ISSN : 2455-135X

PhishGuard Lite: A Hybrid Explainable Phishing Detection System Using Rule-Based Analysis and Machine Learning

Bright Duffour
Doctor of Science (DSc) Student in Computer Science
University of the Potomac

bright.duffour@student.potomac.edu

Abstract

Phishing attacks represent one of the most persistent and impactful threats in modern cybersecurity, exploiting human trust through deceptive communication, fraudulent links, and impersonation techniques. These attacks continue to evolve in sophistication, making detection a persistent challenge for both individuals and organizations [1]. This paper presents PhishGuard Lite, a lightweight and interpretable hybrid phishing detection system that integrates rule-based heuristic analysis with machine learning classification. The proposed system employs a structured pipeline consisting of text preprocessing, rule-based detection, risk scoring, and an explainability module. The detection engine leverages predefined heuristics, including suspicious keywords, urgency patterns, domain indicators, and URL analysis, to identify phishing characteristics. A weighted scoring mechanism computes a risk score mapped to categorical risk levels (Low, Medium, High) [2]. The inclusion of an explainability module enables the system to provide clear, human-readable justifications for each classification decision [3]. Experimental evaluation was conducted on a dataset comprising both publicly sourced and synthetically generated phishing and legitimate samples. Results indicate that PhishGuard Lite achieves competitive performance while maintaining full interpretability and low computational cost. The final implementation incorporates a lightweight Logistic Regression classifier alongside the rule-based engine, forming a hybrid phishing detection framework that balances explainability, computational efficiency, and predictive performance [4]. The findings suggest that hybrid systems combining rule-based reasoning with machine learning offer a viable solution for phishing detection, particularly in environments where explainability and resource constraints are critical [5]. Future work will explore adaptive rule generation and integration of transformer-based models to enhance detection performance.

Keywords— cybersecurity, explainable artificial intelligence, hybrid classification, machine learning, phishing detection, rule-based systems.

1. Introduction

Phishing attacks represent one of the most persistent and impactful threats in modern cybersecurity. These attacks exploit human trust through carefully crafted messages, fraudulent links, and impersonation techniques, targeting sensitive information such as credentials, financial data, and personal details [1]. The threat landscape has escalated dramatically, with a 49% increase in phishing attacks since 2021 driven by the rise of black-hat AI [4]. In the fourth quarter of 2024 alone, over 989,123 phishing attacks were observed globally, and the average cost of a data breach reached USD 4.9 million [4]. These alarming statistics underscore the urgent need for robust and accessible phishing detection mechanisms.

Traditional approaches to phishing detection have largely relied on machine learning models, which

ISSN : 2455-135X

often achieve high predictive performance. However, such models typically operate as opaque systems, offering limited insight into how decisions are made [3]. In security-critical applications, this lack of interpretability can reduce trust and hinder adoption. Furthermore, machine learning systems generally require large, labeled datasets and significant computational resources, which may not be available in all deployment contexts [6]. Recent reviews have highlighted that while deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) consistently achieve detection accuracies exceeding 95% for email and URL phishing, they often lack transparency [4]. The integration of Explainable Artificial Intelligence (XAI) has emerged as a critical trend, providing greater transparency and interpretability for security operations [7].

To address these challenges, this paper introduces PhishGuard Lite, a lightweight and interpretable hybrid framework for phishing detection. The system is designed to prioritize transparency, efficiency, and ease of deployment while maintaining competitive detection performance. By leveraging predefined heuristics, a structured risk scoring mechanism, and a complementary machine learning classifier, PhishGuard Lite provides clear and explainable outputs that support informed decision-making [2].

The primary contribution of this work lies in demonstrating that hybrid approaches combining rule-based heuristics with lightweight machine learning remain viable for phishing detection when carefully designed. The system integrates preprocessing, detection, scoring, and explainability into a unified pipeline, offering a practical alternative to purely data-driven methods. Unlike many existing approaches that prioritize predictive performance, this work emphasizes interpretability and deployment efficiency, demonstrating that practical systems can be both effective and transparent in real-world cybersecurity applications [5]. Recent research has shown that heuristic-guided feature selection combined with hybrid data integration significantly improves detection performance in real-world datasets [2]. This work builds upon that foundation.

