AI-Based Phishing Detection System for Detecting Malicious Websites and Emails

Vitthal B. Kamble[1], Manthan Sanap[2], Aditya Patil[3], Smith Shinde[4]
[1234]*Department of Computer Engineering, Cusrow Wadia Institute of Technology, Pune, India*

***Abstract*—Phishing attacks represent one of the most persistent and damaging forms of cybercrime, in which malicious actors impersonate legitimate institutions through fraudulent websites and deceptive email messages to steal sensitive user credentials and financial information. As digital communication continues to expand globally, the frequency and sophistication of these attacks have escalated considerably, overwhelming conventional detection mechanisms that rely on static databases and manually defined rules. This paper presents an Artificial Intelligence-based phishing detection system designed to identify malicious websites and phishing emails with greater accuracy and adaptability than traditional approaches. The proposed system integrates Machine Learning algorithms trained on a diverse feature set extracted from URL structure, domain registration characteristics, and email content metadata. By automating both the feature extraction and classification stages of the detection pipeline, the system can analyze incoming content in real time and produce reliable predictions regarding its legitimacy. Experimental analysis demonstrates that the AI-based approach achieves substantial gains in detection accuracy, precision, and recall compared to blacklist-based and rule-driven baselines, thereby providing organizations and individual users with an intelligent, scalable, and proactive layer of cybersecurity protection.**

***Index Terms*—Phishing Detection, Machine Learning, Malicious URL Analysis, Cybersecurity, Email Classification, Feature Extraction, Artificial Intelligence.**

I. INTRODUCTION

The digital transformation of the global economy has brought with it an unprecedented expansion of online services, ranging from electronic banking and commerce to remote healthcare and government administration. While these developments have created tremendous convenience, they have simultaneously increased the exposure of individuals and institutions to a wide array of cyber threats. Among these, phishing stands out as one of the most damaging and pervasive attack categories.

In a phishing attack, a malicious actor constructs a deceptive environment — most commonly a fake website or a fraudulent email message — designed to convincingly impersonate a trusted entity. Victims who interact with these fabrications are induced to provide confidential information such as login credentials, credit card numbers, bank account details, or social security identifiers, which the attacker subsequently exploits for financial gain or further intrusion.

The mechanisms through which phishing is conducted have grown increasingly sophisticated. Website-based phishing typically involves the registration of domain names that closely resemble those of well-known organizations, combined with the visual cloning of legitimate web pages to create a compelling illusion of authenticity. Email-based phishing employs personalized messaging, spoofed sender addresses, and psychologically manipulative language to prompt hasty action from recipients.

Conventional phishing detection systems have historically depended on blacklisted databases of known malicious domains and URLs, supplemented by rule-based filters that screen for suspicious characteristics in email headers and message bodies. These approaches are inherently reactive, requiring that a threat be identified and catalogued before it can be blocked. Artificial Intelligence and Machine Learning have emerged as promising solutions to these shortcomings. The present paper introduces an AI-based phishing detection system leveraging these capabilities for both malicious websites and phishing emails, demonstrating measurable improvements in detection accuracy, reduced false positive rates, and stronger adaptability compared to traditional methods.

II. LITERATURE REVIEW

Phishing detection has attracted substantial research attention over the past two decades. Early defensive

systems relied primarily on curated blacklists of known phishing domains. Prakash et al. [1] concluded that while blacklist lookups are computationally efficient, they are inherently limited by their retrospective nature. This motivated researchers to pursue proactive, feature-based detection strategies. Garera et al. [2] proposed a logistic regression classifier trained on hand-crafted URL features.

Mohammad et al. [3] conducted a systematic evaluation of multiple classifiers applied to a dataset enriched with website structural features. Random Forest classifiers consistently achieved the highest accuracy, exceeding 97% on benchmark test sets. Basnet et al. [4] demonstrated that Support Vector Machines trained on HTML-derived and hyperlink-based features could generalize well across heterogeneous phishing samples.

Fette et al. [5] introduced PILFER, a machine learning system trained on features derived from email metadata achieving high detection accuracy. Ma et al. [6] proposed an online learning framework for malicious URL detection capable of processing streaming URL data with incremental model updates. Despite these advances, many proposed systems exhibit notable accuracy degradation when confronted with zero-day phishing pages, underscoring the need for a more comprehensive, adaptive, AI-driven detection framework.

V.B Kamble [11] proposed an AI-based approach for enhancing UPI fraud detection, demonstrating the effectiveness of machine learning models in identifying anomalous financial transactions in real-time digital payment environments.

III. PROBLEM STATEMENT

The detection of phishing attacks presents a fundamentally dynamic challenge that static and rule-based systems are structurally ill-equipped to address. Phishing adversaries operate in an environment of continuous innovation, rapidly iterating on their tactics to evade detection tools as they become known. A new phishing domain can be registered, populated with cloned content, and deployed within hours, while existing blacklists and rule databases lag behind by days or longer.

