Machine Learning Based Phishing URL Detection System

Disha Badgujar[1], Yashika Thakur[2], Anjali Tiwari[3], Meghansh Saxena[4], Rajdeep Shrivastava[5]
[1,2,3,4] *Students, Department of Electronics and Communication Engineering, Lakshmi Narain College of Technology Excellence, Bhopal, India*
[5] *Associate Professor, Department of Electronics and Communication Engineering, Lakshmi Narain College of Technology Excellence, Bhopal, India*

Abstract -Phishing attacks have emerged as one of the most dangerous cybersecurity threats in the digital era. Attackers use deceptive websites and malicious URLs to steal confidential information such as usernames, passwords, and banking credentials. Traditional phishing detection systems based on blacklists and heuristic methods are often unable to identify newly generated phishing websites. This research paper presents a machine learning based phishing URL detection system capable of classifying URLs as legitimate or phishing using extracted lexical and domain-based features. Various machine learning algorithms including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting were trained and evaluated. Among all the models, Gradient Boosting achieved the highest accuracy of approximately 97%. The trained model was integrated into a Flask-based web application to provide real-time phishing detection. The proposed system offers a scalable, efficient, and intelligent approach for improving cybersecurity and protecting users from online threats.

I.INTRODUCTION

The internet has become an essential part of modern life, enabling online banking, social networking, e-commerce, and digital communication. However, the increasing dependency on online services has also led to a rapid rise in cyberattacks. Phishing is one of the most common forms of cybercrime where attackers impersonate trusted organizations to trick users into revealing sensitive information. Traditional phishing detection techniques such as blacklist-based systems are limited because they cannot effectively detect zero-day attacks or newly generated websites.

Machine learning techniques provide a more dynamic and intelligent approach for phishing detection. Instead of relying on predefined rules, machine learning models learn patterns from datasets containing phishing and legitimate URLs. This paper proposes a phishing URL detection system that extracts URL-based features and classifies websites using supervised machine learning algorithms.

II. LITERATURE REVIEW

Several researchers have contributed to phishing website detection using machine learning and content-based approaches. Choon Lin Tan et al. proposed a URL assisted brand name weighting system to improve phishing website detection. Zhang et al. introduced the Cantina system, which used webpage content analysis for detecting phishing websites. Reports published by the Anti-Phishing Working Group (APWG) highlighted the rapid growth of phishing attacks and the limitations of traditional blacklist-based detection systems. Recent studies demonstrate that machine learning algorithms such as Random Forest, SVM, and Gradient Boosting significantly improve phishing detection accuracy.

III. PROPOSED METHODOLOGY

The proposed system follows a machine learning pipeline consisting of data collection, feature extraction, data preprocessing, model training, evaluation, and deployment. The system accepts a URL as input and extracts multiple lexical and domain-based features. These features are then passed to a trained machine learning model that predicts whether the URL is phishing or legitimate.

III.1 Feature Extraction

Feature	Description
URL Length	Measures total characters in the URL
HTTPS Usage	Checks whether the website uses HTTPS
Special Characters	Detects suspicious symbols such as @ and -
Number of Dots	Identifies excessive subdomains
Domain Age	Analyzes trustworthiness of the domain
DNS Record	Checks existence of valid DNS information

III.2 Machine Learning Algorithms

Multiple supervised learning algorithms were used for phishing URL classification:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting

The dataset was divided into training and testing sets using an 80:20 ratio.

IV. RESULTS AND ANALYSIS

The performance of different machine learning models was evaluated using accuracy, precision, recall, and F1-score. Gradient Boosting achieved the best performance with approximately 97% accuracy. The confusion matrix analysis indicated a high true positive rate and low false negative rate, making the model suitable for real-world phishing detection applications.

