IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

Prevention of Phishing Websites using Machine Learning: An Intelligent Approach for Secure Web Browsing

K. Moksha Sri[1] and Swathi Terli[2]
MCA Scholar, Sir C. R. Reddy College (Autonomous), Eluru[1]
MCA, M. Tech, Assistant Professor, Department of MCA, Sir C. R. Reddy College (Autonomous), Eluru[2]

Abstract: *The rapid growth of internet usage and online transactions has significantly increased cyber threats, particularly phishing attacks. Phishing websites imitate legitimate websites to deceive users into revealing confidential information such as usernames, passwords, banking credentials, and personal data. Traditional phishing detection methods primarily rely on blacklist databases and signature-based techniques, which are ineffective against newly created phishing websites. Consequently, there is an increasing demand for intelligent systems capable of identifying phishing websites in real time.*

This study proposes a machine learning-based phishing website detection system using the Random Forest algorithm. The system analyzes various website and URL characteristics, including URL length, HTTPS protocol, domain age, IP address usage, redirection count, special characters, SSL certificate status, and suspicious keywords. Based on these features, the model predicts whether a website is legitimate or phishing. The application is developed using Python, Flask, Scikit-learn, HTML, CSS, JavaScript, and SQLite.

The proposed model provides users with an instant prediction, security risk score, and recommendations for safe browsing. The system demonstrates high prediction accuracy while maintaining computational efficiency, making it suitable for deployment as a web application. The research highlights the potential of Artificial Intelligence in strengthening cybersecurity and protecting internet users from phishing attacks.

Keywords: Phishing Detection, Machine Learning, Cyber Security, Random Forest, Artificial Intelligence, Flask

I. INTRODUCTION

The Internet has become an indispensable part of modern society, enabling online banking, e-commerce, education, healthcare, communication, and digital governance. While these advancements have improved convenience and accessibility, they have also increased exposure to cyber threats. Among various cybercrimes, phishing remains one of the most common and dangerous attacks, causing substantial financial losses and compromising personal information.

Phishing websites are fake web pages designed to resemble trusted websites. Cybercriminals use these websites to trick users into entering sensitive information such as passwords, credit card details, and One-Time Passwords (OTPs). Since phishing websites are continuously modified, traditional blacklist-based detection systems struggle to identify newly created malicious websites.

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as effective tools for addressing this challenge. Machine learning algorithms learn patterns from historical data and classify websites based on multiple characteristics without relying solely on previously identified phishing URLs. These algorithms can detect unknown phishing websites with higher accuracy and adaptability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

This research develops an intelligent phishing detection system using the Random Forest classification algorithm. The application analyzes website features, predicts phishing probability, generates a security risk score, and provides users with recommendations to avoid cyber fraud. The system aims to improve cybersecurity awareness and enhance online safety.

II. REVIEW OF LITERATURE

Cybersecurity researchers have proposed various techniques for detecting phishing websites using machine learning and data mining approaches. Early phishing detection systems depended primarily on manually maintained blacklists and rule-based filtering. Although effective against known phishing websites, these methods failed to detect zero-day phishing attacks.

Fette, Sadeh, and Tomasic (2007) introduced one of the earliest machine learning-based phishing detection systems by extracting features from emails and websites. Their study demonstrated that supervised learning algorithms significantly outperform traditional rule-based approaches.

Ma et al. (2009) proposed malicious URL detection using lexical and host-based features. Their research showed that URL characteristics alone can effectively identify suspicious websites without downloading webpage content, thereby improving detection speed.

Breiman (2001) developed the Random Forest algorithm, which combines multiple decision trees to improve classification performance while minimizing overfitting. Random Forest has since become one of the most widely adopted algorithms for phishing website detection due to its robustness and high prediction accuracy.

Recent studies have integrated Deep Learning, Natural Language Processing, browser extensions, and cloud-based threat intelligence into phishing detection systems. However, many existing solutions require high computational resources or depend on continuously updated phishing databases. Therefore, lightweight machine learning models remain highly relevant for real-time phishing detection.

III. PROBLEM OF THE STUDY

Phishing attacks continue to increase despite advancements in cybersecurity technologies. Attackers frequently create fraudulent websites that closely resemble legitimate banking, e-commerce, and social media websites. Internet users often fail to recognize these fake websites, resulting in identity theft, financial fraud, and unauthorized access to confidential information.

Traditional phishing detection mechanisms primarily depend on blacklist databases, which are ineffective against newly generated phishing websites. Furthermore, many existing systems do not provide immediate risk assessments or user-friendly interfaces.

