

Malicious Webpage Detection Using Machine Learning: A Comprehensive Review

Mayank Singh, Md Absarul Haque, Dr. InderPreet Kaur

Sharda School of Engineering and Technology, Sharda University, Greater Noida, India

Abstract- Traditional, signature-based methods for detecting malicious webpages, such as blacklisting, are increasingly ineffective against the dynamic and sophisticated nature of modern web threats. These methods struggle to identify newly generated or obfuscated malicious URLs, creating a significant security gap. This review paper provides a comprehensive analysis of machine learning (ML) based approaches, which have emerged as a powerful alternative for proactive and accurate malicious webpage detection. We explore the evolution of these systems, from classical algorithms like Random Forest and Support Vector Machines to advanced deep learning models that provide robust, real-time classification. This paper surveys the critical role of feature engineering, categorizing features into lexical, content-based, behavioral, and network-related types, and examines their impact on model performance. A significant portion of this review is dedicated to the critical challenges of model generalizability across different datasets, class imbalance, and the growing threat of adversarial attacks designed to evade ML-based detectors. We synthesize findings from a broad range of studies to identify the current state-of-the-art, pinpoint existing research gaps, and suggest future directions for developing more effective, scalable, and adversarially robust detection systems. A conceptual hybrid framework is presented as a next-generation solution that addresses many of the limitations of existing single-model systems.

Keywords: malicious webpage detection, machine learning, phishing detection, URL classification, feature engineering, deep learning, adversarial attacks, random forest, support vector machine, cybersecurity, web security, convolutional neural network, LSTM, ensemble methods, blacklist evasion

I. INTRODUCTION

The exponential growth of the internet has made it an indispensable tool for communication, commerce, and information access. However, this proliferation has been accompanied by a surge in web-based cyberattacks, including phishing, malware distribution, and drive-by-downloads [1]. Malicious webpages serve as the primary vector for these attacks, posing a significant threat to user security and privacy.

Historically, the primary defense mechanism has been the use of blacklists—curated lists of known malicious URLs. While simple to implement, blacklist-based methods are inherently reactive. They are limited in their ability to detect zero-day threats, newly generated malicious URLs, or pages using cloaking techniques [2]. This limitation has driven the research community to seek more dynamic and intelligent solutions.

The application of machine learning (ML) has marked a paradigm shift in this domain [3]. ML models can learn to distinguish between benign and malicious webpages by analyzing a wide array of features, enabling the detection of novel threats without prior knowledge [4]. Early systems utilized classical ML algorithms, which proved significantly more effective than traditional methods. More recently, deep learning and ensemble techniques have further advanced the field, achieving exceptionally high accuracy rates [5]. These advanced systems leverage complex feature sets—ranging from URL lexical patterns to webpage content and network traffic data—to build robust and adaptable detection models [6].

This review paper aims to provide a comprehensive overview of the field of ML-based malicious webpage detection. We will examine the key methodologies, feature engineering techniques, and prevalent challenges, positioning a proposed hybrid framework within this evolving landscape.

II. METHODOLOGY FOR LITERATURE SEARCH

A systematic literature search was conducted to identify relevant studies on machine learning for malicious webpage detection. The search was performed across several academic databases, including IEEE Xplore, ACM Digital Library, ArXiv,

and Google Scholar, supplemented by the Consensus search engine which covers sources like Semantic Scholar and PubMed [7].

The search terms included, but were not limited to: "malicious webpage detection," "phishing detection using machine learning," "malicious URL detection," "feature engineering for web security," "deep learning for cybersecurity," and "adversarial attacks on web classifiers."

The following inclusion criteria were applied to screen the articles for relevance:

- The study had to be published in English.
- The study had to focus on detecting malicious webpages or URLs using an ML or deep learning component.
- The study had to address one or more of the following: feature engineering, model architecture, comparative analysis, real-world deployment, or adversarial robustness.

