A Comprehensive Review Of Phishing Detection Techniques Based On Machine Learning

Himaniben Bhagirath Pandya[1] , Dr. Khyati Zalawadia[2]

[1]Parul Institute of Engineering & Technology Parul University Vadodara, India.
Himaniben.pandya24277@paruluniversity.ac.in
[2]Parul Institute of Engineering & Technology Parul University Vadodara, India.
Khyati.Zalawadia29490@paruluniversity.ac.in

Abstract-

Phishing attacks continue to pose significant security threats to individuals and organizations worldwide, resulting in financial losses and compromised sensitive information. This comprehensive review examines various machine learning (ML) techniques employed for detecting phishing attempts across multiple vectors, including websites, URLs, and emails. By analysing recent literature, we explore feature selection methodologies, prominent algorithms, dataset characteristics, and performance metrics. Our findings indicate that supervised machine learning approaches, particularly Random Forest and Convolutional Neural Networks, demonstrate superior detection accuracy, often exceeding 97%. Traditional ML algorithms combined with effective feature selection techniques provide practical solutions with reasonable computational requirements, while deep learning approaches offer higher accuracy at the cost of increased complexity. Notable research gaps include limited attention to zero-day attacks, insufficient multimodal phishing detection techniques, and ethical considerations surrounding privacy and consent. This review provides valuable insights for security researchers and practitioners seeking to advance the state-of-the-art in phishing detection through machine learning.

Index Terms—Phishing detection, Machine learning, Deep learning, Random Forest, Convolutional Neural Networks, Feature selection, URL and email security, Supervised learning, Zero-day attacks, Privacy and ethics.

I. Introduction

Phishing attacks represent one of the most prevalent cybersecurity threats in today's digital landscape, targeting millions of users and organizations globally. These attacks typically involve deceptive tactics to trick victims into revealing sensitive information such as login credentials, financial details, and personal data. The dynamic and evolving nature of phishing techniques has made traditional detection methods increasingly ineffective, necessitating more sophisticated approaches to combat this persistent threat.

Machine learning has emerged as a promising solution for combating phishing attacks due to its ability to identify patterns, adapt to new threats, and process vast amounts of data efficiently. The application of ML in phishing detection has gained significant traction in recent years, with researchers exploring various algorithms, feature extraction techniques, and model optimization strategies to enhance detection accuracy and reduce false positives [1]. As phishing attacks grow in sophistication, leveraging intelligent systems capable of recognizing subtle indicators becomes increasingly crucial.

This comprehensive review aims to synthesize current knowledge on ML-based phishing detection techniques, providing researchers and security professionals with a structured understanding of the field.

The paper examines a broad spectrum of approaches, from traditional supervised learning methods to advanced deep learning models, assessing their relative strengths, weaknesses, and practical implications [2]. By analyzing recent advancements and identifying research gaps, we seek to guide future efforts in developing more robust and efficient phishing detection systems.

The review is organized as follows: Section II provides background information on phishing attacks and traditional detection methods. Section III details the methodologies employed in ML-based phishing detection, including feature selection, algorithm selection, and dataset considerations. Section IV presents an analysis of performance metrics and comparative effectiveness of different approaches. Section V identifies research gaps in current literature, while Section VI proposes potential directions for future work. Finally, Section VII concludes the paper with a summary of key findings and implications for the cybersecurity community.

II. Background

Phishing attacks have evolved significantly since their emergence in the mid-1990s, becoming increasingly sophisticated and difficult to detect. These attacks typically involve creating deceptive websites, emails, or messages that mimic legitimate entities to trick users into divulging sensitive information. The term "phishing" itself draws an analogy to fishing, where attackers cast out bait (deceptive communications) to catch unsuspecting victims [3].

Several distinct categories of phishing attacks have been identified in the literature. URL-based phishing involves creating fraudulent websites with URLs that closely resemble legitimate domains but contain subtle modifications [4]. Email phishing targets victims through deceptive messages containing malicious links or attachments. Website phishing incorporates visual and structural elements that mimic trusted platforms to establish credibility with potential victims. Additionally, spear phishing represents a more targeted approach, where attacks are customized for specific individuals or organizations based on gathered intelligence.

