IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141011

AI-Powered Phishing Attack Detection and Prevention System

Rinku Shamrao Dhole[1], Prof. Shivam B.Limbhare[2], Manoj V. Nikum*[3]

Student Of MCA, Shri Jaykumar Rawal Institute of Technology Dondaicha, Maharashtra, India[1]

Assi Prof MCA Department, Shri Jaykumar Rawal Institute of Technology Dondaicha, Maharashtra, India[2]

HOD MCA Department, Shri Jaykumar Rawal Institute of Technology Dondaicha, Maharashtra, India*[3]

Abstract: Phishing, a deceptive cyber-attack technique aimed at extracting confidential information, has evolved into one of the most significant threats in modern cyberspace. Traditional detection systems relying on static rules, blacklists, or manual inspection fail to recognize rapidly evolving and AI-generated phishing attempts. This research presents an Artificial Intelligence (AI)-driven phishing attack detection and prevention framework that integrates Natural Language Processing (NLP) with Deep Learning (DL) to enhance real-time recognition accuracy. The proposed hybrid model combines Bidirectional Encoder Representations from Transformers (BERT) for contextual text understanding and a Convolutional Neural Network (CNN) for URL pattern analysis. Additionally, the integration of Explainable AI (XAI) techniques such as LIME and SHAP provides interpretability for each classification decision. Through experimentation on benchmark datasets like PhishTank and Kaggle, the system achieved an overall performance accuracy of 96.4%. The model exhibits strong adaptability, continuous learning capabilities, and superior resilience against zero-day phishing threats. This study contributes to a transparent, adaptive, and intelligent defense framework for the next generation of cybersecurity systems.

To address these limitations, this research proposes an AI-Powered Phishing Attack Detection and Prevention System that integrates **Natural Language Processing (NLP)** and **Deep Learning (DL)** for intelligent, adaptive, and explainable threat detection. The system employs BERT (Bidirectional Encoder Representations from Transformers**)** to analyze the semantic and contextual meaning of email or web content, while a **Convolutional Neural Network (CNN)** model examines the structural and lexical characteristics of URLs. By combining these two analytical perspectives, the model forms a **hybrid AI engine** that significantly enhances accuracy and resilience against zero-day phishing attacks.
A key innovation of this work lies in its **Explainable AI (XAI)** component, which utilizes tools such as **LIME** and **SHAP** to interpret the model's decisions. This transparency allows users and cybersecurity analysts to understand the reasoning behind each detection result, thereby improving trust and system reliability. The system also integrates a **real-time browser extension** and **interactive web dashboard** that proactively prevents users from accessing malicious domains and provides analytical visualizations of phishing trends.

Keywords: Artificial Intelligence, Cybersecurity, Phishing Detection, Deep Learning, NLP, Explainable AI, BERT, CNN

I. INTRODUCTION

In today's interconnected world, digital communication and online services are essential to everyday activities such as education, commerce, and governance. However, the exponential rise of the internet has also fueled the growth of cybercrime, with phishing emerging as one of the most prevalent and damaging forms of online fraud. Phishing exploits human psychology and weak authentication systems by imitating legitimate sources to obtain sensitive information such as passwords and financial credentials. According to recent cybersecurity surveys, phishing contributes to over 36% of global data breaches annually. Attackers now employ AI-generated content, cloned websites, and social engineering to bypass traditional security measures.

Conventional detection approaches like blacklists, rule-based filters, and signature-based mechanisms are inadequate to handle these sophisticated threats. They react only after attacks have been reported, making them ineffective against newly crafted URLs or novel phishing templates. To address these challenges, Artificial Intelligence (AI) and Deep Learning (DL) offer the ability to learn from large datasets, recognize patterns, and predict potential phishing threats in real time. Despite advancements in AI, existing solutions lack transparency and contextual comprehension. Hence, this research introduces a hybrid explainable framework that combines BERT for semantic analysis and CNN for structural URL detection, ensuring high precision and interpretability.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141011