The references of selected articles were also scanned to identify additional relevant studies. This iterative process resulted in a comprehensive collection of over 60 papers that form the basis of this review. The final selection was then categorized based on primary focus to facilitate a structured analysis.

III. COMPARATIVE ANALYSIS OF FEATURE ENGINEERING TECHNIQUES

The performance of any ML-based detection system is critically dependent on the quality and relevance of the features extracted from the webpage or its URL [8]. These features can be broadly categorized into four types.

- **Lexical Features:** These features are extracted directly from the URL string without accessing the webpage's content. They are computationally inexpensive and effective for real-time analysis. Common lexical features include URL length, number of dots, presence of special characters ('@', '-'), use of IP addresses in the domain name, and characteristics of the hostname and path [9].
- **Content-Based Features:** These features are derived from the raw HTML, JavaScript, and text of a webpage. They provide deep insight into the page's purpose and functionality. Examples include the use of suspicious HTML tags (e.g., <iframe>), obfuscated JavaScript code, keyword frequency analysis using TF-IDF or Word2Vec, and the number of external links [10]. While powerful, these features are more resource-intensive to extract.
- **Behavioral/Network Features:** These features relate to the hosting environment and network interactions of the webpage. They include domain registration information (e.g., domain age from WHOIS records), IP address reputation, geographical location of the server, DNS records, and HTTPS certificate details [11]. These features are highly effective at identifying transient or newly set-up malicious domains.
- **Visual Features:** A more recent approach involves using computer vision to analyze screenshots of webpages. Techniques like perceptual hashing or CNN-based feature extraction can identify visual similarities between a suspicious page and a legitimate one, which is particularly effective for detecting phishing attacks that clone popular websites.

Recent studies consistently demonstrate that combining multiple feature types leads to the most robust and accurate models, as it provides a more holistic view of the potential threat [12].

IV. MACHINE LEARNING MODELS FOR MALICIOUS WEBPAGE DETECTION

A. Classical Machine Learning Models

A wide range of classical supervised algorithms have been successfully applied to this problem. Commonly used models include Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), Decision Trees, Logistic Regression, and Naïve Bayes [13].

Comparative studies have repeatedly shown that ensemble methods, particularly Random Forest, consistently achieve high accuracy and are robust to noisy data [14]. Their ability to handle high-dimensional feature spaces makes them well-suited for this task.

B. Deep Learning Models

With the availability of large datasets, deep learning has emerged as a state-of-the-art approach. Deep Neural Networks (DNNs) can automatically learn complex patterns and feature interactions from raw data, often outperforming classical models [15].

Specific architectures have been adapted for different feature types: Convolutional Neural Networks (CNNs) are used for analyzing webpage screenshots (visual features), while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are effective for processing sequential data like URL strings and HTML code.

C. Ensemble and Multi-Modal Approaches

To further enhance performance, researchers have focused on ensemble and multi-modal frameworks. Ensemble methods combine the predictions of multiple individual models to improve overall accuracy and reduce variance [16]. Multi-modal deep learning integrates different types of features (e.g., lexical, content, and visual) into a unified model, allowing it to capture a more comprehensive representation of the webpage and achieve superior detection rates [17]. These sophisticated approaches often report accuracies exceeding 98% on benchmark datasets [18].

V. CHALLENGES AND ADVERSARIAL THREATS

Despite high reported accuracies, several significant challenges persist in the field.

- **Model Generalizability:** A major issue is the poor generalizability of models across different datasets [19]. A model trained on a specific data distribution may underperform significantly when deployed in a new environment with different web traffic patterns or attack types [20]. This highlights the need for more diverse and standardized public datasets for benchmarking [21].
- **Class Imbalance:** In the real world, benign webpages vastly outnumber malicious ones. This severe class imbalance can bias ML models towards the majority class, leading to a high number of false negatives (missed threats), which is a critical security failure [22]. Techniques like oversampling (e.g., SMOTE) or undersampling are required to mitigate this issue.
- **Adversarial Attacks:** As ML-based detectors become more common, attackers are actively developing techniques to evade them. Adversarial attacks involve creating carefully crafted, malicious webpages that are misclassified as benign by the model [23]. These "adversarial examples" often involve subtle modifications to the webpage's features that are imperceptible to humans but sufficient to fool the classifier [24]. Research into adversarial training and more robust model architectures is crucial to counter this evolving threat.
- **Concept Drift:** The web landscape is constantly changing, with new attack vectors and obfuscation techniques emerging daily. This phenomenon, known as "concept drift," can render static ML models obsolete over time. Continuous monitoring and periodic retraining are necessary to maintain high detection accuracy.

