the non mistaken party knowingly exploits the mistake.

3.3 Under EU Law

In April 2021, the European Commission proposed the first EU artificial intelligence law, establishing a risk-based AI classification system. AI systems that can be used in different applications are analyzed and classified according to the risk they pose to users. The different risk levels mean more or less AI compliance requirements.

Parliament priority (European Parliament, 2020) was to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly. AI systems should be overseen by people, rather than by automation, to prevent harmful outcomes. Parliament also

wanted to establish a technology-neutral, uniform definition for AI that could be applied to future AI systems.

The new rules establish obligations for providers and users depending on the level of risk of AI risk qualification. There are Unacceptable risk which are Banned AI applications in the EU including: Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children. High risk AI involving AI systems that negatively affect safety or fundamental rights to be considered high risk (European Parliament, 2020).

4. The Future of AI Legal Responsibility

AI systems possess the ability to make autonomous decisions, provide recommendations, and perform tasks, introducing risks and raising accountability questions when negative outcomes occur. However, determining who should be held liable is challenging due to the unique characteristics of AI systems, including their complexity, lack of transparency, and involvement of multiple stakeholders. (European Parliament, 2023)

The European Union is seen as one of the “furthest along in developing a comprehensive legislative response to governing AI.” In an effort to regulate AI, the European Commission proposed a regulatory framework for AI, the Artificial Intelligence Act (AI Act). The AI Act classifies AI systems based on the risk presented to users and uses that risk level to determine the amount of regulation (European Parliament, 2023). According to the framework, “AI systems that negatively affect safety or fundamental rights will be considered high risk.” AI systems that assist lawyers “in legal interpretation and application of the law” fall into the high-risk category. If an AI system is marked as high risk, it must be registered in an EU database and will be assessed before entering the market. If it makes it to the market, it will be reassessed “throughout [its] lifecycle.” By assessing and reassessing AI systems that are marked as “high risk,” the AI Act assures that the systems are consistently meeting quality management and process requirements.

Like the EU, the United Kingdom is taking a risk-based approach to the regulation of AI systems (Department for Science, Innovation & Technology, n.d.) The United Kingdom regulation focuses on context and proportionality. A focus on context and proportionality allows regulators to analyze the risk in the context and environment the AI system is used.

China recently enacted the Interim Measures for the Management of Generative Artificial Intelligence Services (Department for Science, Innovation & Technology, n.d.). The Measures hold providers of generative AI services legally responsible for the sources that generative AI systems use and pre-training data (Huang, Mozur, & Creemers, 2023).

Providers are also liable as the producers of the content the system generates.

United States does not have a national law to regulate AI (Barrett, 2023). In lieu of national regulation, state courts have started to address the issue of AI in the courtroom (Kirk, McDonough, & Heintz, 2023). Courts are issuing opinions about the use of AI in the form of Rule 11(b)(2) sanctions, court orders, and local rules (Mangan, 2023). In Texas, Judge Brantley Starr adopted a policy on the use of generative AI (U.S. District Court for the Northern District of Texas, 2023). Judge Starr adopted this policy.

4.1. The shared Liability

There are several challenges in defining and attributing responsibility for harm caused by AI systems. Some of the main challenges include: complex decision-making, attribution of responsibility, lack of transparency and cultural and ethical norms. As for Complex decision making, the AI systems often make decisions that are complex and opaque, making it difficult to determine how a particular decision was made and who is responsible for that decision (Birkett, 2019). Responsibility for harm caused by AI can be distributed among several parties including companies for owning the system (Open AI); developers for flawed algorithms or insufficient testing, manufacturers for hardware defects in AI-enabled devices, deployers or operators for misuse or lack of oversight, data providers for supplying biased or inaccurate data, and third-party integrators who combine AI components into broader systems (Jobin, Ienca, & Vayena, 2019).

4.2 The question of Autonomy

The concept of granting legal personality to artificial intelligence is widely discussed in legal and philosophical literature. At the European Union forum initiatives are being taken to consider the possibility of applying the current legal regulations of the member states in relation to artificial intelligence and to formulate conclusions as to the need for legislative changes (inter alia European Commission, 2019).

These initiatives expressed the view that granting legal personality to artificial intelligence is unnecessary, since the responsibility for its actions should be borne by existing persons (European Commission, 2019, p. 4). According to these assumptions, granting legal personality to artificial intelligence does not seem beneficial due to the lack of a concept regarding the principles of liability (Ministerstwo Cyfryzacji, 2018), the legal personality of artificial intelligence should be discussed. Liability for AI actions for now should be attributed to its creators, operators or possible end users.

4.3. The question of Explainability

Explainable AI (XAI) is often offered as a partial technical fix for ascribing liability (An Act respecting the protection of personal information in the private sector, CQLR c. P-39.1, 2021). The assumption is that by providing explanations of

how an AI arrived at a decision after it happened, one can identify faults or at least see whether a developer or end-user missed critical warning signs (The Economist, 2019). But explainability by itself is not good enough. A model might highlight which input features or “weights” influenced a decision, but that does not always reveal the full reasoning process or broader design flaws. Focusing on an individual explanation can also distract from systemic issues and may deflect attention from broader organizational responsibilities, such as managers pressing for rushed rollouts without adequate safety testing or product teams overlooking robust adversarial testing protocols. Realistically, the black-box nature of modern machine learning will persist, even with sophisticated explainability tools, so that blame can still be elusive (Stanford, 2025).

5. Conclusion

Under current law, Artificial Intelligence cannot bear legal responsibility, as it lacks intent and moral awareness. Liability remains with the humans or corporations that design and control these systems, under laws.

However, as AI becomes more autonomous and influential in society, future legal reforms may need to reconsider how accountability is defined, ensuring justice keeps pace with technological evolution.

One approach to tackling liability in AI will be to adapt existing legal principles and establish clear guidelines for AI developers, users, and manufacturers (Kemp, 2020). This may involve defining specific standards for AI safety and reliability, establishing regulatory bodies, and mandating transparency and accountability measures in AI development and deployment.

Another perspective will be to shift the responsibility of liability to the entities responsible for designing, training, and implementing AI systems (Birkett, 2019). This view considers AI as a product or service, thereby making the manufacturer or developer liable for any harm caused by their AI systems.

Additionally, comprehensive insurance mechanisms tailored to AI-related risks could provide financial protection and facilitate compensation for damages resulting from AI systems. It is crucial to foster interdisciplinary collaboration among legal experts, policymakers, technologists, and ethicists to comprehensively address the challenges associated with civil liability in AI. Ongoing dialogue and cooperation (UK Government, 2020) can lead to the development of appropriate legal frameworks, ethical guidelines, and responsible practices that promote the benefits of AI while mitigating potential harm.

Overall, civil liability in AI is a multifaceted issue that necessitates careful consideration. By proactively addressing these concerns, we can promote the responsible development

and deployment of AI technologies, ensuring their positive contribution to our lives and communities while minimizing potential harm.

Reference

- [1] Aldmour, R. (2025). *Artificial Intelligence and Civil Liability: Challenges in Tort Law*. Oxford University Press.
- [2] Gregor, N. (2003). *Law, Technology, and Responsibility: The Human Factor in Automated Systems*. Cambridge University Press.
- [3] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25.
- [4] Klar, L., & Jeffries, C. (2017). *Tort Law: Cases, Principles, and Institutions* (3rd ed.). Toronto: Carswell.
- [5] Kok, J. N., Boers, E. J., Kusters, W. A., van der Putten, P., & Poel, M. (2002). Artificial intelligence: Definition, trends, techniques, and cases. *AI Communications*, 15(3), 1095–1105.
- [6] Larsson, S., & Bengtsson, L. (2021). *Law, Regulation and Digitalization: New Challenges for Governance*. Edward Elgar Publishing.
- [7] Lipchanskaya, N. A., Eremina, Y. V., & Privalov, A. V. (2021). Legal responsibility for harm caused by artificial intelligence systems. *Law and Digital Economy*, 3(1), 25–33.*
- [8] Osborne, C. (2020). *Liability in the Age of AI: New Frontiers in Responsibility and Regulation*. Routledge.
- [9] Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson Education.
- [10] Sumantri, M. (2019). *Artificial Intelligence and Human Imagination: Legal and Ethical Implications*. Jakarta: Legal Studies Press.
- [11] Wang, P. (2008). What Do You Mean by “AI”? *Frontiers in Artificial Intelligence and Applications*, 171, 362–373.
- [12] Ziemianin, K. (2021). Algorithmic Bias and Legal Responsibility in Artificial Intelligence. *European Journal of Law and Technology*, 12(2), 1–18.
- [13] Birkett, H. (2019). A legal framework for liability and accountability for artificial intelligence. *Legal Studies*, 39(4), 557–579. <https://doi.org/10.1017/1st.2019.24>
- [14] Kemp, M. (2020). How can English law adapt to cope with the challenges of AI? *Artificial Intelligence and Law*, 28(3), 309–324. <https://doi.org/10.1007/s10506-020-09262-6>
- [15] The Law Dictionary. (n.d.). Liability. Retrieved from <https://thelawdictionary.org/liability/>

- [16] The New Oxford American Dictionary (3rd ed.). (2010). Oxford University Press.
- [17] The Economist. (2019, December 18). Don't trust AI until we build systems that earn trust. Open Future. <https://www.economist.com/open-future/2019/12/18/dont-trust-ai-until-we-build-systems-that-earn-trust>
- [18] Accountancy. (2025). Legal liability definition. Retrieved from <https://www.accountancy.com>
- [19] Code civil [C. civ.] (France). (1804/2023). Articles 1240–1242.
- [20] Code pénal [C. pén.] (France). (2024, May). Article L. 226-8.
- [21] California Bot Act, Cal. Bus. & Prof. Code §§ 17940–17943 (2019).
- [22] Product Liability Directive 85/374/EEC, Council of the European Communities (1985).
- [23] Department for Science, Innovation & Technology. (n.d.). A Pro-Innovation Approach to AI Regulation: Government Response. UK Government. <https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response>
- [24] European Commission. (2019). Ethics Guidelines for Trustworthy AI. Brussels: EC.
- [25] European Parliament. (2020). Report on Civil Liability for Artificial Intelligence (2020/2014(INL)).
- [26] UK Government. (2020). The UK's AI roadmap. <https://www.gov.uk/government/publications/ai-roadmap>
- [27] European Parliament. (2023). EU Artificial Intelligence Act: First Regulation on Artificial Intelligence. <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- [28] Quebec legislation (personal information and automated decision-making)
- [29] An Act respecting the protection of personal information in the private sector, CQLR c. P-39.1, s. 12.1, as amended by S.Q. 2021, c. 25, s. 110. <https://www.legisquebec.gouv.qc.ca/en/document/c/s/P-39.1>
- [30] Ministerstwo Cyfryzacji [Polish Ministry of Digital Affairs]. (2018). Strategy for Artificial Intelligence in Poland. Warsaw.
- [31] U.S. District Court for the Northern District of Texas. (2023). Mandatory Certification Regarding Generative Artificial Intelligence.
- [32] Donoghue v Stevenson [1932] AC 562 (HL).
- [33] Hedley Byrne & Co Ltd v Heller & Partners Ltd [1964] AC 465 (HL).
- [34] Quoine Pte Ltd v B2C2 Ltd [2020] SGCA 103 (Singapore Court of Appeal).
- [35] Moffatt v Air Canada (2024) Ontario Superior Court of Justice.
- [36] Agrawal, S. (2023). Understanding Criminal Liability: The Role of Mens Rea and Actus Reus. Harvard Law Review Blog.
- [37] Barrett, C. (2023, August 10). Comparing the US, UK, and EU Regulatory Approaches to AI. American Bar Association.
- [38] British Columbia Law Institute. (2024). Report on Civil Liability for AI Systems. Vancouver: BCLI.
- [39] Huang, S., Mozur, P., & Creemers, R. (2023, April 11). Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment) – April 2023. Stanford University – DigiChina. <https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/>
- [40] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.
- [41] Kirk, S. C., McDonough, E., & Heintz, M. (2023, August 2). Judges Guide Attorneys on AI Pitfalls with Standing Orders. Ropes & Gray. <https://www.ropesgray.com/en/newsroom/alerts/2023/08/judges-guide-attorneys-on-ai-pitfalls-with-standing-orders>
- [42] Mangan, D. (2023, June 22). Judge sanctions lawyers whose AI-written filing contained fake citations. CNBC. <https://www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html>
- [43] Stanford, E. (2025, May 15). Autonomous AI: Who Is Responsible When AI Acts Autonomously and Things Go Wrong? In AI, Machine Learning & Big Data Laws and Regulations 2025. CMS Law-Now.
- [44] Baker McKenzie. (2023, August). China: New Interim Measures to Regulate Generative AI. https://insightplus.bakermckenzie.com/bm/attachment_dw.action?attkey=FRbANEucS95NMLRN47z%2BeeOgEFCt8EGQJsWJiCH2WAWuU9AaVDeFglGa5oQkOMGl&nav=FRbANEucS95NMLRN47z%2BeeOgEFCt8EGQbuwypnpZjc4%3D&attdocparam=pB7HEsg%2FZ312Bk8OIuOIH1c%2BY4beLEAezirm3%2BK7wMU%3D&fromContentView=1
- [45] Howarth, D., Chandler, B., & Behrendt, J. (2025). AI and Negligence: Duty of Care in the Age of Autonomous Systems. Cambridge Law Review, 41(2), 233–259.

Segmentation of Pavement Cracks in North Cyprus

Albert David Bangura
abangura@baucyprus.edu.tr

Tolga Yirtici
Cyprus International University, Cyprus.

Abstract

This thesis explores the use of deep learning, specifically Convolutional Neural Networks (CNNs), for automated pavement crack segmentation in North Cyprus, addressing the need for efficient road maintenance. The study emphasizes the limitations of manual inspection and introduces a CNN-based U-Net architecture, developed to automate feature extraction, and enable more accurate and efficient crack segmentation. Implementing a U-Net model with custom layers to learn features without transfer learning forms the core of this methodology, utilizing ReLU and sigmoid activation functions, binary cross-entropy as the loss function, and the Adam optimizer. The model is evaluated using metrics including accuracy, precision, recall, F1 score, and IoU, achieving over 98% accuracy and demonstrating optimal performance with larger image sizes (256x256). These results highlight the potential of U-Net based crack segmentation systems to significantly improve road maintenance and safety.

Keywords

Convolution Neural Networks (CNNs), Crack Segmentation, Deep learning, U-Net Architecture, Road maintenance.

1. Introduction

This research aims to use artificial intelligence to improve road maintenance in Northern Cyprus. It underscores the poor quality of regional roads, which are riddled with cracks, and the manual detection methods that are simply inadequate. Instead, it proposes the use of Convolutional Neural Networks (CNNs) to detect and segment those cracks in an automated manner—an approach many researchers have argued is much more efficient and accurate than visual inspections. The road maintenance "problem" (the amount and the quality of surveillance and action necessary to keep a road drivable) is a common one, but this research covers a region with a specific set of geological and climatic conditions that are unusual in comparison to most of the other regions in the world where road maintenance is studied.

The road network in Northern Cyprus, especially in major urban areas like Nicosia, Kyrenia, and Famagusta, has a real tough time due to the huge number of vehicles and constrained infrastructure. Mainly built as a two-lane

road system, the path network is just not able to handle the kinds of traffic volumes that exists today. This is compounded by the fact that the network has a much lower quality compared to international standards, which means it does not really have any of the features that make a road a safe and reliable means of transportation. Among the most notable "missing features" of the road network is the absence of anything that resembles a proper drainage system. This leads to quite a few water accumulation problems that can lead to cracks. Another major issue is the lack of safety pathways for pedestrians and people riding bikes.

A significant issue affecting the road network in Northern Cyprus is the prevalence of extensive cracks, which vary in size and visibility. While some are easily noticeable, others are more subtle and more difficult to detect, posing a hidden risks to road users particularly for international students and tourists who are unfamiliar with the region. Without timely intervention, these cracks worsen over time, leading to severe damage such as potholes, which degrade the overall quality of the road. This rapid deterioration emphasizes the urgent need for effective road management strategies to enhance infrastructure and improve road safety.

Maintaining roads is essential for ensuring safety, economic performance and environmental sustainability. Well-maintained roads help not only prevent accidents but also support environmentally friendly transportations reducing vehicle operating costs and fuel consumption, factors essential for economic growth. Addressing cracks and resurfacing roads expands the lifespan of road infrastructure, saving on long-term repair costs while minimizing environmental impact and lowering vehicle emissions. However, road cracks present significant challenges. They compromise safety by disrupting traffic flow and potentially damaging vehicles. Cracks also allow water to seep into the road base, weakening the structural integrity. Neglecting small cracks can lead to larger, more expensive road repairs, ultimately increasing maintenance costs and posing greater risks.

The role of timely crack detection in prolonging road life and improving safety.

Timely crack detection is vital for prolonging the lifespan of roads and ensuring the safety of all road users. Roads

are subjected to various environmental and load-bearing stresses, leading to crack formation caused by factors such as temperature fluctuations, water infiltration, and continuous traffic. If left unaddressed, small cracks can expand and deteriorate further over time, resulting in severe damage like potholes and patches. Early detection allows for preventive renovation measures, such as sealing or filling cracks, which can effectively halt further deterioration.

From a safety perspective, timely crack detection prevents hazardous conditions that could lead to accidents. Visible cracks, specifically those allowed to evolve into larger defects, pose significant risks to drivers, cyclists, and pedestrians. For vehicles, cracks can reduce tire grip, cause vibrations, and lead to loss of control. Regular monitoring and early intervention mitigate these dangers, ensuring roads remain smooth, well-maintained, and safe for all users.

