IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

PHISHING WEBSITE IDENTIFICATION SYSTEM

Mubeena Banu[1], Chandana K[2], Lavanya N L[3], Nayanashree D[4]

Student, Dept. of Computer Science and Engineering, East West College of Engineering, Bengaluru, India[1]

Student, Dept. of Computer Science and Engineering, East West College of Engineering, Bengaluru, India[2]

Project Guide, Dept. of Computer Science and Engineering, East West College of Engineering, Bengaluru, India[3]

Project Guide, Dept. of Computer Science and Engineering, East West College of Engineering, Bengaluru, India[4]

Abstract: The Phishing Website Identification System is designed to detect and prevent fraudulent websites that attempt to steal sensitive user information. Phishing attacks have become a major cybersecurity threat in the digital world. These attacks trick users into visiting fake websites that resemble legitimate platforms. The proposed system uses Machine Learning techniques to identify phishing websites. It analyzes various URL features, domain information, and website characteristics. The extracted features are processed and provided to a trained machine learning model. Algorithms such as Random Forest, Decision Tree, and Support Vector Machine can be used for classification. The system classifies the submitted website as either legitimate or phishing. It provides quick and automated detection of suspicious websites. The proposed approach aims to improve detection accuracy and reduce false-positive rates. It also helps identify potentially harmful websites and alerts users about security risks. This system can be integrated into web applications or browser-based security tools. Overall, the project aims to provide a reliable and intelligent solution for improving online safety and protecting users from phishing attacks.

Keywords: Phishing Detection, Machine Learning, Cybersecurity, URL Analysis, Website Classification, Random Forest, Malicious Websites, Online Security.

I. INTRODUCTION

The internet plays an important role in everyday life through online banking, shopping, education, and communication. However, the increasing use of the internet has also created various cybersecurity threats, especially phishing attacks. Phishing websites are fake websites designed to steal users' personal and confidential information.

These websites often imitate genuine websites to trick users into sharing passwords and banking details. Identifying phishing websites manually is difficult because they closely resemble legitimate websites. Traditional security methods may fail to detect newly created phishing websites.

To overcome these problems, the proposed system uses Machine Learning techniques for phishing website identification. It analyzes website URLs and extracts important features to identify suspicious activities. Features such as URL length, special characters, domain information, and suspicious keywords are considered.

The extracted features are processed using trained Machine Learning algorithms. Algorithms such as Random Forest, Decision Tree, and Support Vector Machine can be used. The system classifies the submitted website as either legitimate or phishing.

It provides quick and automated results to help users identify potentially harmful websites. The system aims to improve phishing detection accuracy and reduce false predictions. It also helps protect users from online fraud, identity theft, and unauthorized access.

The proposed approach provides an effective solution for improving website security. It can also be enhanced through browser extensions and advanced detection techniques. Overall, this project aims to create a safer and more secure internet browsing experience.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

II. BACKGROUND CONCEPTS

A. Internet Security

Internet security protects users and online information from cyber threats and unauthorized access. It helps maintain the privacy and safety of personal and financial information. It plays an important role in preventing online attacks and fraudulent activities.

B. Phishing Attacks

Phishing is a cyberattack in which attackers create fake websites to steal sensitive information. These websites imitate genuine websites to trick users into sharing passwords and banking details. Phishing attacks can cause financial loss and identity theft.

C. URL Analysis

URL analysis is the process of examining website addresses to identify suspicious patterns. It checks URL length, special characters, domain names, and suspicious keywords. These features help identify potentially malicious websites.

D. Machine Learning

Machine Learning enables computers to learn patterns from existing website data. It helps identify differences between legitimate and phishing websites. Algorithms are used to classify websites automatically.

E. Feature Extraction

Feature extraction collects important information from website URLs and domain properties. It converts website characteristics into useful data for classification. These features help improve the performance of the detection system.

F. Website Classification

Website classification is the process of categorizing websites based on their features. The system identifies whether a submitted website is legitimate or phishing. It provides classification results to help users make safer browsing decisions.

G. Domain Analysis

Domain analysis examines website registration details and domain-related information. It considers domain age, DNS records, and SSL certificate details. These properties provide additional information for identifying suspicious websites.