VI. REAL-WORLD DEPLOYMENT AND PERFORMANCE METRICS

Deploying ML models in a real-world setting introduces practical challenges related to scalability and performance. For systems that analyze webpage content, the latency of feature extraction and classification is a major concern, as it directly impacts the user's browsing experience [25]. Therefore, many practical solutions, such as browser extensions or DNS firewalls, use a tiered approach where lightweight lexical models are used for initial, real-time screening, followed by more intensive content analysis for suspicious cases [26].

Key performance metrics go beyond simple accuracy. The False Positive Rate (FPR) is critical, as incorrectly blocking legitimate websites can be highly disruptive. The False Negative Rate (FNR) represents missed threats and is a direct measure of the system's security effectiveness. A balance between these metrics, often captured by the F1-Score, is essential for a practical and reliable detection system.

VII. A PROPOSED HYBRID DETECTION FRAMEWORK

To address the limitations of existing systems, we propose a conceptual hybrid framework designed for robustness, scalability, and adaptability.

A. System Architecture

The proposed system has a multi-layered architecture:

- **Data Collection Layer:** This layer uses web crawlers to gather URLs from various sources, including email spam traps, social media feeds, and web traffic logs.
- **Feature Extraction Layer:** This layer operates in parallel to extract lexical, content-based, and network features. A lightweight module extracts lexical features in real-time, while a more comprehensive module performs deeper content and network analysis asynchronously.
- **Hybrid Detection Layer:** This is the core of the framework. It employs a two-stage detection process. Stage 1 (Rapid Filtering): A highly optimized Random Forest model uses lexical features to quickly classify the majority of incoming URLs. URLs classified as clearly benign or malicious are finalized. Stage 2 (Deep Analysis): URLs flagged as suspicious or uncertain are passed to a multi-modal deep learning model that integrates content, network, and visual features for a more definitive classification. This ensemble approach improves accuracy while managing computational load [27].
- **Adaptive Learning Layer:** The framework incorporates an online learning module. It periodically retrains the models on newly labeled data and includes adversarial training techniques to enhance robustness against evasion attacks.

VIII. CHALLENGES AND FUTURE DIRECTIONS

Despite significant progress, several challenges remain. The primary areas for future research are:

- **Adversarial Robustness:** Developing intrinsically robust models against adversarial attacks is the most pressing challenge. Future work should focus on adversarial training, certified defenses, and anomaly detection to identify manipulated inputs [28].
- **Cross-Dataset Generalization:** There is a critical need for techniques that improve model generalizability. Transfer learning and domain adaptation could help models adapt to new web environments with minimal retraining [29].
- **Explainable AI (XAI):** Most advanced ML models operate as "black boxes." Incorporating XAI can provide insights into why a webpage is flagged as malicious, which is valuable for security analysts and for reducing model bias.
- **Standardized Benchmarking:** The field lacks large-scale, diverse, and publicly available datasets for standardized benchmarking [30]. The creation of such resources would significantly accelerate research and allow for more meaningful comparison of different approaches.

IX. CONCLUSION

Machine learning has fundamentally transformed the approach to malicious webpage detection, offering a dynamic and proactive defense mechanism that far surpasses traditional blacklist-based methods [31]. High-performing systems based on deep learning and ensemble models, leveraging a rich fusion of features, can achieve outstanding accuracy. However, the path

to deploying truly effective real-world solutions is fraught with challenges, most notably the lack of generalizability and the persistent threat of adversarial evasion [32].

