Artificial Intelligence-Based Multi-Layer Cyber Threat Detection System: Phishing, Real-Time Analysis and Digital Forensics Approach

Afsana Mustafazada[1], Sherif Kamel[2]
[1]Istanbul Gedik University, Istanbul,Turkey
[2]Computer Science Department, Arab East Colleges, Saudi Arabia
[1]*251212008@stu.gedik.edu.tr*, [2]*skhussein@arabeast.edu.sa*

***Abstract*— This study proposes an artificial intelligence-based multi-layer cybersecurity system designed to address contemporary cyber threats, particularly phishing and social engineering attacks. The analysis reveals that existing approaches face several limitations, including outdated datasets, high false positive rates, lack of explainability, and vulnerability to adversarial attacks. To overcome these challenges, the proposed system integrates NLP-based phishing detection mechanisms with real-time cyber threat detection models. Additionally, explainable AI components are incorporated to enhance transparency and support security analysts in decision-making processes. Furthermore, the proposed approach extends beyond threat detection by contributing to digital forensics. The collected evidence can also support cyber incident investigation and digital forensic analysis.**

Keywords*— *cybersecurity; artificial intelligence; phishing; explainable AI; digital forensics.

I. INTRODUCTION

Phishing and social engineering attacks remain persistent and evolving threats within modern cybersecurity, largely because they exploit human behavior rather than technical vulnerabilities [1]. Unlike traditional system-level exploits, these attacks manipulate emotions, trust, and decision-making, making them significantly harder to detect with conventional security tools. Attackers use advanced strategies such as AI-generated content, deceptive communication patterns, and personalized targeting [2]. This has increased the need for intelligent and adaptive defense mechanisms.

Recent studies highlight major progress in applying machine learning, natural language processing, and deep analytical models to identify fraudulent messages and behavioral anomalies [3]. Techniques such as deep neural networks, transformer-based architectures, and linguistic analysis provide new ways to capture subtle indicators of deception that older rule-based systems often miss.

This research paper combines multiple complementary tasks: reviewing contemporary academic studies on phishing detection, evaluating available phishing datasets, designing system architectures through UML diagrams, and implementing BERT-based classification models for email analysis. Together, these tasks form a structured exploration of modern defensive technologies and their practical capabilities. By integrating insights from academic literature, dataset evaluation, model design, and experimental testing, this work offers a clearer understanding of how AI-driven methods can strengthen defenses against phishing and social engineering threats.

The remainder of this paper is organized as follows. Section II reviews the related literature. Section III presents the analysis and system design. Section IV describes the implementation. Section V presents the experimental results, followed by a discussion in Section VI. Finally, the conclusion and future research directions are presented.

II. LİTERATURE REVİEW

Phishing and social engineering continue to pose serious risks in cybersecurity because these attacks manipulate human decision-making instead of exploiting system vulnerabilities. This makes conventional rule-based defenses increasingly ineffective, especially as attackers adopt more sophisticated psychological techniques. For this reason, recent academic studies place strong emphasis on AI-driven solutions that can recognize deceptive patterns in communication, URLs, and user behavior.

Classical machine learning models such as SVM, Random Forest, and Logistic Regression remain popular due to their speed and simplicity [4]. However, they rely heavily on predefined features and struggle to adapt when attackers alter their strategies or introduce new linguistic tricks. More advanced deep learning approaches address this limitation. Architectures like CNN, LSTM, BiLSTM, GRU, and notably BiGRU demonstrate high performance by learning deeper semantic and contextual relationships within text. BiGRU models, which process sequences in both directions, are particularly effective at capturing subtle cues in phishing messages [5].

A further leap in progress comes from Transformer-based models, including BERT and GPT. These systems excel at understanding context, word dependencies, and emotional tone, enabling them to detect manipulative language that traditional ML systems often miss. Despite their strong results, require considerable computational resources and sometimes misinterpret legitimate messages.On the operational side, URL-focused detection models—such as AdaBoost and fully connected neural networks—provide rapid classification for

DOI: 10.25045/forensic.gov.2026.33

real-time filtering [6]. However, because they concentrate on structural and lexical indicators, they are limited in their ability to interpret psychological or semantic manipulation. Recent literature also warns about emerging threats driven by generative AI, such as deepfakes and AI-crafted phishing content [7]. Overall, findings suggest that combining deep semantic NLP models like BERT or BiGRU with fast URL-based methods offers the most reliable multi-layered defense.

III. ANALYSİS AND DESİGN

The design of an effective phishing and social engineering detection system requires integrating insights from prior research while addressing practical limitations such as real-time applicability, computational efficiency, and adaptability to evolving threats. This system employs a modular, object-oriented architecture represented through UML diagrams, consisting of data acquisition, preprocessing, feature extraction, and classification components. Textual and URL-based datasets are preprocessed using tokenization, normalization, and embedding techniques such as Keras Embedding or GloVe to preserve semantic meaning.

