Sowmiya R, , 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Preventing Phishing Attacks Through URL Detection Using HML Algorithm in Machine Learning

[1]Sowmiya R, [2]Divakar P, [3]Ragul M, [4]Anbu D

[1]Assistant Professor Department of Information Technology K. S. R College of Engineering Tiruchengode, India
[2,3,4]Department of Information Technology K.S. R College of Engineering Tiruchengode, India

Abstract- Cyber attacks are among the most widespread and dangerous cyber threats, tricking users into revealing sensitive information by imitating legitimate websites. As these attacks become increasingly sophisticated, traditional detection methods often fail to identify them accurately. This project proposes a hybrid machine learning ML -based system that enhances phishing URL detection by analyzing both the structural and semantic features of URLs. The system extracts a rich set of features, including domain names, subdomains, URL paths, query parameters, and overall URL structure, which serve as critical indicators for identifying hidden phishing patterns. To optimize detection performance, the system integrates the capabilities of two robust ML models: Random Forest and Support Vector Machine . RF functions as both a feature selector and classifier, leveraging its ensemble learning mechanism to improve accuracy while minimizing overfitting. SVM, known for its effectiveness in handling high-dimensional data, constructs an optimal hyperplane to separate legitimate URLs from phishing ones. The hybrid approach of combining RF and SVM enhances the system's precision, robustness, and overall detection capability. This dual-model system not only addresses the shortcomings of conventional and single-model techniques but also contributes significantly to preventing data breaches and financial losses. The proposed method demonstrates a scalable and efficient solution for real-world phishing detection by applying advanced machine learning techniques to analyze URL characteristics in depth.

Keywords — Phishing Detection, Machine Learning, URL Analysis, Cybersecurity, Feature Extraction, Cyber Attack Prevention, Hybrid Model.

I. INTRODUCTION

In Today's Interconnected Digital Landscape, Cybersecurity Plays A Vital Role In Safeguarding Personal And Organizational Data. One Of The Most Prevalent Threats Within This Space Is Phishing—An Attack That Tricks Users Into Revealing Sensitive Credentials Via Deceptive Emails And Websites. These Attacks Often Exploit Human Trust And Mimic Legitimate Platforms, Making Traditional Detection Methods Increasingly Inadequate.

Phishing Attacks Primarily Operate By Deploying Fake Websites Or Malicious Urls That Appear Convincingly Authentic. These Urls Lure Users Into Entering Confidential Data, Which Attackers Then Exploit. The Rapid Evolution And Polymorphic Nature Of These Attacks Present Major Challenges For Blacklist-Based And Signature-Based Detection Systems, Which Often Fail To Recognize Novel Or Obfuscated Threats In Real-Time.

To Overcome These Limitations, This Project Introduces A Machine Learning-Based Approach That Leverages Both Support Vector Machine (Svm) And Random Forest (Rf) Algorithms. These Models Have Shown Excellent Performance In Binary Classification Tasks. Svm Identifies Optimal Decision Boundaries In Complex, High-Dimensional Feature Spaces, While Rf Aggregates Decisions From Multiple Trees To Improve Accuracy And Minimize Overfitting.

Combining Svm And Rf Into A Hybrid Model Improves The Combining Svm And Rf Into A Hybrid Model Improves The Robustness Of The Detection System. This Ensemble Approach Ensures That The

System Can Detect A Wide Range Of Phishing Techniques With High Accuracy, Low False Positives, And Better Generalization Across Unseen Data. The System Is Scalable, Efficient, And Well-Suited For Integration Into Browsers, Email Gateways, Or Corporate Firewalls.

In Summary, This Research Proposes A Hybrid, Intelligent Phishing Detection Framework Designed To Meet The Demands Of Modern Cybersecurity Environments. By Integrating Svm And Rf Within A Unified System, It Offers A Proactive Defense Mechanism Against Phishing Threats While Contributing To Scalable And Adaptive Solutions That Can Be Refined Through Continuous Learning.