Traditional detection methods initially relied on static approaches such as blacklisting known phishing domains, which proved inadequate due to the short lifespan of phishing websites and the constant creation of new domains [5]. Heuristic-based approaches emerged as an improvement, examining website characteristics such as URL structure, domain age, and HTML content to identify suspicious sites. However, these rule-based systems often struggle to adapt to novel phishing techniques and frequently generate false positives.

The limitations of traditional approaches have led to increasing interest in more adaptive and intelligent detection methods. Machine learning offers significant advantages in this context, including the ability to identify complex patterns across multiple features, adapt to evolving threats through retraining, and process high-dimensional data efficiently [2]. Early ML applications in phishing detection primarily utilized supervised learning approaches with manually engineered features, while recent advances have incorporated deep learning techniques capable of automatic feature extraction from raw data.

The progression from static lists to heuristic rules and ultimately to ML-based detection represents a natural evolution in response to increasingly sophisticated phishing tactics. As attackers continue to refine their methods, detection techniques must similarly advance, prompting the current focus on optimizing and enhancing ML approaches for phishing identification. Understanding this historical context is essential for appreciating the current state of research and identifying promising directions for future work in the field.

III. Methodologies

The application of machine learning to phishing detection encompasses diverse methodological approaches, each with distinct characteristics and implementation considerations. This section examines the key methodological components of ML-based phishing detection systems, including feature extraction and selection, algorithm selection, dataset characteristics, and evaluation metrics.

a. Feature Engineering and Selection

Feature selection represents a critical aspect of phishing detection, as the quality and relevance of features significantly impact model performance. Research by Mewada and Dewang introduced a novel feature selection method to extract highly correlated features from phishing datasets, enhancing classifier accuracy [1]. Common features extracted for URL-based phishing detection include lexical characteristics (URL length, special character count), domain-specific attributes (domain age, WHOIS information), HTML and JavaScript features, and host-based information (IP address, geographic location) [3]. For email phishing, features typically include header information, linguistic characteristics, and attachment properties [4].

Several feature selection techniques have been employed to identify optimal feature subsets. Chi-squared (Chi-2) testing has shown promising results when combined with Random Forest classifiers, achieving accuracy up to 96.99% [1]. The Optimal Feature Vectorization Algorithm (OFVA) extracts intra-URL features from raw URLs, providing a robust feature set for classification tasks [3]. Studies indicate that feature selection not only improves model accuracy but also reduces computational overhead and training time, which is particularly important for real-time detection systems.

b. Machine Learning Algorithms

A wide range of ML algorithms have been applied to phishing detection, with supervised learning approaches predominating in current research. According to a systematic literature review analyzing 80 scientific papers, Random Forest classifiers are the most commonly employed algorithm (31 studies), followed by Support Vector Machines (SVM), Logistic Regression, and Decision Trees [2]. These traditional ML approaches demonstrate strong performance when coupled with effective feature selection.

Deep learning techniques have gained increasing attention, with Convolutional Neural Networks (CNNs) showing particularly impressive results. In some studies, CNNs achieved up to 99.98% accuracy in detecting phishing websites [2]. The inherent ability of deep learning models to automatically extract hierarchical features from raw data offers advantages for processing complex phishing indicators, though at the cost of increased computational requirements and reduced interpretability [5].

Recent research has also explored ensemble methods and hybrid approaches that combine multiple algorithms to enhance detection capabilities. Techniques such as bagging, boosting, and stacking have shown promising results, leveraging the complementary strengths of different classifiers to improve overall system performance [3]. The implementation of these ensemble methods often provides a balance between accuracy and computational efficiency, making them suitable for various deployment scenarios.

c. Datasets and Evaluation

Dataset quality and composition significantly influence the development and evaluation of phishing detection models. Research indicates that PhishTank is the most commonly used source for phishing datasets (53 studies), while Alexa is frequently employed for legitimate website data (29 studies) [2]. The size and diversity of training datasets vary considerably across studies, from small collections of a few thousand samples to comprehensive datasets containing hundreds of thousands of instances [3].