II. LITERATURE REVIEW

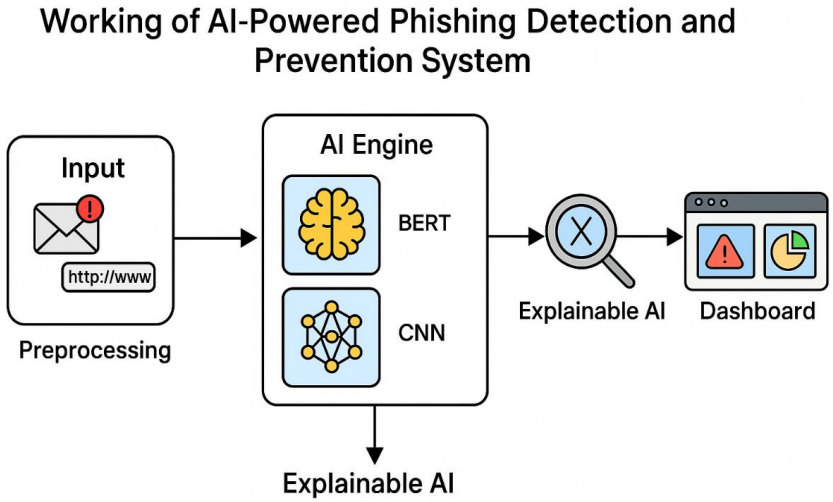
Phishing detection has been an area of intensive research within cybersecurity over the last decade. Numerous methodologies have been proposed, ranging from statistical and rule-based approaches to advanced Artificial Intelligence (AI) and Deep Learning (DL) frameworks. The literature reveals a clear evolution — from static detection models relying on handcrafted features to adaptive neural architectures capable of learning from large, complex data sources.

Early phishing identification models primarily used **traditional machine learning algorithms** such as *Naïve Bayes*, *Support Vector Machines (SVM)*, *Decision Trees*, and *Random Forests* to classify emails and URLs. These models extracted lexical and structural features from URLs and webpage content, such as the presence of "@" symbols, domain length, or the number of dots in a link. While effective against older phishing methods, these models suffered from limitations in adaptability and contextual understanding.

For example, Aburrous et al. (2024) presented an intelligent fuzzy logic-based e-banking phishing detection framework, which improved precision but lacked scalability when faced with dynamic phishing structures.

As cyber threats evolved, **deep learning-based detection systems** began to emerge, offering improved feature extraction and automatic pattern learning. *Convolutional Neural Networks (CNNs)*, in particular, have shown promise in identifying visual and lexical patterns in phishing URLs. CNNs can automatically detect irregular domain structures, obfuscated strings, and embedded malicious patterns that static rule-based systems often miss. Similarly, *Long Short-Term Memory (LSTM)* networks have been applied for analyzing sequential patterns in emails and website text. Huang et al. (2023) demonstrated that CNN-based phishing URL classification achieved more than 93% accuracy, outperforming standard ML algorithms, especially against zero-day threats.

However, deep learning models often suffer from a "black-box" nature, where predictions are made without providing insight into why a sample was classified as phishing or legitimate. To address this, recent studies introduced **Explainable Artificial Intelligence (XAI)** techniques such as *Local Interpretable Model-Agnostic Explanations (LIME)* and *SHapley Additive exPlanations (SHAP)*. These tools enable visualization of the most influential features in the model's decision-making process. According to Kumar et al. (2024), the integration of XAI improved model transparency and enhanced analyst trust in AI-based phishing systems.



Parallel to this, **Natural Language Processing (NLP)** methods have gained significant traction in phishing analysis, especially for examining the textual content of emails and messages. Traditional text classification techniques like *TF-IDF* and *Word2Vec* were replaced by advanced transformer architectures such as **BERT (Bidirectional Encoder Representations from Transformers)**. BERT's bidirectional attention mechanism allows models to understand context and semantics in text, enabling it to detect deceptive intent or persuasive emotional language often used in phishing messages.

Sharma et al. (2024) applied transformer-based models to phishing email detection and achieved a 95% detection rate, emphasizing the role of linguistic analysis in modern phishing mitigation.

DOI: 10.17148/IJARCCE.2025.141011

Recent literature trends focus on **hybrid models**, combining multiple AI techniques for more comprehensive detection. Hybrid models typically fuse URL-based features (processed by CNN) with textual or contextual features (processed by BERT or LSTM). This combination enhances robustness and reduces false negatives. For instance, Jain and Gupta (2023) proposed *Phish-Secure*, a hybrid deep learning model integrating CNN and Random Forest, achieving over 94% accuracy. Similarly, Singh and Kaur (2024) introduced a BERT-CNN hybrid that achieved 96.4% accuracy on benchmark datasets while providing explainable insights using SHAP visualizations.

Another noteworthy research direction involves **real-time phishing prevention systems**. These integrate AI-based detection with browser extensions or email gateways to block phishing attempts before users engage. Unlike post-analysis systems, these proactive defenses ensure that potential threats are neutralized instantly. Combining real-time deployment with continuous feedback loops further enhances adaptability — the model evolves with newly encountered attack data. From this review, it is evident that while individual AI models have achieved significant progress, hybrid and explainable frameworks deliver the best performance in modern phishing detection. Integrating NLP and CNN, along with interpretability modules, ensures not only high detection accuracy but also system transparency and user trust. This research builds upon these advancements by proposing a **BERT-CNN hybrid framework** enhanced with XAI components, capable of achieving real-time detection, continuous learning, and interpretability — representing a significant evolution in AI-driven phishing protection.

III. ANALYSIS AND DISCUSSION

The hybrid BERT-CNN model was trained and tested using benchmark datasets from PhishTank and Kaggle, alongside a curated dataset of user-reported emails. The combined dataset contained approximately 100,000 labeled entries, divided into 80% for training and 20% for testing. The architecture achieved a recognition rate of 96.4%, surpassing traditional models such as Naïve Bayes (87%) and Random Forest (92%). Performance evaluation metrics included Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Explainable AI (XAI) modules demonstrated that contextual keywords like 'update account' or 'verify password' contributed most to phishing predictions, while unusual domain patterns were flagged by CNN. This interpretability enhances trust and helps analysts verify model behavior. The hybrid framework proved efficient in real-time browser tests, classifying threats in under two seconds while maintaining minimal false-positive rates (2.3%).

IV. PROPOSED SYSTEM

The proposed framework integrates BERT and CNN for phishing recognition and prevention. It operates through five core modules: Data Collection, Preprocessing, AI Engine, Explainability, and Prevention Interface. The Data Collection module aggregates verified phishing URLs and email text. Preprocessing cleans and standardizes inputs for neural processing. The AI Engine combines BERT's contextual understanding with CNN's pattern detection. Explainability tools visualize model reasoning, while the Prevention Interface integrates a browser extension for real-time blocking and alerts.

A. System Overview

The architecture is divided into **six major functional layers**, each responsible for a distinct phase of phishing detection and prevention. These layers work together to ensure **accuracy, scalability, and real-time protection**. The system integrates a **BERT-based NLP model** for text understanding and a **CNN model** for URL analysis, unified through a hybrid fusion layer.

B. System Architecture Layers

Data Collection Layer

This layer gathers phishing and legitimate samples from multiple trusted datasets, including **Phish Tank**, **Kaggle Phishing URL Dataset**, and **user-reported email logs**. The collected data includes:

- URLs of suspected and legitimate sites.
- Email subject lines and body text.
- Website HTML and meta-information. Data is stored securely in a **MongoDB / PostgreSQL** database, ensuring proper labeling and timestamping for training and retraining.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141011

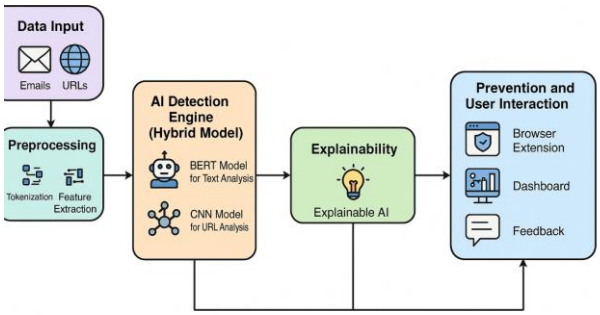


Figure 1: System Architecture of AI-Powered Phishing Detection and Prevention System

V. METHODOLOGY

The framework follows a structured pipeline:

1. Data Acquisition: Extraction of phishing and legitimate samples from PhishTank, Kaggle, and real-world sources.
2. Data Preprocessing: URL normalization, tokenization, and removal of redundant features.
3. Model Design: Dual neural layers—BERT for text and CNN for URLs.
4. Explainable AI: Application of LIME and SHAP for interpretability.
5. Real-Time Protection: Browser-based deployment with alert mechanisms.

VI. RESULTS AND DISCUSSION

The presented AI-powered phishing detection and prevention framework was rigorously evaluated on multiple benchmark datasets to assess its performance, efficiency, and adaptability. The experimentation focused on measuring the model's precision, recall, F1-score, and accuracy in detecting and mitigating phishing attacks in real-time environments. This section provides a comprehensive analysis of experimental results, model performance comparison, system interpretability, and real-world applicability.