H. Phishing Prevention

Phishing prevention helps protect users from fraudulent websites and online scams. The system displays warnings when potentially harmful websites are detected. It improves user awareness and promotes safer internet browsing.

III. LITERATURE REVIEW

The literature survey reviews existing research papers, methods, and techniques related to phishing website identification. It helps understand the existing approaches, their advantages, limitations, and possible improvements. Various researchers have proposed different techniques using Machine Learning, Deep Learning, URL analysis, and website feature extraction to identify phishing websites.

A. Safi and Singh (2023)

The authors presented a systematic literature review on phishing website detection techniques. They studied blacklist-based, heuristic, visual similarity, Machine Learning, and Deep Learning approaches. Their study identified Machine Learning as one of the widely used techniques for phishing detection.
Source: ScienceDirect

B. Keshav Rao et al. (2022)

The researchers presented a literature survey on Machine Learning approaches for phishing website detection. They discussed different classification algorithms and their applications in identifying fraudulent websites. The study highlighted the importance of selecting suitable algorithms for accurate phishing detection.
Source: Taylor & Francis Online

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2026.151114

C. Kytidou et al. (2025)
The authors reviewed Machine Learning techniques used for phishing detection. Their research discussed existing detection methods, technical challenges, and future research directions. The study emphasized the need for improved and adaptive phishing detection systems.
Source: SAGE Journals

D. Feature Engineering Review (2025)
This research focused on feature engineering techniques for phishing website detection using Machine Learning. It examined URL-based, domain-based, and webpage content features. The study highlighted the importance of feature selection and combining different features to improve detection performance.
Source: IEEE Xplore

E. Machine Learning-Based Phishing Website Identification (2025)
The researchers discussed phishing website identification using algorithms such as Random Forest, Decision Tree, and Support Vector Machine. Their study considered URL features, domain information, and webpage characteristics for classification. It highlighted the importance of automated and real-time phishing website detection.
Source: Mat Journals

F. URL-Based Phishing Detection
Researchers explored URL-based detection techniques to identify suspicious website addresses. These methods analyze URL length, special characters, domain names, and suspicious keywords. The approach helps identify potentially malicious websites without depending entirely on webpage content.

G. Deep Learning-Based Phishing Detection
Several studies have investigated Deep Learning techniques for identifying phishing websites. These approaches analyze complex patterns in website URLs and webpage content. They can support automated detection and improve the identification of sophisticated phishing websites.
H. Random Forest-Based Classification

3.2 Research Gap
From the above literature survey, it is observed that existing phishing detection techniques have certain limitations, such as dependence on outdated datasets, difficulty identifying newly created phishing websites, and false-positive predictions. Some traditional systems depend mainly on blacklists, while other methods use limited website features. These approaches may not always identify advanced phishing websites.
Therefore, the proposed system focuses on analyzing multiple website features and applying Machine Learning algorithms to improve phishing website identification. It aims to provide automated classification, faster detection, and better protection against online threats.

IV. COMPARATIVE ANALYSIS AND RESEARCH GAP

Different sign-recognition approaches have different strengths and limitations. Table I compares their general characteristics with the intended functionality of Sign-Speak.

TABLE I. COMPARATIVE ANALYSIS OF REVIEWED SYSTEMS

Approach	Main characteristics	Limitations
Blacklist-Based Detection	Identifies phishing websites using databases of known malicious URLs.	Cannot easily detect newly created phishing websites.
Whitelist-Based Detection	Allows access to trusted and verified websites.	Unknown legitimate websites may be restricted.
Heuristic-Based Detection	Uses predefined rules and suspicious URL patterns.	Attackers can bypass predefined rules.
Decision tree	Uses decision-making rules to classify websites.	May overfit the training dataset.
Random Forest	Uses multiple decision trees to improve classification performance.	Requires more computational resources than a single decision tree.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