Evaluation metrics typically include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). While accuracy remains the most commonly reported metric, researchers increasingly recognize the importance of balanced evaluation approaches that consider both false positives and false negatives [6]. Cross-validation techniques, particularly k-fold cross-validation, are widely employed to assess model generalizability and prevent overfitting [4].

Recent studies have employed increasingly large datasets to improve model robustness. For instance, research by Ali et al. utilized a dataset of 274,446 URLs (134,500 phishing and 139,946 legitimate) to train and evaluate their models [3]. This trend toward larger, more diverse datasets reflects the understanding that comprehensive training data is essential for developing models capable of detecting the wide variety of phishing techniques employed by attackers.

IV. Analysis

This section provides a comparative analysis of various machine learning approaches for phishing detection, examining performance metrics, strengths, weaknesses, and implementation considerations across different techniques.

a. Performance Comparison

Performance metrics reveal significant variations across different ML approaches for phishing detection. Traditional supervised learning methods demonstrate strong overall performance, with Random Forest classifiers consistently achieving accuracy rates between 96-97% when combined with appropriate feature selection techniques [1]. Support Vector Machines (SVM) and Logistic Regression also perform well, though typically with slightly lower accuracy compared to ensemble methods like Random Forest [7].

Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have achieved the highest reported accuracy rates in some studies, reaching up to 99.98% for website phishing detection [2]. This superior performance can be attributed to CNN's ability to automatically extract hierarchical features from raw data, capturing subtle patterns that might be missed by traditional feature engineering approaches. However, this advantage comes with increased computational requirements and reduced model interpretability [8].

Heuristic-based machine learning approaches demonstrate promising results across different phishing vectors. Recent research reports accuracy rates of 97.2% for URL phishing detection, 97.4% for email phishing detection, and 98.1% for website phishing detection [4]. These results suggest that tailoring ML approaches to specific phishing vectors may enhance detection performance compared to generic models. The effectiveness of these specialized approaches highlights the importance of considering the unique characteristics of different phishing channels.

Feature selection methods significantly impact model performance. Studies indicate that Chi-squared testing combined with Random Forest classifiers improves accuracy by up to 5% compared to using the full feature set [1]. The Optimal Feature Vectorization Algorithm (OFVA) has also demonstrated effectiveness in enhancing model performance through the extraction of relevant features from raw URLs [3]. These findings underscore the critical role of feature engineering in building effective phishing detection systems.

b. Trade-offs and Considerations

Several important trade-offs emerge when comparing different ML approaches for phishing detection. Traditional supervised learning methods offer advantages in terms of computational efficiency, interpretability, and ease of implementation. These qualities make them suitable for deployment in resource-constrained environments and scenarios where understanding model decisions is important for security analysts [9].

Deep learning approaches, while potentially more accurate, require substantial computational resources for training and deployment. This limitation may prevent their adoption in certain contexts, such as browser-based detection or mobile applications. Additionally, the "black box" nature of deep learning models presents challenges for explaining detection decisions, which may be problematic in security-critical contexts where transparency is valued [8].

Real-time detection capabilities represent another important consideration. Models with lower computational complexity can make faster predictions, enabling real-time protection against phishing attempts. Research suggests that optimized Random Forest implementations and lightweight neural network architectures offer a good balance between accuracy and prediction speed [3]. This balance is particularly important for practical deployment scenarios where immediate detection is necessary to prevent user interaction with phishing content.

Dataset characteristics significantly influence model performance and generalizability. Models trained on larger, more diverse datasets typically demonstrate better performance on novel phishing attacks [3]. However, the rapid evolution of phishing techniques means that even models trained on comprehensive datasets may struggle with zero-day attacks that employ previously unseen patterns or tactics [10]. This challenge highlights the need for continuous model updating and adaptation in operational settings.

