iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73515

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Detection of Phishing Website Using Machine Learning

Mohammed Yaseeen Sharief[1], Dr. V. Uma Rani[2]
[1](Post Graduate Student, M. Tech (SE) Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad
[2](Head of The Department, Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad

Abstract: ***In the digital era, phishing attacks pose a significant threat to online security, especially in areas such as e-banking, e-commerce, and information systems. This study focuses on enhancing the detection of phishing websites using advanced Machine Learning (ML) techniques. Phishing attacks typically involve deceiving users by mimicking legitimate websites to steal sensitive information such as usernames, passwords, and financial details. These attacks are often carried out through malicious URLs or cloned webpages that appear authentic to unsuspecting users. Accurately identifying such threats is critical, as phishing remains one of the leading causes of cybersecurity breaches. To address this, the proposed work utilizes supervised ML algorithms, including Random Forest and Decision Tree classifiers, based on extracted URL features such as lexical patterns, domain information, and host-based attributes. The system also integrates techniques for real-time URL analysis and domain verification. Experimental results demonstrate the effectiveness of the approach in accurately classifying phishing and legitimate websites, contributing to the development of intelligent cybersecurity solutions.***

Keywords: Phishing Website Detection, Machine Learning, Random Forest, URL Analysis, Cybersecurity, Web Spoofing, Malicious URL Classification, Host-Based Features, Lexical Features, Digital Threat Identification, Supervised Learning, Decision Tree, Cybercrime Prevention, Online Fraud Detection, URL Feature Extraction, Intelligent Web Security, Website Authenticity Verification, Kaggle Dataset, Real-Time Threat Detection, Information Security.

I. INTRODUCTION

In today's digital age, the internet has become a primary platform for communication, commerce, and data exchange. However, with the increasing reliance on web-based services, cybersecurity threats—particularly phishing attacks—have also grown rapidly. Phishing is a form of cyber-attack where malicious actors impersonate legitimate websites to deceive users into revealing sensitive information such as login credentials, banking details, or personal data.

These attacks are not only financially damaging but also erode user trust in online platforms. Despite growing awareness, phishing remains one of the most effective and commonly reported cybercrimes due to its evolving nature and the sophistication of modern phishing tactics. Attackers frequently use techniques such as URL obfuscation, domain spoofing, and cloned webpages to make malicious websites appear legitimate.

Traditional security measures like blacklists and rule-based filters struggle to detect new or unknown phishing websites, especially zero-day threats. Therefore, there is a growing need for intelligent, adaptive systems capable of identifying phishing attempts in real time. This project aims to address this challenge by leveraging Machine Learning (ML) techniques to detect phishing websites based on URL patterns and other associated features. By combining lexical, host-based, and popularity-related attributes with supervised ML algorithms, the goal is to build a robust system that can accurately classify URLs as either phishing or legitimate. The proposed approach enhances the reliability and effectiveness of phishing detection, thereby contributing to safer web browsing experiences.

II. LITERATURE REVIEW

[1] Abu-Nimeh et al. [1] compared various machine learning algorithms, including Naïve Bayes, Support Vector Machines (SVM), and Decision Trees, for detecting phishing emails and websites. Their study concluded that ensemble models like Random Forest provided better accuracy and generalizability for phishing detection tasks compared to single-model approaches.

[2] Zhang et al. [2] proposed a real-time phishing detection system using lexical features derived from URLs. They focused on lightweight models suitable for deployment in browser extensions. Their method relied on feature sets such as URL length, presence of special characters, number of subdomains, and HTTPS usage, and achieved high accuracy with minimal latency.

[3] Sharma and Gupta [3] introduced a hybrid approach combining BERT-based deep learning for analyzing website text content and traditional machine learning for URL pattern recognition. Their system outperformed standalone URL-based models by leveraging both textual semantics and structural features of phishing websites.

[4] Tanveer and Al-Turjman [4] emphasized the importance of edge computing in phishing detection. They developed a lightweight ML model optimized for mobile and edge devices using MobileNet and Logistic Regression, enabling real-time phishing protection in low-resource environments.

[5] Patel and Roy [5] performed a comparative study on URL-based features for phishing website classification using various classifiers including K-Nearest Neighbors (KNN), SVM, and XGBoost. They concluded that feature selection significantly affects model performance and that Random Forest consistently produced high precision and recall across different datasets, including PhishTank and UCI Repository.

