International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

AI-BASED PHISHING WEBSITE DETECTION

Ravi Kumar Mahone[1], Nishanshu Deshmukh[2], Nishant Sahu[3], Sahil Darokar[4], Sangam jain[5]

Abstract:
Phishing remains one of the most persistent and detrimental cyber threats in the modern digital landscape. Traditional defence mechanisms, such as blacklists and manual reporting, are critically ineffective against the rapid evolution and deployment of sophisticated, zero-day phishing sites. This review paper explores the transition to Artificial Intelligence (AI) and Machine Learning (ML) as a robust, adaptive solution to this problem. We outline a systematic methodology involving feature engineering—analysing URL characteristics, domain metadata, and page content—followed by the application of classification algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM). Through critical analysis, we compare the performance trade-offs between simple classifiers and ensemble methods, highlighting the superior accuracy and generalisation capability of ensemble models like Random Forest. The paper provides an extensive survey of existing literature, tracing the evolution from simple ML techniques to state-of-the-art Deep Learning and hybrid systems. Finally, we identify key research gaps, notably model vulnerability to adversarial attacks and the need for standardised, real-time deployment strategies, outlining the future scope for developing continuous learning and hybrid detection mechanisms.

Keywords: Phishing detection, Artificial Intelligence (AI), Machine Learning (ML), Feature engineering, URL analysis, Domain metadata, Page content analysis, Classification algorithms, Decision Tree, Random Forest, Support Vector Machine (SVM).

1. INTRODUCTION

In an era defined by global connectivity, the internet has become an indispensable tool for commerce, communication, and governance. This dependency, however, exposes users and organisations to increasingly complex cyber threats. Among these, phishing—the fraudulent attempt to obtain sensitive information like usernames, passwords, and credit card details by disguising itself as a trustworthy entity—stands out as a primary vector for identity theft and massive financial loss. The sheer volume and increasing sophistication of phishing attacks, often mimicking legitimate brand identity with near-perfect replication, necessitate a fundamental shift in defensive strategy.

Historically, phishing detection relied primarily on **signature-based methods**, such as blacklists and whitelists, which maintain databases of known malicious and trusted URLs. While simple and fast, these methods are inherently reactive. A newly created phishing website, often referred to as a "zero-day" phishing threat, can operate for hours or even days before being reported and added to a blacklist, exploiting countless victims in the interim. Furthermore, attackers constantly employ techniques like URL obfuscation, subtle domain name variations (typo squatting), and rapid domain migration to evade these static defences.

To overcome these critical limitations, the cybersecurity community has pivoted toward Artificial Intelligence and Machine Learning. AI-based systems offer a **proactive and scalable solution** by training models to discern subtle, hidden patterns that are invisible to the human eye or simple rule sets. These patterns reside not just in the content of the page, but in the structural and lexical features of the URL, the age of the domain registration, and the presence of suspicious code elements like pop-up scripts. The central objective of this research is to review the current state-of-the-art in this domain, focusing on the methodologies, algorithms, and performance metrics employed by researchers to develop a robust, real-time AI-based phishing website detection system capable of strengthening cybersecurity and safeguarding

users from online fraud. This review aims to consolidate knowledge on effective ML techniques, provide a critical comparative analysis of common algorithms, and chart a clear course for future research directions.

2. LITERATURE SURVEY

The body of research on phishing detection has undergone a rapid evolution, moving from rudimentary manual checks to highly sophisticated, automated AI systems. This survey traces the key methodological milestones that have defined this progression, categorising the approaches by their underlying detection principles.

[1]. Aburrous et al. (2010) - "Intelligent Phishing Detection System for E-banking"

This paper proposes an early intelligent system specifically designed to protect electronic banking users from phishing. It moves beyond simple blacklisting by introducing a rule-based expert system that analyzes website features. The system identifies key characteristics in the URL, domain, and content to calculate a risk score. This heuristic approach demonstrated improved accuracy and was a foundational step toward feature-based detection methods that later leveraged machine learning.