The analysis of current literature reveals that no single ML approach universally outperforms all others across all contexts and evaluation metrics. Rather, the optimal approach depends on specific requirements, constraints, and objectives of the detection system. This observation highlights the importance of carefully considering various factors when designing and implementing ML-based phishing detection solutions.

V. Research Gaps

Despite significant advances in ML-based phishing detection, several important research gaps remain unaddressed in the current literature. Identifying these gaps is crucial for guiding future research efforts and enhancing the effectiveness of phishing detection systems.

a. Limited Focus on Zero-Day Attack Detection

A significant limitation of existing research is the inadequate attention to detecting zero-day phishing attacks—those employing previously unseen techniques or patterns. Most studies evaluate models using historical datasets, which may not reflect the dynamic and evolving nature of phishing tactics [3]. While high accuracy rates are reported for known attack patterns, there is limited evidence regarding how well these models generalize to novel phishing approaches [8]. Models that perform well on established datasets may fail when confronted with innovative attack strategies, creating a critical vulnerability in real-world deployment scenarios.

b. Insufficient Multimodal Approaches

Current research predominantly focuses on single-vector phishing detection (URL, email, or website), with limited exploration of integrated approaches that consider multiple attack vectors simultaneously [4]. Sophisticated phishing campaigns often employ multiple channels, creating a coordinated attack that may evade detection systems focused on a single vector. The lack of multimodal approaches represents a significant gap in the literature, potentially leaving detection systems vulnerable to complex, multi-channel phishing campaigns [7]. This siloed approach fails to capture the interconnected nature of modern phishing attacks, which frequently leverage multiple communication channels and technical vulnerabilities.

c. Underexploration of Contextual Factors

Most existing models neglect contextual factors that might influence the likelihood and nature of phishing attacks, such as user behavior patterns, organizational context, or temporal trends. This gap limits the ability of detection systems to adapt to specific contexts and user profiles, potentially resulting in higher false positive rates or missed detections [3]. Contextual intelligence could significantly enhance detection performance by incorporating factors such as typical user communication patterns, organizational email practices, and seasonal variations in phishing tactics. The absence of such contextual awareness represents a missed opportunity for improving detection accuracy and reducing false alarms.

d. Limited Attention to Adversarial Attacks

The vulnerability of ML models to adversarial attacks—deliberate manipulations designed to evade detection—remains underexplored in phishing detection literature. As ML-based detection becomes more widespread, sophisticated attackers may employ adversarial techniques to circumvent these systems [2]. Few studies systematically evaluate the robustness of phishing detection models against adversarial manipulations or propose defense mechanisms to mitigate such threats [10]. This gap represents a potential vulnerability that could undermine the effectiveness of ML-based detection in practical deployments, particularly as attackers become more sophisticated in their evasion techniques.

e. Ethical and Privacy Considerations

While some research acknowledges the importance of ethical considerations in phishing detection, comprehensive frameworks for addressing privacy concerns, data protection requirements, and potential biases in ML models remain limited [4]. The collection and analysis of user data for phishing detection raises important questions regarding consent, data minimization, and compliance with regulations such as GDPR. Future research should more explicitly address these ethical dimensions, developing approaches that balance effective detection with privacy preservation and ethical data usage [6]. The absence of well-defined ethical frameworks may impede the adoption of ML-based phishing detection in privacy-sensitive contexts.

f. Insufficient Standardization

The lack of standardized evaluation methodologies, datasets, and metrics makes it difficult to directly compare different approaches and assess progress in the field. Studies employ diverse datasets, feature sets, and evaluation protocols, complicating the interpretation of reported performance metrics [2]. This heterogeneity makes it challenging to determine which approaches truly represent advances in the state-of-the-art and which merely reflect differences in evaluation methodology. Establishing standardized benchmarks and evaluation frameworks would facilitate more meaningful comparisons between approaches and accelerate progress in the field.