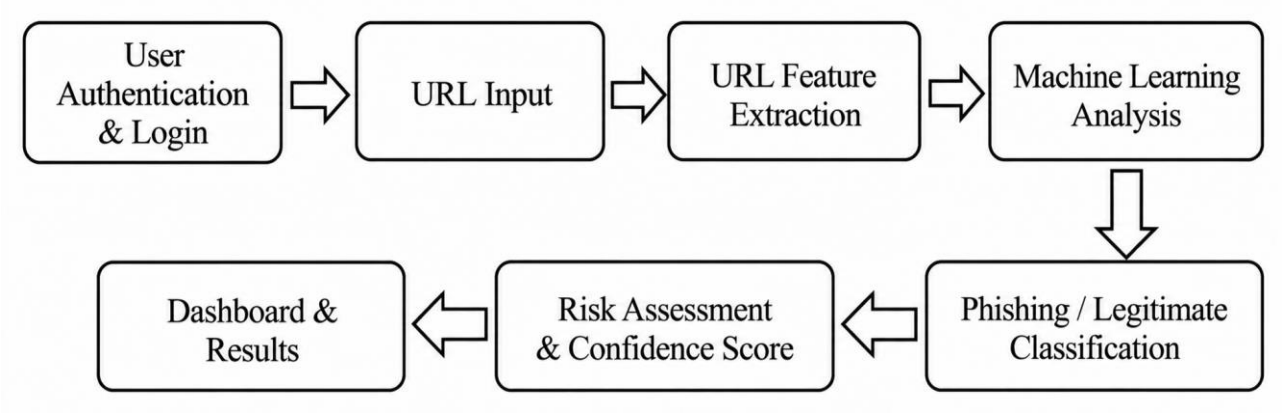
A. Identified Research Gaps

A. Identified Research Gap: Existing phishing website detection systems mainly depend on blacklist-based methods and traditional detection techniques. These approaches may fail to identify newly created phishing websites that are not yet registered in databases. Many existing systems also use limited URL features and do not consider important factors such as domain age, DNS records, SSL certificates and webpage content. Additionally, some approaches lack real-time detection and instant warning notifications.

B. Proposed Research Solution: The proposed system aims to overcome these limitations by developing an advanced phishing website detection system using Random Forest and XGBoost algorithms. It integrates URL-based, domain-based and webpage-based features to improve detection performance. The system also provides real-time URL checking, phishing classification and instant warning notifications to help users identify and avoid malicious websites.

V. PROPOSED SYSTEM ARCHITECTURE

A. User Input: The user enters a website URL into the system through the user interface. The system accepts the URL and checks whether it is valid and properly formatted before further processing.



B. Feature Extraction: The system extracts important features such as URL length, special characters, domain age, HTTPS status, DNS records, SSL certificate details and webpage content. These features help identify suspicious characteristics of phishing websites.

C. Machine Learning: The extracted features are processed using machine learning algorithms such as Random Forest and XGBoost. The models analyse the website characteristics, identify suspicious patterns and classify the website as phishing or legitimate.

D. Result and Alert: Based on the prediction, the system displays the final classification result to the user. If a phishing website is detected, the system generates an instant warning notification and provides information about the potential security

VI. METHODOLOGY

A. Data Collection and Preprocessing: The phishing and legitimate website URLs are collected from reliable datasets such as PhishTank and other trusted sources. The collected data is cleaned, labelled and preprocessed by removing duplicate records and handling missing values to prepare it for model training.

B. Feature Extraction and Selection: Important features are extracted from website URLs and webpage information, including URL length, special characters, domain age, HTTPS status, DNS records, SSL certificates and webpage content. Feature selection is performed to identify the most relevant characteristics for improving detection performance.

C. Model Training and Classification: The extracted features are used to train machine learning algorithms such as Random Forest and XGBoost. The dataset is divided into training and testing sets to evaluate model performance. The trained models analyse website patterns and classify URLs as phishing or legitimate.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151114

D. Prediction and Result Analysis: The trained model predicts whether a given website is phishing or legitimate. The system displays the prediction result along with the phishing risk level. If a suspicious website is detected, an instant warning notification is generated to alert the user and improve online security.

E. Model Evaluation: The trained machine learning models are evaluated using performance metrics such as accuracy, precision, recall and F1-score. These metrics help measure the effectiveness and reliability of the phishing detection system.

F. Real-Time URL Detection: The system allows users to enter website URLs for real-time verification. It analyses the URL and its extracted features to identify potentially malicious websites.