Fig 4.1: Homepage of Phishdetector

Figure 5.5.2: Output Showing Legitimate URL Detection of Phish-Detector

Figure 5.5.3:Output Showing Phishing URL Detection of Phish-Detector

IV.1 Accuracy Comparison

Algorithm	Accuracy
Logistic Regression	91%
Support Vector Machine	94%
Random Forest	96%
Gradient Boosting	97%

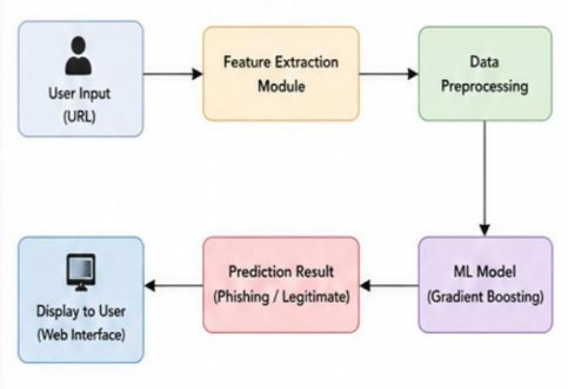
V. SYSTEM DEPLOYMENT

The trained Gradient Boosting model was deployed using the Flask framework. A web interface was developed using HTML and CSS, allowing users to enter URLs and receive real-time phishing predictions. The deployment ensures ease of use and scalability.

VI. CONCLUSION AND FUTURE SCOPE

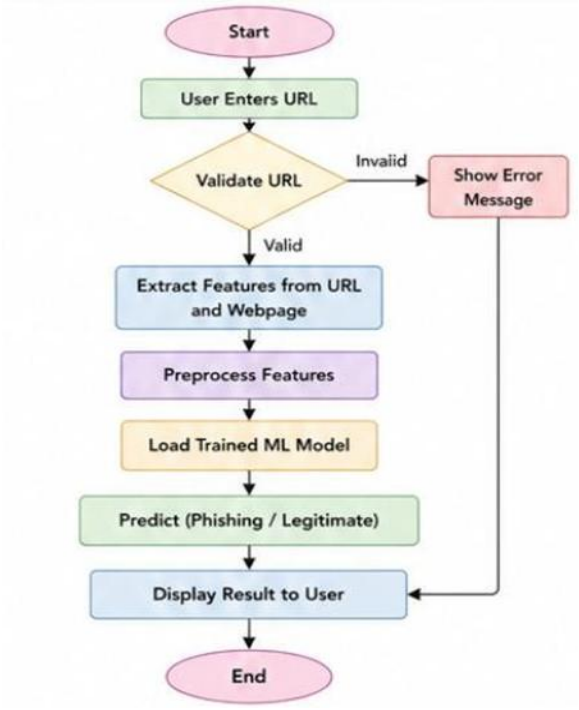
The proposed phishing URL detection system successfully demonstrates the effectiveness of machine learning in identifying malicious URLs. The Gradient Boosting model achieved high accuracy and provided reliable real-time predictions. Future improvements

may include deep learning techniques such as CNN and LSTM, browser extension integration, and real-time threat intelligence APIs to further improve phishing detection capabilities.



System Architecture and Flowchart

Block Diagram of the Proposed Phishing URL Detection System



Flowchart of the Proposed Phishing URL Detection System

REFERENCES

[1] Ali Aljofey et al., "An Effective Detection Approach for Phishing Websites Using URL and HTML Features," Scientific Reports, vol. 12, 2022.

[2] Cagatay Catal et al., "Applications of Deep Learning for Phishing Detection: A Systematic Literature Review," Knowledge and Information Systems, 2022.

[3] Nuria Reyes-Dorta et al., "Detection of Malicious URLs Using Machine Learning," Wireless Networks, vol. 30, 2024.

[4] Hayk Ghalechyan et al., "Phishing URL Detection with Neural Networks: An Empirical Study," Scientific Reports, 2024.

[5] Kousik Barik et al., "Web-Based Phishing URL Detection Model Using Deep Learning Optimization Techniques," International Journal of Data Science and Analytics, 2025.

[6] Marie-Laure E. Alorvor et al., "RealTime Phishing Detection for Brand Protection Using Temporal Convolutional Network-Driven URL Sequence Modeling," Electronics, 2025.

[7] Sneha Baskota, "Phishing URL Detection Using Bi-LSTM," arXiv preprint, 2025.