VI. Future Work

Based on the identified research gaps, several promising directions for future work emerge. Addressing these areas could significantly advance the field of ML-based phishing detection and enhance the effectiveness of defensive measures against increasingly sophisticated phishing attacks.

a. Adaptive Learning for Zero-Day Attack Detection

Future research should focus on developing adaptive learning frameworks capable of detecting previously unseen phishing patterns. This may involve semi-supervised or unsupervised learning approaches that can identify anomalous behaviors without requiring examples of specific attack vectors [3]. Continuous learning systems that incrementally update their knowledge based on new observations could help address the evolving nature of phishing tactics [8]. Additionally, transfer learning techniques might enable models to leverage knowledge gained from known attack patterns to identify conceptually similar but previously unseen phishing attempts. Such approaches would enhance resilience against zero-day attacks, addressing a critical limitation of current systems.

b. Integrated Multimodal Detection Frameworks

Developing comprehensive detection frameworks that simultaneously analyze multiple phishing vectors represents an important direction for future work. Such frameworks could integrate signals from URLs, email content, website characteristics, and network behavior to provide holistic protection against sophisticated phishing campaigns [4]. Research should explore effective architectures for combining these diverse data sources, potentially through multi-branch neural networks, ensemble methods, or

graph-based approaches that capture relationships between different attack components [7]. These integrated approaches would better reflect the complex, multi-channel nature of advanced phishing campaigns and provide more comprehensive protection against coordinated attacks.

c. Context-Aware Phishing Detection

Incorporating contextual intelligence into phishing detection systems offers significant potential for enhancing performance and reducing false positives. Future work should explore methods for modeling user behavior patterns, organizational context, and temporal factors that influence phishing susceptibility and characteristics [3]. This might involve personalized detection models that adapt to individual user profiles, domain-specific models tailored to particular organizational contexts, or dynamic models that adjust their parameters based on emerging trends and patterns [9]. Context-aware approaches could enable more nuanced and accurate phishing detection across diverse scenarios while reducing the burden of false alarms on users and security teams.

d. Adversarial Robustness in Phishing Detection

As ML-based detection becomes more widespread, research should systematically address the vulnerability of these systems to adversarial attacks. This includes developing methods for evaluating model robustness against realistic adversarial manipulations, designing architectures inherently resistant to such attacks, and creating defensive mechanisms that can detect and neutralize adversarial attempts [2]. Techniques such as adversarial training, robust optimization, and ensemble defenses could enhance the resilience of phishing detection systems against sophisticated evasion tactics [10]. This research direction is crucial for ensuring the practical effectiveness of ML-based approaches in adversarial security contexts where attackers actively attempt to circumvent detection.

e. Privacy-Preserving Phishing Detection

Future work should explore privacy-preserving techniques that enable effective phishing detection while minimizing privacy risks and ensuring compliance with data protection regulations. This might involve federated learning approaches that keep sensitive data on user devices, differential privacy mechanisms that provide formal privacy guarantees, or privacy-by-design frameworks that minimize data collection and retention [4]. Research should also address potential biases in training data and detection algorithms that might disproportionately affect certain user groups. Developing ethical guidelines and best practices for ML-based phishing detection represents an important complementary direction for ensuring responsible deployment of these technologies.

f. Standardized Evaluation Frameworks

Establishing standardized benchmarks, datasets, and evaluation protocols would facilitate more meaningful comparisons between different approaches and accelerate progress in the field. Future work should focus on developing comprehensive evaluation frameworks that assess multiple performance dimensions, including accuracy across diverse phishing vectors, robustness against adversarial manipulations, computational efficiency, and false positive rates in realistic deployment scenarios [2]. Creating continuously updated benchmark datasets that reflect emerging phishing tactics would help ensure that reported performance metrics remain relevant to current threats and provide a foundation for meaningful comparison between different detection approaches.

VII. Conclusion

This comprehensive review has examined the application of machine learning techniques to phishing detection, synthesizing findings from recent literature and identifying important research directions. The analysis reveals significant progress in developing effective ML-based approaches for identifying phishing attempts across various vectors, including websites, URLs, and emails.