[2]. Jain and Gupta (2020) - "Phishing Detection: Analysis of Machine Learning Techniques"

This work provides a comprehensive review and comparative analysis of various Machine Learning (ML) techniques applied to phishing detection. It systematically evaluates classifiers such as Decision Tree, Random Forest, SVM, and Naïve Bayes based on their performance metrics like accuracy, precision, and recall. The paper identifies which types of ML models are most effective for different feature sets (URL-based, content-based, etc.). It serves as an excellent resource for researchers looking to choose the optimal ML algorithm for a specific phishing detection scenario.

[3]. Shahzad and Aman (2024) - "Unveiling the Efficacy of AI-based Algorithms in Phishing Attack Detection"

This paper assesses the effectiveness of various AI-based algorithms in combating modern phishing attacks, which are becoming increasingly sophisticated. The research compares a wide range of algorithms, including conventional ML and deep learning models, to determine their respective accuracies and efficiencies. It highlights how certain advanced models, such as Convolutional Neural Networks (CNNs) and ensemble methods, can achieve high detection rates. The study aims to guide the selection of appropriate AI models for building robust and up-to-date anti-phishing solutions.

[4]. Islam et al. (2024) - "Phishing URL Detection via Machine Learning: A Comprehensive Survey"

This paper offers a detailed and systematic literature review focusing exclusively on phishing detection systems that use only URL features with Machine Learning. It explores different methodologies for extracting lexical and structural features from URLs, which are often the first line of defence. The review categorises and summarises the performance of various ML models on standard URL datasets. The work identifies trends, challenges, and future directions for developing lightweight, real-time phishing detection specifically using URL analysis.

[5]. UCI Machine Learning Repository - Phishing Websites Dataset

This is not a research paper, but a widely used public dataset that provides labelled data for training and testing phishing detection models. It typically contains thousands of entries, each representing a website labelled as either legitimate or phishing. Each entry is characterised by various computed features extracted from the URL and website source code. This repository is foundational for enabling reproducible research and comparative studies in the field of AI-based cybersecurity.

[6]. PhishTank - Community-driven Phishing Database

PhishTank is an essential resource and collaborative community effort, not an academic paper. It functions as a free, verified, and continuously updated database of known phishing websites. This platform is primarily used to provide immediate real-time blacklisting services and serves as a crucial source of fresh, ground-truth data for researchers. ML researchers often use PhishTank to collect recent examples of active phishing URLs for training and testing their models' ability to detect the latest attack types.

[7]. Do et al. (2022) - "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions"
This systematic literature review provides an in-depth analysis of Deep Learning (DL) techniques applied to phishing detection. It constructs a taxonomy to categorize different DL architectures, such as CNNs and RNNs, and examines their advantages and disadvantages. Crucially, the paper discusses major current challenges faced by DL models, including the need for large datasets and vulnerability to adversarial attacks. The authors provide recommendations and clear directions for future research in DL-based anti-phishing solutions.
[8]. Al-Sarem et al. (2021) - "An Optimized Stacking Ensemble Model for Phishing Websites Detection"
This research proposes and evaluates an optimized stacking ensemble model to improve the accuracy of phishing website detection. Stacking combines the predictions of multiple diverse base models (like SVM, Random Forest, etc.) using a meta-classifier to make the final prediction. The paper demonstrates that this hierarchical ensemble approach significantly boosts performance, achieving higher accuracy and lower false positive rates compared to using individual classifiers. It confirms the superiority of advanced ensemble learning for achieving highly robust detection systems**.**

3.1 Proposed Methodology

The proposed methodology for developing the detection system is systematically structured into four core phases, ensuring high-quality data and optimised model performance:

1. Decision Tree (DT)

Overview:

The Decision Tree is a supervised learning algorithm that uses a tree-like structure to represent decisions and their possible outcomes. Each internal node represents a feature condition (e.g., “Is the domain age < 30 days?”), branches denote decision outcomes, and leaf nodes represent classification labels—either *phishing* or *legitimate*.