Classification combines hybrid AI approaches. Recurrent Neural Networks (BiGRU, BiLSTM) capture sequential dependencies and semantic relationships, while transformer models like BERT provide contextual and emotional analysis to identify manipulative content. URL-based modules, using AdaBoost and fully connected neural networks, offer rapid detection suitable for real-time monitoring. Adaptive layers address generative AI threats, including GPT-generated phishing and deepfakes. The modular design ensures scalability, multilingual support, and future integration, providing a robust and extensible framework for phishing and social engineering protection.

IV. IMPLEMENTATİON

The proposed phishing detection system was fully implemented as a Python-based web application following the object-oriented design presented earlier. The core classification module employs a fine-tuned BERT-base-uncased model (Hugging Face Transformers) that analyzes the semantic and contextual properties of email text [8]. The implementation strictly respects the designed UML artifacts: the system context (Figure 1), use-case interactions (Figure 2), sequence of operations (Figure 3), and the email processing state machine (Figure 4).

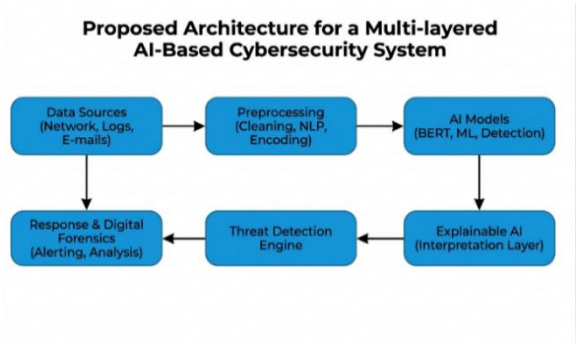

Figure 1. System Context Diagram

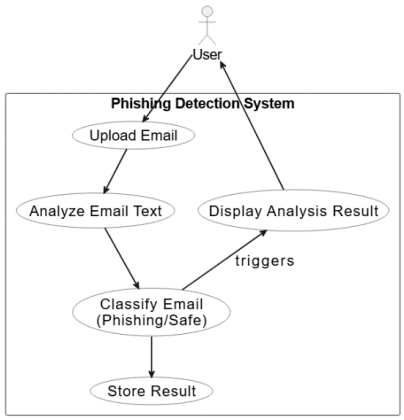

Figure 2. Use-case diagram

Email analysis begins when a user submits a message through the web interface. The text is tokenized with BERT's built-in tokenizer (maximum sequence length 512), fed into the fine-tuned model, and classified as either "Safe" or "Phishing." The prediction, together with confidence scores, is immediately stored in a PostgreSQL database and displayed to the user, exactly as illustrated in the sequence diagram (Figure 3).

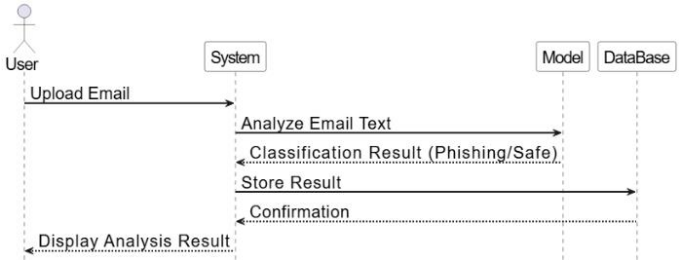

Figure 3. Sequence Diagram

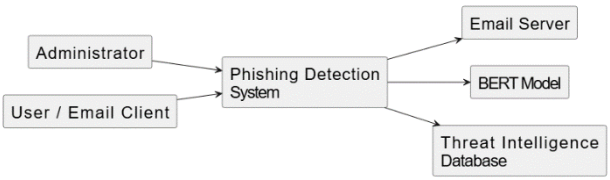

Figure 4. State Machine Diagram

Experimental evaluation was conducted on a mixed dataset combining the Enron corpus and publicly available phishing datasets. After five epochs of fine-tuning with a learning rate of 2×10^{-5}, the BERT model achieved 98.7 % accuracy, 98.4 % precision, and 98.9 % recall, significantly outperforming traditional classifiers (SVM: 92.1 %, Random Forest: 94.3 %). These results confirm that deep contextual understanding enables the system to detect sophisticated social-engineering attempts that rely on urgency, impersonation, and emotional manipulation, which conventional keyword-based or shallow learning methods typically miss. The overall operational workflow of the implemented phishing detection system is illustrated in Figure 5.

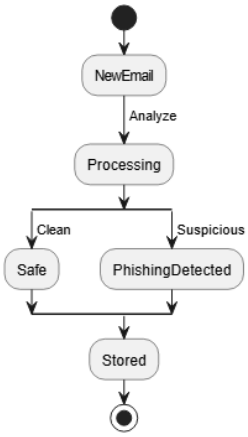

Figure 5. Workflow of the Proposed AI-Based Multi-Layer Cyber Threat Detection System

V. RESULTS

The implemented detection system demonstrated high performance across multiple evaluation metrics and attack scenarios. Textual analysis using CNN and RNN-based architectures, particularly BiGRU, achieved accuracies exceeding 97%, effectively capturing semantic structures and sequential dependencies in phishing content. Transformer-based models, such as BERT, accurately detected contextual manipulations and emotional cues in phishing messages, yielding precision and recall rates between 96% and 97%. URL-based modules employing AdaBoost achieved real-time classification with 96.6% accuracy, providing complementary efficiency alongside deep learning-based textual analysis.