II. RELATED WORKS

A. Ashok and team (2024) conducted a comprehensive comparison of machine learning, deep learning, and boosting algorithms for phishing URL detection. Their target was to identify the most accurate and efficient model for real-time phishing threat classification. By evaluating traditional models like Decision Trees, Support Vector Machines (SVM), and ensemble techniques like XGBoost and AdaBoost, the study aimed to optimize detection rates. Performance metrics such as Accuracy, Precision, Recall, and ROC-AUC were used. Their work concludes that boosting algorithms outperform others in accuracy but require more computational power, providing valuable insights into model selection for real-world phishing detection applications.

Zhang Y. and team, in their 2020 study, presented CANTINA, a content-based approach to detecting phishing websites. Unlike many studies focusing solely on URL or structural features, Zhang's research targeted the actual content of websites to detect phishing attempts. Using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm—a common technique in text mining—the system analyzes web content to identify suspicious textual elements and keywords commonly found in phishing attacks. Their model was able to correctly identify about 95% of phishing websites, demonstrating high accuracy through content-based analysis. However, this method requires substantial computational resources and can sometimes fail when attackers use contextually clever or minimal text. Zhang's study is pivotal because it shifts the focus from external URL features to internal web content, thereby strengthening phishing detection even when URLs appear legitimate. This approach is particularly useful in cases where obfuscated or short URLs are used to mask malicious intent. Their work laid the groundwork for integrating natural language processing into cybersecurity solutions

In 2024, H. Baliyan and A. Rama Prasath proposed a powerful phishing detection framework based on ensemble machine learning models. Their research focused on improving accuracy, scalability, and robustness by combining several base classifiers into a meta-model. The algorithms used included Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Their ensemble techniques involved Voting Classifiers—using both hard and soft voting strategies—as well as Bagging (through RF) and Boosting (via Gradient Boosting and XG Boost). To optimize model performance, they employed Grid Search and k-fold cross-validation. Features such as domain age, HTTPS presence, URL length, and usage of special symbols were used as model inputs. The authors highlighted the advantages of ensemble models in reducing overfitting and improving generalization. However, they also noted that ensemble learning increases computational complexity and model training time. Their research is significant for cybersecurity practitioners aiming to deploy scalable and reliable phishing detection systems that can adapt to evolving threats while minimizing false positives.

In 2024, B. V. Pavani, D. Mahitha, and B. U. Maheswari proposed a phishing detection system combining machine learning (SVM, RF, LR) with explainable AI techniques such as SHAP and LIME. Their objective was to enhance transparency in model decisions while maintaining strong accuracy. The system analyzed URL-based features and applied interpretation tools to explain model outputs. Although these tools added processing time, they

improved trust and accountability. This approach is highly beneficial for enterprise cybersecurity frameworks that demand regulatory compliance and detailed auditing alongside effective phishing prevention.

R. Ferdaws and N. E. Majd (2024) introduced a two-phase hybrid system combining machine learning (SVM, RF, LR) with deep learning models (CNN, LSTM) to detect phishing URLs. Their system used ML for fast filtering and DL for deeper sequential and structural analysis. Trained on over 1.2 million URLs, the CNN and LSTM models achieved over 98% accuracy. While resource-intensive, their hybrid model offered high scalability, precision, and resilience against sophisticated phishing tactics—making it suitable for critical security domains such as finance and national infrastructure protection.
The researchers also utilized Recursive Feature Elimination (RFE) to optimize input features, enhancing model efficiency. Hyperparameter tuning was performed using Grid Search to ensure optimal configuration. This layered system effectively balances speed, depth, and real-time adaptability in phishing detection environments.

III. EXISTING SYSTEM

The existing phishing detection systems primarily rely on traditional machine learning techniques and heuristic-based methods to identify malicious URLs. These systems analyze manually engineered features such as URL length, presence of IP addresses, special characters, HTTP/HTTPS status, and domain age to classify URLs using classifiers like Decision Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF). While these models achieve reasonable accuracy, they often struggle to detect newly emerging threats due to their dependence on historical data and static rule-based detection. Many systems also depend on blacklists that store known phishing URLs, but these are ineffective against zero-day attacks where attackers frequently generate new, previously unseen links. Advanced systems have begun incorporating deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to improve detection by learning from sequential or structural URL patterns, but these require extensive computational resources and large datasets. Ensemble methods like bagging and boosting have been employed to enhance robustness, yet they still face challenges with obfuscated URLs and adversarial evasion techniques. Overall, the current systems lack adaptability, suffer from false positives and false negatives, and require improvements in real-time detection and scalability to counter the evolving nature of phishing attacks effectively.