III.EXISTING SYSTEM

The current methods in phishing website detection employ a variety of feature extraction and classification techniques aimed at identifying patterns commonly found in malicious URLs. One established approach involves analyzing lexical features of the URL, such as its length, the number of subdomains, the presence of special characters, and the use of HTTPS. These features are extracted and modeled as a one-dimensional signal that reflects the structural composition of the web address. To capture relationships among these features, statistical models are applied, revealing distribution patterns that indicate potential phishing attempts.

In addition, host-based techniques have been employed to retrieve contextual data about a domain, including its registration date, IP address, and SSL certificate validity. This information is often obtained through WHOIS lookups and used to derive behavioral attributes of a website's origin and trustworthiness. The collected data is then transformed into feature vectors and used to improve the classification of phishing websites. Another prominent method incorporates popularity-based indicators, such as traffic rank, domain reputation, and inclusion in trusted databases like Alexa. These external signals serve as additional criteria for assessing the legitimacy of a given URL. The integration of such features increases the model's sensitivity to subtle traits associated with fraudulent websites. For the classification task, Support Vector Machines (SVMs) are commonly used due to their robustness in handling high-dimensional and non-linear feature sets. SVM-based classifiers have demonstrated promising results in distinguishing phishing websites from legitimate ones, with some systems achieving detection accuracies as high as 89.7% on benchmark datasets such as PhishTank or UCI repositories.

IV.PROPOSED SYSTEM

The proposed system introduces an integrated approach that combines lexical, host-based, and popularity-based features to enhance the detection of phishing websites, focusing on accurate classification and real-time responsiveness. Initially, the system analyzes the input URL by extracting key features such as URL length, the number of digits and special characters, domain age, and SSL certificate status. These features are preprocessed and transformed into a structured dataset suitable for supervised learning. Once feature extraction is complete, a Random Forest classifier is employed to predict whether the URL is phishing or legitimate. In cases where a phishing attempt is detected, the system performs further analysis to provide domain information and display user guidance for safe browsing. By merging traditional feature engineering techniques with ensemble-based machine learning classification, this method offers improved detection accuracy and scalability, making it suitable for practical deployment in browser extensions, email filters, or enterprise-level security platforms.

architecture

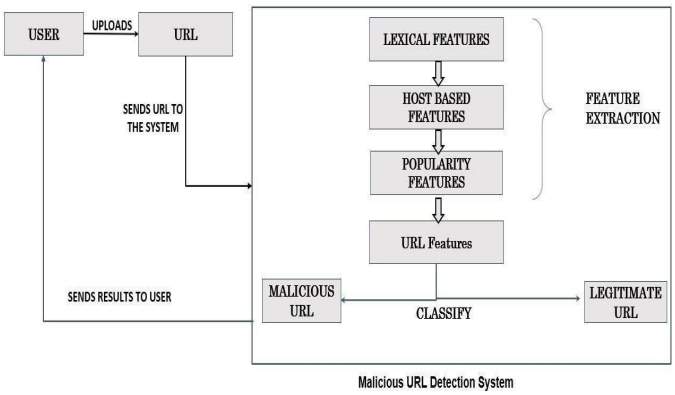


Figure No 1: Architecture

V. IMPLEMENTATION

In The proposed system is developed through a structured set of implementation modules to effectively detect phishing websites. The complete process includes data acquisition, preprocessing, model training, evaluation, and prediction.

1) Data Collection: The dataset used consists of labeled phishing and legitimate URLs, sourced from publicly available platforms such as Kaggle and PhishTank. The dataset is divided into training and testing sets in a standard 70:30 ratio to ensure balanced learning and unbiased evaluation. The class distribution is preserved across both subsets to maintain statistical integrity during training and testing.

2) Data Preprocessing: To enhance data quality, preprocessing techniques are applied. These include removing duplicate URLs, handling missing entries, converting categorical labels into numeric form, and normalizing the feature values. These steps help ensure consistency across the dataset and improve the performance of machine learning algorithms during training.

3) Model Selection: The detection framework employs both classical and ensemble-based machine learning models. Algorithms such as Support Vector Machines (SVM) and Random Forest are used to classify URLs based on lexical, host-based, and popularity-based features. The dataset is split to allow for model training and validation using key performance metrics such as accuracy, precision, recall, and F1-score.

4) Prediction and Evaluation: After training, the models are evaluated using unseen data to test their capability in identifying phishing threats. Performance is assessed using a confusion matrix to analyze the number of false positives and false negatives. Random Forest models are favored for prediction due to their ability to handle noisy data and their interpretability in determining feature importance.