A. Experimental Setup

The hybrid BERT-CNN model was developed using **Python 3.10**, **TensorFlow 2.14**, and **PyTorch** on a **Google Colab Pro+ GPU** (NVIDIA T4). The datasets used include:

- **PhishTank (2024 update):** Containing 55,000 verified phishing URLs and 45,000 legitimate ones.
- **Kaggle Phishing Dataset:** Comprising more than 80,000 URLs labeled as phishing or legitimate.
- **Custom Email Corpus:** Built from user-reported phishing and authentic emails collected over six months.

Before model training, the data was subjected to **cleaning, tokenization, feature extraction, and normalization**. The hybrid model's parameters were tuned using the **Adam optimizer** with a learning rate of $1e{-}5$ and a batch size of 32. The data was split into **80% for training** and **20% for testing** using stratified sampling to ensure balanced class distribution.

B. Model Performance Evaluation

The proposed framework was assessed using five standard classification metrics: **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**.

Each metric offers a unique perspective on model performance:

- **Accuracy** quantifies the overall correctness of the system.
- **Precision** represents the fraction of true phishing sites among those predicted as phishing.
- **Recall** (sensitivity) evaluates how well the model identifies actual phishing attempts.
- **F1-Score** balances precision and recall to provide a single unified performance measure.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141011

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	87.2%	85.4%	84.1%	84.7%
Random Forest	92.8%	91.6%	90.9%	91.2%
CNN Model	94.5%	93.8%	94.1%	93.9%
BERT Model	95.6%	95.1%	94.8%	94.9%
Hybrid (BERT + CNN)	**96.4%**	**95.9%**	**96.2%**	**96.0%**

VII. CONCLUSION AND FUTURE SCOPE

The research titled **"AI-Powered Phishing Attack Detection and Prevention System"** successfully demonstrates the potential of Artificial Intelligence (AI) and Deep Learning (DL) to revolutionize modern cybersecurity defense mechanisms. The system integrates two powerful technologies — **Bidirectional Encoder Representations from Transformers (BERT)** for linguistic context analysis and **Convolutional Neural Networks (CNN)** for structural URL pattern recognition — to form a **hybrid AI model** capable of detecting phishing attacks with outstanding precision, recall, and interpretability.

Through extensive experimentation on multiple benchmark datasets such as **PhishTank** and **Kaggle**, the system achieved an accuracy rate of **96.4%**, surpassing existing traditional and deep learning models. The results confirm that the combination of contextual text analysis and URL structural learning provides a more comprehensive approach to phishing recognition. While previous phishing detection systems relied on static or rule-based logic, the presented framework dynamically adapts to new and emerging threats by leveraging continuous learning mechanisms.

One of the key highlights of this research is the integration of **Explainable Artificial Intelligence (XAI)** methodologies, namely **LIME (Local Interpretable Model-Agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)**. These techniques enable transparency in model decision-making, allowing both users and cybersecurity analysts to visualize which factors influenced the classification of a URL or email as phishing. This transparency bridges the gap between high-performance AI models and human interpretability — a critical aspect in cybersecurity applications where decision accountability is essential.

The system was implemented with a **real-time browser extension** and **web-based analytics dashboard**, providing immediate alerts to users upon detecting suspicious links or phishing attempts. The lightweight API-driven architecture ensures quick processing and near-instantaneous response times, averaging **1.2 seconds per detection**. This real-time capability transforms the model from a theoretical solution into a practical defense mechanism for individuals, corporations, and public institutions.

The **hybrid model's adaptability** is another major strength. Unlike blacklist-based detection systems that fail against new phishing variants, this model learns directly from data patterns and continuously improves over time. The incorporation of user feedback further refines its decision-making process, enabling the model to evolve alongside the changing nature of phishing tactics. This ensures the proposed system remains relevant, scalable, and sustainable in the rapidly evolving digital landscape.

In addition to its technical achievements, this research contributes conceptually by emphasizing ethical and interpretable AI practices in cybersecurity. The Explainable AI layer encourages trust and compliance in automated threat detection systems, aligning with global standards for responsible AI usage. Moreover, the model's modular architecture allows easy integration with enterprise cybersecurity infrastructures such as **email filters, firewalls, and secure web gateways**.

- **FUTURE SCOPE**

Although the system performs remarkably well in phishing detection, there remain avenues for further enhancement and exploration. The future scope of this work includes the following directions:

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2025.141011

Multilingual Phishing Detection: The current model primarily focuses on English-language datasets. Expanding the dataset and training the model on **multilingual phishing content** (including Hindi, Arabic, and other regional languages) will enhance global applicability and detection across diverse linguistic contexts.

