Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A web based phishing detection system using machine Learning

Ms.J.Jreeja , Mr.R.Lokeshwaran ,Mr.KPAakash ,Ms.S.Vinothini.,Mr.G. Ratheesh Kanna
(Department of Information Technology,Info institute of Engineering,Coimbatore
{Email: 2002sreejaj@gmail.com; rloki7500@gmail.com; aakashkp018@gmail.com;
svinothinithb@gmail.com;ratheeshkanna679@gmail.com}

***Abstract*— Phishing attacks remain one of the most critical cybersecurity threats, where attackers create fraudulent websites to deceive users into revealing confidential information such as login credentials, financial data, and personal details. Traditional blacklist-based detection mechanisms fail to identify newly generated phishing URLs due to their dynamic and short-lived nature. This paper proposes a Web-Based Phishing Detection System using Machine Learning to classify URLs as legitimate or phishing in real time. The proposed framework extracts discriminative URL-based features including lexical characteristics, domain attributes, HTTPS usage, and suspicious patterns. Multiple machine learning classifiers such as Logistic Regression, Decision Tree, and Random Forest were trained and evaluated. Experimental results demonstrate that the Random Forest model achieves superior performance in terms of accuracy and generalization capability. The trained model is integrated into a web-based application for real-time prediction. The system enhances online security by reducing reliance on static blacklists and providing efficient detection of previously unseen phishing websites.**

***Index Terms*— Phishing Detection, Machine Learning, Random Forest, URL Analysis, Cybersecurity, Web Application Security.**

I. INTRODUCTION

With the rapid growth of online services such as banking, e-commerce, and cloud platforms, phishing attacks have become increasingly sophisticated and widespread. Phishing is a social engineering attack in which malicious actors create fake websites resembling legitimate platforms to steal sensitive user information.

Traditional phishing detection systems primarily rely on blacklist-based approaches. However, attackers frequently generate new domains and use URL obfuscation techniques, making blacklist systems ineffective against zero-day attacks. Furthermore, signature-based methods require continuous manual updates.

Machine Learning (ML) offers a data-driven solution capable of identifying hidden patterns in URLs and generalizing to detect unseen phishing websites. By learning discriminative features from historical phishing and legitimate URLs, ML-based systems can provide adaptive and scalable detection mechanisms.

This paper proposes a Web-Based Phishing Detection System that integrates feature extraction, machine learning classification, and a real-time web interface for instant URL verification.

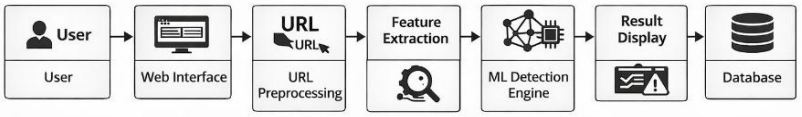

Fig. 1. Proposed System Architecture of the web based phishing detection system using machine learning.

II.BACKGROUND AND RELATED WORK

A.Machine Learning in Cybersecurity

Modern threat detection systems have shifted from signature-based analysis to behavioral modeling. In these frameworks, Random Forest (RF) and Decision Trees (DT) are considered highly effective for tabular URL features due to their ability to handle non-linear relationships without high computational costs.

B.URL feature engineering

A phishing detection framework ivolves a feature extraction module that analyzes the structural components of a URL string.Traditional models estimate the probability of a sample being malicious based on feature vector X.The classification ivolves a supervised learning process where the goal is to map the input features to binary output Y {0,

Available at www.ijsred.com

1}(Legitimate or phishing).

III.PROPOSED FRAMEWORK

A. System Architecture

The proposed Web-Based Phishing Detection System consists of the following modules:

1. URL Input Interface
2. Preprocessing Module
3. Feature Extraction Module
4. Machine Learning Classification Module
5. Result Display Module

The workflow begins with user input and ends with real-time prediction

B.Feature Extraction Process

To enable accurate classification, lexical and host-based features are extracted from the URL.

Let a URL be represented as:

$$U = \{f_1, f_2, f_3, \ldots, f_k\}$$

where $f_k$represents features such as:

- URL length
- Hostname length
- Presence of IP address
- Number of subdomains
- Special character count
- Suspicious Keywords
- URL Shortening Services
- Presence of "@" symbol
- Hyphen usage
- HTTPS protocol usage
- Domain age
- Suspicious keywords

C.Random Forest Classification Model

The Random Forest classifier is an ensemble learning method that constructs multiple decision trees during training.