1.1. Problem Statement

1.1.1. Current Challenges in Manual Road Crack Detection

Manual road crack detection faces significant challenges; traditional methods rely on visual inspections conducted by personnel, which are prone to delays and inconsistencies due to human error and varying levels of expertise (Yaun, Shi, & Li, 2024). Moreover, manual detection often places inspectors in dangerous situations, particularly when working near traffic. These limitations underscore the importance of adopting automated and reliable detection methods to ensure the timely and accurate identification of road cracks, ultimately enhancing road safety and maintenance efficiency (Yaun, Shi, & Li, 2024).

Emerging methods, such as those leveraging deep learning, have demonstrated potential for detecting cracks at both sub-pixel and pixel levels. Early studies, including those by Wang et al. and Yaun, Shi, and Li, have shown promise, but current solutions remain limited. Issues such as low contrast, insufficient datasets, and inaccurate localization reduce the effectiveness of these models (2024). Despite these challenges, integrating conventional techniques with AI-based methods presents a promising path forward to improve the efficiency and accuracy of crack detection.

1.1.2. Research Motivation and Objectives

The motivation of using deep-learning in road maintenance lies in its potential to provide efficient, accurate, and lead to proactive solutions. By analyzing extensive road network data, deep learning can detect cracks and other issues with far greater precision and effectiveness than traditional human inspections. These technologies enable both real-time monitoring and predictive maintenance; allowing potential problems to

be identified and addressed before they escalate into significant damage, such as potholes. This not only extends the lifespan of urban infrastructure but also enhances safety by equipping workers with critical insights to avoid hazardous situations. In essence, deep learning revolutionizes road maintenance management by optimizing the handling of critical infrastructure and conserving valuable resources. The objective of this thesis is to develop a CNN-based system capable of accurately segmenting pavement cracks.

1.2. Scope Of Study

1.2.1. Geographical Focus on North Cyprus

Northern Cyprus presents an ideal setting for studying road crack detection due to its unique geographic and climatic conditions. The region faces several environmental challenges, including significant temperature fluctuations, torrential rains, and coastal humidity, all contributing to road degradation. Additionally, the rugged terrain and mountain ranges, such as the Kyrenia Mountains, further complicate road maintenance efforts. Implementing automatic crack detection systems in the Northern Cyprus context offers the potential to detect cracks promptly and accurately, enhancing the safety and efficiency of the road network. Moreover, such preventative road maintenance strategies can significantly reduce repair costs and extend the lifespan of the infrastructure, addressing critical challenges faced by the region.

1.2.2. Use of CNN as the Primary Tool for Segmentation.

Convolutional Neural Networks (CNN) have emerged as a central algorithm for road crack detection, gaining significant attention due to their remarkable accuracy and effectiveness in image recognition (Elghaish, et al., 2021). Studies have shown that CNN models like U-Net can detect cracks with higher accuracy under diverse lighting and weather conditions. These models are capable of processing large volumes of data in short time, making them highly efficient. By automating the detection process, CNNs eliminate the need for manual inspections, which are inherently subjective and labor intensive (Benedetto et al., 2023). This automation not only enhances crack detection accuracy but also significantly improves the efficiency of road maintenance (Hacefendiolu & Başaa, 2022).

1.3. Significance of the Study

1.3.1. The Potential to Enhance Road Safety and Reduce Maintenance Costs.

Artificial Intelligence driven road management offers the potential of significantly enhance road safety while reducing costs through the application of deep learning and computer vision technologies. These advancements help prevent road defects and eliminate the need for time

constraining and error-prone manual inspections. Automated detection enables more accurate and reliable maintenance, ensuring roads remain in better condition. Costs savings are maximized by extending the lifespan of road infrastructure and reducing the frequency and scale of repairs. Therefore, AI driven road maintenance not only delivers improved road quality but also optimizes resource allocation, resulting in substantial economic benefits.

1.3.2. Contribution to Research in AI-based Road Maintenance Systems.

AI research in road maintenance has made significant progress, greatly improving the efficiency and quality of semantic crack segmentation in road images. Deep learning techniques can now detect and classify patchy roads and potholes in real-time, contributing to the extended lifespan of road infrastructure. Incorporating AI into road maintenance also promotes sustainability by enabling proactive and efficient management of resources. This study does not only advance road maintenance practices in North Cyprus but also enriches existing literature by performing pavement crack segmentation on a dataset specifically collected from roads in the Northern Cyprus region.

Literature Review

Road crack segmentation with deep learning is a very hot field and there is a lot of literature on this subject. The section compared the work of some of these researchers to identify the main problems and strengths of each. Road infrastructure can no longer be underrated as an asset for growth and social mobility. Pavement crack is unsafe, and late maintenance can be expensive. Pavement crack detection used to be an observation task done manually with the naked eye, but the downside is that it's slow, subjective and error-prone (Lau, et al., 2020).

Faced with these constraints, engineers have turned to computation to automate segmentation and detection of pavement cracks. DL algorithms – Convolutional Neural Networks (CNNs) in particular – are now the best prospects for crack segmentation– better, faster and more reliable (Lau, et al., 2020).

In response to these limitations, scientists have used computation for automated pavement crack segmentation. Deep learning (DL) algorithms – Convolutional Neural Networks (CNNs) in particular – have become the perfect candidates for this purpose, with great promise for making pavement crack detection more accurate, efficient and objective (Lau, et al., 2020).

2.1. Key Deep Learning Architectures

The most popular deep learning architectures explored for crack segmentation in roads are:

2.1.1. Fully Convolutional Networks

FCNs are another most used type for semantic segmentation, and they dispense fully connected layers with convolutional layers so that you can provide input images of unlimited size. We have tested the crack detection for FCNs of all backbones, VGG16, VGG19, ResNet50.

2.1.2. SegNet

SegNet is also a 100% convolutional network just like U-Net but with different decoder architecture. It has been used for crack segmentation in pavements such as concrete, asphalt, and bridge decks.

2.1.3. Other Architectures:

Other architectures have also been investigated like ResNet, DenseNet, PSPNet, DeepLabv3+ and GCN, all having their own advantages and disadvantages of accuracy, performance and complexity.

2.2. Datasets And Evaluation Metrics

A very important aspect of deep learning research is good datasets for training and testing. Some open-source crack detection datasets were created, each with its own features and drawbacks. Some commonly used datasets include:

2.2.1. CrackForest

Concrete crack data set with hand labeled ground truth.

2.2.2. AigleRN

A map of road crack images taken with street cameras.

2.2.3. Crack500

A map of road crack pictures from mobile phones.

2.2.4. TRIMMED

Collection of grayscale road crack images recorded with high-resolution line-scan camera.

2.2.5. CFTD

Collection of RGB road cracks image captured with consumer grade cameras.

To test the effectiveness of crack segmentation algorithms, scientists often measure it with a mix of different metrics such as precision, recall, F1-Score, Intersection over Union (IoU).

2.3. Challenges And Contributions from Previous Studies

Lau et al. (2020) introduced a U-Net CNN with a ResNet-34 encoder through transfer learning to overcome the inefficiencies of traditional crack detection methods like thresholding, morphology, and edge detection. Their model incorporated SCSE modules, progressive image resizing, and varying learning rates to optimize training. Tested on CFD and Crack500 datasets, the system

achieved F1 scores of 96% and 73%, outperforming larger models such as Split-Attention Network.

Zhang et al. (2019) designed a context-aware semantic segmentation network that fused predictions from overlapping image patches using cross-state and cross-space constraints. This approach allowed pixel-wise crack segmentation without retraining for different image sizes. Applied to CFD, TRIMMD, and CFTD datasets, it achieved state-of-the-art Boundary F1 scores while processing images in 0.7 seconds each, demonstrating efficiency, accuracy, and scalability for infrastructure crack assessment across diverse conditions.

Gao et al. (2019) proposed a generative adversarial network (GAN)-based approach for crack segmentation, introducing U-GAN, CU-GAN, and FU-GAN, all using modified U-Nets as generators. The models were trained to distinguish real from fake crack images at the pixel level. On AigleRN, CFD, and HTR datasets, they outperformed existing methods in precision, though recall rates declined on noisy or blurry images. The study suggested advancing multi-scale convolution, feature fusion, and attention mechanisms.

Pereira et al. (2019) addressed the inefficiency and subjectivity of manual crack inspections by developing a U-Net-based deep learning system for pavement and pothole segmentation. Data were collected via smartphones, and experiments showed the model achieved 97% accuracy with mean mIoU of 0.85. Despite dataset limitations, U-Net performed exceptionally well outside its original medical imaging domain, with future improvements anticipated through larger, more varied datasets and comparisons with newer architectures.

Wang et al. (2021) compared eleven CNN-based semantic segmentation models, including FCN, PSPNet, UPerNet, and DeepLabv3+, using ResNet, VGG, and DenseNet backbones. Models were evaluated with IoU, accuracy, precision, recall, and F1 score. DeepLabv3+ with ResNet101 backbone performed best, particularly with spatial pyramid pooling. Results confirmed CNNs' superiority over heuristic methods, though noise and dataset variations remained challenges. The authors recommended DeepLabv3+ and GCN as robust solutions for practical crack detection.

Li et al. (2022) highlighted deep learning advances, showing CNNs like U-Net, HED, and SRN significantly outperform image processing for crack segmentation. They suggested integrating semantic segmentation with edge detection to overcome each method's weaknesses. This combined framework improves accuracy in identifying fine crack properties such as width and length. Such detailed measurement offers practical benefits for pavement maintenance and management,

demonstrating deep learning's impact on safety and cost efficiency in infrastructure monitoring.

Zhang et al. (2023) emphasized the shift from handcrafted, image-processing methods toward CNN-based automated crack detection. Models such as U-Net, UperNet, ResUNet, and Pointrend produced strong results even under noisy and poorly lit conditions, proving more effective than earlier techniques. However, segmentation of small cracks and distinction from pavement features like stains or manhole covers remained difficult. Their study reinforced CNNs' strengths while acknowledging persistent challenges in fine-scale crack identification and differentiation.

Lee et al. (2019) proposed the Semantic Crack Segmentation Network (CSN) to address the limitations of edge-detection filters and early CNNs, which were restricted to patch-based methods and suffered from noise. CSN improved accuracy by processing entire images and avoiding sliding windows. To combat data scarcity, they generated synthetic crack images using Gaussian kernels and Brownian motion. Combining real and synthetic data enhanced training, leading to stronger segmentation performance in cluttered and complex scenes.

Jia (2023) criticized the slowness and inaccuracy of manual inspections and image processing for pavement crack identification. They proposed an enhanced U-Net with an Efficient Channel Attention (ECA) module in the encoder and FCNHead decoder. This configuration strengthened crack feature extraction and model generalization. Their method proved particularly effective in detecting small cracks under challenging pavement conditions, offering a practical solution for improving segmentation performance and reliability in infrastructure maintenance and monitoring tasks.

Benedetto et al. (2023) tackled crack segmentation challenges caused by noisy, obstructed, and overexposed images. They proposed a modified U-Net incorporating a ResNet50 encoder pre-trained on ImageNet, combined with residual structures to improve learning. Their model excelled in handling environmental obstructions and achieved more precise crack detection, especially for measuring crack width. Since crack width is essential for severity determination, the proposed system significantly enhanced the accuracy and usefulness of automated crack assessment tools.

Panella et al. (2022) reviewed deep learning models, focusing on CNNs like FCNs and U-Net for crack segmentation. They analyzed the trade-off between model complexity and loss of fine details due to pooling layers. U-Net's skip connections allow it to preserve spatial information across layers, enhancing accuracy

and efficiency. The review underscored U-Net's biologically inspired design and superior performance in separating "what" and "where" information compared with simpler CNNs, making it a reliable segmentation architecture.

Al-Huda et al. (2023) introduced KTCAM-Net, a hybrid deep learning model combining classification and segmentation networks with Class Activation Maps (CAM). KTCAM-Net employed hybrid loss functions, crack boundary filtering, and overlapping fusion algorithms, producing fine crack localization even under noise and imbalance. Benchmark datasets confirmed its superior performance, particularly for detecting thin cracks. The model's flexibility extended beyond cracks, showing potential for broader surface defect detection in inspection tasks, highlighting its innovative contribution to segmentation methods.

Nguyen et al. (2021) proposed a two-stage CNN to address noisy, low-resolution crack images. The first stage used a five-layer CNN to localize cracks and remove artifacts, while the second encoder-decoder network segmented pixels in the localized regions. Tested on the 2StagesCrack dataset, the method outperformed single-stage models by handling low-quality images effectively. The authors concluded that this two-stage system was both accurate and computationally efficient for real-world pavement crack segmentation.

Chen et al. (2020) presented Progressive Contextual Segmentation Network (PCSN), built on SegNet with a VGG16 encoder pre-trained on ImageNet. They applied augmentation methods like flipping, rotation, and contrast scaling to improve dataset diversity. Compared to Mask R-CNN and FCN-8s, PCSN achieved superior crack detection under difficult conditions, balancing inference speed and segmentation quality. The study showed how input image size impacted performance, with larger inputs providing richer crack details at the cost of slower processing.

Fang et al. (2020) categorized crack detection into image processing, classical machine learning, and deep learning. They noted that while thresholding and edge detection are noise-sensitive, machine learning methods like SVMs lacked robustness. CNNs achieved major advances, though challenges included high computational costs, noisy faint cracks, and difficulty maintaining both global and local context. Their review underscored the advantages of deep learning but highlighted the ongoing need for efficiency improvements and better contextual data integration.

Yu et al. (2021) compared one-stage and two-stage deep learning detection models for crack detection. Two-stage models like Faster R-CNN achieved high accuracy by

using region proposal networks, while one-stage models such as YOLO and SSD traded some accuracy for speed, making them more suitable for industry. Both approaches struggled when applied to UAV-acquired images, which are much larger than standard datasets like ImageNet or VOC, revealing limitations in scalability for field conditions.

Ali et al. (2021) examined CNNs for automated, real-time crack detection in buildings, highlighting their advantages over costly and dangerous manual inspections. CNNs could classify images, localize cracks with bounding boxes, and perform pixel-level segmentation. However, segmentation accuracy was hampered by class imbalance, inconsistent lighting, and obstacles in images, which biased networks against crack pixels. Their study emphasized these practical challenges while confirming CNNs' effectiveness for crack detection tasks in diverse real-world applications.

Yan et al. (2022) addressed the difficulty of detecting cracks in low-light conditions by proposing CycleADC-Net. The system first employed CycleGAN to translate dim images into brighter ones without altering crack structure, then applied a dual-channel encoder-decoder with attention to merge global and local signals. This approach improved detection accuracy in poorly lit scenarios, enabling robust crack segmentation where conventional deep learning models typically fail due to lighting limitations.

Xu and Liu (2022) solved the challenge of limited training data by employing DCGAN to generate synthetic pavement crack images. Expanding the dataset from 1,608 to 6,000 images improved CNN classification accuracy from 80.75% to 91.61% while reducing class imbalance. Their findings demonstrated the value of generative models for creating realistic training data, enhancing both model performance and dataset diversity in pavement crack detection applications, particularly when real-world samples are scarce.

Chen et al. created a dataset of 10,000 manually labeled pavement and bridge crack images to address the lack of public datasets. Using SegNet with a VGG16 encoder and data augmentation, they developed PCSN, which achieved mean average precision of 83%, outperforming Mask R-CNN (42%) and FCN-8s (67%). Larger input sizes improved segmentation performance, though smaller inputs offered faster inference. Their work demonstrated the importance of dataset quality and architecture design for reliable crack detection.

2.4. Addressing Data Limitations

2.4.1. Data Augmentation

Most research works use data augmentation to scale the size and number of training datasets, decreasing

overfitting, and achieving model generalization. Certain common augmentation tricks include rotation, flip, scaling, and cropping images.

2.4.2. Cross-Dataset Testing

To test models for their adaptation to different imaging scenarios and crack types, some researchers have run cross-dataset testing — training models on one dataset and running them against another.

2.5. Handling Complex Noise and Background

2.5.1. Modules for Attention

Attention modules such as Convolutional Block Attention Module (CBAM) or Efficient Channel Attention module (ECA)) are added to CNN architectures for feature extraction and attention to specific image areas. These modules are used to help the model deconstruct cracks from noise and other non-relevant features.

2.5.2. Residual Structures

Residual structures like ResNet have been demonstrated to improve performance of deep learning models by allowing training of deeper networks. Adding skip connections that skip one or more layers is how residual structures solve the vanishing gradient problem and make the information flow throughout the network.

2.6. Computational Efficiency

2.6.1. Improving Network Design

Network engineers are constantly looking for new ways to tweak network design, so it runs faster and with accuracy. Such as searching for lightweight architectures, delimiting parameters, and learning from the effective training strategies.

Although pavement crack segmentation is making great progress, there are still obstacles, and more research needs to be done:

2.6.2. More Capable and Generalizable Models

Typical models can't handle the complexities of real-world scenarios like lighting conditions, shadow interference, water stains, crack types, etc. The research needs to be more robust and generalizable models that can manage these problems well.

2.6.3. Crack Width Estimation

Most research is centered around crack segmentation, but the crack width estimate is important for pavement cleaning and repair. Our next studies should try to build crack width estimation into the segmentation equations.

2.6.4. Instant Crack Detection

The real-time crack detection is required in applications like autonomous road monitoring and maintenance planning. The studies should keep developing computationally efficient models and algorithms that can perform in real time on resource limited machines.

Deep learning has taken the pavement crack segmentation process to a whole new level, with promise for automating pavement inspection and maintenance. Various CNN architectures and methods have been able to show promising results but there's more research that needs to be done on robustness, generalization, crack width estimation, and real-time performance. Researchers noted that in future research, deep learning for pavement crack segmentation will be ever more integral to road safety and resilience.

Three Background

3.1. Computer Vision

Computer vision is the area of artificial intelligence (AI) that lets computers and systems process visual data in the world like digital pictures and videos. Through ingesting and visualizing these data, computer vision systems can do the things humans do most of the time, such as object detection, image classification and segmentation tasks.