Modern phishing websites are often visually indistinguishable from their legitimate counterparts, employing valid SSL certificates, familiar brand imagery, and URLs that differ from genuine addresses by only a single character or subdomain. Similarly, phishing emails increasingly leverage personalization derived from prior data breaches, rendering generic keyword-based filters ineffective. There is a clear and pressing need for an intelligent, data-driven detection system capable of learning from historical attack patterns and generalizing to novel phishing variants.

IV. OBJECTIVES OF THE STUDY

1. Phishing Website Detection: To develop and evaluate a Machine Learning model capable of accurately classifying web pages as phishing or legitimate based on URL and content-derived features.
2. Phishing Email Identification: To build a classification pipeline for detecting phishing emails through analysis of message metadata, header attributes, and body content features.
3. Feature Engineering and Selection: To identify the most discriminative features from URL structure, domain registration data, HTML content, and email metadata for use in model training.
4. Detection Accuracy Improvement: To demonstrate measurably higher precision, recall, and F1-score compared to conventional blacklist and rule-based approaches.
5. False Positive Rate Reduction: To optimize classifier configurations to minimize erroneous blocking of legitimate websites and emails.
6. Scalable Cybersecurity Framework: To design the system architecture with extensibility in mind, enabling future integration with browsers, email servers, and enterprise security platforms.

V. PROPOSED SYSTEM

The proposed AI-based phishing detection system is designed as a multi-stage pipeline that processes raw input data — in the form of URLs or email messages — through sequential layers of feature extraction, preprocessing, machine learning classification, and result reporting. The architecture is modular, allowing the URL detection subsystem and the email detection subsystem to operate independently or in combination as part of a unified security gateway.

A. Input Layer

The system accepts two categories of input. For website detection, the input is a URL string that may be submitted directly by a user, extracted from an email body, or intercepted from browser activity. For email detection, the input consists of the full email object including the message header, sender metadata, and body content.

B. Feature Extraction Module

For URL-based detection, features are extracted across three categories: (i) lexical features, including URL length, number of special characters, presence of IP addresses, directory path depth, and use of URL-shortening services; (ii) host-based features, including domain age from WHOIS records, DNS resolution behavior, and SSL certificate validity; and (iii) content-based features, including ratio of external hyperlinks, presence of sensitive-input forms, and link destination mismatches.

C. Classification Engine

The preprocessed feature vector is passed to a trained Random Forest classifier, selected for its demonstrated performance in prior phishing detection research and its robustness to noisy or irrelevant features. The classifier outputs a binary prediction — phishing or legitimate — along with a confidence score.

VI. METHODOLOGY

The methodology adopted in this study follows a structured machine learning pipeline encompassing data collection, preprocessing, feature extraction, model training, testing, and performance evaluation.

A. Data Collection

Training and evaluation data were sourced from publicly available phishing and legitimate website datasets. The PhishTank repository and the OpenPhish database provided verified phishing URLs, while the Alexa Top Sites list and the Common Crawl corpus supplied legitimate URL samples. For email classification, the SpamAssassin public corpus and the Nazario Phishing Email Dataset were used. The combined dataset comprised approximately 25,000 labeled samples across both detection tasks.

B. Model Training

Four classifier types were trained and compared: a Decision Tree, a Support Vector Machine with a radial basis function kernel, a Naive Bayes classifier, and a Random Forest ensemble with 200 decision trees. Each model was trained using a stratified 80/20 train-test split with hyperparameters tuned via five-fold cross-validation.

C. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). The false positive rate was specifically monitored, as misclassification of legitimate content carries meaningful operational costs in deployed security environments.

VII. IMPLEMENTATION EXAMPLE

Consider a user who receives an email appearing to originate from a major online banking institution, providing a hyperlink reading: “http://secure-bankofamerica-login.verification-portal.com/account/confirm.” The feature extraction module immediately identifies high-signal lexical features: URL length is 72 characters; two hyphens appear within the domain segment; subdomain depth is three levels; and the domain was registered only four days prior. No valid SSL certificate is found.

The email feature extractor independently identifies the absence of a valid DKIM signature, urgency language in the subject line, a mismatch between the displayed sender name and the actual sending server domain, and a body-to-link ratio indicating the message consists almost entirely of a single call-to-action hyperlink. Of 200 constituent decision trees, 193 vote to classify the input as phishing, yielding a confidence score of 96.5%. The entire analysis pipeline completes in under 180 milliseconds.

VIII. RESULTS AND DISCUSSION

The experimental evaluation produced quantitatively strong results across both the URL-based and email-based classification tasks. For URL-based phishing detection, the Random Forest classifier achieved an overall accuracy of 97.4% on the held-out test set, with precision and recall scores of 97.1% and 97.8% respectively, yielding an F1-score of 97.4%. The AUC-ROC value of 0.991 indicates near-perfect separability between the phishing and legitimate classes.