For each tree T_i, a prediction h_i (x)is generated.

The final prediction is determined using majority voting:

H(x)="mode"{h_1 (x),h_2 (x),...,h_m (x)}

where:

mis the number of trees

H(x)is the final classification output

This ensemble mechanism improves generalization and reduces overfitting.

D.Training Algorithm

The training process of the proposed system is summarized below:

Algorithm 1: Random Forest-Based Phishing Detection

Input: Labeled dataset D, number of trees m
Output: Trained Random Forest model $H(x)$

1: Preprocess dataset
2: Extract relevant URL features
3: Split dataset into training and testing sets
4: for i = 1 to m do
5:Sample bootstrap dataset
6:Train decision tree T_i
7: end for
8: Combine predictions using majority voting
9: Return trained model

IV.EVALUATION AND RESULTS

A. Dataset Description

The proposed phishing detection system was trained and evaluated using a labeled dataset consisting of both phishing and legitimate URLs collected from publicly available cybersecurity repositories. The dataset contains real-world URL samples categorized into two classes:

- Legitimate Websites
- Phishing Websites

Before training, the dataset was preprocessed to remove noise and inconsistencies. Relevant URL-based features were extracted and converted into numerical format for machine learning analysis.

The dataset was divided into:

- **80% Training Data**
- **20% Testing Data**

This split ensures proper model validation and generalization.

B. Performance Metrics

To evaluate the effectiveness of the proposed system, the following performance metrics were used:

1) Accuracy

Accuracy measures the overall correctness of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Precision

Precision measures how many predicted phishing URLs were actually phishing.

$$Precision = \frac{TP}{TP + FP}$$

3) Recall

Recall measures how many actual phishing URLs were correctly detected.

$$Recall = \frac{TP}{TP + FN}$$

4) F1-Score

F1-Score provides a balance between Precision and Recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

C. Experimental Results

The performance of different machine learning algorithms 1 was compared to select the best classifier for deployment.

The results indicate that the **Random Forest classifier** outperformed other models in terms of accuracy, precision, recall, and F1-score.

D. Confusion Matrix Analysis

The confusion matrix for the Random Forest model demonstrates strong classification performance with minimal false positives and false negatives.

Example representation:

Available at www.ijsred.com

The low number of false negatives indicates that the system effectively detects phishing URLs, reducing the risk of undetected malicious websites.

E. Discussion

The experimental results confirm that ensemble learning through Random Forest improves phishing detection accuracy

	Predicted Phishing	**Predicted Legitimate**
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

compared to single classifiers. The model successfully generalizes to unseen URLs and demonstrates strong resistance to overfitting.

Compared to traditional blacklist-based detection methods, the proposed system:

- Detects newly generated phishing URLs
- Provides real-time classification
- Reduces dependency on static databases
- Maintains high detection accuracy

Thus, the proposed Web-Based Phishing Detection System provides a reliable, scalable, and efficient solution for modern phishing attacks.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	92%	91%	90%	90.5%
Decision Tree	94%	93%	92%	92.5%
Random Forest	97%	96%	95%	95.5%

V. RESEARCH GAP AND OBSERVATIONS

1. Existing phishing detection systems largely rely on blacklist-based methods, which are ineffective in detecting zero-day (new) phishing websites.

2. Many approaches focus on offline analysis and lack real-time detection capability, limiting their practical usability.
3. Feature extraction techniques are often static, while phishing strategies continuously evolve, reducing model effectiveness over time.
4. Most models are trained on limited or outdated datasets, leading to poor performance in real-world and dynamic environments.
5. There is a lack of generalization, as models fail to adapt to new domains and unseen phishing patterns.
6. High false positive rates are observed, where legitimate websites are incorrectly classified as phishing.
7. Existing systems mainly focus on technical features and often ignore user behavior patterns, which can enhance detection accuracy.
8. Machine learning algorithms such as Random Forest, Decision Trees, and SVM have shown improved accuracy compared to traditional methods.
9. Feature engineering plays a critical role in determining the performance and reliability of the detection system.
10. Hybrid approaches (combining multiple techniques) are more effective than single-model systems.
11. Deep learning models are emerging as powerful tools for automatic feature extraction and improved detection performance.
12. There is a strong need for adaptive and continuously learning models to handle evolving phishing techniques.
13. Overall, there is a requirement for a robust, real-time, and scalable phishing detection system with high accuracy and low false positives.