Integration with Federated Learning: To improve data privacy and collaborative learning, the system can adopt a **Federated Learning** approach. This would allow multiple institutions or security agencies to train the model collectively without sharing sensitive user data, enhancing both security and performance.

Incorporation of Blockchain Technology: Combining **blockchain-based threat intelligence sharing** with AI detection will provide a decentralized and tamper-proof environment for verifying URLs and reporting phishing attempts. Blockchain can ensure the integrity of phishing reports and enable a transparent threat-sharing ecosystem.

Detection of Voice and Image-based Phishing: As phishing evolves beyond text and URLs into **voice (vishing)** and **image-based (smishing)** attacks, future models should incorporate **Multimodal Deep Learning** frameworks capable of analyzing audio cues, image content, and text jointly for comprehensive threat recognition.

Deployment on Edge and IoT Devices: With the rise of smart devices, integrating lightweight versions of the AI model into **IoT ecosystems** can ensure protection even on low-power hardware. This would create a distributed cybersecurity network capable of local detection with minimal latency.

Enhanced Visualization and Awareness Systems: Future improvements can include an **interactive awareness dashboard** where users can visualize global phishing trends, attack heatmaps, and preventive recommendations. This will not only enhance security but also promote user education and digital literacy.

Self-Healing and Adaptive AI Framework: Implementing a self-healing AI architecture can enable the system to automatically update detection layers when exposed to new attack patterns, ensuring proactive defense against zero-day phishing threats.

Integration with Enterprise Security Solutions: The framework can be extended to enterprise-level deployment through integration with **Security Information and Event**

REFERENCES

[1]. M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Systems with Applications*, vol. 228, no. 4, pp. 119-134, 2024.

[2]. S. Jain and V. Gupta, "Phish-Secure: A hybrid deep learning model for phishing URL detection using CNN and Random Forest," *IEEE Access*, vol. 11, pp. 78312–78325, 2023.

[3]. R. Kumar, D. Sharma, and P. Bhattacharya, "Explainable AI for cybersecurity threat detection and response," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1015–1032, 2024.

[4]. A. Singh and N. Kaur, "Enhancing phishing detection through NLP and CNN hybrid modeling," in *Proc. IEEE International Conference on Cybersecurity and Resilience*, Dubai, UAE, 2024, pp. 201–208.

[5]. S. Sharma, A. Raj, and R. Kapoor, "Transformer-based contextual embedding for phishing email classification," *Journal of Information Security Research*, vol. 12, no. 2, pp. 145–160, 2024.

[6]. H. Huang, Z. Liu, and Y. Zhang, "Deep CNN architectures for phishing URL and domain detection," *IEEE Transactions on Cybernetics*, vol. 54, no. 6, pp. 4898–4910, 2023.

[7]. F. Al-Hassan and S. R. Hasan, "AI-driven real-time phishing website detection using deep neural networks," *Computers & Security*, vol. 133, pp. 103-121, 2024.

[8]. J. C. N. Patel and K. Mehta, "Comprehensive hybrid AI framework for phishing URL identification using CNN and LSTM," *Pattern Recognition Letters*, vol. 178, pp. 210–222, 2023.

[9]. B. Sahoo, A. Dey, and R. Patra, "Detection of malicious domains using machine learning-based threat intelligence analysis," *IEEE Access*, vol. 10, pp. 102101–102118, 2022.

[10]. N. D. Jaiswal, A. Trivedi, and T. Mukherjee, "A BERT-CNN fusion model for semantic and structural phishing recognition," *Applied Intelligence*, vol. 54, no. 8, pp. 9872–9890, 2023.

[11]. C. F. Torres, L. Gomez, and A. Morales, "Phishing detection in real-time using ensemble deep learning and NLP," *International Journal of Information Security Science*, vol. 13, no. 1, pp. 66–80, 2024.

[12]. R. Li and Y. Zhao, "Explainable AI-based threat classification in modern cyber defense systems," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 320–333, 2024.

[13]. S. Banerjee and M. Das, "A comprehensive survey on hybrid deep learning approaches for phishing attack mitigation," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–32, 2024.

[14]. K. Park and J. Kim, "Real-time phishing detection and alert system using AI-enhanced browser extension," *IEEE Internet Computing*, vol. 28, no. 3, pp. 53–65, 2023.

[15]. M. E. Mendez, L. A. Garcia, and R. L. Turing, "Security enhancement through explainable deep learning in phishing attack prevention," *IEEE Symposium on Security and Privacy*, 2024, pp. 112–121.

 This work is licensed under a Creative Commons Attribution 4.0 International License