This study addresses these issues by developing a machine learning-based phishing website detection system capable of identifying phishing websites in real time and providing users with security recommendations before they interact with suspicious websites.

IV. OBJECTIVES

The objectives of the study are:

- To develop a machine learning-based phishing website detection system.
- To analyze website URLs and extract security-related features.
- To classify websites as legitimate or phishing using the Random Forest algorithm.
- To generate a website risk score.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

V. HYPOTHESES

Null Hypothesis (H_0)

There is no significant relationship between website characteristics and the detection of phishing websites using machine learning algorithms.

Alternative Hypothesis (H_1)

There is a significant relationship between website characteristics and the detection of phishing websites using machine learning algorithms, resulting in improved phishing detection accuracy.

VI. RESEARCH GAP

The literature review reveals several limitations in existing phishing detection systems.
Most systems rely on blacklist databases that cannot detect newly created phishing websites.
Many studies consider only URL features while ignoring SSL certificates, webpage behavior, and domain characteristics.
Existing systems often require continuous database updates.
Few web-based systems provide both prediction results and security recommendations.
Limited research focuses on lightweight machine learning models suitable for real-time deployment.
The proposed system addresses these gaps by combining machine learning, risk scoring, and a user-friendly web interface into a single phishing detection platform.

VII. RESEARCH METHODOLOGY

Feature	Description
URL Length	Length of the website URL
HTTPS Status	Whether HTTPS protocol is used
Domain Age	Age of the registered domain
Number of Dots	Number of "." symbols in the URL
Number of Subdomains	Total number of subdomains
IP Address Usage	Whether an IP address is used instead of a domain name
URL Prefix/Suffix	Presence of '-' symbol in the domain
SSL Certificate	Availability of a valid SSL certificate
Redirection Count	Number of webpage redirections
Website Traffic	Popularity and traffic ranking of the website
DNS Records	Availability of DNS records
Target Class	Legitimate Website (0) or Phishing Website (1)

IX. DATA ANALYSIS

The collected data were analyzed using the **Random Forest Classification** algorithm. Data preprocessing included feature extraction, data cleaning, label encoding, and normalization where necessary. The model was trained using 80% of the dataset and tested using the remaining 20%.

The performance of the model was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

- Confusion Matrix

The Random Forest classifier effectively distinguished phishing websites from legitimate websites by analyzing multiple website characteristics. The model demonstrated high prediction accuracy and can be used for real-time phishing website detection.

X. RESULTS AND DISCUSSION

Experimental evaluation showed that the Random Forest classifier effectively predicts phishing websites with high accuracy. Compared to traditional blacklist-based approaches, the proposed system successfully identifies newly created phishing websites by analyzing multiple website characteristics.

The developed web application provides users with:

- Instant website classification.
- Security risk score.
- Safe browsing recommendations.
- Prediction history.

The system is computationally efficient and suitable for deployment in educational institutions, organizations, and personal cybersecurity applications.

XI. CONCLUSION

The research successfully developed an intelligent phishing website detection system using machine learning techniques. The Random Forest algorithm demonstrated reliable performance in distinguishing phishing websites from legitimate ones using multiple URL and website features.

The web-based application provides real-time predictions, risk scores, and security recommendations, thereby enhancing user awareness and reducing the risk of phishing attacks. Future improvements may include browser extension integration, real-time URL scanning, deep learning models, and cloud-based threat intelligence services.

The study concludes that Artificial Intelligence has significant potential to improve cybersecurity by providing accurate, efficient, and scalable phishing website detection systems.

REFERENCES

[1]. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

[2]. Fette, I., Sadeh, N., & Tomasic, A. (2007). *Learning to Detect Phishing Emails*. Proceedings of the 16th International Conference on World Wide Web, 649–656.

[3]. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). *Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs*. Proceedings of the ACM SIGKDD Conference.

[4]. UCI Machine Learning Repository. *Phishing Websites Dataset*. https://archive.ics.uci.edu/

[5]. PhishTank. *Community Site for Phishing Data*. https://phishtank.org/

[6]. Kaggle. *Phishing Website Detection Dataset*. https://www.kaggle.com/

[7]. OWASP Foundation. *Phishing Prevention Cheat Sheet*. https://owasp.org/

[8]. Python Software Foundation. *Python Documentation*. https://docs.python.org/3/

[9]. Scikit-learn Developers. *Scikit-learn Documentation*. https://scikit-learn.org/

[10]. Flask Documentation. https://flask.palletsprojects.com/