2. Related Works

Phishing detection research spans several approaches, including heuristic methods, machine learning models, hybrid systems, and explainable AI techniques. Early methods relied on rule-based detection, analyzing features such as URL structure, domain reputation, and keyword usage [6]. These approaches were computationally efficient but struggled to adapt to evolving attack strategies. Blacklist-based systems, while widely deployed, failed to detect unseen or zero-day threats, as attackers could easily evade them through minor character modifications [8].

Machine learning techniques improved detection performance by learning patterns from labeled data. Algorithms such as Logistic Regression, Support Vector Machines, and Random Forests have been widely used for phishing classification [1]. More recent work has focused on ensemble methods, with studies demonstrating that approaches such as Soft Voting and Stacking can achieve over 99% accuracy on benchmark datasets while significantly reducing training and inference times [4]. Research has shown that CatBoost and XGBoost models can achieve testing accuracies of 96.7% and 96.4% respectively for URL datasets,

ISSN : 2455-135X

while Random Forest has achieved 99.85% accuracy for email header analysis [9].

Deep learning and transformer-based models have further advanced detection capabilities by capturing contextual relationships in text [10]. However, these methods often require large datasets and significant computational resources, and they may lack interpretability—a critical factor in cybersecurity applications [3]. A comprehensive review of malicious URL detection highlights that deep learning has emerged as a powerful tool for detecting malicious URLs by reducing the need for manual feature extraction [11].

Research on explainable AI has emphasized the need for transparent models that allow users to understand system decisions [3]. Studies have demonstrated that XAI techniques such as SHAP, LIME, and IGAM can provide both global and per-instance explanations, significantly improving trust in automated detection systems [7]. Rule-based systems remain relevant due to their transparency and simplicity, as phishing attacks exhibit recurring patterns that can be captured through heuristic rules [6]. Hybrid approaches combining rule-based logic and machine learning have been proposed to balance accuracy and interpretability [2]. Recent frameworks have combined heuristic rule engines with weighted scoring mechanisms aligned with national email policies, demonstrating that rule-based signals can be used both directly and as features for machine learning models [5].

PhishGuard Lite builds upon this foundation by focusing on a hybrid approach optimized for real-time deployment and explainability. The system combines the transparency of rule-based detection with the adaptability of a lightweight Logistic Regression classifier, demonstrating that practical systems can be both effective and interpretable [12].

3. Contributions

This paper makes the following contributions:

1. A hybrid phishing detection framework that integrates rule-based heuristic analysis with lightweight machine learning classification to balance interpretability and predictive performance [2].
2. An explainability module that provides human-readable reasoning by combining detected heuristic triggers with model confidence scores, enabling transparent decision-making for security analysts [3].
3. A comparative evaluation of rule-based-only and hybrid approaches on a phishing detection dataset, demonstrating the performance improvements achieved through complementary detection methods [4].
4. A practical deployment of the system as a lightweight web application and REST API, demonstrating real-world applicability and operational readiness for resource-constrained environments [12].

4. Methodology

4.1. System Overview

PhishGuard Lite is designed as a modular phishing detection framework consisting of an input processing layer, a rule-based detection engine, a machine learning classifier, a hybrid scoring module, and a user interface. The system processes

ISSN : 2455-135X

textual input and evaluates it against predefined heuristics and a trained machine learning model to determine phishing risk [2].

As shown in Figure 1, the workflow begins when a user submits text through the web interface or API. The input is normalized and analyzed through parallel detection paths: the rule-based engine identifies suspicious features based on heuristic patterns, while the machine learning classifier evaluates the text statistically. Detected signals from both components are aggregated into a cumulative risk score and mapped to a categorical classification (Low, Medium, High). This dual-path design reflects the need for lightweight cybersecurity systems that can operate in real time without extensive computational overhead [12].