Proposed System

The "Preventing Phishing Attacks Through URL Detection Using HML Algorithm" is introduces a hybrid machine learning model combining Support Vector Machine (SVM) and Random Forest (RF) algorithms to detect phishing URLs with improved accuracy and efficiency. Unlike traditional methods, this model extracts key features from URL such as length, use of IP addresses, HTTPS presence, and domain age and classifies them based on learned patterns. SVM offers precision in defining clear decision boundaries, while RF adds robustness by aggregating results from multiple decision trees to reduce overfitting. This hybrid approach ensures better generalization across evolving phishing techniques. The system is lightweight, scalable, and suitable for real-time implementation in web browsers or email filters, making it highly effective in preventing phishing attacks in dynamic cybersecurity environments.

IV. SYSTEM DESIGN

System Architecture

The system architecture for phishing URL detection is designed to efficiently identify malicious websites in real time. It begins with collecting labeled URL datasets from trusted sources like Phish Tank. Key features such as URL length, presence of IP addresses, HTTPS usage, and domain age are extracted and preprocessed. These features are input into two machine learning models: Support Vector Machine (SVM) and Random Forest (RF). SVM identifies optimal decision boundaries, while RF improves accuracy through ensemble learning. The final prediction is made through voting. Evaluation

metrics like accuracy, precision, and recall are used to assess model performance and effectiveness.

Feature Extraction

The system identifies phishing attempts by analyzing critical attributes derived from URLs, focusing on both address bar-based and domain-based features. Address-level indicators include URL length, use of IP addresses, presence of special characters like "@" and "//", subdomain depth, and shortened URL usage. Domain-related features involve checking the availability of DNS records, domain registration age, expiration time, and web traffic data. These features are chosen for their effectiveness in capturing deceptive patterns commonly found in phishing attacks. Once extracted and preprocessed, the features are encoded into a structured format, enabling accurate classification through machine learning models such as Support Vector Machine and Random Forest.

Key Components

The project includes a curated dataset of phishing and legitimate URLs, feature extraction techniques focusing on URL length, IP address presence, special characters, and domain age. It uses a hybrid machine learning model combining Support Vector Machine (SVM) and Random Forest (RF) for classification. Python and libraries like scikit-learn are used for implementation. Evaluation metrics such as accuracy, precision, and recall assess performance. The system is designed for real-time deployment in browsers or email filters..

Model selection

The proposed system utilizes Support Vector Machine (SVM) and Random Forest (RF) to ensure high accuracy and resilience in phishing detection. SVM excels in handling high-dimensional data and identifying optimal decision boundaries for binary classification tasks. RF, an ensemble learning method, constructs multiple decision trees and combines their outputs to improve generalization and reduce the risk of overfitting. This hybrid approach leverages SVM's precision and RF's robustness, enabling the system to effectively classify URLs as phishing or legitimate. The combination enhances detection performance, lowers false positives, and ensures adaptability to evolving phishing strategies.

Dataset Module

The dataset module plays a crucial role in training and evaluating the phishing detection system. It utilizes a labeled dataset comprising both phishing and legitimate URLs, sourced from reliable platforms such as Phish Tank and Kaggle. Each URL is represented by structured features including address-bar attributes e.g., presence of '@', IP address, domain-based features e.g., domain age, DNS record, and lexical indicators e.g., URL length, hyphen count. Feature extraction transforms raw URLs into numerical formats suitable for machine learning algorithms. The dataset undergoes preprocessing steps such as normalization, handling missing values, and binary encoding. This structured and cleaned dataset is then split into training and testing sets to develop robust models using Support Vector Machine and Random Forest, ensuring accurate and generalizable phishing detection.