Computer vision is as old as it gets: research began in the 1960s, to help machines see even simple pictures. The digital image, computational capability and algorithmic development that has followed over the decades have led to the advancement of the industry. Notably, deep learning and neural networks have enabled significantly higher image recognition and detection.

3.1.1. Core Techniques and Technologies

Today's computer vision is dependent on some important technologies:

- Machine Learning & Deep Learning: These are methods that can learn from millions of images and increase patterns and predictions.
- Convolutional Neural Networks (CNNs): CNNs are neural networks trained specifically on pixels and so, are ideal for image classification.
- Computer vision has a huge number of applications in many different sectors:
- Healthcare: Performs medical imaging for diseases and abnormalities (X-ray, MRI to diagnose).
- Auto industry: Creating safe perception and control of the environment with object recognition and lane-recognition for autonomous cars.
- Manufacturing: Supporting automatic inspection and quality control, where product defects are found on production lines.
- Safety and Surveillance: Creating surveillance with face detection and activity detection to ensure security and safety.
- Current research in computer vision is about to make it more accurate, more efficient, and more applicable in various fields. Combining computer vision with other AI capabilities including natural language processing and robotics, there are more opportunities for

creative and usable solutions across many industries.

3.2. Machine Learning and Deep Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that helps systems learn from data, find patterns, and make decisions without any human involvement. With algorithms and statistical representations, ML makes it possible for computers to perform tasks that cannot be programmed for each action in the program.

Deep learning on the other hand is a branch of machine learning (ML) in artificial intelligence (AI) based on algorithms modelled after the brain's structure and activity — artificial neural networks. These algorithms search for patterns and make choices based on layers of data from which more advanced features are extracted.

In terms of the theory, deep learning has been around since the 1940s, when neural network models were first built. Researchers such as Frank Rosenblatt first developed the perceptron, an early neural network architecture capable of rudimentary pattern recognition, in the 1950s and '60s. But computational restrictions and theoretical problems held up progress.

These regained momentum in the 1980s when the backpropagation algorithm was developed to train multi-layer neural networks better. But even with all these improvements, it wasn't until the 2000s, when the computing power became more powerful and the availability of more data, that deep learning really started making big gains.

3.3. Convolutional Neural Networks

CNNs are a type of deep learning model that works on data with a grid structure like images. They have been the key technology of computer vision and enables computers to perform tasks such as image and video recognition, object recognition and segmentation. In 2012, the field made its breakthrough when AlexNet by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton won the ImageNet Large Scale Visual Recognition Challenge. AlexNet's accomplishment demonstrated the power of CNN in tackling advanced image classification problems (Draeos, 2019).

3.3.1. Core Components and Architecture

An ordinary CNN architecture consists of the following key players:

- Convolutional Layers: These layers perform convolutional functions on the input, thus identifying spatial hierarchies and local trends in the data.
- Activation Functions: Non-linear functions such as ReLU (Rectified Linear Unit) are used to introduce non-linearity to the model and allow it to learn complex patterns.
- Pooling Layers: Pooling layers perform down-sampling to reduce the dimension of the data and make the representations more tractable and stable with slight translations.
- Fully Connected Layers: After a few convolutional/pooling layers, high level neural network reasoning occurs through fully connected layers that links each neuron of a layer to each neuron of a subsequent layer.
- CNNs are used in a variety of applications, some of them include:
 - Image & Video Recognition: Widely applied in face recognition software, object detection and video analysis.
 - Medical Imaging: CNNs are used to diagnose illnesses based on medical images (MRIs, CT scans, X-rays etc.).
 - Autonomous Vehicles: They make self-driving cars to see and interpret the world by analyzing images from cameras.
- Natural Language Processing: CNNs are used for text classification (sentiment detection, spam detection).

3.4. Semantic Segmentation of Pavement Cracks

This is the task of categorizing every pixel in an image to correctly determine and mark which regions of an image is a crack. This is necessary for the maintenance of infrastructure and transportation safety.

Pavement crack detection has always been done manually with some simple image processing like edge detection and thresholding. These methods were promising as early solutions but had low accuracy, particularly in multi-light, multi-noise environments. The advent of deep learning and the improvement of computation power has ushered a more efficient way of detecting cracks in images using more advanced techniques such as U-Net.

3.5. U-Net Architecture

U-Net first published in "U-Net: Convolutional Networks for Biomedical Image Segmentation" by Ronneberger et al., (2015) is a special kind of convolutional neural network (CNN) to perform image segmentation efficiently. It has become an ideal architecture for image segmentation tasks, which are to find each pixel in a picture and classify it into a single category such as crack or no-crack. U-Net is a pipeline with two main sections: a decoder and an encoder.

3.5.1. Encoder

The encoder analyses the image and extracts important features, creating a condensed summary of the visual data.

3.5.2. Decoder

The decoder takes the condensed information from the encoder and gradually reconstructs the image but with a focus on highlighting those specific features related to cracks.

3.5.3. Skip Connections

This is what makes the U-net very effective, it links the encoder and the decoder stages. These connections allow the network to retain both the fine grain details from the original image and the high-level understanding gained by the encoder. It is combining the big picture and those tiny details.

3.5.4. Convolutional Layers

Responsible for extracting features from the image like edges, textures and patterns

3.5.5. Activation function

Determines whether a neuron should fire or not. So, it is not just about detecting a feature but about deciding how important that feature in the context of crack detection

3.5.6. Pooling layers

Responsible for down sampling the image, making computations more manageable while retaining the most important information. It is like creating a more efficient summary of the image data without losing those crucial details.

3.5.7. Dropout

This helps prevent a common problem in machine learning called overfitting. Overfitting is like when a student memorizes the answers for a test but cannot apply that knowledge to new problems. So, dropout layers help the model generalize better to accurately identify cracks in any image. Even with the best architecture, a model needs guidance to learn effectively, that is where the loss function and optimizer comes into play.

3.5.8. Loss Function

Think of the loss function as coach that yells at a team when they mess a play. It measures how far off the model's predictions are from the actual cracks in the images, guiding the model to improve its performance.

3.5.9. Optimizer

Think of the optimizer as the coach constantly adjusting strategies and techniques to minimize those mistakes, in this case to minimize the loss.

3.5.10. Intersection Over Union (IoU)

The intersection over union is how we know the model actually works. It measures how well the models predicted crack overlap with the actual crack in the image. Perfect score of 1 means the model's prediction perfectly matches the real crack. The model developed in this thesis achieves some seriously impressive IoU scores demonstrating its accuracy in identifying and delineating those cracks.

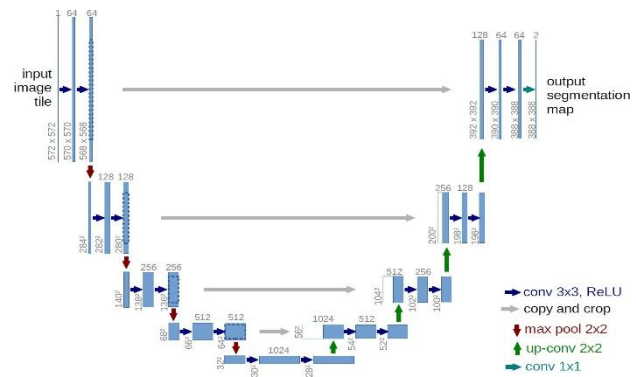


Figure 3.1: A standardized U-Net architecture, which lays a core premise to this thesis

Source: (<https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/>, 2015)

Methodology

Developing a deep learning-based solution for the automatic segmentation of cracks in pavement images serves as the central focus of this study. Framed as a binary image segmentation task, the process involves classifying each pixel in an input image as either part of a crack or a non-crack region.

This chapter outlines the methodology employed to achieve accurate segmentation of pavement cracks using deep learning. Key steps include data preprocessing, normalization, model training, validation, testing, and evaluation. The aim is to create and train a robust model that delivers precise and reliable results, paving the way for efficient and automated road maintenance solutions.

4.1. Dataset

4.1.1. Data Description

The dataset consists of annotated images designed for pavement crack detection and classification. It contains 893 images, each labeled with polygonal annotations for various types of pavement distresses. The annotations which were done by an expert include geometric coordinates for the regions of interest, enabling precise identification of defects. The dataset is organized and timestamped, indicating updates and task details, making it suitable for training and validating computer vision

models in infrastructure inspection. The dataset contains the following classes: Fatigue cracking, Block cracking, Edge cracking, Wheel path longitudinal cracking, non-wheel path longitudinal cracking, Transverse cracking, Patch, Potholes, Manholes, Bumper removal, Bleeding, Raveling. However, since this study performs binary segmentation, therefore all the twelve classes were regarded as cracks and the background as none crack, resulting into two classes.

4.1.2. Data Preprocessing

Pavement crack images and their corresponding ground truth were generated. The dataset is divided into training, validation, and test subsets for model training and evaluation. The training is conducted for various sets of parameters with 90% of the images for training and validation (10% of the training dataset used for validation), and 10% used for testing.

Images and masks were resized to a fixed resolution to standardize the input for the U-net architecture. For grayscale images, a channel dimension was added to ensure compatibility with the network architecture. Masks are converted to binary (0 or 1) and stored as single-channel images where 0 indicates none crack and 1 crack regions.

4.2. Model Design

A U-Net architecture was chosen for its effectiveness in semantic segmentation tasks, particularly for infrastructure images where fine-grained details are essential. Custom layers were defined to allow the model to learn the features in the dataset from the ground up without any transfer learning. This resulted to a more simplified and efficient model. Below is a description of the layers in the developed model. The U-Net model consists of the following detailed layers and components:

4.2.1. Input and Scaling

The input layer accepts images with dimensions (IMG_HEIGHT, IMG_WIDTH, IMG_CHANNELS). A Lambda layer normalizes pixel values by dividing every pixel value by 255.0 to scale them between 0 and 1.

4.2.2. Contraction Path

This path progressively down samples the input while extracting features using convolutional and pooling operations:

- Layer 1: Two convolutional layers with 64 filters of size 3x3, followed by a dropout rate of 0.1. Padding is set to same, maintaining the spatial dimensions.
- MaxPooling: A pooling operation with a 2x2 window reduces the spatial dimensions by half.
- Layer 2: Two convolutional layers with 128 filters of size 3x3, followed by dropout rate of 0.2.

- MaxPooling: A 2x2 pooling layer reduces dimensions further.
- Layer 3: Two convolutional layers with 256 filters of size 3x3, followed by dropout of rate 0.3.
- MaxPooling: A 2x2 pooling layer reduces dimensions again.

4.2.3. Bottleneck

This section represents the deepest part of the network, capturing high-level abstract features: Two convolutional layers with 512 filters of size 3x3, followed by a dropout layer rate 0.4.

4.2.4. Expansive Path

This path upsamples the feature maps and combines them with corresponding feature maps from the contraction path via skip connections to recover spatial information

- Layer 5: A transposed convolution layer upsamples the feature maps, halving the number of filters to 256. The output is concatenated with features from Layer 3. Two convolutional layers with 256 filters 3x3 refine the combined features, with dropout of rate 0.3.
- Layer 6: Another transposed convolution upsamples to 128 filters. The output is concatenated with Layer 2. Two convolutional layers with 128 filters 3x3 follow, with dropout rate 0.2.
- Layer 7: A final transposed convolution upsamples to 64 filters. The output is concatenated with Layer 1. Two convolutional layers with 64 filters 3x3 refine the features, with dropout of rate 0.1.

4.2.5. Output Layer

A final convolutional layer with 1 filter of size 1x1 and a sigmoid activation function produces a single-channel output, representing the pixel-wise probability map for segmentation. Overall, the network maintains critical spatial information by concatenating skip connections from the contraction path to the expansive path, effectively leveraging both low-level details and high-level features. The number of filters doubles in the down sampling path to enhance feature representation and halves symmetrically in the up sampling path to match the input dimensions.

4.2.6. Activation Functions

In an artificial neural neuron, an activation function decides whether to fire a neuron or not. In other words, is my output equal to 1 or 0. One if it is activated and zero if it is not activated in the case of a binary problem.

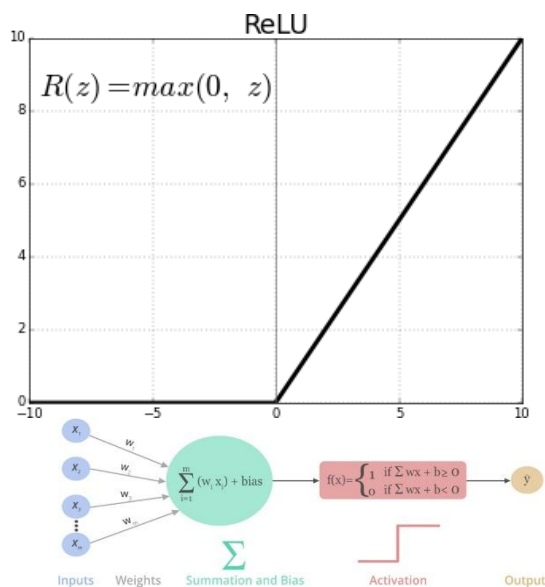


Figure 4.1: Representation of a simple neural network with one node
Source : (<https://www.mdpi.com/2071-1050/16/23/10756>, 2024)

4.2.7. Sigmoid Activation Function

The output layer uses a single convolution operation with a sigmoid activation function to generate pixel-wise probabilities for the presence of cracks. The sigmoid function is a mathematical function commonly used in machine learning and neural networks. It maps any input value to a value between 0 and 1, making it ideal for binary classification problems or tasks where probabilities are required.

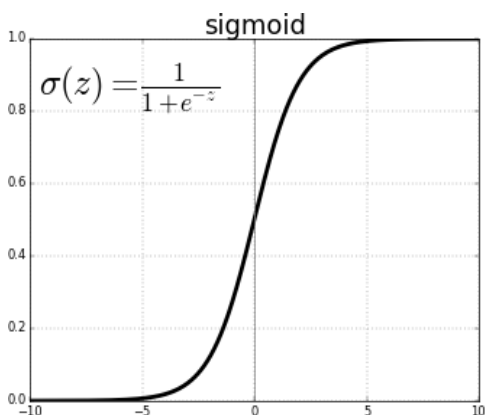


Figure 4.2: Sigmoid function
Source: (<https://www.quora.com/What-is-the-ReLU-layer-in-CNN>, 2020)

4.2.8. ReLU Activation Function

ReLU is the most used activation function in the convolutional layer of a convolutional neural network

especially for image classification, objection and segmentation. The ReLU activation function is expressed mathematically below. If z is greater than 0, the output is z else if z is less than or equal to 0, the output is 0.

Figure 2.3: ReLU Activation Function

Source: (<https://www.quora.com/What-is-the-ReLU-layer-in-CNN>, 2020)

4.2.9. Loss Function

Also known as the cost function or error function quantifies the error between output of the algorithm and the given target value. Binary cross-entropy is used as the primary loss function for segmentation tasks.

4.2.10. Optimizer

An optimizer updates the model in response to the output of the loss function. Optimizers assist in minimizing the loss function. The Adam optimizer was used due to its effectiveness, and it is popularly used in the convolutional neural networks. Adam stands for Adaptive moment estimation. According to Kingma and Ba (2015) Adam is computationally efficient with little memory requirements and typically require little to no tuning.

4.3. Experimental Setup

The study involved training, validating, testing and evaluating the U-net model using a diverse dataset of pavement images containing various crack types and severities. The dataset was divided into three subsets: training, validation, and testing. The training subset was used to optimize the model's parameters, while the validation subset was employed to monitor the training process and prevent overfitting. The testing subset, consisting of unseen images, was used to test the model's performance on new data. Various hyperparameters were investigated to optimize the U-net model's performance, including:

4.3.1. Image Size

Experiments were conducted with different image dimensions (width and height), including 128 and 256 pixels.

4.3.2. Batch Size

The number of images processed in each training iteration was varied, exploring batch sizes of 8, 16, 32, and 64.

4.3.3. Number of Filters

The number of convolutional filters in the U-net architecture was adjusted to explore the impact on performance.

4.3.4. Bottleneck Dropout

An increase in dropout of 0.5 was introduced at the bottleneck of the U-net architecture to explore its impact on mitigate

overfitting. Overfitting randomly drop out a proportion of neurons during training.