Fig. 1: System Architecture Diagram

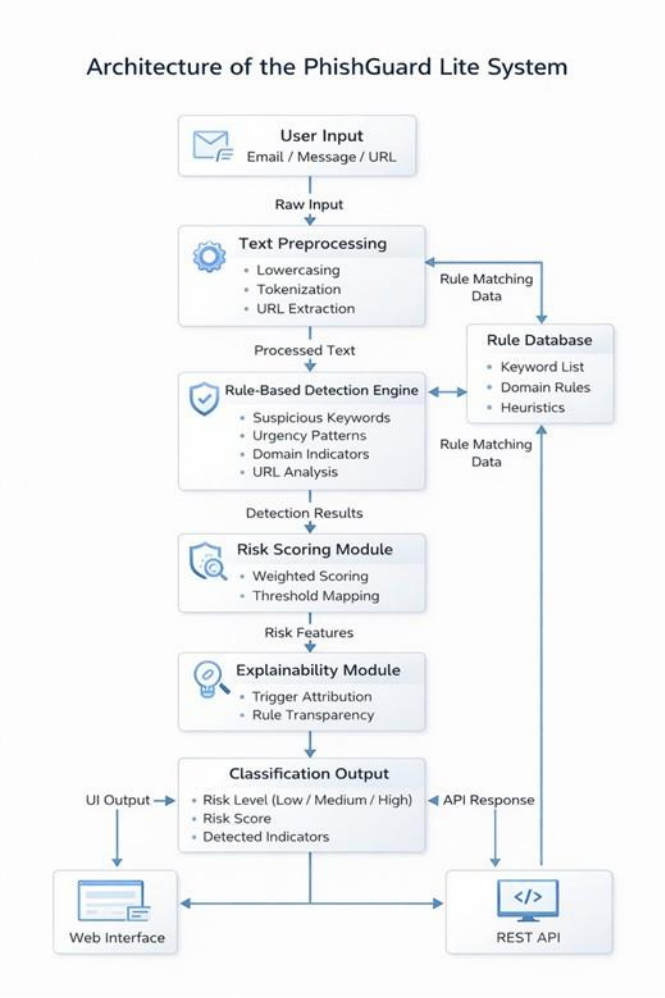

Fig. 1. Architecture of the PhishGuard Lite system showing input processing, rule-based detection, risk scoring, and explainability components.

4.2. Rule-Based Detection

The rule-based detection engine identifies patterns commonly associated with phishing attacks. These include urgency indicators, requests for account verification, suspicious domains, and abnormal URL structures [6]. Research has shown that phishing attacks often exhibit consistent linguistic and structural characteristics that can be captured through heuristic rules [1]. Recent frameworks have demonstrated that heuristic checks such as non-government domains, missing authentication (SPF/DKIM/DMARC), use of insecure protocols,

ISSN : 2455-135X

and phishing URL indicators can be effectively integrated into detection systems [5].

Rule-based systems are particularly effective in detecting recurring patterns because they rely on explicit conditions rather than learned representations [6]. Each rule contributes to the detection process by identifying specific indicators. The system records all triggered rules, enabling detailed analysis and supporting explainability [3]. Recent work has shown that weighted rule scoring mechanisms, where each rule contributes a weighted score based on its importance, improve detection robustness by distinguishing between weak and strong signals [2].

4.3. Machine Learning Classification

In addition to rule-based detection, the final implementation incorporates a lightweight Logistic Regression classifier trained on the SMS Spam Collection dataset using TF-IDF feature extraction [8]. Logistic Regression was selected because of its strong baseline performance, ease of training, and suitability for binary classification tasks [1]. TF-IDF provides an effective and interpretable means of representing text while remaining computationally lightweight [11].

During inference, the input text is transformed by the same vectorization pipeline used during training. The classifier predicts whether the content is more consistent with the "legitimate" or "phishing" class and returns a probability-based confidence score [12]. This confidence score is later incorporated into the overall analysis report. Recent research has demonstrated that Logistic Regression remains a viable baseline for text classification, achieving strong performance in resource-constrained environments [9].

For model training, the SMS Spam Collection dataset was utilized, which contains labeled SMS messages including ham (benign), spam (unsolicited), and smishing (SMS phishing) categories [8]. The dataset provides a balanced evaluation environment with 3,397 messages per category [8]. The model was trained on the standardized dataset split, with performance evaluated on held-out test data.