Flow Chart and HML Algorithm Analysis

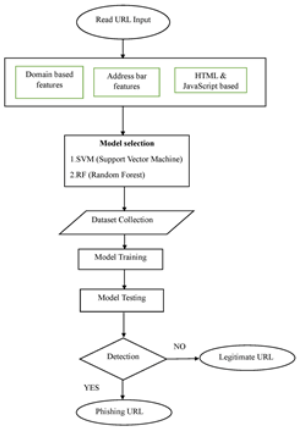

Figure 1. Block diagram

The phishing detection system utilizes a hybrid machine learning architecture integrating Support Vector Machine (SVM) and Random Forest (RF) algorithms. The architecture encompasses URL input, feature extraction (address-bar, domain-based, lexical), and model training. It enables efficient classification of URLs as phishing or legitimate using ensemble learning for improved accuracy and robustness. With real-time analysis, the

system offers scalable, adaptable, and proactive defense against evolving phishing threats, making it suitable for deployment in browsers and email security platforms.

Step 1: The system receives input in the form of a URL, either manually entered by a user or extracted from a dataset. This URL serves as the basis for subsequent feature extraction and analysis.
Step 2: Feature Extraction The system derives relevant features from the URL to aid in classification. These include address bar-based features (e.g., URL length, use of "@", presence of IP address), domain-based features (e.g., DNS record availability, domain age), lexical features (e.g., token patterns, number of hyphens), and content-based or script features (e.g., presence of JavaScript redirects or hidden fields).
Step 3: Model Selection Two machine learning models are utilized: Support Vector Machine (SVM) for its precision in binary classification and Random Forest (RF) for its robustness and generalization through ensemble learning.
Step 4: Dataset Collection and Model Training
a dataset of labeled phishing and legitimate URLs is collected. preprocessing steps like normalization and feature scaling are applied before training the SVM and RF models.
Step 5: Testing and Evaluation The models are tested on unseen URLs and evaluated using metrics such as accuracy, precision, recall, and F1-score to determine detection performance.
Step 6: URL Classification The system classifies each URL as phishing or legitimate based on the trained model's output.
Step 7: Action Execution If the URL is classified as legitimate, the system allows access. If identified as phishing, it blocks access, warns the user, and logs the event.
Step 8: Continuous Learning The model is periodically updated with new phishing data and threat intelligence to enhance adaptability and detection accuracy over time.

G. Result Analysis

Parameter	Existing System	Proposed System
Overall Detection Accuracy (%)	90.23	98.12
True Positive Rate (%)	91.38	96.33
True Negative Rate (TNR / Specificity) (%)	50.77	85.83
False Negative Rate (FNR) (%)	45.62	77.5

Table 1. Overall Phishing Detection Ability Comparison Table

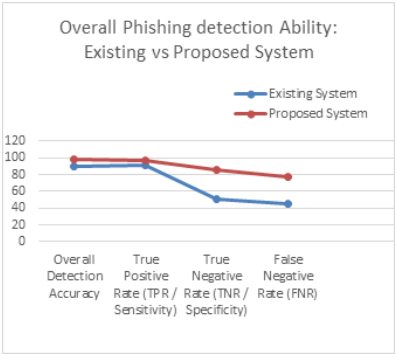

Figure 2. Comparison graph

This study presents a comparison of machine learning models for phishing website detection, emphasizing accuracy and robustness. The proposed hybrid model, which combines Support Vector Machine (SVM) and Random Forest (RF), outperforms traditional models like Logistic Regression (LR) and Decision Tree (DT). SVM is effective at drawing clear decision boundaries, enabling accurate identification of phishing sites, while RF strengthens the model by using multiple decision trees to reduce the impact of noisy or incomplete data. The graphical results in the report indicate that SVM delivers high precision, and RF enhances overall reliability by lowering false negatives—phishing sites wrongly classified as legitimate. Together, these models compensate for each other's weaknesses, resulting in a system that performs well under various conditions. Unlike existing models that may falter when dealing with complex or constantly changing phishing tactics, the

hybrid approach adapts effectively and maintains high accuracy. Additionally, RF's ability to highlight the importance of different features aids in improving the model's transparency and performance. Overall, the hybrid SVM-RF model offers a more accurate, robust, and scalable solution that is well-suited for real-time phishing detection and can significantly enhance cybersecurity defenses.

V. CONCLUSION

The proposed phishing detection system effectively leverages machine learning techniques, specifically Support Vector Machine and Random Forest, to identify and classify malicious URLs with high accuracy. By extracting and analyzing key features from the URL and domain, the system demonstrates strong performance in detecting phishing attempts while minimizing false positives. The hybrid model combines the strengths of both algorithms, ensuring improved generalization and robustness. Through comprehensive evaluation and structured implementation, the system proves to be scalable, efficient, and suitable for real-time deployment. This work contributes to enhancing cybersecurity measures by offering a proactive solution against evolving phishing threats.

Future Work

The system can be improved by incorporating real-time threat intelligence feeds and online learning mechanisms to detect zero-day phishing attacks more effectively. Advanced deep learning models such as LSTM and CNN may be explored to enhance classification accuracy and handle complex obfuscation techniques. Expanding the feature set to include visual similarity metrics and email header analysis could offer a more holistic defense strategy. Deployment as a lightweight browser extension or integration with enterprise-level security tools is a practical next step. Reducing computational complexity and improving real-time responsiveness will make the system more efficient while ensuring compliance with privacy regulations like GDPR.

REFERENCES

1. A. Ashok, D. Rathis, R. Raghavendra, and V. Umadevi, "A Comparative Analysis of Traditional Machine Learning, Deep Learning and Boosting Algorithms on Phishing URL Detection," 2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), 2024.
2. B. V. Pavani, D. Mahitha, and B. U. Maheswari, "Enhancing Online Safety: Phishing URL Detection Using Machine Learning and Explainable AI," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024.
3. H. Baliyan and A. Rama Prasath, "Enhancing Phishing Website Detection Using Ensemble Machine Learning Models," 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, 2024.
4. R. Ferdaws and N. E. Majd, "Phishing URL Detection Using Machine Learning and Deep Learning," 2024 IEEE World AI IoT Congress (AIIoT), 2024.
5. S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," IEEE Access, vol. 11, 2023.
6. S. Liu, H. Wu, G. Cheng, and X. Hu, "Real-Time Phishing Detection Based on URL Multi-Perspective Features: Aiming at the Real Web Environment," ICC 2023 - IEEE International Conference on Communications, 2023
7. M. A. Al Ahasan, M. Hu, and N. Shahriar, "OFMCDM/IRF: A Phishing Website Detection Model based on Optimized Fuzzy Multi-Criteria DecisionMaking and Improved Random Forest," 2023 Silicon Valley Cybersecurity Conference (SVCC), 2023.
8. Kara, M. Ok, and A. Ozaday, "Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods," IEEE Access, vol. 10, 2022.
9. Y. Xu, G. Chen, Q. Liu, W. Xu, L. Zhang, J. Wu, and X. Fan, "A Phishing Website Detection and Recognition Method Based on Naive Bayes," 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC), 2022.

10. R. Raj and S. S. Kang, "Spam and Non-Spam URL Detection using Machine Learning Approach," 2022 3rd International Conference for Emerging Technology (INCET), 2022.
11. H. Faris and S. Yazid, "Phishing Web Page Detection Methods: URL and HTML Features Detection," Proceedings of the 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), IEEE, 2021.
12. H. Faris and S. Yazid, "Phishing Web Page Detection Methods: URL and HTML Features Detection," Proceedings of the 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), IEEE, 2021.
13. . S. Venugopal, S. Y. Panale, M. Agarwal, R. Kashyap, and U. Ananthanagu, "Detection of Malicious URLs through an Ensemble of Machine Learning Techniques," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021.
14. M. Sameen, K. Han, and S. O. Hwang, "PhishHaven—An Efficient RealTime AI Phishing URLs Detection System," IEEE Access, vol. 8, 2020.
15. 15. Y. Huang, J. Qin, and W. Wen, "Phishing URL Detection Via Capsule-Based Neural Network," Proc. IEEE 13th Int. Conf. Anti-counterfeiting, Security, and Identification (ASID), 2019.