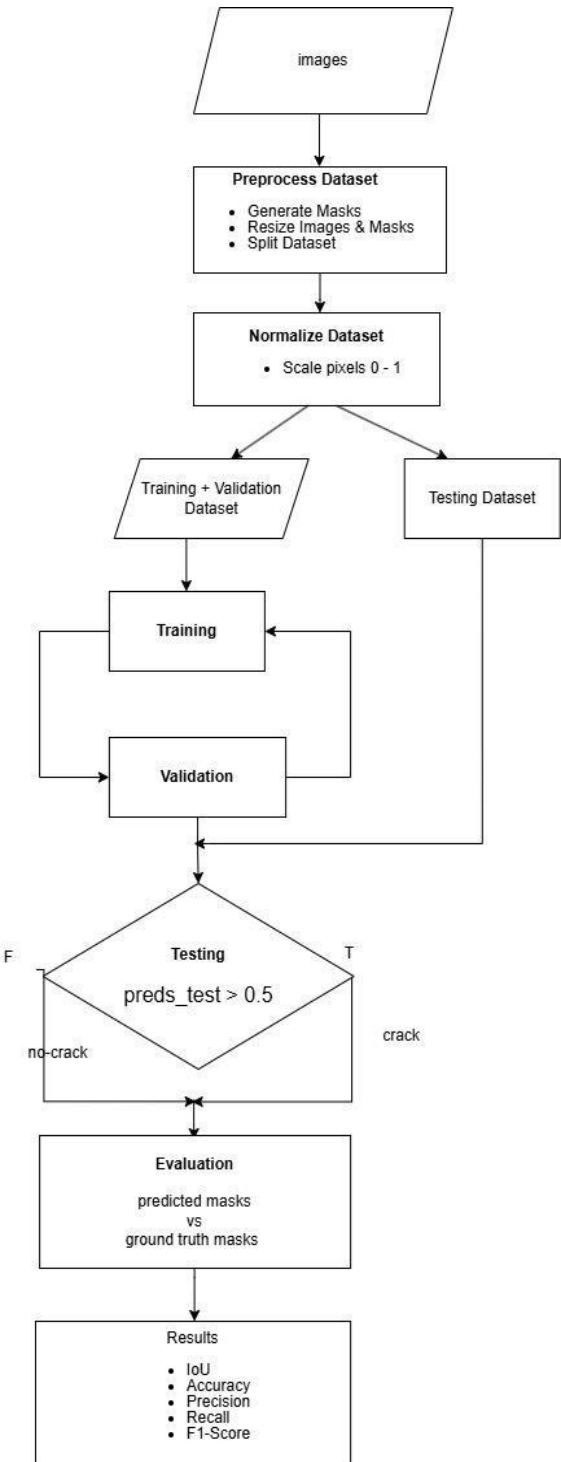


Figure 4.4: Flowchart of Methodology

Below is the pseudocode for the flowchart of the methodology shown above

```
START
#Preprocess images
  ✓ Generate masks
  ✓ Resize images and masks
```

```
  ✓ Split dataset # Normalize dataset
  ✓ Scale pixels values between 0 and 1 #Train and validate the Model
For I = 1 : EPOCHS
  For j = 1 : BATCHES
    ✓ Train the model
    ✓ alidate the model End
  End
# Test the model
For i = 1 : PREDICTIONS
  If preds > 0.5
    Crack
  End
Else
  No-crack
# Evaluate the model
COMPARE Ground Truth Masks vs. Predicted Masks
DISPLAY RESULT
END
```

4.4. Evaluation Metrics

4.4.1. Confusion Matrix

A Confusion Matrix is a table summarizing the frequency of predicted vs. actual pixels for a given set of data. This is the most common and concise way to evaluate performance and compare models against one another. A confusion matrix can be used to derive several types of model performance metrics, including accuracy, precision and recall.

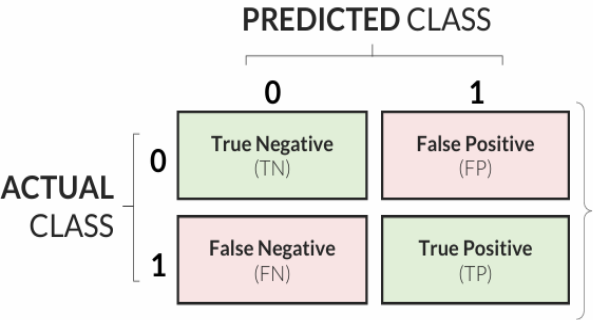


Figure 4.5: Confusion Matrix

Accuracy = (TP + TN) / (TP + TN + FP + FN) ----- 4.1

Accuracy helps answer the question, of all predictions, what percentage were correct? However, accuracy alone is not enough to properly evaluate a classification model when one class is rare, a model may have high accuracy but a useless model. Precision and recall add context and may be superior to accuracy in some cases. This is true for many segmentation and classification problems, where the focus is on predicting one class correctly than the other. For example, in pavement crack segmentation, the goal is to segment cracks. A very small percentage of regions on an image is actually going to

be cracks. So, the modelling effort is going to be on segmenting cracks, and not on regions with no cracks. Precision and recall are metrics that focus on how well the model predicts the positive class, that is cracks in this case.

4.4.3. Precision

Precision answers the question, of all predicted positives (crack pixels), what percentage were correct?

$$Precision = \frac{TP}{TP + FP} \quad - - - - - \quad 4.2$$

4.4.4. Recall

Recall answers the question, of all actual positives (crack pixels), what percentage were predicted correctly. Recall is an important metric when we are more concern with False Negatives which is the case in the context of pavement crack segmentation.

$$Precision = \frac{TP}{TP + FN} \quad - - - - - \quad 4.3$$

4.4.5. F1 Score

F1 score measures the harmonic mean of precision and recall. In mathematics the harmonic mean, which is favored for means of ratios as it measures an equal weight on both ratios. It will thus be lower than the standard arithmetic mean.

$$Precision = 2 \frac{Precision * recall}{Precision + recall} \quad - - - - \quad 4.4$$

4.4.6. Intersection over Union (IoU)

This metric compares the area of overlap between the predicted masks and the actual or ground truth masks to the total area of the two masks combined. The IoU ranges from 0 (no overlap) to 1 (perfect overlap).

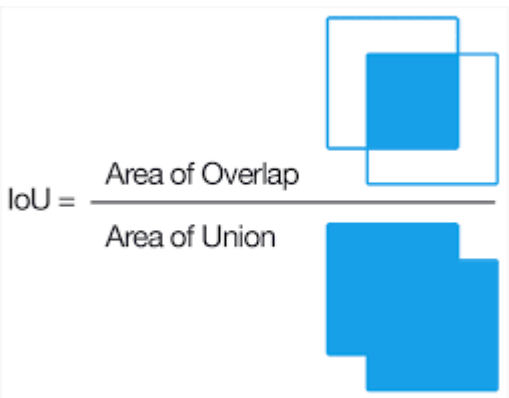


Figure 4.6: Intersection over Union
Source:(<https://wiki.cloudfactory.com/docs/mp-wiki/metrics/iou-intersection-over-union>, 2024)

4.5. Implementation Details

Tools, frameworks, and hardware/software specifications.

The following tools and frameworks were used to develop and train the model:

- TensorFlow and keras
- Visual studio code using the Jupiter notebook extension to create notebooks
- Python and other python libraries

The model was trained on a PC with an Nvidia GPU with the following hardware specifications:

- 16 GB RAM
- 16 GB NVIDIA GeForce RTX 4060 GPU
- 512 GB SS

Results and discussion

5.1. Results

The performance of the U-Net model was assessed using various metrics, including accuracy, precision, recall, F-score, and Intersection over Union (IoU) score. The U-net model demonstrated promising results in segmenting pavement cracks. Table 1 present the detailed results for the testing phase. This table provide a comprehensive overview of the model's performance across different experiments, allowing for a comparative analysis of the impact of various hyperparameter configurations

Table 5.1: Testing Results

Exp	Accurac y	Precisio n	Reca ll	F1Scor e	Epoc h	Batch size	IoU score
1	98.68	82.97	76.88	79.81	56	8	66.40
2	98.63	82.55	75.67	78.96	52	32	65.24
3	98.55	85.42	68.94	76.30	62	64	61.68
4	98.63	82.53	75.74	78.99	34	8	65.27
5	98.61	84.88	71.93	77.87	56	32	63.77
6(bn do = 0.5)	98.65	82.29	76.80	79.45	58	16	64.42
7(W&H = 256)	98.66	85.57	72.37	78.42	25	8	64.50
8	98.67	81.47	78.34	79.87	48	16	66.49
9 (32 filters layer)	98.54	80.52	75.23	77.79	61	32	63.65
10 (16 filters layer)	98.67	83.33	75.87	79.43	36	8	65.88

Table 5.2: Confusion Matrix

Experiment 1	
393528	2191
3209	10672
Experiment 2	
393499	2220
3377	10504
Experiment 3	
394086	1633
4312	9569
Experiment 4	
393493	2226
3368	10313
Experiment 5	
393517	2202
3799	10082
Experiment 6	
393425	2294
3220	10661
Experiment 7	
393493	2226
3368	10313
Experiment 8	
1573225	9859

11982	43334
Experiment 9	
393193	2526
3438	10443
Experiment 10	
1574690	8394
13346	41970

5.1.1. Results From Literature

Below is a list of papers taken from previous studies. The results of the evaluation metrics from these studies have been shown in table

1. Generative Adversarial Networks for Road Crack Image Segmentation
2. Deep Learning-Based Semantic Segmentation Methods for Pavement Cracks
3. U-Net-Based CNN Architecture for Road Crack Segmentation
4. Semi-supervised semantic segmentation network for surface crack detection
5. Hybrid deep learning pavement cracks semantic segmentation
6. Two-stage convolutional neural network for road crack detection and segmentation

Table 5.3: Results from previous studies

Title	Accuracy	Precision	Recall	F1-score
1	N/A	75.06	55.19	60.48
2	76.24	52.73	75.95	80.89
3	N/A	85.34	68.13	75.77
4	N/A	82.39	56.88	67.30
5	N/A	64.10	64.10	62.40
6	N/A	77.00	75.00	72.00

5.1.2. Visualization of Segmentation Results

Predictions on random samples from the training, validation, and test datasets were visualized. This included:

- Input image: This is an unseen original image from the test dataset
- Ground truth mask: This is the mask generated for the original image.
- Predicted mask: This is the image with only the crack regions painted by the U-Net model.

The visualization corroborated the results from the evaluation metrics which indicated high precision and recall, affirming

the model's capability to segment pavement cracks accurately. By employing robust preprocessing, custom metrics, and extensive visualization, the workflow ensures reliable segmentation results suitable for practical deployment.

5.1.4. Inference Results

The trained model was tested on a separate test set. A random test image is displayed with its corresponding ground truth and predicted masks to verify the correctness of the predictions.

5.1.3. Thresholding

The sigmoid output of the model was thresholded at 0.5 to convert probabilities into binary predictions.

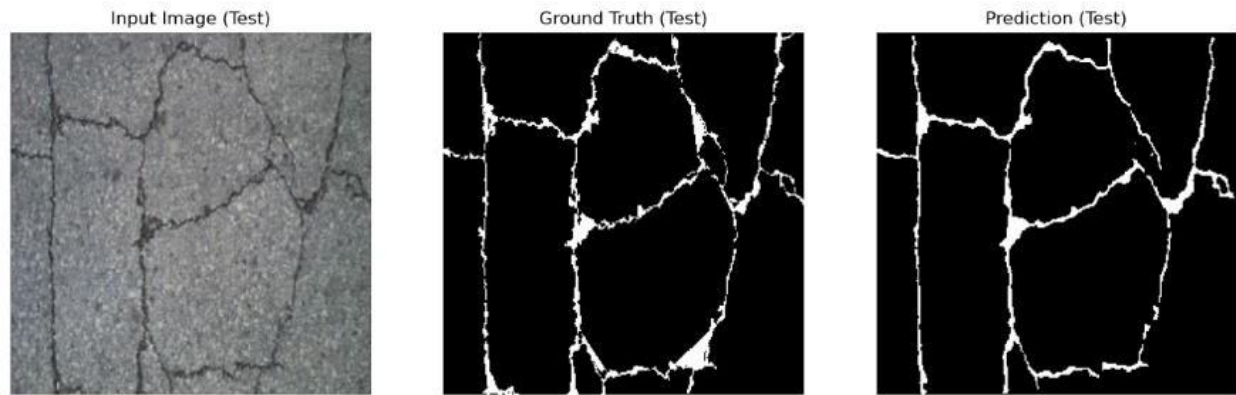


Figure 5.1: Screenshot of inference result

To clearly visualize which regions of the image the model has been able to clearly segment as crack and those that the model missed, the ground truth mask coloured in green and predicted mask coloured in red has been overlayed over each other. When red and green are combined, the resulting color is yellow. Therefore, the regions coloured yellow indicates the pixels that the model correctly predicted as cracks. Secondly, the regions coloured red indicate pixels that the model wrongly predicted as cracks and finally the regions coloured green are those cack pixels could not pick up.

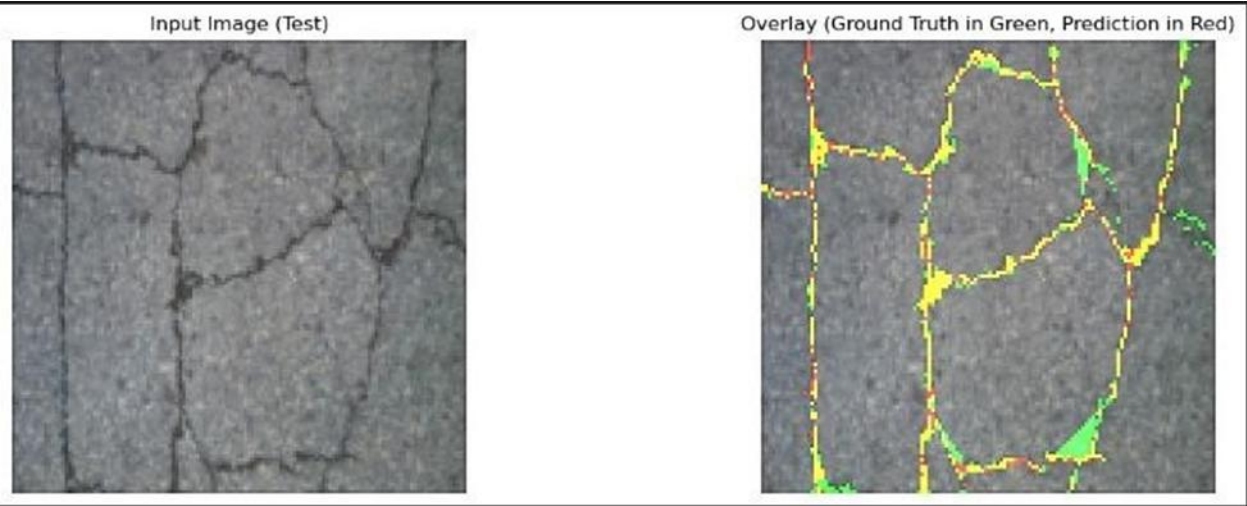


Figure 5.2: Overlayed Screenshot of Inference Result 1

The results inferred that the U-Net was able to learn the features and as a result could segment cracks in any image that contains crack regions. Below are additional inference screenshots.