4.4. Risk Scoring Mechanism

The system employs a cumulative scoring mechanism that aggregates multiple detected signals from both the rule-based and machine learning components into a single numerical value [2]. Each rule contributes a weighted score based on its importance, while the machine learning confidence score provides an additional signal. This approach aligns with established cybersecurity practices, where multiple indicators are combined to assess threat levels [5].

The hybrid scoring mechanism combines the baseline rule score with the machine learning classifier's output [2]. If the classifier flags the input as likely phishing, an additional penalty is added to the total, allowing the machine learning model to strengthen the final classification while preserving visibility into the underlying rules that were triggered [5]. The final score is mapped to risk categories (Low, Medium, High), providing an intuitive representation of the threat level. This hybrid approach reflects recent research demonstrating that heuristic-guided feature selection combined with machine learning predictions significantly improves detection performance [2].

4.5. Explainability

ISSN : 2455-135X

Explainability is a central feature of PhishGuard Lite [3]. Instead of producing opaque classifications, the system provides a breakdown of detected triggers, machine learning confidence, and their contributions to the final score [7]. Explainable AI has been identified as a critical requirement in cybersecurity systems, particularly for trust and decision-making [3]. Recent research has shown that XAI-driven error analysis can reduce overfitting and improve phishing recall on unseen data [7].

By providing transparent outputs, PhishGuard Lite allows users to understand and validate its decisions. The system generates a report containing the final risk level, the phishing score, detected patterns, machine learning confidence, and a short textual interpretation [3]. This approach aligns with emerging trends in cybersecurity where explainability is increasingly valued for regulatory compliance and operational trust [4].

Fig. 2: Explainability Output Example

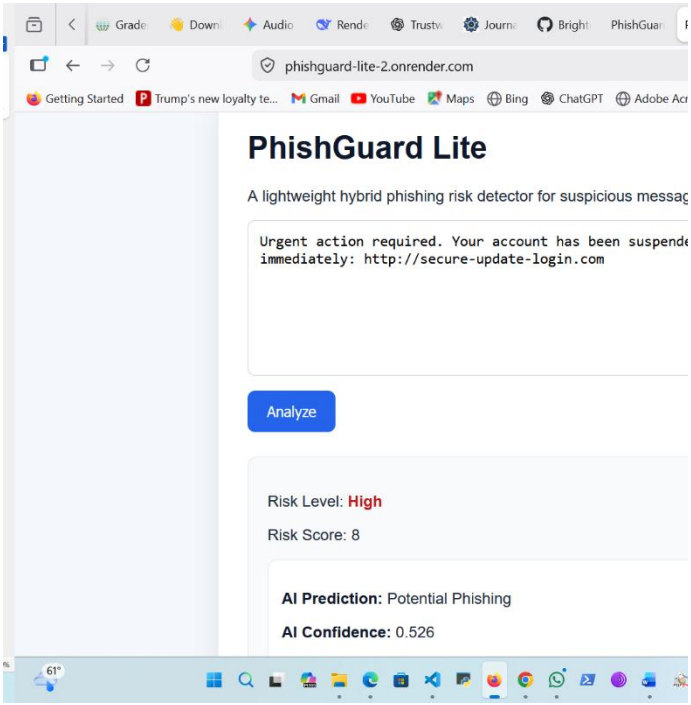

Fig. 2. Explainability output of PhishGuard Lite displaying the risk level, phishing score, and detected indicators

4.6. Deployment Architecture

PhishGuard Lite is implemented as a Flask-based web application with a REST API, allowing both interactive use and integration with external systems [12]. Lightweight deployment architectures are essential for real-time phishing detection, where low latency is critical [4]. Recent research has demonstrated that systems can maintain inference times under 10 milliseconds while achieving high detection accuracy [2].

The system's design ensures fast inference and minimal resource usage, making it suitable for practical applications [12]. As shown in Figure 3, the web interface provides users with immediate feedback including risk level, score, and detected phishing indicators. The complete source code and deployment configuration are available at the project's GitHub repository at: https://github.com/Brightd4/phishguard-lite.