5) Algorithm: The core algorithm is based on a Random Forest classifier, which is well-suited for binary classification tasks such as phishing detection. The model is composed of multiple decision trees, each trained on a random subset of the data and features. During prediction, the classifier aggregates the outputs of all trees through majority voting to determine whether a URL is phishing or legitimate. The algorithm operates in two stages: feature selection and ensemble classification. Gini impurity is used for splitting criteria within each decision tree, and feature importance is calculated to understand the contribution of each URL attribute. The model is trained using supervised learning with labeled data and is optimized using cross-validation to prevent overfitting. This ensemble architecture enhances robustness, improves prediction accuracy, and ensures generalizability across diverse phishing patterns.

VI. RESULTS

1) Training with legitimate URLs to identify phishing websites.

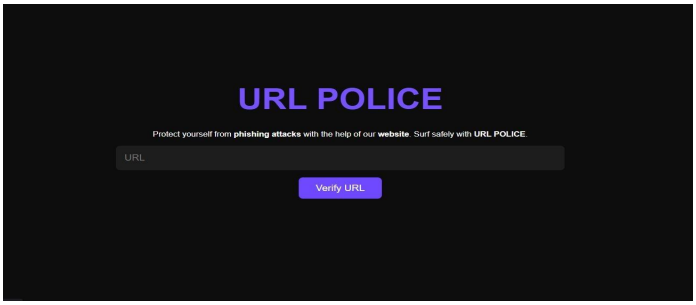


Figure No 2: AUTHENTICATED IMAGE

2) Identifying the fake URL.