Several key findings emerge from this review. First, supervised machine learning approaches, particularly Random Forest classifiers and Convolutional Neural Networks, demonstrate strong performance in phishing detection tasks, with accuracy rates frequently exceeding 97% [1] [2]. The effectiveness of these approaches depends significantly on feature selection and engineering, with optimal feature sets substantially improving model performance. Second, the integration of heuristic approaches with machine learning techniques offers promising results for specific phishing vectors, enabling more targeted and effective detection strategies [4].

Despite these advances, important challenges remain. The dynamic and evolving nature of phishing attacks presents ongoing difficulties for detection systems, particularly regarding zero-day attacks employing previously unseen techniques [3]. Additionally, the fragmented approach to research—focusing on individual phishing vectors rather than integrated, multimodal detection—limits the ability to address sophisticated phishing campaigns that leverage multiple channels simultaneously [4]. Ethical and privacy considerations also require greater attention, ensuring that detection systems comply with regulatory requirements and respect user privacy [6].

Future research should address these challenges through several approaches: developing adaptive learning frameworks capable of detecting novel phishing patterns, creating integrated multimodal detection systems, incorporating contextual intelligence into detection models, enhancing adversarial robustness, implementing privacy-preserving techniques, and establishing standardized evaluation frameworks. Progress in these areas would significantly advance the state-of-the-art in phishing detection and enhance defensive capabilities against this persistent threat.

The continued evolution of ML-based phishing detection represents an important front in the ongoing battle against cybersecurity threats. By leveraging advances in machine learning while addressing current limitations and research gaps, security researchers and practitioners can develop more effective countermeasures against increasingly sophisticated phishing attacks. This review provides a foundation for such efforts, offering a structured analysis of current approaches and identifying promising directions for future work in this critical domain.

VIII. References

[1] M. Mewada and S. Dewang, "An approach for efficient and accurate phishing website prediction using improved ML classifier performance for feature selection," IEEE Transactions on Information Forensics and Security, vol. 19, no. 1, pp. 2364–2378, Jun. 2024.

[2] A. Kumar, P. Singh, and R. Kumar, "A systematic literature review on phishing website detection techniques," Journal of Cybersecurity and Privacy, vol. 3, no. 1, pp. 55–79, Feb. 2023.

[3] S. Ali, M. Ahmed, and K. Khan, "Unveiling suspicious phishing attacks: Enhancing detection with an optimal feature vectorization algorithm and supervised machine learning," Frontiers in Computer Science, vol. 6, pp. 1–18, Jul. 2024.

[4] V. Rajan, S. Kumar, and P. Mehta, "Heuristic machine learning approaches for identifying phishing threats across web and email platforms," Frontiers in Artificial Intelligence, vol. 7, no. 2, Art. no. 1414122, Oct. 2024.

[5] S. Salihovic, J. Chao, and M. Chen, "A comprehensive review of phishing attack detection using machine learning techniques," IEEE Access, vol. 12, pp. 114562–114590, Oct. 2024.

[6] L. Wu, X. Du, and J. Wu, "Effective machine learning based detection of phishing websites," Journal of Information Security and Applications, vol. 60, no. 2, Art. no. 102866, Feb. 2021.

[7] G. Tariq, A. Almomani, and A. Sultan, "Phishing website detection using convolutional neural networks with minimal feature engineering," Computer Networks, vol. 218, Art. no. 109418, Apr. 2023.

[8] F. Nagaraj, P. Rao, and A. Pais, "PhishGuard: A deep learning approach for zero-day phishing attack detection," Security and Communication Networks, vol. 2022, Art. no. 1–15, Mar. 2022.

[9] Y. Sahingoz, A. Buber, and O. Demir, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, Mar. 2019.

[10] H. Aldakheel, A. Adebowale, and K. Nagaraj, "DLPhish: A novel deep learning framework for effective phishing webpage detection," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 5, pp. 4456–4470, Sep. 2023.