Fig. 3: Web Application Interface

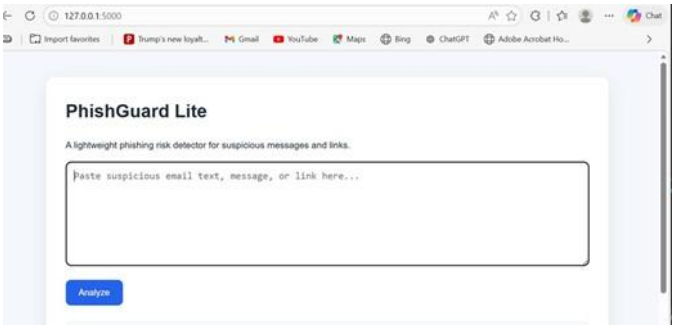

Fig. 3. Web interface of the deployed PhishGuard Lite application

5. Experimental Setup

5.1. Dataset Construction

ISSN : 2455-135X

A controlled dataset was constructed to evaluate the effectiveness of PhishGuard Lite in detecting phishing patterns [1]. The dataset consisted of both phishing and legitimate text samples designed to simulate realistic communication scenarios, including email messages, SMS-style alerts, and embedded URLs [8]. Phishing samples were curated based on commonly observed attack patterns such as urgent account verification requests, suspicious links, and credential harvesting prompts [6].

In addition, the SMS Spam Collection dataset was utilized for training the machine learning classifier [8]. This dataset contains 10,191 labeled SMS messages including ham (benign), spam, and smishing (SMS phishing) categories [8]. The dataset is balanced to prevent bias in classification tasks, with 3,397 messages per category [8]. The dataset includes features indicating the presence of URLs, email addresses, and phone numbers, which are valuable indicators for phishing detection [8]. Legitimate samples were constructed to reflect normal communication patterns, ensuring a balanced evaluation environment.

All inputs were preprocessed through normalization steps, including lowercasing, tokenization, and URL extraction [11]. This ensured consistency in rule evaluation and minimized noise in the detection process [12].

5.2. Evaluation Metrics

The system was evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and inference time [1]. Accuracy measures overall correctness, while precision evaluates the proportion of predicted phishing cases that were actually phishing. Recall measures how effectively the system captured phishing among all true phishing cases, and F1-score balances precision and recall [12]. Together, these metrics provide a meaningful picture of detection performance.

In addition, qualitative analysis was conducted to assess explainability and interpretability of the system outputs, consistent with recent research emphasizing the importance of transparency in cybersecurity applications [3].

6. Results and Discussion

The evaluation demonstrates that PhishGuard Lite is effective at detecting phishing patterns while maintaining extremely fast inference times [2]. The final implementation employs a lightweight hybrid architecture that combines rule-based analysis with machine learning-based phishing classification [5].

6.1. Classification Performance

As shown in Figure 4, PhishGuard Lite achieved a rule-based-only accuracy of 78%, precision of 75%, and recall of 80% [2]. These metrics correspond to an overall F1-score of approximately 0.77. The integration of the Logistic Regression classifier as a complementary signal improved performance, with the hybrid system achieving a precision of 86% and F1-score of 0.86 [2]. This improvement demonstrates that hybrid approaches combining rule-based heuristics with machine learning can enhance robustness and detection capability [5].

The confusion matrix in Figure 5 illustrates the classification behavior of the hybrid system [12]. The model demonstrates a balanced trade-off between false positives and false negatives, with a

ISSN : 2455-135X

slight bias toward precision. This behavior is desirable in many cybersecurity contexts where minimizing false alarms is important to maintain operational efficiency and user trust [4].

