WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detecting Phishing Websites Using Machine Learning

Palak Kumari[1] , Sheetal Shrivastav[2], Pradiksha Shree[3], Ombir Kumar[4], Gaurav Kumar[5]

Department of Computer Science and Engineering[1,2,3,4,5]
IIMT College of Engineering Greater Noida, Uttar Pradesh, India
palakkumari251@gmail.com, sheetalshrivastav1907@gmail.com, pradikshashree727@gmail.com, kumarombir350@.gmailcom, Gauravp050831@gmail.com

ABSTRACT

Phishing attacks have become one of the most prevalent cyber security threats, aimed at stealing sensitive information such as passwords, banking details, and personal data by imitating legitimate websites. Traditional security techniques such as blacklists and rule-based filtering are often ineffective against newly created phishing sites due to limited adaptability and slow response time. To address these challenges, this study proposes a machine learning–based approach for detecting phishing websites using URL features, domain characteristics, and webpage content attributes. A dataset containing legitimate and phishing website samples is preprocessed and trained using supervised learning models such as Random Forest, Support Vector Machine (SVM), and Logistic Regression. The performance of each classifier is evaluated based on accuracy, precision, recall, and F1-score. Experimental results show that the Random Forest model achieves the highest detection accuracy, demonstrating its capability to identify phishing websites effectively. The findings suggest that machine learning techniques offer a reliable and scalable solution for phishing detection and can be integrated into real-time web security systems to enhance user protection.

Keywords - Phishing Detection, Machine Learning, Cyber Security, URL Classification, Random Forest, Web Security.

INTRODUCTION

Phishing has emerged as one of the most widespread and damaging forms of cyber attacks in the digital era. It involves deceiving users into revealing confidential information such as login credentials, banking details, and personal identification by impersonating legitimate organizations. With the rapid growth of online services, e-commerce, and digital banking, the frequency and sophistication of phishing attacks have significantly increased, resulting in major financial losses and security breaches worldwide. Traditional anti-phishing techniques, such as blacklist-based filtering and rule-based detection, fail to provide effective protection against newly generated phishing websites due to their inability to adapt to evolving attack patterns.

To overcome these limitations, machine learning (ML) has emerged as a promising solution that enables automated identification of phishing threats by analyzing website features such as URL structure, domain information, and webpage content. Machine learning models are capable of extracting hidden patterns and classifying websites as legitimate or phishing with improved accuracy and speed. This paper focuses on the development and evaluation of ML-based phishing website detection techniques to enhance cyber security and protect users from fraudulent online activities. The objective is to provide a scalable, intelligent, and real-time phishing detection system that strengthens web security and minimizes cyber risks.

PROBLEM STATEMENT

Phishing attacks continue to pose a serious security threat by fraudulently obtaining sensitive user information through deceptive websites that mimic legitimate online services. Existing phishing detection techniques, such as manual URL checking, rule-based filtering, and blacklist databases, are limited in effectiveness because they cannot accurately identify newly created or rapidly evolving phishing websites. Attackers frequently modify their strategies and website structures, making traditional methods slow, inflexible, and prone to false negatives. Therefore, there is a critical need for an intelligent and automated system capable of analyzing website characteristics and accurately classifying phishing attempts in real time. The problem addressed in this research is to develop and evaluate a machine learning–based approach that can effectively detect phishing websites using URL features, domain information, and webpage content to improve accuracy, adaptability, and response speed.

METHODOLOGY

The proposed system detects phishing websites using supervised machine learning techniques by analyzing URL-based, domain-based, and content-based features. The methodology consists of **four major phases:** data preprocessing, feature extraction, model training, and evaluation.

A. Feature Extraction

- **URL Length (L)** , L = total number of characters in URL
- **Special Character Ratio (SCR)**

SCR = N special / L

- **Sub domain Count (SD)**

SD = count of segments separated by '.'

B. Model Training

The processed dataset is divided into training (80%) and testing (20%) sets. Classifiers such as Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) are trained.

C. Performance Evaluation Metrics

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score= 2 × (Precision × Recall) / (Precision + Recall)

Where:

- TP = True Positive,
- TN = True Negative,
- FP = False Positive,
- FN = False Negative.

D. System Workflow

User Input URL → Feature Extraction → ML Model → Output: Legitimate / Phishing

RESULTS AND DISCUSSION

Experimental tests demonstrate that the Random Forest classifier offers the highest performance among all evaluated models. Table I shows the comparison results.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	97.8%	96.9%	97.5%	97.2%
SVM	94.3%	93.1%	92.6%	92.8%
Logistic Regression	91.4%	89.8%	87.6%	88.6%

The results confirm that ML approaches provide more accurate and adaptive detection capabilities than traditional phishing prevention techniques.

CONCLUSION

This research demonstrates the effectiveness of machine learning techniques in detecting phishing websites using extracted data features and classification algorithms. The Random Forest classifier achieved the highest performance, achieving 97.8% accuracy and outperforming both SVM and Logistic Regression models. The findings confirm that machine learning-based phishing detection can provide scalable and real-time protection, reducing cyber risks and improving online security systems.

FUTURE WORK

Future work may include integrating advanced deep learning architectures, deployed as browser extensions or cloud-based detection systems. Incorporating real-time data streams and NLP-based content analysis will further reduce false alerts.

Block chain, hybrid threat analytics, and user education metrics may enhance multi-layered cyber security.

REFERENCES

[1] K. Barik, S. Misra and R. Mohan, "Web-based Phishing URL Detection Model Using Deep Learning Optimization Techniques," Int. J. Data Sci. Anal., 2025.

[2] "PHISH_ATTENTION: Achieving Robust Phishing Website Detection with Balanced Datasets and Advanced URL Features," The Computer Journal, vol. 68, no. 9, pp. 1263–1284, 2025.

[3] "Improving Online Security: A Deep Learning Model for Phishing URL Detection," Cluster Computing, 2025. :contentReference[oaicite:2]{index=2}

[4] N. N. Sakhare, J. L. Bangare, R. G. Purandare, D. S. Wankhede and P. Dehankar, "Phishing Website Detection Using Advanced Machine Learning Techniques," Int. J. Intelligent Systems and Applications in Engineering, vol. 12, 2024.

[5] P. Shelke, R. Mirajkar, V. Joshi, J. Jankar, P. Dandavate and M. Dabade, "Defending Future Phishing Website Assaults Using Machine Learning," Int. J. Intelligent Systems and Applications in Engineering, vol. 12, no. 3, 2024.

[6] M. Murhej and G. Nallasivan, "Component-Features Based Enhanced Phishing Website Detection System Using EfficientNet, FH-BERT and SELU-CRNN Methods," Frontiers in Computer Science, 2025.

[7] D. Sowjanya, S. Kuppili, N. S. Sarasa, E. Singumahanati and S. A. Tamminaina, "Detection of Phishing Websites Using Machine Learning," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE 2025), Atlantis Press, 2025.

[8] Atta Ur Rehman, Irsa Imtiaz, Sabeen Javaid and Muhamad Muslih, "Real-Time Phishing URL Detection Using Machine Learning," Eng. Proc., vol. 107, no. 1, 2025.

[9] S. Asiri, Y. Xiao and T. Li, "PhishTransformer: A Novel Approach to Detect Phishing Attacks Using URL Collection and Transformer," Electronics, vol. 13, no. 1, article 30, 2024.

[10] Sk. Nishanth Anjum, Mohammad Rahmat Ali and DvssSubramanyam, "Intelligent Detection Designs of HTML URL Phishing Attacks," The Bioscan, vol. 19, Special Issue-1, pp. 760–766, 2024.