Working Principle : The algorithm recursively splits the dataset based on feature values that best separate classes using measures such as *Information Gain* (Entropy) or *Gini Index*.

Advantages:

- Simple to understand and interpret (transparent decision logic).
- Works well with both numerical and categorical data.
- Requires minimal preprocessing.

Limitations:

- High tendency to **overfit** training data, reducing generalization.
- Sensitive to small variations in data, which can lead to unstable trees.

Use in Phishing Detection:

In phishing detection, Decision Trees provide interpretable results and quick insights into which URL or domain features contribute most to classification. However, standalone trees may fail to generalize across unseen or evolving phishing patterns.

2. Random Forest (RF)

Overview:

Random Forest is an **ensemble learning** method that combines the predictions of multiple Decision Trees to improve accuracy and robustness. Each tree is trained on a random subset of the dataset and features, and the final classification is determined through majority voting.

Working Principle:

- Uses *Bootstrap Aggregation (Bagging)* to create diverse Decision Trees.
- Each tree contributes an independent prediction, reducing variance.

- Aggregation of multiple trees ensures stable and generalized performance.

Advantages:

- **High accuracy** and resistance to overfitting compared to a single DT.
- Handles large feature sets effectively, including non-linear relationships.
- Provides feature importance scores, useful for interpretability and optimization.

Limitations:

- Computationally more expensive than DT due to multiple tree generation.
- Reduced interpretability compared to a single Decision Tree.

Use in Phishing Detection:

Random Forest is considered one of the most effective algorithms for phishing URL detection. Its ensemble structure allows it to identify subtle correlations between URL structure, domain metadata, and content-based indicators. It achieves superior **accuracy**, **recall**, and **generalization** across diverse phishing datasets.

3. Support Vector Machine (SVM)

Overview:

Support Vector Machine (SVM) is a powerful classification algorithm based on finding the **optimal hyperplane** that maximally separates classes in a high-dimensional feature space. It is particularly effective for complex, non-linear problems when combined with kernel functions.

Working Principle:

- Finds the hyperplane that maximizes the margin between phishing and legitimate samples.
- Uses *kernel tricks* (linear, polynomial, RBF) to transform data into higher dimensions where separation becomes easier.
- Focuses on *support vectors*—data points closest to the boundary—that most influence classification.

Advantages:

- Performs well in **high-dimensional spaces** typical of URL and content-based phishing features.
- Robust to outliers and effective even with smaller training datasets.
- Strong generalization capability due to margin maximization principle.

Limitations:

- Computationally intensive for large datasets.
- Kernel and parameter selection can significantly affect performance.
- Less interpretable than tree-based models.

Use in Phishing Detection:

SVM excels at identifying subtle, high-dimensional feature interactions—making it suitable for detecting sophisticated phishing URLs. It provides a strong decision boundary between legitimate and malicious sites, often achieving **high precision** though sometimes at the cost of longer training time.

Comparative Study of Algorithms:

Algorithm	Model Type	Accuracy	Interpretability	Overfitting Risk	Computation Time	Best Use Case
Decision Tree (DT)	Single model	Moderate	High	High	Low	Baseline analysis, feature insight

Random Forest (RF)	Ensemble (Bagging)	High	Moderate	Low	Moderate	Real-world deployment, robust detection
Support Vector Machine (SVM)	Margin-based classifier	High	Low	Low	High	Complex, high-dimensional datasets

3.2 Algorithms Used

The project selected **Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)** to classify websites as legitimate or malicious. This selection balances the need for interpretability with high predictive accuracy.

4.1 Research Gap

Despite the remarkable progress achieved by AI in phishing detection, several critical challenges remain, forming the primary focus for future research.