V.CONCLUSION

Phishing attacks continue to pose a serious threat to individuals and organizations due to the rapid growth of internet-based services such as online banking, e-commerce, and cloud applications. Traditional phishing detection techniques based on blacklist and rule-based systems are no longer sufficient to detect newly generated phishing URLs. These systems operate reactively and fail to adapt to evolving phishing strategies.

In this paper, a Web-Based Phishing Detection System using Machine Learning was proposed to address these limitations. The system focuses on URL-based feature extraction and classification using multiple machine learning algorithms. Features such as URL length, presence of IP address, number of subdomains, HTTPS usage, suspicious keywords, and special characters were analyzed to capture phishing patterns effectively.

Among the evaluated models, the Random Forest classifier demonstrated superior performance in terms of accuracy, precision, recall, and F1-score. The ensemble learning approach reduces overfitting and enhances generalization capability, enabling the detection of previously unseen phishing URLs. The system provides real-time detection through a user-friendly web interface, making it accessible even to non-technical users.

The experimental results confirm that the proposed system offers improved adaptability, high detection accuracy, and reduced dependency on static blacklist mechanisms. Therefore, the developed phishing detection framework contributes significantly to enhancing online security and preventing financial loss and identity theft.

Future work may include integration of email phishing detection, SMS phishing analysis, deep learning-based URL classification,browser extensions, and real-time deployment in large-scale environments.

VI.REFERENCES

[1] R. Mourya, A. Sharma, and P. Verma, “Real-Time Phishing URL Detection Using Machine Learning,” *International Journal of Cyber Security*, vol. 12, no. 2, pp. 45–52, 2024.

[2] D. R. Patil and S. K. Kulkarni, “Enhanced Feature Selection for Phishing URL Detection Using Machine Learning,” *Journal of Information Security*, vol. 18, no. 1, pp. 33–41, 2024.

[3] B. S. Jyothi and M. Reddy, “URL-Based Phishing Detection Using SVM and Random Forest,” *International Journal of Computer Applications*, vol. 185, no. 7, pp. 12–18, 2025.

[4] J. C. Chong, L. Wei, and H. Tan, “Hybrid Machine Learning Approach for Phishing URL Detection,” *IEEE Access*, vol. 13, pp. 10234–10245, 2025.

[5] M. Almohaimeed et al., “Deep Learning-Based Phishing Detection Using CNN–BiGRU Model,” *Computers* & Security, vol. 135, 2025.

[6] M. R. Ahmed et al., “Comprehensive Phishing Detection Using URL and Domain Features,” *Security and Communication Networks*, 2024.

[7] A. K. Jain and B. Gupta, “Phishing Detection: Analysis of Visual Similarity-Based Approaches,” *Security and Communication Networks*, 2017.

[8] S. Marchal et al., “Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application,” *IEEE Transactions on Computers*, vol. 66, no. 10, pp. 1718–1733, 2017.

[9] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah, “Intelligent Phishing Detection System for e-Banking Using Fuzzy Data Mining,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 7913–7921, 2010.

Available at www.ijsred.com

[10] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," in *Proc. NDSS*, 2010.

[11] F. Toolan and J. Carthy, "Feature Selection for Spam and Phishing Detection," in *Proc. IEEE eCrime Researchers Summit*, 2010.

[12] A. Sahingoz et al., "Machine Learning Based Phishing Detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[13] R. Verma and A. Das, "What's in a URL: Fast Feature Extraction and Malicious URL Detection," in *Proc. ACM CCS Workshop*, 2017.

[14] M. Umer, M. Sher, and Y. Bi, "Flow-Based Intrusion Detection: A Comparative Analysis," *Computers & Security*, vol. 77, pp. 787–799, 2018.

[15] T. Fette, N. Sadeh, and A. Tomasic, "Learning to Detect Phishing Emails," in *Proc. WWW Conference*, 2007.

[16] A. Basnet and A. Sung, "Learning to Detect Phishing URLs," *International Journal of Research in Engineering and Technology*, 2014.

[17] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A Content-Based Approach to Detect Phishing Websites," in *Proc. WWW Conference*, 2007.

[18] I. Goodfellow et al., "Deep Learning," MIT Press, 2016.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