For email-based phishing detection, the Random Forest model achieved an accuracy of 96.8%, with precision of 96.4% and recall of 97.2%. The false positive rate was 2.9%, a meaningful improvement over the rule-based baseline, which exhibited a false positive rate of 11.3%. In a targeted evaluation using zero-day phishing URLs, the Random Forest classifier correctly identified 89.7% of samples, whereas the blacklist-based reference system detected only 34.2%.

Classifier	Task	Acc (%)	Prec (%)	Recall (%)	F1 (%)
Random Forest	URL	97.4	97.1	97.8	97.4
Decision Tree	URL	94.2	93.8	94.6	94.2
SVM (RBF)	URL	95.6	95.2	96.0	95.6
Naive Bayes	URL	88.3	87.9	88.7	88.3
Random Forest	Email	96.8	96.4	97.2	96.8
Decision Tree	Email	92.1	91.7	92.5	92.1
SVM (RBF)	Email	94.3	94.0	94.6	94.3
Naive Bayes	Email	86.5	86.0	87.0	86.5

TABLE I. Comparative Performance of Classifiers

The system's processing latency averaged 165 milliseconds per sample in a single-threaded test environment, and preliminary parallel processing experiments suggest that throughput can be scaled substantially through batching and multi-core execution. These performance characteristics confirm that the proposed system is operationally viable for real-time deployment in both browser extension and server-side email filtering contexts.

IX. CONCLUSION

Phishing attacks remain a critical and evolving threat in the cybersecurity landscape, exploiting human psychology and the visual complexity of digital interfaces to deceive users into compromising their own security. This paper has presented an AI-based phishing detection system that addresses these limitations through the application of machine learning to a rich set of features derived from URL structure, domain metadata, and email content.

The experimental results demonstrate that the proposed system, anchored by a Random Forest classification engine, achieves high detection accuracy for both malicious websites and phishing emails, with substantially reduced false positive rates compared to conventional blacklist and rule-based approaches. The system's ability to generalize to previously unseen phishing instances — evidenced by strong performance on zero-day URL evaluations — represents a meaningful practical advantage.

X. FUTURE WORK

Several promising directions for future research emerge from the current study. First, the integration of deep learning architectures — particularly Long Short-Term Memory networks and transformer-based language models — may further improve classification performance. Second, real-time browser plugin deployment warrants dedicated investigation. Third, direct integration with organizational email servers would enable automatic pre-delivery scanning of incoming messages. Fourth, the extension of the system to mobile platforms addresses a growing attack vector, as mobile users are increasingly targeted through SMS-based phishing (smishing). Finally, the application of federated learning techniques could enable multiple organizations to collaboratively improve detection models without sharing sensitive raw data.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering for providing the resources and environment necessary to conduct this research. We also acknowledge the open-source repositories — PhishTank, OpenPhish, SpamAssassin, and the Nazario Phishing Email Dataset — whose publicly available datasets made this experimental evaluation possible.

REFERENCES

[1] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks," in Proc. IEEE INFOCOM, San Diego, CA, USA, 2010, pp. 1–5.

[2] S. Garera, N. Provos, M. Chew, and A. D. Rubin, “A Framework for Detection and Measurement of Phishing Attacks,” in Proc. ACM Workshop on Recurring Malcode (WORM), Fairfax, VA, USA, 2007, pp. 1–8.

[3] R. M. Mohammad, F. Thabtah, and L. McCluskey, “Predicting Phishing Websites Based on Self-Structuring Neural Network,” Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, Aug. 2014.

[4] R. Basnet, S. Mukkamala, and A. H. Sung, “Detection of Phishing Attacks: A Machine Learning Approach,” in Soft Computing Applications in Industry, B. Prasad, Ed. Berlin, Germany: Springer, 2008, pp. 373–383.

[5] I. Fette, N. Sadeh, and A. Tomasic, “Learning to Detect Phishing Emails,” in Proc. 16th Int. Conf. World Wide Web (WWW), Banff, AB, Canada, 2007, pp. 649–656.

[6] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,” in Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245–1254.

[7] A. Jain and B. B. Gupta, “Phishing Detection Using Machine Learning Techniques: A Systematic Literature Review and Future Directions,” IEEE Access, vol. 10, pp. 71639–71668, 2022.

[8] Y. Cao, W. Han, and Y. Le, “Anti-Phishing Based on Automated Individual White-List,” in Proc. 4th ACM Workshop on Digital Identity Management, Alexandria, VA, USA, 2008, pp. 51–60.

[9] M. Zareapoor and P. Shamsolmoali, “Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier,” Procedia Computer Science, vol. 48, pp. 679–685, 2015.

[10] V. B. Kamble, “Enhancing UPI Fraud Detection Using Artificial Intelligence and Machine Learning Techniques,” International Journal of Innovative Research in Technology, 2024.