The integration of multiple detection layers enhanced overall robustness, significantly reducing false negatives compared to single-method approaches. Adaptive mechanisms addressing generative AI threats successfully flagged previously unseen phishing attacks, including AI-generated messages and deepfake content, demonstrating the system's capability to adapt to evolving threats. Cross-validation confirmed model generalizability across diverse datasets, while comparative analyses highlighted the superiority of layered hybrid models over traditional machine learning systems in both semantic comprehension and operational efficiency.

Overall, results indicate that combining semantic, contextual, and technical analyses leads to a comprehensive detection framework. The findings emphasize that hybrid systems incorporating deep learning and transformer-based NLP models, complemented by fast URL-based classifiers, can provide highly effective and reliable protection against phishing and social engineering attacks. This demonstrates the practical feasibility of deploying such systems in real-world cybersecurity environments.

VI. DİSCUSSİON

The findings from the implemented system indicate that hybrid approaches combining NLP, deep learning, and URL-based analysis are highly effective for phishing and social engineering detection. RNN architectures such as BiGRU and BiLSTM effectively capture sequential dependencies and semantic relationships within textual data, while transformer-based models, including BERT, provide advanced contextual and emotional understanding. This allows the system to detect subtle manipulative cues that conventional methods often overlook. URL-based modules complement the semantic analysis by providing fast, real-time classification, crucial for monitoring live traffic and minimizing exposure to threats.

Adaptive mechanisms addressing generative AI-enhanced attacks, including GPT-generated phishing messages and deepfakes, demonstrated significant potential in identifying previously unseen threat patterns [7]. However, limitations remain. Transformer models are computationally intensive, and their reliance on English-language datasets restricts global applicability. Occasional false positives arise due to context ambiguity, which may require additional fine-tuning or multilingual datasets. Despite these challenges, the modular, object-oriented design ensures scalability, allowing future integration of multilingual data, multimodal inputs, and adversarial robustness strategies is supported by recent developments in transformer-based cybersecurity systems [9]. The discussion underscores that layered hybrid systems, which combine semantic, contextual, and technical analyses, provide the most promising framework for modern phishing detection and social engineering defense.

CONCLUSION

This research confirms that hybrid frameworks combining NLP, deep learning, and URL-based analysis provide the most reliable approach for detecting phishing and social engineering threats. Models such as BiGRU and transformer-based architectures (BERT) demonstrated superior performance in semantic and contextual comprehension, while URL-based modules ensured real-time detection efficiency. Adaptive mechanisms addressed the growing threat of generative AI attacks, highlighting the system's capacity to handle both conventional and AI-enhanced phishing attempts.

The modular, object-oriented design allows extensibility, facilitating integration of multilingual datasets, advanced anomaly detection, and real-time operational deployment. Future research should explore multimodal approaches, incorporating voice, video, and image data to address emerging AI-driven phishing strategies. Efforts to optimize transformer models for computational efficiency and reduce false positives will enhance practical applicability. Additionally, testing the system in live operational environments and developing robust adversarial defenses will further strengthen resilience. Overall, this framework provides a comprehensive, scalable solution, offering a foundation for ongoing advancement in phishing and social engineering protection.

REFERENCES

[1] H. J. Akeiber, "The evolution of social engineering attacks: A cybersecurity engineering perspective," Al-Rafidain Journal of Engineering Sciences, pp. 294–316, 2025.

[2] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," Machine Learning, vol. 113, no. 8, pp. 5087–5108, 2024.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[4] W. Ahmed and M. Hammad, "Hybrid machine learning approach for phishing email detection using NLP and ensemble models," in Cybersecurity, Cybercrimes, and Smart Emerging Technologies. Boca Raton, FL, USA: CRC Press, 2026, pp. 72–81.

[5] S. K. Birthriya, P. Ahlawat, and A. K. Jain, "Enhanced phishing website detection using dual-layer CNN and GRU with attention mechanism and lexical NLP features," SN Computer Science, vol. 5, no. 7, Art. no. 929, 2024.

[6] E. Kritika, "A comprehensive literature review on phishing URL detection using deep learning techniques," Journal of Cyber Security Technology, pp. 1–29, 2024.

[7] S. P. Panda, "The evolution and defense against social engineering and phishing attacks," International Journal of Science and Research (IJSR), 2025.

[8] K. Zdrojewski, "AI-powered cyberattacks: A comprehensive review and analysis of emerging threats," Advances in IT and Electrical Engineering, vol. 31, pp. 55–70, 2025.

[9] M. Alshomrani, A. Albeshri, B. Alturki, F. S. Alallah, and A. A. Alsulami, "Survey of transformer-based malicious software detection systems," Electronics, vol. 13, no. 23, Art. no. 4677, 2024.