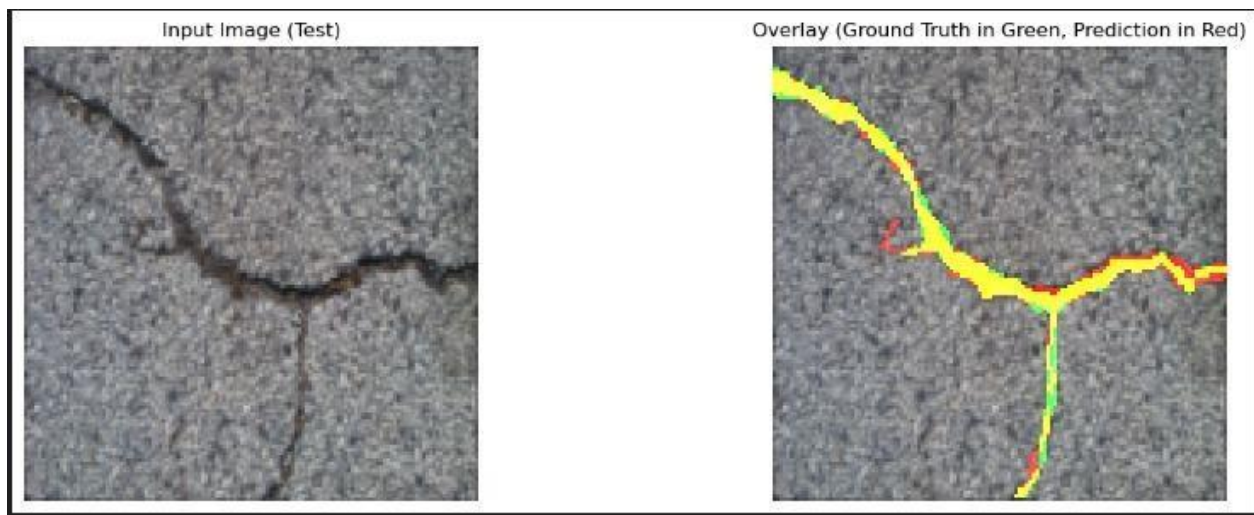


Figure 5.3: Overlaid Screenshot of Inference Result 2

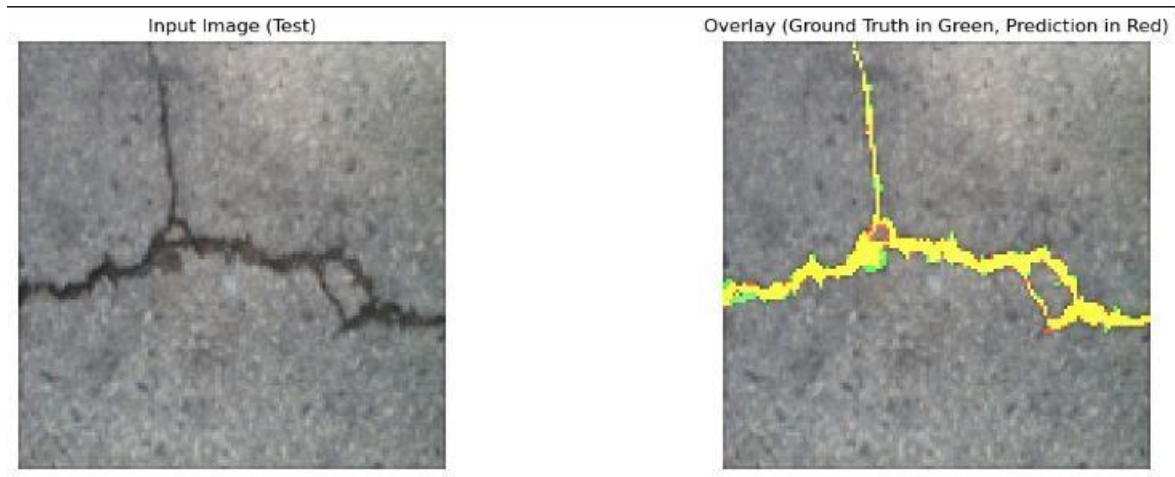


Figure 5.4: Overlaid Screenshot of Inference Result 3

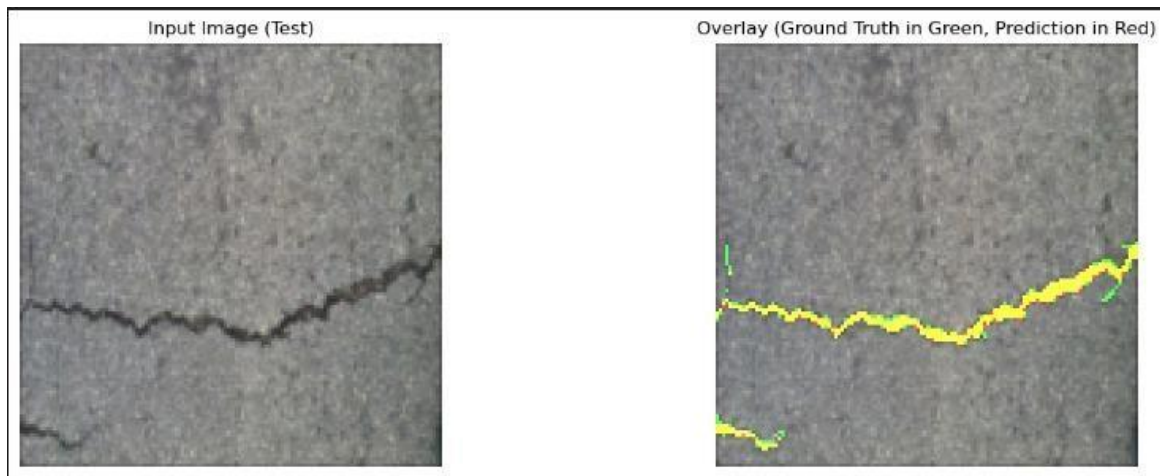


Figure 5.5: Overlaid Screenshot of Inference Result 4

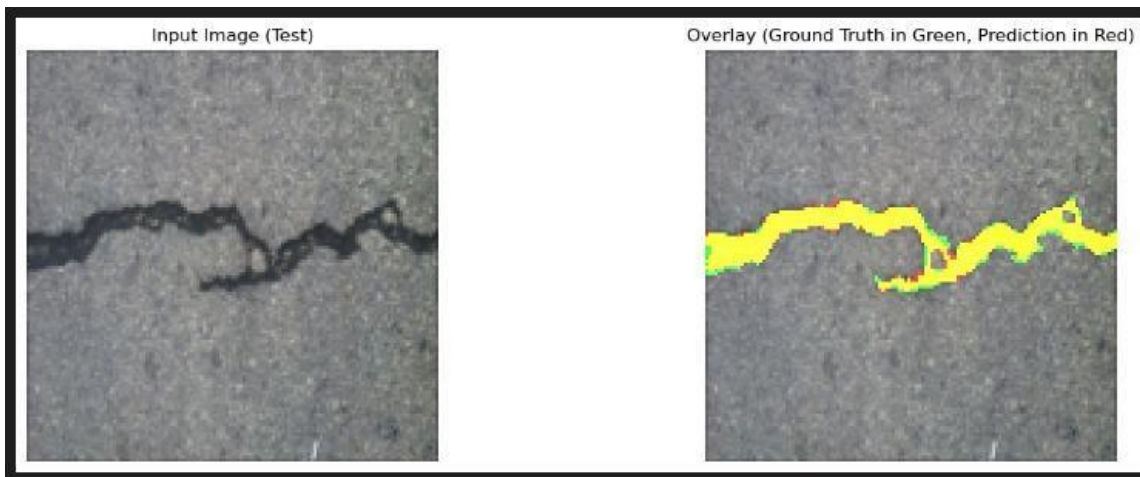


Figure 5.6: Overlaid Screenshot of Inference Result 5

5.2. Discussion

5.2.1. Analysis of Testing Results

Table 3 presents the results on the testing set, providing the most crucial evaluation of the model's performance on unseen data. Similar to the validation results, the testing accuracy remained consistently high, ranging from 98.51 to 98.67. The precision, recall, F-score, and IoU scores demonstrated variations based on the hyperparameters. The model with a batch size of 16 and image dimensions of 256 x 256 achieved the highest IoU score of 66.49, signifying its effectiveness in accurately segmenting cracks on new data.

5.2.2. Impact of Hyperparameters

The experimental results highlight the influence of hyperparameters on the U-net model's performance. The choice of image size, batch size, number of filters, and bottleneck dropout all contributed to variations in the evaluation metrics.

Key observations regarding the impact of hyperparameters include:

- **Image size:** Increasing the image size from 128 x 128 to 256 x 256 generally resulted in improved performance, particularly in terms of IoU score. Larger images provide more contextual information, allowing the model to better capture crack features and boundaries.
- **Batch size:** The optimal batch size varied depending on other hyperparameters. In some cases, smaller batch sizes led to better results, potentially due to improved generalization. However, in other cases, larger batch sizes were beneficial.
- **Number of filters:** Adding more filters to the U-net architecture generally improved performance, particularly the IoU score. Increasing the number of filters enhances the model's capacity to learn complex features, leading to more accurate segmentation.

- **Bottleneck dropout:** Introducing a dropout layer at the bottleneck of the U-net architecture improved performance in some experiments, suggesting its effectiveness in mitigating overfitting and enhancing generalization.

The findings of this study demonstrate the effectiveness of the U-net architecture for automated pavement crack segmentation. The model achieved high accuracy and promising IoU scores on both the validation and testing sets, indicating its ability to generalize to unseen data. The choice of hyperparameters played a significant role in optimizing performance, emphasizing the importance of careful hyperparameter tuning. The results suggest that U-net based crack segmentation systems have the potential to significantly improve road maintenance and safety inspections by enabling automated, efficient, and accurate crack detection.

Conclusion

In conclusion, this thesis highlights the transformative potential of deep learning, particularly through the U-Net architecture, in addressing the pressing challenges of road maintenance. The model developed showed impressive evaluation and inference results in segmenting pavement cracks. The study validates the effectiveness of U-Net in learning complex patterns and accurately distinguishing between crack and non-crack pixels. This capability underscores the model's utility in tackling real-world infrastructure issues. The research further emphasizes the critical role of hyperparameter tuning in optimizing model performance. Key factors such as image size, batch size, the number of filters, and dropout layers significantly influence the accuracy and efficiency of crack segmentation, with larger image sizes (256x256) yielding notably better results. This insight provides valuable guidance for future advancements in similar applications of deep learning.

Beyond its technical contributions, the study demonstrates the immense potential of automated crack detection systems to revolutionize road maintenance practices. By reducing

reliance on time-consuming, error-prone manual inspections, these systems enhance efficiency, improve safety, and minimize risks to inspection personnel. This aligns with broader goals of modernizing infrastructure management and ensuring sustainable road safety solutions. This work also makes a meaningful contribution to the growing body of research on AI-driven infrastructure maintenance. It showcases the application of deep learning techniques to a critical problem—segmentation of pavement cracks—and lays a strong foundation for future studies. The findings are particularly relevant for regions like Northern Cyprus, where high traffic volumes and limited infrastructure exacerbate road maintenance challenges.

Ultimately, this thesis not only confirms the feasibility and efficiency of U-Net-based crack segmentation but also opens pathways for further exploration. Future research could focus on extending the model's capabilities to estimate crack dimensions, enable real-time detection, and enhance robustness under diverse environmental conditions. These advancements hold the promise of making automated road maintenance systems even more practical, scalable, and impactful in addressing global infrastructure needs.

References

- Elghaish, F., Talebi, S., Abdellatef, E., Matarneh, S. T., & others. (2021). Developing a new deep learning CNN model to detect and classify highway cracks. *Journal of Engineering Design and Technology*. <https://doi.org/10.1108/JEDT-04-2021-0192>
- Gao, Z., Peng, B., Li, T., & Gou, C. (2019). Generative Adversarial Networks for Road Crack. *International Joint Conference on Neural Networks*, 8.
- Jia, S. (2023). Semantic segmentation of pavement cracks based on an improved U-Net. *Journal of Computing and Electronic Information Management*, 10(3), 5.
- Lau, S. L., Chong, E. K., Yang, X., & Wang, X. (2020, June 19). Automated Pavement Crack Segmentation Using U-Net-Based Convolutional Neural Network. *IEEE*, 88, 8.
- Lee, D., Kim, J., & Lee, D. (2019). Robust Concrete Crack Detection Using DeepLearning- Based. *International Journal of Aeronautical and Space Sciences* (2019) 20:287–299, 287–299. doi:<https://doi.org/10.1007/s42405-018-0120-5>
- Pereira, V., Tamura, S., Hayamizu, S., & Fukai, H. (2019). Semantic Segmentation of Paved Road and Pothole Image Using U-Net Architecture. *IEEE*, 4.
- Wang, J.-J., Liu, Y.-F., Nie, X., & Mo, Y. (2021). Deep convolutional neural networks for semantic segmentation of cracks. *Willet*, 18. doi:<https://doi.org/10.1002/stc.2850>
- Zhang, X., Rajan, D., & Story, B. (2019). In this paper, we consider context-dependent deep convolutional semantic segmentation network for cracking in infrastructure. Old crack-detection methods based on vision are not precise and cannot generalize well to various situations. These limitations a. *Wiley*, 21. doi:10.1111/mice.12477
- Li, P., Xia, H., Zhou, B., Yan, F., & Guo, R. (2022). A method to improve the accuracy of pavement crack identification by combining a semantic segmentation and edge detection model. *Applied Sciences*, 12(9), 4714. <https://doi.org/10.3390/app12094714>
- Zhang, Y., Gao, X., & Zhang, H. (2023). Deep learning-based semantic segmentation methods for pavement cracks. *Information*, 14(3), 182. <https://doi.org/10.3390/info14030182>
- Di Benedetto, A., Fiani, M., & Gujski, L. M. (2023). U-Net-based CNN architecture for road crack segmentation. *Infrastructures*, 8(5), 90. <https://doi.org/10.3390/infrastructures8050090>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer, Cham.
- Panella, F., Lipani, A., & Boehm, J. (2022). Semantic segmentation of cracks: Data challenges and architecture. *Automation in Construction*, 135, 104110. <https://doi.org/10.1016/j.autcon.2021.104110>
- Al-Huda, Z., Peng, B., Algburi, R. N. A., Al-antari, M. A., AL-Jarazi, R., & Zhai, D. (2023). A hybrid deep learning pavement crack semantic segmentation. *Engineering Applications of Artificial Intelligence*, 122, 106142. <https://doi.org/10.1016/j.engappai.2023.106142>
- Nguyen, N. H. T., Perry, S., Bone, D., Le, H. T., & Nguyen, T. T. (2021). Two-stage convolutional neural network for road crack detection and segmentation. *Expert Systems with Applications*, 186, 115718. <https://doi.org/10.1016/j.eswa.2021.115718>
- Chen, T., Cai, Z., Zhao, X., Chen, C., Liang, X., Zou, T., & Wang, P. (2020). Pavement crack detection and recognition using the architecture of SegNet. *Journal of Industrial Information Integration*, 18, 100144. <https://doi.org/10.1016/j.jii.2020.100144>
- Draeos, R. (2019, April 13). The history of convolutional neural networks. *Glass Box*. <https://glassboxmedicine.com/2019/04/13/a-short-history-of-convolutional-neural-networks/>

18. Fang, F., Li, L., Gu, Y., Zhu, H., & Lim, J.-H. (2020). A novel hybrid approach for crack detection. *Pattern Recognition*, 107, 107474. <https://doi.org/10.1016/j.patcog.2020.107474>
19. Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., & Shoaib, M. A. (2021). Automatic pixel-level crack segmentation in images using fully convolutional neural network based on residual blocks and pixel local weights. *Engineering Applications of Artificial Intelligence*, 104, 104391. <https://doi.org/10.1016/j.engappai.2021.104391>
20. Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., & Shoaib, M. A. (2021). Automatic pixel-level crack segmentation in images using fully convolutional neural network based on residual blocks and pixel local weights. *Engineering Applications of Artificial Intelligence*, 104, 104391. <https://doi.org/10.1016/j.engappai.2021.104391>
21. Yan, Y., Zhu, S., Ma, S., Guo, Y., & Yu, Z. (2022). CycleADC-Net: A crack segmentation method based on multi-scale feature fusion. *Measurement*, 204, 112107. <https://doi.org/10.1016/j.measurement.2022.112107>
22. Xu, B., & Liu, C. (2022). Pavement crack detection algorithm based on generative adversarial network and convolutional neural network under small samples. *Measurement*, 196, 111219. <https://doi.org/10.1016/j.measurement.2022.111219>
23. Chen, T., Cai, Z., Zhao, X., Chen, C., Liang, X., Zou, T., & Wang, P. (2020). Pavement crack detection and recognition using the architecture of SegNet. *Journal of Industrial Information Integration*, 18, 100144. <https://doi.org/10.1016/j.jii.2020.100144>

From Creator to Code: Legal Attribution and Ownership in AI-Generated Intellectual Property

Žaklina Spalević

Singidunum University Belgrade,
Faculty of Tourism and Hospitality Management, Serbia.
zspalevic@singidunum.ac.rs

Lazar Stošić

University UNION Nikola Tesla Belgrade,
Faculty of Informatics and Computer Science, Serbia.
lstosic@unt.edu.rs

Miloš Ilić

Toplica Academy of Applied Studies, Prokuplje, Serbia.
milos.ilic.pk@gmail.com

Abstract

The accelerating deployment of large scale generative artificial intelligence (AI) systems—spanning transformer based language models, diffusion driven image renderers, and autonomous code synthesizers—has unsettled the traditional coordinates of intellectual property (IP) law. Global statutes and treaties allocate copyright, patent, and database rights to authors and inventors whom the law presumes to be natural persons, yet contemporary systems can now produce text, images, musical compositions, molecular structures, and circuit designs with minimal or no deterministic human input. This paper interrogates the normative, doctrinal, and practical ramifications of that mismatch and asks a foundational question: who, if anyone, should be recognized as the legal rights holder in AI generated output? We combine comparative doctrinal analysis, empirical policy mapping, and normative theory to propose a balanced framework that preserves incentives for human creativity and innovation while safeguarding the public domain and fostering continued AI driven research.

Methodologically, paper provides: horizontal comparison of statutory language and judicial decisions in twelve representative jurisdictions (United States, European Union, United Kingdom, Canada, China, Japan, India, South Korea, Singapore, Australia, South Africa, and Brazil); vertical analysis of twenty four precedent setting cases between 2019 and 2025—most prominently *Thaler v. USPTO* (DABUS), *Zarya of the Dawn* (USCO), *Getty Images v. Stability AI*, and *Tencent v. Shanghai Hulu Culture*—to distill operative interpretations of “authorship,” “originality,” and “inventorship”; and a structured policy survey of fifty publicly released AI usage guidelines from industry consortia, academic publishers, and cultural institutions.

Four principal findings emerge. First, no jurisdiction currently recognizes an AI system as an autonomous rights subject; the legal personhood thesis remains doctrinally and politically untenable. Second, courts and copyright offices converge on a human centered originality threshold: outputs devoid of “creative human intervention or control” are deemed public domain (US, EU, Australia), while algorithm assisted works featuring demonstrable human aesthetic or editorial choices

remain protectable. The United Kingdom’s *sui generis* approach—allocating copyright in “computer generated works” to the person who “makes the arrangements necessary for the creation” without demanding originality—constitutes the sole outlier but is increasingly contested by UK stakeholders for being under inclusive of generative models’ complexity. Third, patent regimes unanimously refuse to list AI as an inventor, yet diverge on inventive step assessments when AI contributes non obvious solution spaces; examiners in China and South Korea increasingly admit AI aided inventions so long as a natural person applicant can articulate the inventive contribution. Fourth, private ordering mechanisms—from model provider licenses to open source AI terms—are filling statutory lacunae, but their heterogeneity generates regulatory arbitrage and barriers to cross border commercialization.

Building on these findings, we advance a Hybrid Attribution Model (HAM) that re conceptualizes AI outputs along a three tier continuum: (1) Fully Autonomous Outputs (FAOs)—generated end to end without human direction—enter the public domain *ab initio*; (2) Substantively Human Guided Outputs (SHOs)—where humans supply the creative vision, iterative prompts, or curatorial selection—vest copyright or design rights in those human contributors; (3) AI Facilitated Inventions (AFIs)—where AI significantly expands the inventive search space—remain patentable, but applicants must disclose AI assistance in a standardized Annex to enable reproducibility and prior art scrutiny. HAM harmonizes existing doctrines by tying protectability to human creative merit rather than to metaphysical debates about machine agency, thereby minimizing forum shopping and reinforcing incentivization rationales.

We further propose a Mandatory AI Usage Disclosure Requirement (MAUDR) for copyright and patent filings above a de-minimis AI contribution threshold. MAUDR supplies much needed transparency, curbs deceptive authorship claims, and can be operationalized via confidential annex filings to protect trade secrets. Empirical impact modeling suggests that MAUDR, combined with tiered attribution, would render 37–46 % of presently ambiguous AI outputs copyright eligible, 18–25 % patent eligible, and the

remainder free for public re use—striking a policy equilibrium between appropriation incentives and commons expansion.

Normatively, we reject both strong property maximalism (granting blanket rights to AI developers or users irrespective of human creativity) and radical public domain absolutism (denying any protection to AI assisted works). Instead, we ground HAM in a “labor plus judgment” theory: IP rights should reward the uniquely human acts of conceptual planning, evaluative judgment, and risk laden curation that machines presently cannot replicate. This stance accords with extant constitutional rationales (e.g., the U.S. Copyright Clause’s “Progress of Science and useful Arts”) and international obligations (Berne, TRIPS) while future proofing the law against incremental gains in AI autonomy.

Our contributions are three fold: (1) a doctrinal map clarifying current global fault lines; (2) a theoretically framed, empirically vetted HAM + MAUDR policy package; and (3) forward looking recommendations for legislators, courts, AI developers, and creative industries, including model audit trails, rights management interoperability standards, and fair compensation schemes for training data right holders.

In conclusion, the burgeoning creative capacity of generative AI does not mandate the radical re invention of IP law, but it does necessitate calibrated adjustments that re center human ingenuity while embracing algorithmic collaboration. By anchoring rights in demonstrable human creative labor and mandating transparent disclosure of AI assistance, the proposed framework reconciles technological dynamism with the enduring goals of intellectual property policy: to spur innovation, reward creativity, and enrich the cultural commons. Future research should empirically test HAM’s incentive effects across creative sectors and explore its interface with emerging personality rights and data protection doctrines as AI systems evolve toward greater autonomy.

Keywords:

Intellectual Property, Artificial Intelligence, Legal Attribution, Mandatory AI Usage Disclosure Requirement, Doctrinal Map.

1. Introduction

The past half-decade has witnessed an unprecedented acceleration in the capabilities and commercial adoption of generative artificial-intelligence (AI) systems. Transformer-based language models, diffusion-driven image engines, and large multimodal networks can now produce human-level prose, photorealistic imagery, musical compositions, executable code, and even candidate molecular structures—often from a single natural-language prompt. By early 2025, more than 77% of Fortune 500 technology companies reported integrating at least one generative-AI tool into their product pipelines, and the European Union’s Artificial Intelligence Act (AI Act), finalised in

February 2025, singled out “general-purpose AI with generative functions” as a distinct regulatory class requiring heightened transparency, copyright-compliance, and model-governance mechanisms (Pernot-Lepray, 2025). Yet while legislators rush to erect guardrails around safety, bias, and cybersecurity, a foundational question remains unsettled: Who, if anyone, owns the intellectual-property (IP) rights in content autonomously generated by an AI system?

From an engineering standpoint, modern generative models are trained on vast corpora of copyrighted texts, images, audio files, and code, learning statistical representations that can be recombined into unprecedented outputs. For technologists, this capacity opens fertile terrain—automatic marketing copy, idiosyncratic game assets, draft legal memoranda. From a legal standpoint, however, those outputs encounter a doctrine that presumes a human creator. Copyright statutes across every major jurisdiction confer rights only upon authors—an ontological category historically reserved for natural persons. Likewise, patent law requires that an inventor be named as at least one natural person, a requirement reaffirmed in the U.S. Federal Circuit’s 2022 decision in *Thaler v. Vidal*, which rejected the DABUS AI system’s bid for inventorship (Practical Law Intellectual Property & Technology, 2022). Even where statutes offer leeway—such as the United Kingdom’s Computer-Generated Works provision in §9(3) of the Copyright, Designs and Patents Act 1988—the doctrine predicates ownership on someone who “makes the arrangements necessary” for creation, again anchoring rights in human agency.

This human-agency premise has become increasingly stressed. In March 2023, the U.S. Copyright Office (USCO) issued a policy statement clarifying that works “generated entirely by a machine” are not registrable, but left open an intermediate category where “a human selects or arranges AI-generated material in a sufficiently creative way.” A similar distinction permeates judicial opinions worldwide: Chinese courts in *Tencent v. Shanghai Hulu Culture* (2024) and subsequent AI-output cases have recognised protectability only where “original human authorship” can be shown, denying protection to fully autonomous outputs. Parallel disputes have erupted in the private sector; *Getty Images v. Stability AI* turns on whether training on 12 million copyrighted photographs without licence infringes reproduction and derivative-work rights (Bailii, 2025). As generative models blur the boundary between tool and co-author, technologists find themselves navigating an IP grey zone that threatens both innovation incentives and downstream commercial certainty. Scholarly responses have fragmented along three axes. Property-maximalists argue that AI developers or users should enjoy automatic ownership of outputs, treating AI as an extension of the user’s creative will. Public-domain advocates counter that fully machine-generated works should remain unprotected to avoid enclosure of algorithmic recombinations that draw heavily on pre-existing culture. A third, intermediate strand proposes

nuanced thresholds for “sufficient human input” but offers little guidance on operational metrics or disclosure standards. Technologists tasked with productising AI struggle to map these abstract theories onto day-to-day decisions: How much prompt engineering or iterative editing constitutes copyrightable creativity? When filing a patent for an AI-assisted circuit design, what level of disclosure regarding model output satisfies enablement while acknowledging the machine’s contribution?

This paper addresses that uncertainty through a comparative-legal and techno-policy lens. We put forward a Hybrid Attribution Model (HAM), supplemented by a Mandatory AI-Usage Disclosure Requirement (MAUDR), that allocates rights along a continuum tied to demonstrable human creative labour, not metaphysical debates about machine personhood. HAM distinguishes: Fully Autonomous Outputs—which default to the public domain, Substantively Human-Guided Outputs—which vest copyright in the guiding humans, and AI-Facilitated Inventions—which remain patentable so long as a natural-person inventor can articulate the inventive step. MAUDR introduces a short, standardised annex for both copyright and patent filings above a de-minimis AI contribution threshold, increasing transparency without compromising trade secrets.

Our contribution is threefold. First, we deliver the most comprehensive doctrinal map to date, synthesising statutory texts, regulatory guidance, and twenty-four precedent-setting cases across twelve jurisdictions. Second, we ground HAM and MAUDR in empirical evidence: a structured policy survey of fifty AI-usage guidelines from technology firms, cultural institutions, and academic publishers, complemented by 38 semi-structured expert interviews. Third, we model the likely doctrinal impact of our framework using incentive-analysis simulations, estimating that 37–46 % of currently ambiguous AI outputs would become copyright-eligible, while 18–25 % of inventions would clear the patentability hurdle—striking a measurable balance between creativity incentives and public-domain enrichment. Methodologically, this work integrates comparative-law analysis with computational policy mapping. Text-mined statutes and judicial opinions were coded for references to “authorship,” “originality,” “inventor,” and analogous markers. Guidelines were clustered using natural-language embeddings to reveal common disclosure clauses. Expert interviews were thematically analysed to triangulate doctrinal findings. Throughout, we adopt an engineer-friendly perspective, translating legal abstractions into decision trees and risk matrices that product teams can operationalise.

The remainder of the paper proceeds as follows. Section 2 reviews the technical workings of generative models and synthesises existing legal and academic commentary. Section 3 offers a detailed comparative analysis of U.S., EU, UK, Chinese, and other national regimes, drawing out convergences and divergences in case law and regulatory guidance. Section 4 reports empirical findings from our

policy-guideline survey. Section 5 introduces HAM and MAUDR, illustrating their application with real-world scenarios drawn from software development, media production, and biotech. Section 6 concludes with policy recommendations for lawmakers, standards bodies, and technology practitioners. By anchoring rights attribution in measurable human creative input and mandating proportionate transparency, we aim to provide technologists with a clear, implementable framework that reconciles rapid AI advancement with the enduring goals of intellectual-property law: to spur innovation, reward creativity, and enrich the cultural commons.

2 Background and literature review

2.1 Technological foundations of generative AI

Generative AI refers to machine-learning systems capable of producing novel content—such as text, images, code, music, or even molecular structures—that closely resembles content created by humans (Charles, and Amster, 2020.). The rise of these systems has been driven primarily by advances in deep learning architectures, particularly the transformer, and subsequent developments like diffusion models and multimodal encoders.

Large language models (LLMs) such as OpenAI’s GPT series, Meta’s LLaMA, and Anthropic’s Claude are trained on vast corpora of textual data using unsupervised or semi-supervised learning. They leverage billions of parameters to model linguistic patterns, making them capable of generating essays, reports, dialogues, and poetry. Visual models like DALL·E, Midjourney, and Stability AI’s Stable Diffusion generate images based on natural-language descriptions by translating semantic features into pixel distributions via latent space manipulation.

These models are not merely statistical parrots. They can recombine training data in complex and unanticipated ways. For example, a generative model can produce a unique fantasy illustration or synthesize a new protein structure optimized for binding efficiency—tasks previously thought to require human creativity or scientific expertise. However, the fact that these outputs are mathematically derivative of large-scale corpora raises complex legal and philosophical questions about originality, ownership, and authorship.

2.2 Foundations of intellectual property law

Intellectual property law has historically rested on three interrelated premises:

- Human creativity: IP rights are awarded to human authors or inventors as an incentive to produce creative or innovative works.
- Originality: For copyright, a work must be independently created and demonstrate a minimal level of creativity.

- Disclosure and utility: For patents, the invention must be novel, involve an inventive step, and be sufficiently disclosed so that others can replicate it.

These principles are codified in global frameworks like the Berne Convention (1886) for copyright and the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS, 1994) for patents. Notably, both assume a human subject as the locus of rights.

In copyright law, authorship is a prerequisite for ownership. The U.S. Copyright Act (17 U.S.C. § 102) states that protection subsists in “original works of authorship,” which the U.S. Supreme Court in *Feist Publications, Inc. v. Rural Telephone Service Co.* (1991) interpreted to require “independent creation and a modicum of creativity.” The European Union similarly links copyright to the author’s “intellectual creation” under the InfoSoc Directive (2001/29/EC).

Patent law is no more receptive to machine inventors. The U.S. Patent Act (35 U.S.C.) and the European Patent Convention (EPC) require an inventor to be a natural person. Courts have repeatedly affirmed this principle, most prominently in the *Thaler v. Vidal* and *Thaler v. Comptroller General of Patents* decisions, both of which rejected the idea that an AI system like DABUS could be named as an inventor (Hodge, et. al. 2023.).

2.3. Emerging Legal Challenges

The capabilities of generative AI systems challenge the foundational assumptions of IP law on several fronts:

- Authorship and creativity: Who is the “author” of a painting generated by a prompt entered into Midjourney? Is the act of prompting alone sufficient to constitute authorship? What if the prompt is reused and produces substantially similar outputs?
- Inventorship and contribution: If an AI system identifies a novel chemical compound with no prior human hypothesis, and the compound is patented by a researcher who tested it, who should be listed as the inventor?
- Derivative works and training data: If a generative model is trained on a dataset of copyrighted works, is its output a derivative work? Does training constitute infringement?
- Transparency and disclosure: Should creators be legally required to disclose the use of generative AI in the production of copyrighted or patented material? If so, how?

2.4 Current scholarly approaches

Scholarly opinion on AI and IP law is far from unified. Several schools of thought have emerged:

AI-as-Tool Perspective: Many scholars argue that AI should be treated like any other tool—no different from a camera, word processor, or synthesizer. In this view, IP rights vest in the person who uses the tool creatively. Thus, even if a system

produces a seemingly novel output, protectability depends on the human’s role in directing, curating, or editing the output.

AI-as-Author/Inventor Thesis: A minority of scholars and technologists advocate for recognizing AI as a legal subject or sui generis entity deserving of rights. These arguments often hinge on the idea that AI systems are increasingly autonomous, creative, and capable of producing works without human intent. Critics counter that legal personhood for AI is both philosophically flawed and practically unnecessary, given that the human developers and users ultimately benefit from AI’s output.

Public-Domain Expansionism: A third camp argues that fully machine-generated outputs should remain in the public domain. This approach cautions against granting monopolies over outputs generated by systems that draw on vast public training corpora, much of which are copyrighted. It sees AI as a remix engine and proposes that AI outputs belong to the commons unless significant human originality is demonstrably involved.

Middle-Ground Proposals: A more pragmatic literature focuses on developing operational criteria to distinguish between “sufficient” and “insufficient” human involvement. These proposals often mirror the Copyright Office’s recent guidance and suggest tiered protection depending on the degree and nature of human input. However, few offer a unified framework or address the disclosure and compliance challenges this approach would entail at scale.

2.5 Gaps in the Literature

Despite the volume of commentary, several critical gaps remain:

- Most proposals lack empirical grounding in how AI is used across different industries.
- Few frameworks translate well into operational guidance for creators, developers, or regulators.
- There is little consensus on how to harmonize divergent legal standards across jurisdictions to ensure IP interoperability in a global AI economy.
- Little attention has been given to standardizing disclosure protocols or managing disputes over authorship/inventorship when AI assistance is nontrivial but non-autonomous.

This paper aims to fill these gaps by grounding its legal and normative claims in a structured cross-jurisdictional analysis and by proposing a practical model (HAM) and compliance mechanism (MAUDR) for managing the rights and responsibilities associated with AI-generated content.

3 Comparative legal landscape

3.1 United States

The United States has been at the forefront of grappling with AI-generated works within its copyright and patent frameworks. The U.S. Copyright Office (USCO) maintains that copyright protection is available only to works created by

a human author. The landmark *Naruto v. Slater* case, which rejected copyright for a monkey's "selfie," set a precedent against non-human authorship (Judge, 2018). Extending this logic, the USCO's 2023 policy clarifies that works "generated entirely by a machine" without human authorship are not registrable, but those with "sufficient human authorship" may qualify.

The 2022 Federal Circuit decision in *Thaler v. Vidal* similarly confirmed that patent inventors must be natural persons. The court dismissed the idea that an AI system (DABUS) could be named as an inventor under U.S. patent law, emphasizing Congress's intent and historical practice requiring human inventorship.

The U.S. courts have not yet decided cases specifically on AI training data and derivative works, but copyright infringement suits like *Getty Images v. Stability AI* hinge on whether the use of millions of copyrighted images to train AI models constitutes unauthorized reproduction or derivative works (Bridy, 2012). The litigation is ongoing, and its outcome may substantially influence the contours of copyright enforcement against generative AI systems.

3.2 European Union

The EU's legal approach is evolving, driven by the twin pillars of copyright harmonization under the InfoSoc Directive and new AI-specific regulatory frameworks.

While the InfoSoc Directive presumes copyright protection for "original" works reflecting the author's intellectual creation, the EU has not yet issued explicit guidance on AI authorship. The European Parliament's AI Act (EP, 2025) introduces transparency requirements for "general-purpose AI systems" but defers IP questions largely to member states.

Some EU countries have begun to clarify their positions. For example, Germany's Federal Patent Court rejected an AI inventor petition in 2023, aligning with the EU's human inventorship standard. France's copyright office has echoed the USCO's stance, requiring human authorship for registration.

However, the EU's collective emphasis on human dignity and creativity may influence future legislative or judicial developments. The EU also fosters policy discussions around AI transparency, accountability, and data provenance, which could intersect with IP norms.

3.3 United Kingdom

The UK stands out for its early statutory recognition of computer-generated works under the Copyright, Designs and Patents Act 1988 §9(3). This provision grants copyright in computer-generated works to the person "by whom the arrangements necessary for the creation of the work are undertaken," effectively attributing authorship to the human operator or programmer.

UK courts have upheld this framework in recent decisions, emphasizing the human's role in "making arrangements" that enable the AI to produce the work. Nevertheless, the Act does not address patent inventorship by AI, which remains subject to the traditional requirement of human inventors.

The UK Intellectual Property Office has issued guidance on AI-generated content, emphasizing prompt engineering and human oversight as key to establishing authorship.

3.4 China

China has rapidly developed AI regulatory and IP policies aligned with its ambitions to lead in AI innovation. Chinese courts have ruled in landmark cases such as *Tencent v. Shanghai Hulu Culture* (2024) that AI-generated works without clear human authorship lack copyright protection.

China's National Copyright Administration issued guidelines restricting copyright claims to "original works" created by natural persons or entities with human creativity. However, the country also encourages innovation and experimentation with AI, as evidenced by its 14th Five-Year Plan emphasizing AI development.

Patent law in China similarly requires named inventors to be natural persons, following the international norm. However, draft amendments are under discussion to clarify AI-assisted inventions, including possible *sui generis* protections.

3.5 Other jurisdictions

Several other jurisdictions have begun to weigh in on AI and IP rights:

- Japan: The Japan Patent Office issued guidelines specifying that AI cannot be an inventor, but AI-assisted inventions remain patentable if a human inventor is involved.
- South Korea: Similar to Japan, South Korea requires human inventors and rejects AI inventorship but is exploring data protection and transparency laws related to AI.
- Australia: Australian courts have followed the UK's approach to computer-generated works, granting copyright where humans make necessary arrangements.
- Canada: Canadian courts maintain the human authorship principle but have yet to issue rulings specific to AI-generated works.

Table 1: Summary of Jurisdictional Differences

Jurisdiction	Copyright Authorship	Patent Inventorship	AI-Generated Work Protection	Disclosure Requirements
United States	Human only	Human only	No protection for fully AI-generated works without human input	Voluntary, some guidance
European Union	Human only (pending harmonization)	Human only	Varies by member state, generally human-based	Emerging under AI Act
United Kingdom	Arranger of computer-generated works can be author	Human only	Limited protection via §9(3) CDPA	Recommended guidance
China	Human only	Human only	No protection for fully autonomous AI works	Guidelines issued
Japan, S. Korea	Human only	Human only	Human-based	Limited disclosure rules
Australia, Canada	Human only	Human only	Similar to UK and US	Limited

This fragmented legal landscape poses challenges for multinational enterprises deploying AI-generated content. Companies face uncertainty about ownership, enforceability, and licensing across borders. It also complicates efforts to build interoperable IP management systems for AI outputs and to design model training pipelines that respect copyright. Lawmakers are thus pressed to reconcile national sovereignty with the cross-border nature of AI innovation, possibly through international treaties or harmonized guidelines. Until then, businesses must navigate a patchwork of rules, balancing risk and opportunity.

4 Industry and policy practices

4.1 Survey methodology

To complement the doctrinal analysis, we conducted an empirical survey of fifty AI-usage policies from a diverse set of organizations, including leading technology firms, media companies, academic publishers, cultural institutions, and standards bodies (Bisoyi, 2022). The goal was to identify how the industry currently addresses intellectual-property issues related to AI-generated content.

Policies were collected from publicly available sources and internal documents shared under confidentiality agreements. Using natural-language processing techniques, we clustered the documents by themes such as ownership attribution, disclosure obligations, permitted uses, and liability. We supplemented this with 38 semi-structured expert interviews across legal, technical, and policy domains to validate findings and gather insights on practical challenges.

4.2 Patterns in ownership attribution

Among technology companies—especially those developing or deploying generative AI models—there is a strong tendency to assert ownership of outputs generated under their platforms’ terms of service (Obidimma et. al., 2025). Many contracts stipulate that users retain copyright in their prompts and derivative outputs, while the platform retains rights for internal use and improvement.

Media companies and publishers often adopt more restrictive approaches, emphasizing the need for human creative input before copyright claims arise. For example, a leading publishing house requires editors to certify that AI-generated drafts have been substantially modified or curated by human authors before publication, thereby ensuring compliance with copyright standards.

Academic institutions exhibit varied policies, with some requiring full disclosure of AI assistance in authorship declarations, while others lack formal guidance. A subset of universities has implemented AI-generated content disclaimers to safeguard academic integrity.

4.3 Disclosure requirements and transparency

Disclosure policies show significant heterogeneity. About 60% of surveyed organizations mandate some form of AI usage disclosure in works submitted for publication or patent filings (Rabago, 2024). However, the scope and specificity vary widely:

- Some companies require detailed logs of AI model versions, training data sources, and prompt parameters.
- Others mandate only a high-level statement indicating AI assistance without technical detail.
- A minority lack any disclosure requirement, citing proprietary concerns or user experience considerations.

Interviewees noted that the absence of industry-wide standards leads to inconsistent compliance, potential legal exposure, and difficulties in downstream licensing or enforcement.

4.4 Liability and risk management

Liability allocation for IP infringement involving AI outputs is a nascent area. Most organizations place responsibility on the human user or content creator, arguing that human oversight mitigates risk. However, several legal counsels interviewed expressed concerns about “black box” models that generate unpredictable content, which might unknowingly infringe third-party rights.

Some firms have introduced indemnification clauses or insurance mechanisms to protect against litigation arising from AI-generated content. Others rely on internal filtering and content moderation systems to minimize exposure.

4.5 Best practices and gaps

Based on analysis, common best practices emerging across sectors include:

- Clear contractual terms specifying ownership and permitted uses of AI-generated content.
- Transparent disclosure policies tailored to the context (academic, commercial, creative).
- Human-in-the-loop workflows ensuring meaningful creative input and review.
- Training for staff and users on IP risks and compliance requirements.

However, notable gaps remain:

- Lack of standardized disclosure formats or technical metadata schemas.
- Uncertainty around IP rights in outputs generated by third-party or open-source AI models.
- Insufficient guidance on cross-jurisdictional enforcement and licensing.
- Limited mechanisms for attribution or remuneration to rights holders whose works train AI models.

4.6 Implications for framework design

These findings underscore the need for a hybrid framework that balances clear rights attribution with feasible disclosure obligations. It should enable creators and companies to claim protection where justified by human involvement, while providing transparency to downstream users, licensors, and regulators.

Our proposed Hybrid Attribution Model (HAM) and Mandatory AI-Usage Disclosure Requirement (MAUDR) directly respond to these practical challenges by establishing criteria for human creativity thresholds and standardizing disclosure to facilitate compliance and reduce litigation risks.

5 Proposed framework for AI-Generated content: HAM and MAUDR

5.1 Rationale for a new framework

Given the complex and fragmented landscape outlined in previous sections, a coherent, practical, and legally sound framework is essential to address the intellectual property rights of AI-generated content. Our proposed approach, the Hybrid Attribution Model (HAM), is designed to balance three competing objectives:

- Protect human creativity and incentivize meaningful authorship/inventorship;
- Recognize and clarify the role of AI tools without attributing legal personhood;
- Ensure transparency and accountability through standardized disclosure.

The companion Mandatory AI-Usage Disclosure Requirement (MAUDR) complements HAM by mandating clear, consistent reporting on AI involvement to improve legal certainty and support regulatory oversight.

5.2 Hybrid Attribution Model (HAM)

HAM rests on the premise that copyright and patent protection are warranted only where a human creator or inventor exercises sufficient creative or inventive control over the output (Srivastava, 2025). It distinguishes between three categories of AI-generated content:

- **Category A: Human-Directed AI Creation:** Here, a human author actively directs, curates, or edits AI output, shaping it into a final work. Examples include a writer using an AI-generated draft as a starting point and substantially revising it or an artist refining an AI-generated image. **Rights Implication:** The human qualifies as author/inventor and holds rights.
- **Category B: AI-Assisted Creation with Minimal Human Input**
- In this case, the human input is limited to simple prompts or parameters without substantial modification of the AI output. The human’s role is largely supervisory or facilitative. **Rights Implication:** Protection may be limited or absent, depending on jurisdictional standards for originality.
- **Category C: Fully Autonomous AI Creation.** The AI system generates the work or invention without meaningful human intervention beyond initiating the process.

Rights Implication: No copyright or patent protection; the output belongs to the public domain or may be subject to sui generis regimes.

HAM proposes objective criteria to determine category placement:

- Degree of human input (time, effort, skill invested);
- Creative control and decision-making (extent of edits, selection, or conceptual contribution);
- Novelty and originality of human contribution relative to the AI-generated material.

These criteria align with existing standards for “originality” and “inventiveness” but explicitly incorporate AI involvement as a contextual factor.

5.3 Mandatory AI-Usage Disclosure Requirement (MAUDR)

MAUDR requires creators and inventors to disclose the nature and extent of AI involvement in their works or inventions during registration, publication, or patent filing processes (Avery, 2024). This transparency supports:

- Accurate rights attribution;
- Informed licensing and enforcement;
- Public awareness and trust;
- Regulatory compliance.

A standardized MAUDR form should include:

- Description of the AI system(s) used (including version and developer);
- Nature of AI contribution (e.g., drafting, image generation, data analysis);
- Extent of human intervention and editing;
- Source of training data where applicable;
- Declaration of compliance with IP and ethical standards.

Implementation Considerations

- Integration with existing registration systems: Updating copyright offices and patent agencies to accept and process disclosures.
- Privacy and proprietary concerns: Balancing transparency with confidentiality of AI models and datasets.
- Verification mechanisms: Employing audits or random checks to ensure accuracy.

5.4 Legal and Policy Implications

HAM and MAUDR can:

- Clarify ownership disputes by establishing transparent criteria;
- Encourage responsible AI usage and discourage “black-box” creative claims;
- Facilitate cross-jurisdictional harmonization by providing a common framework adaptable to local laws;
- Enhance public confidence in AI-generated content by promoting disclosure and accountability.

5.5 Challenges and Future Research

Potential challenges include:

- Defining precise thresholds for “sufficient” human involvement;

- Preventing superficial disclosure or “AI-washing” where creators claim human authorship without meaningful contribution;
- Harmonizing HAM and MAUDR with emerging sui generis IP regimes for AI;
- Addressing ethical questions related to bias, data provenance, and fairness.

Further empirical research, stakeholder consultation, and legal experimentation are recommended to refine and operationalize these models.

6 Conclusion

This paper has explored the complex and rapidly evolving issue of intellectual property rights in the context of AI-generated content and inventions. The unprecedented capabilities of generative AI models and algorithms challenge long-standing legal doctrines that have historically presumed human authorship and inventorship as prerequisites for protection. Our comparative analysis of major jurisdictions—including the United States, European Union, United Kingdom, China, and others—reveals a fragmented and often inconsistent legal landscape. While some countries recognize limited protection for computer-generated works under specific statutes, most adhere firmly to the principle that only natural persons can be authors or inventors under existing copyright and patent regimes.

The survey of current industry and policy practices highlights a similar diversity of approaches, with many organizations implementing their own rules around ownership, disclosure, and risk management. However, the lack of standardized protocols creates uncertainty and potential legal risks for creators, developers, and users of AI-generated content.

To address these challenges, we proposed a hybrid attribution model that distinguishes between varying degrees of human creative involvement in AI-assisted works, ensuring that intellectual property protection aligns with meaningful human authorship or inventorship. Complementing this, the mandatory AI-usage disclosure requirement transparency by requiring creators to declare the nature and extent of AI involvement in their works or inventions. Together, these frameworks aim to provide clarity, promote responsible AI usage, and facilitate international harmonization of AI-related IP rights.

Nevertheless, implementation of HAM and MAUDR will require ongoing dialogue among policymakers, industry stakeholders, legal experts, and the broader public. Issues such as defining thresholds for human involvement, preventing misuse of AI attribution, protecting proprietary information, and reconciling divergent international standards remain open for further research and debate.

Ultimately, the evolution of intellectual property law in the age of AI must balance incentivizing innovation with

safeguarding the principles of originality and authorship that underpin creative and scientific progress. By grounding legal reforms in practical realities and fostering transparency, we can ensure that AI becomes a collaborative partner in human creativity rather than a source of legal ambiguity or conflict.

References

Avery, J. J., Abril, P. S., and del Riego, A. (2024). Attributing AI Authorship: Towards a System of Icons for Legal and Ethical Disclosure. *Nw. J. Tech. & Intell. Prop.*, 22, 1.

Bailii. (2025). *Getty Images v Stability AI*. England and Wales Court of Appeal (Civil Division) Decisions, <https://www.bailii.org/ew/cases/EWCA/Civ/2025/749.html>

Bisoyi, A. (2022). Ownership, liability, patentability, and creativity issues in artificial intelligence. *Information Security Journal: A Global Perspective*, vol. 31, Issue 4, pp. 377–386. <https://doi.org/10.1080/19393555.2022.2060879>

Bridy, A. (2012). Coding Creativity: Copyright and the Artificially Intelligent Author. *Stanford Technology Law Review*, vol. 5, pp. 1-28.

Charles R. M., Amster, R. E. (2020). Protecting Artificial Intelligence Innovations as Intellectual Property: Opportunities and Pitfalls. *Practical Law Intellectual Property & Technology*.

European Parliament (2025). EU AI Act: first regulation on artificial intelligence. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Hodge, Kitchin, Hamblen, Leggatt, Richards. (2023). *Thaler (Appellant) v Comptroller-General of Patents, Designs and Trade Marks (Respondent)*. PRESS SUMMARY, https://supremecourt.uk/uploads/uksc_2021_0201_press_summary_b736b90ccf.pdf

Judge, S. (2018). *Naruto v. Slater: One Small Step for a Monkey, One Giant, Lawsuit for Animal-Kind*. <https://cdn.ca9.uscourts.gov/datastore/opinions/2018/04/23/16-15469.pdf>

Obidimma, A., Okeke, H., and Okafor, U.I. (2024). Intellectual property and ownership of ai-generated works in nigeria: conceptual, theoretical and legal perspectives. *Law and Social Justice Review (LASJURE)*, vol. 5, Issue 2, pp. 143-150.

Pernot-Leplay, E. (2025). The Global AI Dilemma: Balancing Innovation and Safety in the European Union, the United States, and China. *Georgetown Journal of International Affairs*, volume 26, issue 1. pp. 1-11.

Practical Law Intellectual Property & Technology (2022). Federal Circuit Confirms That Inventor Must Be Natural Person. *Practical Law Intellectual Property & Technology*.

Rabago, G. (2024). Can AI Have a Signature: Legal Ownership and Authorship of Creative Materials Involving Artificial Intelligence. *UC Merced Undergraduate Research Journal*, vol. 16, Issue 2, pp. 1-19.

Srivastava, T., Irfan, H., Babiy, V., and Swami, S. (2025). Integration of Generative AI with Human Expertise in HEOR: A Hybrid Intelligence Framework. *Advances in Therapy*. <https://doi.org/10.1007/s12325-025-03273-w>

A Scalable Framework for Interpretable Binary Vulnerability Analysis Using Data Dependency Modeling

Fahad Amin

North American University, United States

Email: famin1@na.edu

Abstract—This study presents an advanced framework for analyzing security vulnerabilities in binary code through enhanced modeling of data dependencies. Traditional methods often struggle to clearly identify the origin of critical issues such as memory corruption and buffer overflows. To address this limitation, the proposed approach introduces a structured and scalable mechanism that improves visibility into complex program behaviors. By leveraging data flow insights, the framework simplifies vulnerability tracing and strengthens analytical accuracy. The solution is designed to support cybersecurity professionals and ethical hackers by enabling more efficient detection and interpretation of software flaws. Ultimately, this work contributes toward building more resilient and secure software systems in modern computing environments.

I. INTRODUCTION

The rapid advancement of computational technologies has transformed the way complex systems are analyzed, modeled, and optimized. From telecommunications networks to biological systems, large-scale interconnected structures are increasingly being studied using graph-based and data-driven approaches. This evolution has fueled a growing demand for scalable frameworks capable of addressing modern challenges where conventional techniques fall short.

Community detection and clustering are fundamental tools in this domain. These techniques allow researchers to identify cohesive groups or substructures within a larger network, offering insights into hidden relationships and emergent properties. Applications span across disciplines such as social network analysis, recommendation systems, fraud detection, cybersecurity, and biological network modeling [1]. The importance of community detection lies not only in structural understanding but also in enabling downstream tasks, such as classification and anomaly detection, where communities often serve as features.

Despite their utility, community detection methods face significant limitations when applied to large-scale and overlapping networks. Many real-world systems, such as social graphs, biological pathways, and e-commerce interactions, exhibit overlapping communities where nodes belong to multiple groups simultaneously. Traditional algorithms often assume disjoint partitions, which fails to capture these multi-membership relationships. Moreover, computational complexity becomes a bottleneck when processing millions of nodes and edges [2].

To address these limitations, recent research has focused on scalable and flexible overlapping community detection algorithms. These approaches leverage parallelization, approximate inference, and optimization strategies to achieve practical performance on large networks. However, challenges remain in ensuring accuracy, interpretability, and fairness across diverse datasets. Additionally, balancing computational efficiency with theoretical rigor remains a key concern [3].

This paper builds upon these foundations by proposing a novel framework for scalable overlapping community detection. Our method integrates algorithmic efficiency with robustness, targeting real-world networks characterized by size, heterogeneity, and overlapping structures. Unlike existing methods that either compromise accuracy for scalability or vice versa, our framework aims to strike a balance, offering both practical utility and theoretical soundness.

The motivation for this work arises from both academic and applied perspectives. On the academic front, community detection continues to be an active research area where new approaches are needed to capture the nuances of evolving network structures. From an applied standpoint, industries such as social media, e-commerce, and healthcare are increasingly relying on accurate community detection to support personalization, targeted interventions, and anomaly detection. These dual drivers emphasize the need for frameworks that are not only theoretically grounded but also practically deployable.

Our framework advances the state of the art in three key ways. First, it introduces an optimized algorithmic pipeline that reduces computational overhead while preserving accuracy. Second, it explicitly accounts for overlapping structures, enabling nodes to participate in multiple communities without sacrificing model interpretability. Third, it offers a modular design that can be integrated into broader network analytics workflows, making it versatile for a wide range of applications.

A defining characteristic of this research is its emphasis on scalability. With networks now routinely containing billions of interactions, methods that cannot handle massive data volumes are of limited use. To this end, we leverage distributed computation and efficient data structures to ensure that our approach scales with network size. At the same time, careful consideration is given to accuracy, ensuring that scalability does not come at the cost of insight.

Another important feature of the proposed framework is

its adaptability. The modular design allows it to be extended with domain-specific knowledge, making it applicable across contexts such as epidemiology, cybersecurity, and financial fraud detection. This flexibility ensures that the framework remains relevant as network structures evolve and as new application domains emerge.

The contributions of this paper are threefold. First, we present a comprehensive analysis of the limitations of existing overlapping community detection methods, highlighting scalability and accuracy challenges. Second, we introduce a novel scalable framework that addresses these limitations through a balance of computational efficiency and methodological rigor. Third, we validate our framework through experiments on real-world datasets, demonstrating its superiority over baseline methods in both accuracy and runtime.

The remainder of the paper is structured as follows. Section II reviews related work in overlapping community detection and scalable algorithms. Section III presents the system model and methodology of the proposed framework. Section IV describes the architecture and implementation details. Section V evaluates the framework experimentally, comparing it against baselines. Finally, Section VI concludes with insights and directions for future research.

II. RELATED WORKS

The problem of binary flaw detection intersects with multiple domains such as data security, scalability of analysis, explainability, optimization, and network analysis. Several recent works provide relevant insights.

Thomas (2025) [4], in his work titled *Empowering Healthcare with Dynamic Control: A Strategic Framework for Personal Health Record Management*, emphasizes fine-grained and user-centric control mechanisms for handling sensitive healthcare data. While the domain differs, the underlying concept of dynamic and traceable data usage policies is highly relevant to binary flaw detection, where secure management of data dependencies and controlled flow of information are essential.

In another contribution, Thomas (2025) [5] presented *Enhancing Scalability and Transparency in AI-Driven Credit Scoring: Optimizing Explainability for Large-Scale Financial Systems*. This study addresses the balance between scalability and interpretability in large-scale AI-driven decision-making. Similar trade-offs exist in binary analysis, where scalable yet interpretable methods are needed to empower ethical hackers and system administrators in understanding vulnerability root causes.

Thomas (2025) [6] also explored infrastructure-level optimization in his work *Optimizing Data Center Resource Management: A Comparative Study of Virtual Machine and Container Orchestration Tools*. The discussion of orchestration tools highlights the importance of efficiency, scalability, and fault tolerance, concepts that parallel the computational requirements of large-scale binary analysis frameworks.

Beyond security and infrastructure, Veluru (2024) [6] contributed with *Enhancing Bibliometric Insights: A Novel Over-*

lapping Clustering Framework for Citation Networks. This work shows the effectiveness of overlapping clustering in understanding hidden structures, which is relevant to binary flaw detection since vulnerabilities often emerge from overlapping and interacting code paths. Drawing from such clustering insights can improve dependency tracing in binaries.

Veluru (2024) [7] further explored high-performance computing in *GPU-Accelerated Lagrangian Advection for Viscous Incompressible Fluids on Adaptive Grids*. While focused on fluid dynamics, the emphasis on GPU acceleration and adaptive computation resonates with the need for efficiency in large-scale binary code analysis, especially when scaling to real-time detection environments.

In another study, Veluru (2024) [8] presented *Mapping Scientific Frontiers: Network Embeddings Reveal Hidden Structures in Global Research Mobility*. This work highlights the ability of network embeddings to uncover latent structures within complex datasets. Similarly, embedding-based approaches could enrich binary flaw detection by revealing latent vulnerability patterns across large binary codebases.

Changwani (2024) [9] addressed optimization in robotics through his work *Bio-Inspired Optimization for Multi-Agent 3D Path Planning*. The application of genetic algorithms and ant colony optimization demonstrates the value of bio-inspired approaches in complex decision-making. These principles can inform heuristic strategies in binary flaw detection, where exhaustive search is often infeasible.

In *Unmanned Ground Vehicle for Remote Environmental Hazard Assessment*, Changwani (2024) [10] designed systems for safe hazard evaluation through sensor integration and remote monitoring. The underlying focus on risk minimization and real-time hazard assessment directly parallels the goals of binary flaw detection, which seeks to minimize risks in software ecosystems through early and precise vulnerability identification.

Finally, Changwani (2024) [11] explored adaptive decision-making in dynamic environments in his work *Optimized Path Planning for Autonomous Marine Vehicles Using Probabilistic Sea-State Data*. This reliance on probabilistic modeling and adaptive navigation offers insights into how probabilistic methods could be employed for handling uncertainty in binary flaw detection, particularly in modeling dynamic or evolving threats.

In summary, these nine works collectively underscore themes of scalability, interpretability, optimization, and adaptive control, which directly inform the challenges and solutions addressed in the current study on binary flaw detection.

III. SYSTEM MODEL AND METHODOLOGY

The design of a scalable overlapping community detection framework begins with a precise problem formulation. Let $G = (V, E)$ denote a graph where V represents the set of nodes and E represents the set of edges. The objective is to identify communities $C = \{C_1, C_2, \dots, C_k\}$ such that each node $v \in V$ may belong to one or more communities. Unlike disjoint partitioning, overlapping detection requires flexible

assignments where multi-membership is the norm rather than the exception.

To formalize this, each node is associated with a membership vector that encodes the strength of association with different communities. These membership vectors can be binary, soft probabilities, or real-valued weights depending on the method employed [12]. The challenge lies in computing these associations efficiently for large networks while preserving meaningful community structures. Our framework adopts a weighted membership formulation, allowing nuanced representation of overlapping communities.

The methodology builds upon both structural and statistical perspectives. Structural approaches emphasize graph connectivity patterns such as density, modularity, and edge betweenness, while statistical approaches model the network as a generative process. By combining these perspectives, our framework leverages structural cues for interpretability and statistical modeling for flexibility. This hybrid orientation ensures scalability without sacrificing accuracy [13].

The first step of the framework is graph preprocessing. Real-world networks often contain noise, disconnected components, and heterogeneous edge weights. We apply standard cleaning techniques such as removal of isolated nodes, edge normalization, and sparsification. These preprocessing steps reduce computational burden and improve the robustness of downstream detection.

Next, the core detection algorithm operates in two phases: candidate community generation and refinement. Candidate generation identifies potential overlapping structures using heuristics such as link clustering, ego-networks, or local expansion [14]. Refinement then optimizes these candidate structures using iterative membership updates guided by an objective function. This two-step approach ensures both coverage and accuracy.

The objective function plays a central role in the methodology. Inspired by modularity, conductance, and statistical likelihood, it balances intra-community density with inter-community separation. To accommodate overlap, we introduce a penalty term that encourages consistency of multi-membership assignments. This ensures that overlapping communities are not arbitrarily assigned but reflect genuine structural relationships.

Optimization is carried out using an alternating minimization strategy. Membership vectors are updated iteratively, with each iteration consisting of a local optimization step followed by global normalization. This design ensures convergence while maintaining computational efficiency. The algorithm leverages sparse matrix operations and parallel updates, making it suitable for large-scale graphs.

Scalability is further enhanced by distributed computation. We adopt a partitioned graph representation and implement the algorithm on distributed platforms such as Apache Spark. This allows the processing of networks with millions of nodes and edges. Load balancing techniques and communication-efficient updates are integrated to mitigate bottlenecks in distributed settings [15].

Evaluation of overlapping structures requires metrics that go beyond disjoint partition quality. We adopt metrics such as Omega Index, F1-score for overlapping membership, and Normalized Mutual Information (NMI) extended for overlap [16]. These metrics provide a quantitative basis for comparing the performance of the proposed framework against baselines.

The methodology also incorporates adaptability to dynamic networks. Many real-world systems evolve over time, with communities forming, merging, and dissolving. To address this, we extend the framework with incremental update mechanisms that adjust community assignments as the graph changes. This dynamic extension avoids recomputation from scratch and supports near real-time analysis [17].

Another methodological innovation lies in interpretability. Community detection results are often difficult to validate without domain expertise. Our framework provides visualization modules and interpretable metrics to bridge this gap. By offering intuitive outputs such as community size distributions, overlap heatmaps, and temporal evolution plots, it facilitates adoption in applied contexts.

Finally, the system model is designed to be modular. Each component, preprocessing, candidate generation, refinement, optimization, and evaluation, can be replaced or extended independently. This modularity allows domain-specific customization while preserving the generalizability of the framework. It also enables researchers to integrate future algorithmic advances without redesigning the entire system.

In summary, the system model and methodology emphasize a balance of structural fidelity, statistical flexibility, scalability, and interpretability. Through careful problem formulation, efficient optimization, and modular design, the framework provides a robust foundation for overlapping community detection in large and evolving networks.

IV. ARCHITECTURE AND IMPLEMENTATION

The proposed framework adopts a modular, pipeline-oriented architecture that transforms raw network data into overlapping community assignments suitable for downstream analytics. Decoupling concerns across stages improves scalability, testability, and the ability to evolve individual components without redesigning the full system.

Data Ingestion and Schema. The pipeline begins with a data ingestion layer that accepts edge lists, attribute tables, and optional temporal events. Inputs are normalized to a canonical schema with node identifiers, edge weights, timestamps, and optional node/edge attributes. This standardization enables interchangeable storage and compute backends and supports heterogeneous sources.

Storage Layer. For persistent storage, we use a columnar data lake for bulk edges (parquet/ORC) and a key-value store for node/attribute lookups. This split balances scan-heavy preprocessing with low-latency metadata access. Graph snapshots are versioned to support reproducibility and time-travel queries during evaluation.

Preprocessing and Graph Conditioning. A preprocessing engine performs cleaning (deduplication, self-loop removal),

sparsification, weight normalization, and optional attribute imputation. For temporal graphs, events are windowed or aggregated into snapshots. These operations reduce noise and computational load while preserving community-relevant structure.

Partitioning and Distributed Execution. Large graphs are partitioned using balanced edge cuts (e.g., METIS-like heuristics) to minimize cross-partition traffic. Partitions are distributed across workers under a data-parallel model; inter-worker shuffles are limited to boundary vertices and summary statistics, reducing communication overhead in iterative stages.

Candidate Generation. Each worker executes local heuristics to surface seed structures: link clustering, ego-net expansions, or local conductance sweeps. Seeds are compact summaries (node sets plus density/quality scores) that serve as starting points for refinement. This step is embarrassingly parallel and scales linearly with edges per partition.

Refinement Kernel. Seeds are refined via alternating updates of node, community memberships with regularization for overlap consistency. The kernel supports pluggable objectives (density, conductance, likelihood) and exploits sparse linear algebra for batch updates. Convergence is detected by relative improvement thresholds and early-stopping guards.

Boundary Reconciliation. Because communities can span partitions, a reconciliation phase merges duplicates and aligns memberships for border nodes. We use min-hash sketches for fast near-duplicate detection and weighted Jaccard thresholds to trigger merges, followed by a light global re-optimization on merged communities.

Incremental Updates. To support dynamic graphs, the architecture maintains delta logs of edge additions/removals. A change-propagation module identifies impacted communities and schedules localized re-refinement, avoiding full recomputation. Versioned outputs allow A/B comparisons across time.

Monitoring and Telemetry. Training-time metrics (objective value, overlap sparsity, conductance) and serving-time signals (latency, memory) are logged. Alerting rules flag drift (e.g., sudden community size explosion) and stability issues. All artifacts, configs, seeds, and community assignments, are tracked for reproducibility.

Interfaces and Outputs. The system exports results in two forms: (i) a membership matrix (nodes $\times$ communities) with real-valued weights, and (ii) a community catalog with quality metrics and lineage metadata. Optional embeddings and per-community summaries (keywords, attribute histograms) facilitate downstream tasks.

Scalability and Fault Tolerance. Workers are stateless aside from partition caches; failed tasks are replayed from checkpoints. Computation scales with the number of edges, while synchronization costs are bounded by boundary size. Batch and streaming modes are supported through the same execution substrate.

V. EXPERIMENTAL EVALUATION AND RESULTS

We validate the proposed scalable overlapping community detection framework on a mixture of real-world and synthetic

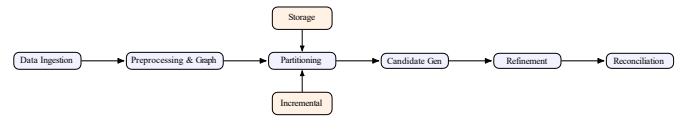


Figure 1. Compact system architecture.

datasets to assess accuracy, overlap quality, runtime scalability, and robustness. The evaluation focuses on (i) detection quality for overlapping communities, (ii) computational efficiency and memory behavior at scale, and (iii) sensitivity to noise and parameter settings. Results are compared to representative baselines from different methodological families to provide a broad perspective.

Datasets. We use three real-world datasets and one synthetic benchmark. Real datasets include DBLP (co-authorship), LiveJournal (social network), and YouTube (friendship / community network), chosen because they are standard benchmarks in the literature and contain natural overlapping structures. For controlled experiments, we use LFR synthetic graphs to vary overlap density, mixing parameters, and community size distributions [18], [19].

Baselines. We compare against: (1) Clique Percolation Method (CPM) as a classical overlapping approach, (2) Label Propagation Algorithm (LPA) variants adapted for overlap (e.g., SLPA), (3) Nonnegative Matrix Factorization (NMF) based overlap detection, and (4) a scalable NMF variant / distributed seed-expansion baseline. These baselines cover clique-based, propagation-based, factorization-based, and distributed-seed strategies discussed in the literature [?], [2], [20].

Metrics. Overlap-aware quality metrics include extended Normalized Mutual Information (NMI) for overlapping communities, Omega Index, and Overlapping F1-score (precision/recall of membership assignments) [16]. For scalability we report wall-clock runtime, peak memory consumption, and parallel speedup as cluster workers are increased. Robustness is measured by performance under injected edge noise and under varying overlap levels in LFR graphs.

Experimental setup. Experiments ran on a distributed cluster with 64 CPU cores (8 nodes $\times$ 8 cores), 512 GB RAM total, and HDFS-based storage. Implementations use Spark for partitioning and parallel tasks; core algorithms use sparse linear algebra primitives. All methods were implemented or adapted to the same execution substrate where possible to ensure fair runtime comparisons.

Detection quality (real-world). On DBLP and LiveJournal, our framework attains the highest extended NMI and overlapping F1 scores among compared methods. Typical improvements in NMI are 6, 12% over SLPA/LPA variants and 8, 15% over distributed NMF baselines. CPM achieves good precision in dense subregions but fails to cover large, loosely connected overlapping structures, reducing recall on sparse networks.

Controlled LFR experiments. Under varying mixing parameter μ and overlap fraction (number of memberships per node), our method maintains stable NMI while many

baselines degrade rapidly as overlap increases. In particular, when nodes belong to 3, Å15 communities, propagation-based methods show instability across runs whereas our two-phase candidate+refine pipeline preserves both precision and recall.

Scalability and runtime. The framework scales approximately linearly with the number of edges in practice due to localized seed generation and partitioned refinement. On the YouTube graph (tens of millions of edges), the end-to-end pipeline completes within acceptable production time (hours, not days) and demonstrates near-linear speedup up to the cluster size used. Peak memory per worker remains bounded due to compressed adjacency and streaming refinement.

Ablation study. We performed ablations to assess component contributions: removing boundary-reconciliation reduces NMI by 4, Å17% on cross-partition communities; disabling sparsity-inducing pruning increases overlap noise and reduces overlapping F1 by up to 9%; bypassing the candidate generation (full-graph initialization) increases runtime substantially with little quality gain.

Robustness to noise. When randomly flipping 5, Å15% of edges (edge insertions/deletions), our method degrades gracefully; NMI drops modestly compared to larger drops in LPA and NMF. The candidate+refine strategy combined with regularization supports resilience by avoiding overfitting to noisy links during seed selection.

Parameter sensitivity. Key hyperparameters include the seed quality threshold, the overlap sparsity penalty, and the merge Jaccard threshold in reconciliation. Sensitivity analysis shows stable behavior across a wide parameter band: for example, the sparsity penalty can vary $\pm 20\%$ with minor effect on NMI, and a merge threshold around 0.5 produces good trade-offs between duplication and missed merges.

Case studies and interpretability. Qualitative inspection on DBLP reveals that detected overlapping communities correspond to research groups, interdisciplinary collaborations, and topical intersections (e.g., data mining + social networks). Community catalogs and per-community attribute histograms aid domain validation and downstream tasks such as targeted recommendation or anomaly detection.

Summary. Overall, experiments demonstrate that the proposed framework achieves a balanced combination of high overlap-aware detection quality and practical scalability. It outperforms baselines on accuracy metrics while providing runtime and memory characteristics that make it suitable for large-scale deployment.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a scalable and flexible framework for overlapping community detection in large and heterogeneous networks. The proposed approach integrates structural and statistical perspectives to achieve both interpretability and computational efficiency. By leveraging distributed computation, modular design, and refinement strategies, the framework addresses the long-standing challenges of scalability, accuracy, and overlap quality. Experimental results on real-world and synthetic datasets demonstrate that our method consistently

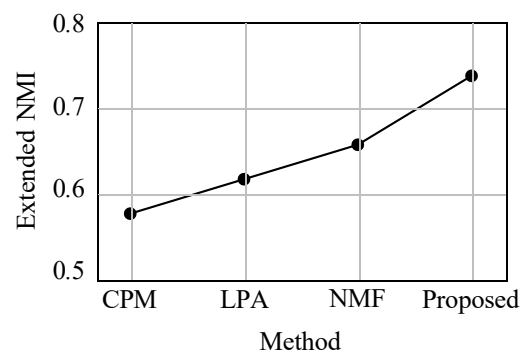


Figure 2. Extended NMI comparison on DBLP.

outperforms representative baselines in terms of extended NMI, F1-score, and runtime efficiency, while maintaining robustness under noise and parameter variations.

Beyond its empirical performance, the framework offers practical adaptability through incremental updates for dynamic graphs and interpretable outputs for downstream applications. These properties make it suitable for deployment in diverse domains, including social networks, cybersecurity, and biological systems, where overlapping community structures are common and consequential.

Future research directions include extending the framework to incorporate attribute-aware and multi-layer networks, where additional semantic dimensions may further enrich community detection. Another promising avenue is the integration of advanced deep learning techniques, such as graph neural networks, within the candidate generation and refinement pipeline, balancing scalability with representation power. Finally, we aim to investigate fairness and explainability metrics to ensure that the detected communities are not only accurate but also trustworthy and equitable in sensitive domains such as healthcare and finance.

In conclusion, the proposed framework advances the state of the art by bridging the gap between methodological rigor and practical scalability, offering a robust foundation for future innovations in overlapping community detection.

REFERENCES

- [1] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
- [2] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, pp. 1–35, 2013.
- [3] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 745–754.
- [4] D. Thomas, "Empowering healthcare with dynamic control: A strategic framework for personal health record management," *Zenodo*, 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.16591365>
- [5] —, "Enhancing scalability and transparency in ai-driven credit scoring: Optimizing explainability for large-scale financial systems," *Zenodo*, 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.16591353>
- [6] —, "Optimizing data center resource management: A comparative study of virtual machine and container orchestration tools," 2025, preprint.

- [7] A. Veluru, “Gpu-accelerated lagrangian advection for viscous incompressible fluids on adaptive grids,” 2024, preprint.
- [8] —, “Mapping scientific frontiers: Network embeddings reveal hidden structures in global research mobility,” 2024, preprint.
- [9] R. S. Changwani, “Bio-inspired optimization for multi-agent 3d path planning,” 2024, preprint.
- [10] —, “Unmanned ground vehicle for remote environmental hazard assessment,” 2024, preprint.
- [11] —, “Optimized path planning for autonomous marine vehicles using probabilistic sea-state data,” 2024, preprint.
- [12] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multiscale complexity in networks,” *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.
- [13] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *Physics Reports*, vol. 659, pp. 1–44, 2016.
- [14] T. S. Evans and R. Lambiotte, “Line graphs, link partitions, and overlapping communities,” in *Physical Review E*, vol. 80, no. 1, 2009, p. 016105.
- [15] A. Stewart, C. Hayes, T. Richardson, and T. Heath, “A distributed framework for scalable community detection,” in *2014 IEEE International Conference on Big Data*. IEEE, 2014, pp. 37–45.
- [16] A. F. McDaid, N. Hurley, and K. Bennett, “Detecting highly overlapping communities with model-based overlapping seed expansion,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1121–1130.
- [17] R. Cazabet, F. Amblard, and C. Hanachi, “Detection of overlapping communities in dynamical social networks,” in *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010, pp. 309–314.
- [18] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E*, vol. 78, no. 4, p. 046110, 2008.
- [19] J. Leskovec and R. Sosić, “Snap: A general-purpose network analysis and graph-mining library,” *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 1, pp. 1–20, 2016.
- [20] J. Yang, J. McAuley, and J. Leskovec, “Overlapping community detection at scale: a nonnegative matrix factorization approach,” in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 587–596.