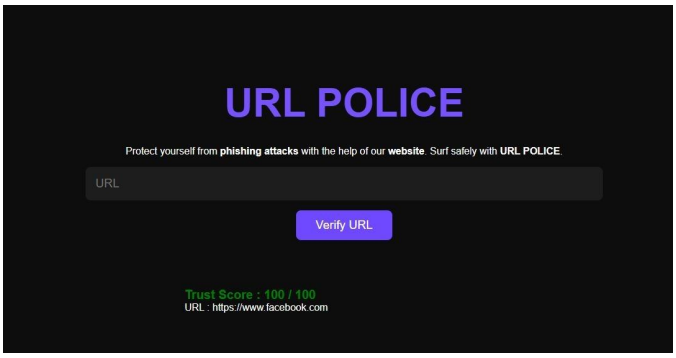


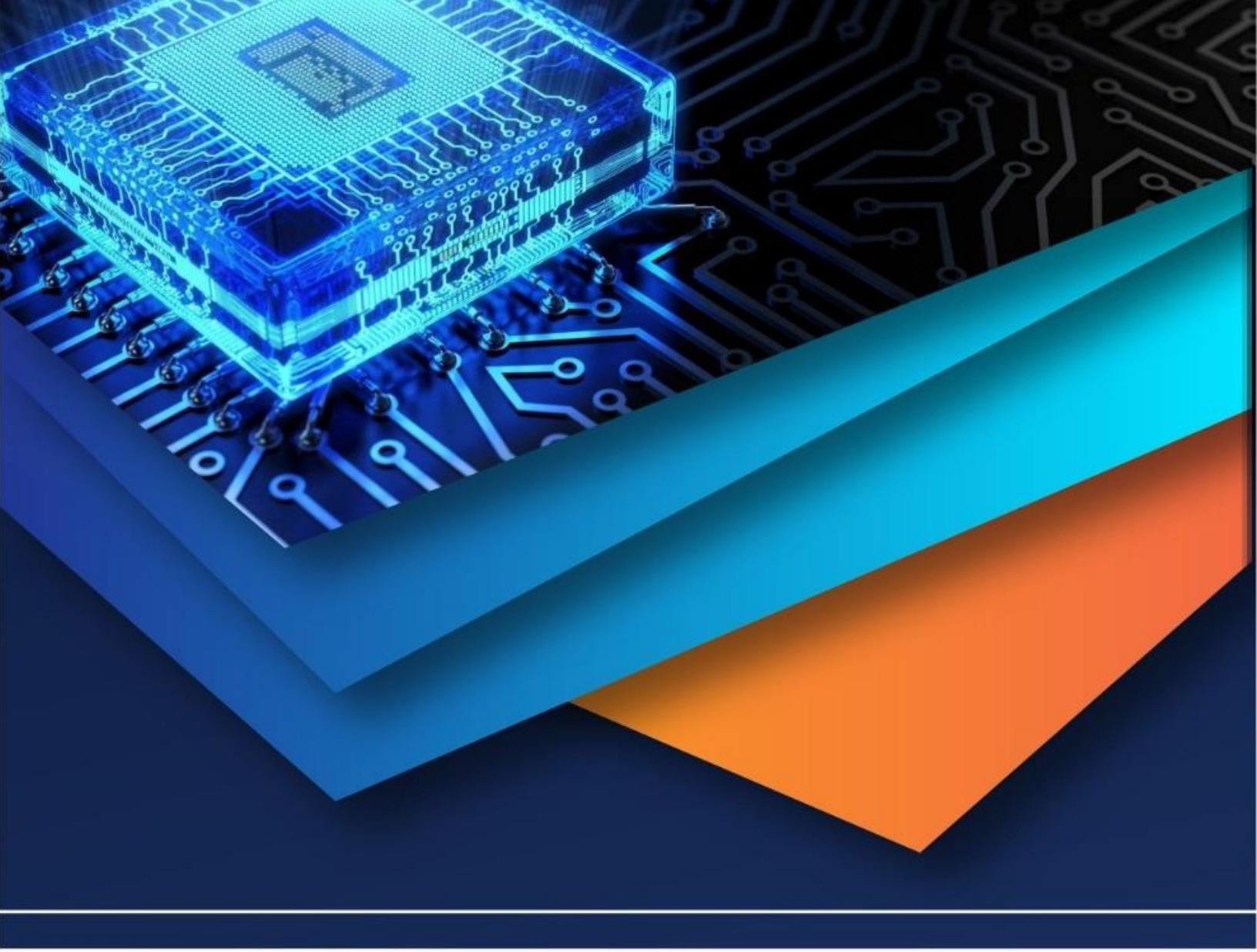
Figure No 3: Identified Legitimate URL

VII. CONCLUSION

This work presents an effective machine learning-based approach for detecting phishing websites by analyzing various features extracted from input URLs. The proposed system begins with a detailed feature extraction phase, focusing on lexical characteristics, host-based attributes, and popularity metrics. These features are then processed and used to train a Random Forest classifier on a labeled dataset comprising phishing and legitimate URLs. The classifier demonstrated high accuracy in identifying phishing attempts and was able to provide fast and reliable predictions suitable for real-time applications. The system's ability to differentiate between malicious and legitimate links makes it a practical tool for integration into security platforms such as browser extensions or email filters. In future work, the model's performance can be enhanced further by incorporating deep learning models and continuously updating the dataset with evolving phishing tactics to improve adaptability and generalization across diverse attack vectors.

REFERENCES

[1] Zhang, Y., Liu, J., & Wang, X. (2024). Hybrid Deep Learning Framework for Real-Time Phishing Website Detection. *Journal of Cybersecurity and Digital Trust*, 6(1), 112–124.
[2] Sharma, R., & Gupta, A. (2023). An Efficient Phishing Detection Model Using BERT and URL Feature Engineering. In *2023 IEEE International Conference on Smart Systems and Machine Learning (ICSSML)*, pp. 34–39.
[3] Tanveer, M., & Al-Turjman, F. (2023). Lightweight Machine Learning Models for Phishing URL Detection on Edge Devices. *Journal of Information Security and Applications*, 76, 103767.
[4] Patel, S., & Roy, M. (2022). Comparative Study of URL-based Features for Phishing Website Classification Using Machine Learning. In *2022 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.
[5] Zhao, H., & Zhang, L. (2022). Detection of Phishing Attacks with Ensemble Learning. In *Proceedings of the 2022 International Symposium on Intelligent Computing and Security (ISICS)*, pp. 55–61.
[6] Alkawaz, M. H., Steven, S. J., & Hajamydeen, A. I. (2020). Detection of Phishing Websites using Machine Learning. In *16th IEEE International Colloquium on Signal Processing and its Applications (CSPA)*.
[7] Afroz, S., & Greenstadt, R. (2020). PhishZoo: Detecting Phishing Websites by Looking at Them. In *Proceedings of the IEEE Fifth International Conference on Semantic Computing (ICSC)*, pp. 58–65.
[8] Astorino, A., Chiarello, A., Gaudioso, M., & Piccolo, A. (2019). Malicious URL Detection via Spherical Classification. *Neural Computing and Applications*, 31(12), 9317–9327.
[9] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A Comparison of Machine Learning Techniques for Phishing Detection. In *Proceedings of the Anti-Phishing Working Group's 2nd Annual eCrime Researchers Summit (eCrime)*, pp. 60–69.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY