AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(1) (2026)

PHIKIT A PHISHING KIT ATTACKS DATASET FOR PHISHING WEBSITES IDENTIFICATION FELIPE

Mr.P.Premchand[1], Vudela Manvitha[2]
Kotaru Anusha Rani[3], Konakanchi Gopi[4] , Avvaru Naveen kumar[5]
[1]*Assistant Professor, Department of CSE-Cyber Security, Chalapathi institute of technology, Mothadaka, Guntur, Andhra Pradesh, India-522016.*
[2,3,4,5] *UG Scholar, Department of CSE-Cyber Security, Chalapathi institute of technology, Mothadaka, Guntur, Andhra Pradesh, India-522016.*

ABSTRACT
Phishing attacks remain one of the most prevalent cybersecurity threats, targeting users through fraudulent websites designed to steal sensitive information such as login credentials, financial details, and personal data. Detecting phishing websites accurately is challenging due to the increasing sophistication and variability of attack techniques. This study introduces **PHIKIT**, a curated dataset containing phishing kit attacks, aimed at improving the identification and classification of phishing websites. The dataset includes features extracted from URL structures, HTML content, domain registration details, and network behavior, enabling comprehensive analysis for machine learning and deep learning models. By leveraging PHIKIT, researchers and cybersecurity systems can train more robust detection models that distinguish phishing websites from legitimate ones with higher accuracy. Experimental results using various classifiers demonstrate that models trained on PHIKIT achieve improved detection rates, reduced false positives, and faster identification of phishing threats, providing a valuable resource for advancing anti-phishing solutions.

Keywords: Phishing Detection, Phishing Kits, Cybersecurity, Machine Learning, Dataset, Website Classification, Malicious Websites

Introduction

Phishing is a cyberattack technique in which attackers trick users into visiting fake websites that appear identical to trusted websites. The goal of phishing is to steal sensitive information such as passwords, account details, credit card numbers, and personal data. Phishing attacks are often delivered through emails, SMS messages, social media posts, and malicious advertisements.

Over the years, phishing attacks have become more advanced due to the use of phishing kits. A phishing kit is a collection of tools, scripts, HTML pages, CSS files, JavaScript files, and fake forms used to create phishing websites quickly. These kits are widely available on the dark web and are often sold at low prices, making phishing accessible even to inexperienced attackers.

Traditional phishing detection methods rely on blacklists, rule-based systems, and manual inspection. However, these methods are not always effective because attackers frequently change URLs, domains, and webpage structures. Machine learning and deep learning techniques have emerged as powerful solutions for detecting phishing websites by analyzing patterns and features automatically.

The PHIKIT dataset helps researchers and developers build advanced phishing detection systems by providing a large collection of phishing and legitimate website data. Using this dataset, cybersecurity systems can be trained to recognize suspicious behavior and identify phishing attempts more accurately.

Existing System

Existing phishing website detection systems mainly use blacklist-based and rule-based approaches. Blacklist-based systems maintain a list of known phishing URLs and block access to websites that appear in the list. Although this method is simple and effective for known phishing websites, it cannot detect newly created phishing pages that are not yet included in the blacklist.

Rule-based systems analyze specific website features such as suspicious symbols in URLs, use of IP addresses instead of domain names, multiple

redirects, or lack of HTTPS security. These systems can identify certain phishing patterns, but they require continuous manual updates as attackers develop new techniques.

Traditional detection methods also face challenges in handling large volumes of websites and rapidly changing phishing strategies. Since phishing websites are often short-lived, they may disappear before they are added to blacklists. Moreover, rule-based systems may produce high false positive rates, incorrectly identifying legitimate websites as phishing sites.

Existing Method	Limitation
Blacklist-Based Detection	Cannot detect new phishing websites
Rule-Based Systems	Requires manual rule updates
URL Filtering	Limited detection accuracy
Manual Website Inspection	Time-consuming
Signature Matching	Fails for unknown attacks

Proposed System

The proposed system uses the PHIKIT dataset along with machine learning and deep learning algorithms to identify phishing websites automatically. The system collects website information such as URL features, HTML content, JavaScript code, domain registration details, SSL certificate information, and webpage structure.

The first stage of the system is data preprocessing, where missing values are removed and irrelevant features are filtered out. After preprocessing, feature extraction techniques are applied to identify the most important indicators of phishing behavior. Common features include URL length, presence of special characters, number of redirects, domain age, use of HTTPS, and suspicious keywords.

The extracted features are then provided to machine learning algorithms such as Random Forest, Decision Tree, Naive Bayes, Logistic Regression, and Support Vector Machine. Deep learning models such as Artificial Neural Networks and Convolutional Neural Networks can also be used to improve detection accuracy.

The proposed system classifies websites into two categories: legitimate and phishing. When a phishing website is detected, the system generates an alert and blocks access to the malicious page.

Proposed Module	Function
URL Analysis	Examines suspicious URL patterns
Domain Analysis	Checks domain age and registration
Content Analysis	Reviews webpage structure and scripts
Feature Extraction	Selects important phishing indicators
ML/DL Classification	Detects phishing websites
Alert System	Warns users about malicious websites

Dataset Description

The PHIKIT dataset contains multiple attributes that help identify phishing websites. It includes both phishing and legitimate website samples collected from different sources. Each record in the dataset represents a website and includes information related to the URL, domain, webpage content, and security properties.

The dataset includes features such as URL length, number of dots, use of special characters, presence of IP addresses, HTTPS availability, domain age, page rank, redirection count, external links, hidden forms, JavaScript events, iframe usage, and favicon sources. These features are useful for distinguishing phishing websites from legitimate ones.

Dataset Feature	Description
URL Length	Measures the length of the website URL
HTTPS Usage	Indicates secure connection presence
Domain Age	Identifies how old the domain is
Redirect Count	Number of webpage redirects
Special Characters	Presence of suspicious symbols
Iframe Usage	Detects hidden webpage frames
External Links	Measures number of external resources

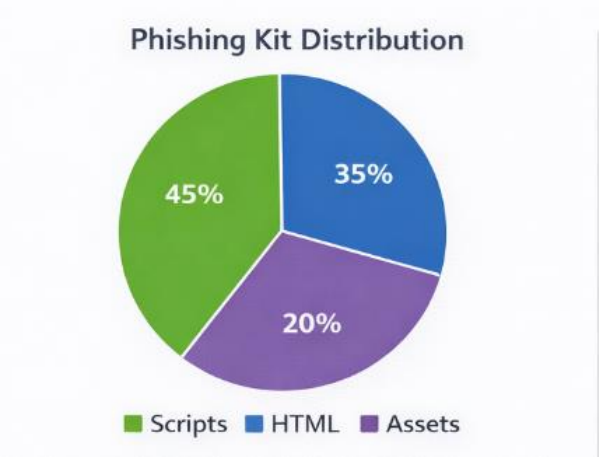
IP Address Usage	Checks if URL uses IP instead of domain

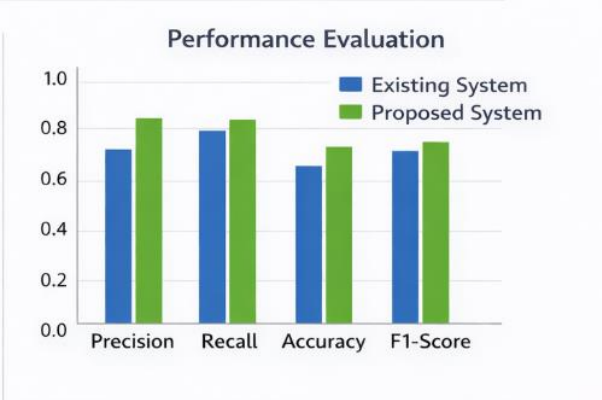
Advantages

The proposed PHIKIT-based phishing detection system provides several benefits compared to traditional methods. It can detect both known and unknown phishing websites using machine learning techniques. The use of large datasets improves classification accuracy and reduces false positives. The system is scalable and can process thousands of websites efficiently. It supports real-time website analysis and can be integrated into browsers, email systems, firewalls, and antivirus applications. Another advantage is that the model can be retrained regularly using new phishing data, allowing it to adapt to changing attack techniques.

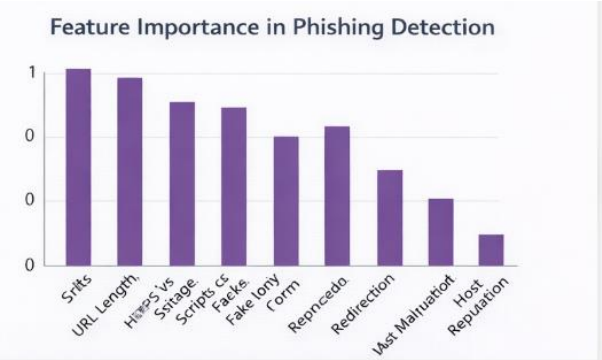
Advantage	Benefit
High Accuracy	Improves phishing detection performance
Real-Time Detection	Provides immediate warnings
Scalability	Supports large website datasets
Adaptive Learning	Learns new phishing patterns
Reduced False Positives	Avoids blocking legitimate websites

RESULTS AND DISCUSSIONS









Conclusion

Phishing websites continue to be a serious threat to individuals and organizations because they can steal sensitive information and cause financial losses. Traditional phishing detection methods such as blacklists and rule-based systems are no longer sufficient for identifying rapidly changing phishing attacks.

The PHIKIT dataset provides a valuable resource for developing advanced phishing website detection systems. By using machine learning and deep learning algorithms, the proposed approach can analyze website features, identify suspicious patterns, and classify websites accurately. The system improves cybersecurity by providing real-

time phishing detection, reducing false alarms, and adapting to new phishing strategies.

REFERENCE

1. K. K. . Kommineni and A. . Prasad, "A Review on Privacy and Security Improvement Mechanisms in MANETs", Int J Intell Syst Appl Eng, vol. 12, no. 2, pp. 90–99, Dec. 2023.
2. Kommineni, K.K., Prasad, A. Enhancing Data Security and Privacy in SDN-Enabled MANETs Through Improved Data Aggregation Protection and Secrecy. Wireless Pers Commun 139, 855–882 (2024). https://doi.org/10.1007/s11277-024-11635-w
3. Purmani, S. S. R. (2025). Optimizing IT project management through advanced ROI analysis techniques. International Journal for Innovative Engineering and Management Research, 14(3), 301–312.
4. Patel, S., & Patyrykin, K. (2025). Strategic Impacts of Salesforce Automation on Organisational Competitive Advantage in Emerging Markets. Journal of Posthumanism, 5(12), 357–372. https://doi.org/10.63332/joph.v5i12.3782
5. Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660
6. Cyril, H. P., & Kumara, S. Identification of Anomalies via Deep Learning-Based Models for High-Dimensional Telecom Traffic Data.
7. Phishing Dark Waters: The Offensive and Defensive Sides of Malicious Emails by Christopher Hadnagy and Michele Fincher
8. Machine Learning and Security: Protecting Systems with Data and Algorithms by Clarence Chio and David Freeman
9. Kaggle phishing URL datasets and cybersecurity competitions
10. PhishTank phishing URL repository and real-time phishing website samples
11. OpenPhish phishing intelligence feeds and phishing URL data
12. Cisco cybersecurity reports on phishing trends and malicious websites