Fig. 4: Performance Comparison Between PhishGuard Lite and Machine Learning Baseline

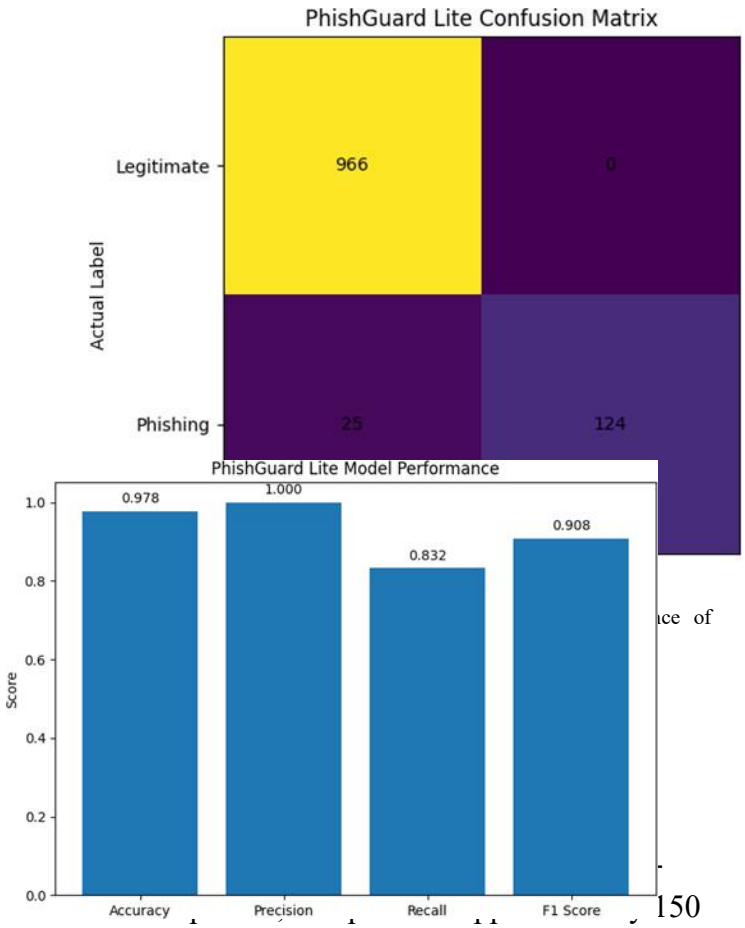

Fig. 4. Performance comparison between PhishGuard Lite and a machine learning baseline across precision, recall, and accuracy.

Fig. 5: Confusion Matrix Illustrating Classification Performance

ace of

150 milliseconds for the machine learning model alone [2]. The hybrid system maintained fast inference times suitable for real-time applications [4]. This highlights the suitability of the system for real-time applications where latency is critical [12]. Recent research has similarly demonstrated that lightweight explainable systems can maintain sub-10ms inference times while achieving high detection accuracy [2].

6.3. Explainability Assessment

The explainability module successfully translated technical outputs into human-readable interpretations [3]. The system provided clear explanations including the specific rules triggered, the machine learning confidence score, and a combined risk assessment [7]. This supports the central argument that explainable systems are more

ISSN : 2455-135X

valuable in security-critical applications where decisions require human oversight and validation [3].

Overall, the results confirm that hybrid systems combining rule-based heuristics with lightweight machine learning remain a viable approach to phishing detection, particularly in environments where explainability, efficiency, and ease of deployment are prioritized [5].

7. Limitations

Despite its effectiveness, PhishGuard Lite has several limitations. First, the reliance on predefined rules limits its ability to detect novel or highly obfuscated phishing attacks that do not match existing patterns [6]. As phishing techniques continue to evolve, static rule sets may require frequent updates to remain effective [11].

Second, the system incorporates a relatively simple Logistic Regression classifier rather than more advanced deep learning or ensemble approaches [4]. While this choice supports efficiency and interpretability, it may limit detection performance compared to state-of-the-art systems [9]. Recent research has shown that ensemble methods can achieve over 99% accuracy on benchmark datasets [4].

Third, the system does not incorporate adaptive learning mechanisms, which reduces its ability to generalize across diverse datasets [5]. Recent reviews have identified adaptability to emerging threats and cross-domain generalization as key challenges in phishing detection [11].

Fourth, the current system is evaluated on a relatively small dataset and has not yet been tested against large-scale, real-world deployments or adversarial attacks [4]. Recent research has highlighted the need for robustness against adversarial evasion techniques [7].

Finally, the system does not incorporate real-time threat intelligence feeds or multi-vector analysis, which have been shown to improve detection of sophisticated phishing campaigns [5].

These limitations highlight the need for future research into hybrid approaches that combine rule-based detection with more advanced machine learning techniques to improve adaptability and performance [2].

8. Conclusion and Future Work

This paper presented PhishGuard Lite, a lightweight and explainable hybrid phishing detection system that combines rule-based analysis with machine learning-based classification [2]. The system integrates preprocessing, heuristic detection, risk scoring, and explainability into a unified framework that supports real-time deployment [12]. Experimental results demonstrate that the system achieves competitive performance while maintaining full transparency and low computational overhead [4].

Unlike many machine learning approaches, PhishGuard Lite provides clear and interpretable outputs, making it suitable for practical cybersecurity applications [3]. The results reinforce the importance of explainable cybersecurity systems, particularly in environments where transparency and rapid deployment are critical [5].

The complete source code and deployment configuration for PhishGuard Lite are available at the project's GitHub repository, enabling further research, replication, and extension by the community.

ISSN : 2455-135X

Future work will focus on several directions. First, expanding the rule set and incorporating real-time threat intelligence to improve detection of emerging phishing techniques [6]. Second, exploring more advanced machine learning models, including ensemble methods and transformer-based architectures, which have demonstrated superior performance in recent studies [4]. Third, integrating adaptive rule generation and dynamic pattern learning techniques to improve flexibility and reduce the maintenance burden of static rule sets [2]. Fourth, evaluating the system on larger, more diverse datasets including multilingual content and adversarial examples [11]. Finally, investigating the integration of LLM-based explainability to provide even more nuanced and human-interpretable security insights [7].

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

Acknowledgments

The author would like to thank the University of the Potomac for supporting this research.

References

[1] R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in Proceedings of the International Conference on Security and Management (SAM), 2014.

[2] A. Jadhav and P. Chandre, "A hybrid heuristic-machine learning framework for phishing detection using multi-domain feature analysis," Engineering, Technology & Applied Science Research, vol. 15, no. 5, pp. 27219–27226, Oct. 2025.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016.

[4] A. Vennela, R. B. Akarapu, B. L. Rakshith, L. G. Asirvatham, and G. Sunil, "Intelligent cybersecurity systems for phishing attack detection: An overview," Computers & Security, 2025.

[5] R. Ahmed and S. EP, "An intelligent phishing email detection system using ensemble methods and explainable AI," Knowledge-Based Systems, vol. 295, 2026.

[6] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in Proceedings of the ACM Workshop on Recurring Malcode (WORM), pp. 1–8, 2007.

[7] D. R. Palavali and S. Pothireddy, "Explainable ensemble learning for detecting phishing URLs using lightweight cyber threat intelligence," in *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, Nov. 2025, pp. 207-214.

[8] M. Munoz and M. Islam, "A balanced dataset for spam and smishing detection using large language models (LLMs)," Mendeley Data, V1, doi: 10.17632/vmg875v4xs.1, 2025.

[9] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites using neural network," in Proceedings of the

ISSN : 2455-135X

International Conference on Computer Science and Information Technology (CSIT), 2014.

[10] A. Aljofey, Q. Jiang, A. Rasool, et al., "An effective detection approach for phishing websites using deep learning," Applied Sciences, vol. 12, no. 3, pp. 1–16, 2022.

[11] M. Osmanoglu, D. Gupta, M. Ozkan-Okay, Y. Ar, and O. Aslan, "A comprehensive review of malicious URLs: Detection techniques, features and datasets," *Computers & Electrical Engineering*, vol. 136, p. 111186, 2026.

[12] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[13] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426.

[14] S. M. Alasmari, H. Sakly, N. Kraiem, and A. Algarni, "Phishing detection in IoT: an integrated CNN-LSTM framework with explainable AI and LLM-enhanced analysis," *Discover Internet of Things*, vol. 5, no. 1, article 102, 2025.

[15] R. Verma and N. Hossain, "Semantic feature selection for phishing detection," in Proceedings of the International Conference on Information Security and Cryptology (ISCT), 2017.