One of the most pressing research gaps is the issue of **Adversarial Attacks**. Sophisticated attackers can use minor, computationally optimised changes to a malicious URL (e.g., adding a benign sub-domain or subtly altering a keyword) that are invisible to the user but are specifically designed to misclassify the URL as legitimate by the ML model. Current models, particularly deep learning architectures, are proving vulnerable to these calculated perturbations. Future work must focus on developing **Adversarial Training** techniques to make models more resilient and robust against such targeted attacks.

Another significant gap lies in **real-time scalability and standardisation**. Many state-of-the-art models, especially complex Deep Learning systems, require substantial computational resources and long inference times, making them difficult to deploy globally in low-latency environments like browser extensions or cloud-based filtering services. Future research should prioritise developing **lightweight architectures** and optimising deployment strategies for real-time performance.

4.2 Future Scope

The **Future Scope** of this field is clearly directed toward **Hybrid Detection Systems**. These systems will move beyond relying solely on features, combining the speed of traditional list-based checking, the interpretability of heuristic rules, and the high accuracy of advanced AI models. Key areas include:

1. **Deep Learning Integration:** Implementing advanced neural networks (CNNs, LSTMs, and ultimately, Transformer-based models) to extract deeper semantic meaning from page content and language.
2. **Continuous Learning (CL):** Developing automated retraining pipelines where newly reported phishing data immediately feeds back into the model, allowing the system to adapt and evolve automatically as attack patterns change.
3. **Client-Side Defence:** Deploying the optimised detection models as a **Browser Extension** or client-side application to provide immediate, real-time user alerts while browsing, effectively cutting off the attack chain before the user can interact with the malicious site.

5. CONCLUSION

The threat posed by phishing websites necessitates a continuous and intelligent defence strategy, and this review affirms that AI and Machine Learning provide the most effective modern tools. By shifting from reactive, list-based methods to proactive, pattern-recognition models, detection systems can now

effectively combat sophisticated, rapidly evolving threats. The established methodology—from meticulous feature engineering across URL, domain, and content layers, to the application of robust classification algorithms—has proven its efficacy. Critically, the adoption of ensemble techniques like Random Forest over simpler classifiers such as the Decision Tree has provided the necessary stability and high accuracy required for real-world cybersecurity applications. While we have highlighted challenges, particularly concerning model resilience to adversarial attacks and the complexity of real-time deployment, the future is promising. The transition to hybrid, continuous learning systems and the deployment of advanced deep learning models in lightweight architectures will be the keys to ensuring that anti-phishing defences can successfully maintain pace with the ingenuity of cybercriminals, ultimately contributing to a safer and more secure digital experience for all users.

REFERENCES:

1. Aburrous, M., M.A. Hossain, K. Dahal, and F. Thabtah. "Intelligent Phishing Detection System for E-banking." *Expert Systems with Applications*, vol. 37, no. 12, 2010, doi:10.1016/j.eswa.2010.02.068.
2. Jain, A.K., and B.B. Gupta. "Phishing Detection: Analysis of Machine Learning Techniques." *Security and Privacy*, vol. 3, no. 5, 2020, doi:10.1002/spy2.93.
3. Shahzad, T., & Aman, K. (2024). "Unveiling the Efficacy of AI-based Algorithms in Phishing Attack Detection." *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 116–133.
4. Islam, Jahirul, et al. "Phishing URL Detection via Machine Learning: A Comprehensive Survey." *International Journal on Artificial Intelligence Tools*, vol. 33, no. 5, 2024.
5. *UCI Machine Learning Repository - Phishing Websites Dataset*. Kaggle, The University of California, Irvine. Web.
6. *PhishTank - Community-driven Phishing Database*. OpenDNS, LLC. Web.
7. Do, N. Q., et al. "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions." *IEEE Access*, vol. 10, 2022, pp. 36431-36449.
8. Al-Sarem, M., et al. "An Optimized Stacking Ensemble Model for Phishing Websites Detection." *Electronics*, vol. 10, no. 11, 2021, p. 1285.