The future of web security will depend on the development of the next generation of ML systems that are not only accurate but also robust, adaptable, and scalable. By addressing the open research questions surrounding adversarial defense, generalization, and explainability, the research community can build a more resilient and secure web infrastructure for all users.

REFERENCES

- [1] Aljabri, M., et al. (2022). Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions. IEEE Access.
- [2] Sahoo, D., Liu, C., & Hoi, S. (2017). Malicious URL Detection using Machine Learning: A Survey. ArXiv.
- [3] Reyes-Dorta, N., et al. (2024). Detection of malicious URLs using machine learning. Wireless Networks.
- [4] Liaquathali, S., & Kadirvelu, V. (2025). Integration of natural language processing methods and machine learning model for malicious webpage detection. IAES International Journal of Robotics and Automation.
- [5] Vanhoenshoven, F., et al. (2016). Detecting malicious URLs using machine learning techniques. IEEE Symposium Series on Computational Intelligence (SSCI).
- [6] Kazemian, H., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications.
- [7] Shin, S., Ji, S., & Hong, S. (2022). A Heterogeneous Machine Learning Ensemble Framework for Malicious Webpage Detection. Applied Sciences.
- [8] Hani, R., et al. (2024). Malicious URL Detection Using Machine Learning. 15th International Conference on Information and Communication Systems (ICICS).
- [9] Hou, Y., et al. (2010). Malicious web content detection by machine learning. Expert Systems with Applications.
- [10] Oshingbesan, A., et al. (2022). Detection of Malicious Websites Using Machine Learning Techniques. ArXiv.
- [11] Tabassum, T., et al. (2023). A Review on Malicious URLs Detection Using Machine Learning Methods. Journal of Engineering Research and Reports.
- [12] Nguyen, L. A. T., et al. (2020). PSI-rooted subgraph: A novel feature for IoT botnet detection using classifier algorithms. IEEE Access.
- [13] Buber, E., Diri, B., & Sahingoz, O. K. (2017). NLP based phishing attack detection from URLs. International Conference on Computer Science and Engineering (UBMK).
- [14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
- [15] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
- [16] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
- [17] Tian, K., Jan, S., Hu, H., Yao, D., & Wang, G. (2018). Needle in a haystack: Tracking down elite phishing domains in the wild. ACM Internet Measurement Conference.

- [18] Rao, R. S., et al. (2019). Jail-Phish: An improved search engine based phishing detection system. *Computers & Security*.
- [19] Marchal, S., et al. (2016). Know your phish: Novel techniques for detecting phishing sites and their targets. *IEEE ICDCS*.
- [20] Liang, G., et al. (2016). Cracking classifiers for evasion: A case study on the Google's phishing pages filter. *WWW 2016*.
- [21] Ma, J., et al. (2009). Identifying suspicious URLs: an application of large-scale online learning. *ICML*.
- [22] Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- [23] Goodfellow, I. J., et al. (2015). Explaining and harnessing adversarial examples. *ICLR*.
- [24] Corona, I., et al. (2017). DeltaPhish: Detecting phishing webpages in compromised websites. *ESORICS*.
- [25] Stringhini, G., et al. (2013). Shady paths: Leveraging surfing crowds to detect malicious web pages. *ACM CCS*.
- [26] Cova, M., Kruegel, C., & Vigna, G. (2010). Detection and analysis of drive-by-download attacks and malicious JavaScript code. *WWW 2010*.
- [27] Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169–198.
- [28] Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. *ICLR*.
- [29] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.
- [30] Patil, D. R., & Patil, J. B. (2018). Malicious web pages detection using feature selection techniques and machine learning. *International Journal of Information Technology*.
- [31] Sahingoz, O. K., et al. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*.
- [32] Montazer, G. A., & ArabYarmohammadi, S. (2015). Detection of phishing attacks in Iranian e-banking using a fuzzy-rough hybrid system. *Applied Soft Computing*.