VII. IMPLEMENTATION DETAILS

A. Frontend Development: A user-friendly web interface is developed using HTML, CSS and JavaScript. It allows users to enter website URLs and view phishing detection results easily.

B. Backend Development: Python with the Flask framework is used to handle user requests, process URLs and connect the frontend with the machine learning models.

C. Dataset Preparation: Phishing and legitimate website URLs are collected from sources such as PhishTank and other trusted datasets. The collected data is cleaned, labelled and prepared for model training.

D. Feature Extraction: Important website features such as URL length, domain age, HTTPS status, special characters, DNS records and SSL certificate information are extracted for analysis.

E. Model Implementation: Random Forest and XGBoost algorithms are implemented using Python libraries such as Scikit-learn and XGBoost to detect phishing websites.

F. Model Testing and Evaluation: The trained models are tested using unseen data. Accuracy, precision, recall and F1-score are used to evaluate their performance.

G. Result and Alert Generation: The system displays whether the submitted URL is phishing or legitimate. If a phishing website is detected, a warning message is generated to alert the user.

VIII. RESULTS AND EVALUATION

Phishing Website Detection: The proposed system analyses the submitted website URL and classifies it as either phishing or legitimate using machine learning algorithms.

Model Performance: Random Forest and XGBoost models are evaluated to measure their ability to identify phishing websites accurately and reliably.

Performance Metrics: The system is evaluated using accuracy, precision, recall and F1-score to measure the effectiveness of the detection process

.

Real-Time Detection: The system provides real-time URL checking and displays the prediction results to users, helping them identify suspicious websites.

Alert Notification: When a phishing website is detected, the system generates a warning notification to alert users about potential security threats.

Overall Evaluation: The proposed system aims to improve phishing detection, reduce false predictions and provide better protection against malicious websites.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

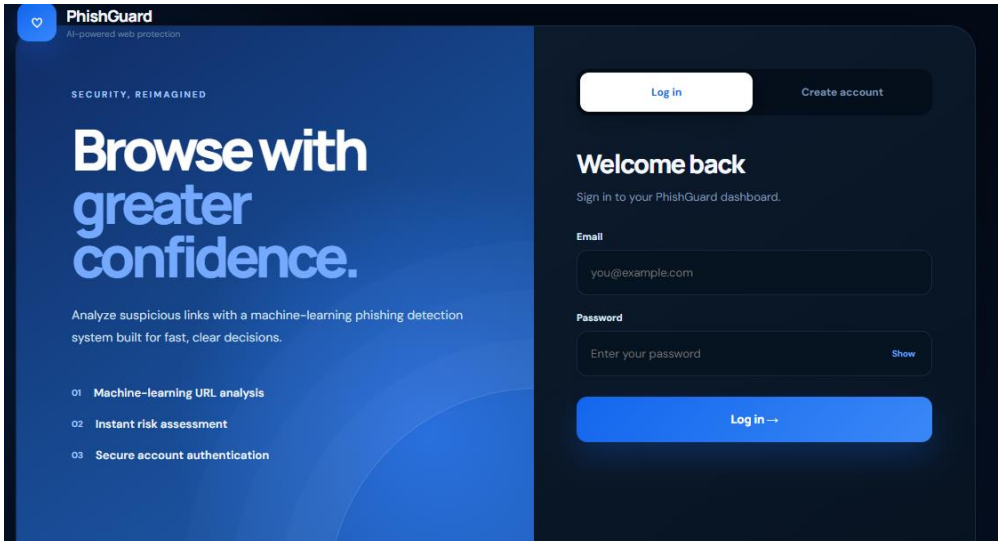


Fig. 2. Live camera Input

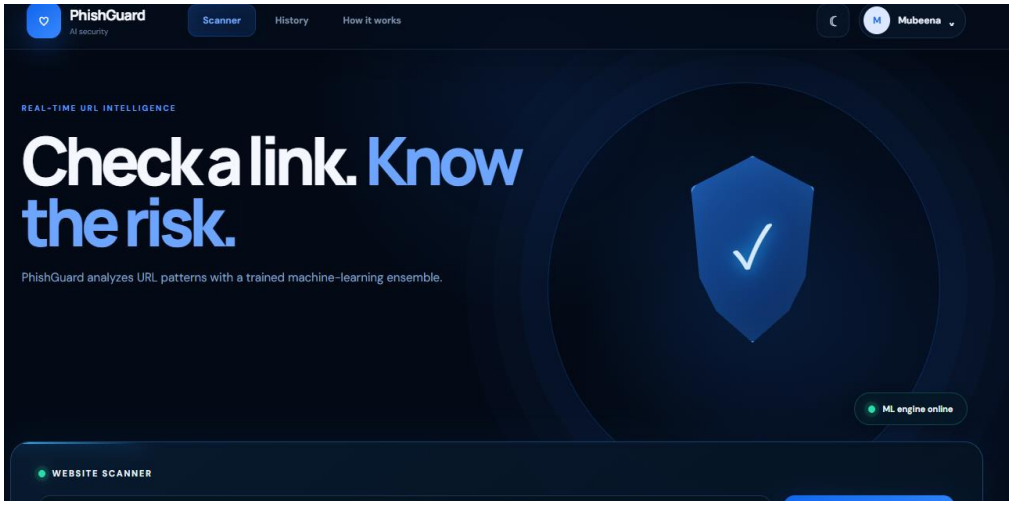


Fig. 3. Recorded Video Input

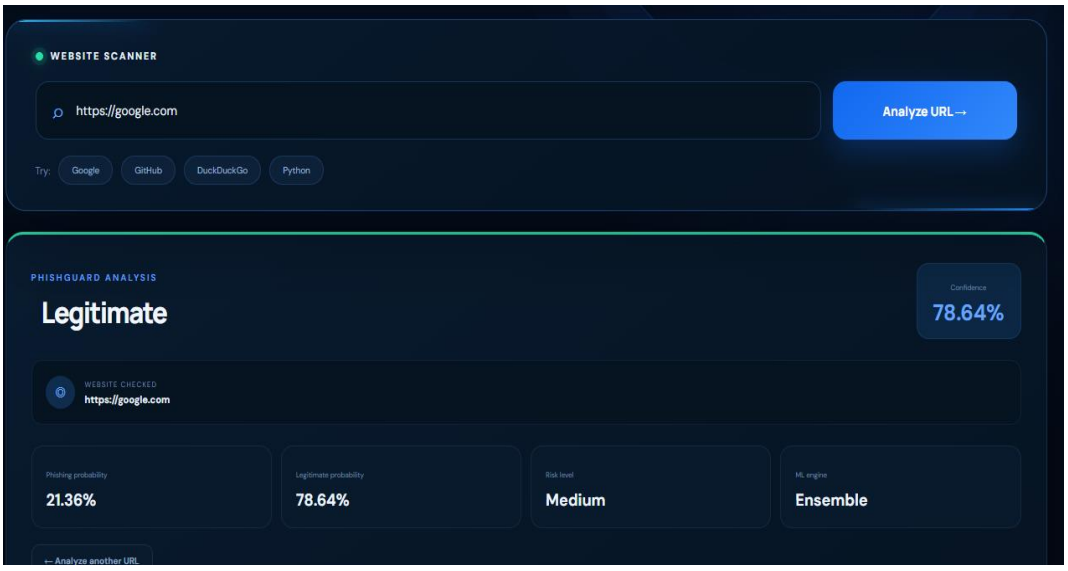


Fig. 4. Prediction of sign

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

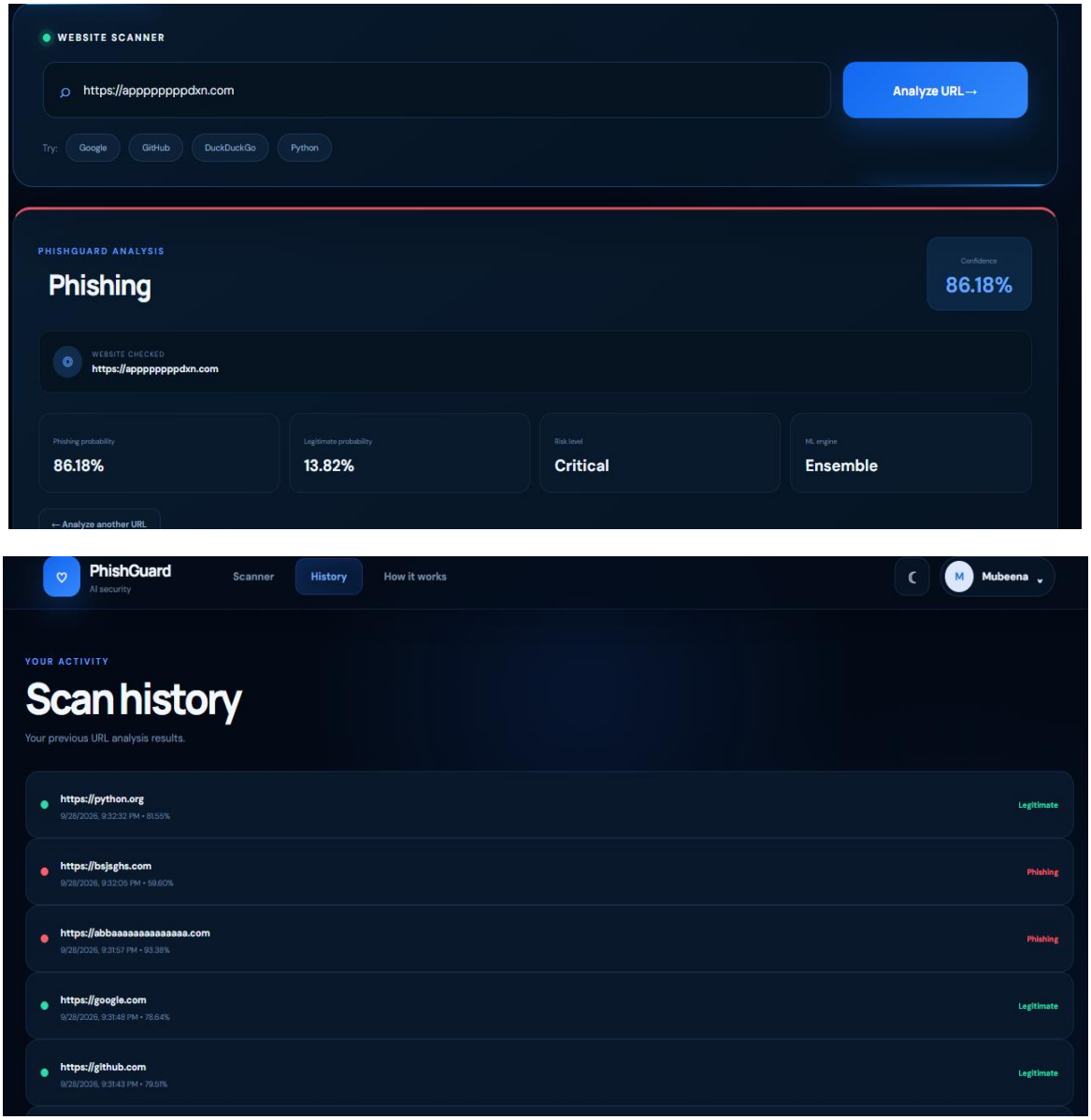


Fig. 5. Recorded video output

IX. DISCUSSION

Detection Effectiveness: The proposed system uses machine learning algorithms to identify phishing and legitimate websites by analysing multiple URL and website features. It considers important characteristics such as URL length, special symbols, domain age and HTTPS status. This helps the system detect suspicious patterns and improve phishing website identification.

Model Comparison: Random Forest and XGBoost algorithms are implemented to compare their detection performance. Both models analyse the extracted features and classify websites based on learned patterns. Their performance is compared using accuracy, precision, recall and F1-score to identify the more effective model.

Improved Security: The system analyses URL characteristics, domain information, DNS records and SSL certificates to identify potentially malicious websites. It helps users recognize suspicious links before accessing them. This approach can reduce the chances of users becoming victims of phishing attacks.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151114

Real-Time Analysis: The proposed system supports real-time URL checking and provides quick detection results. Users can enter a website URL and receive its classification without manually checking the website. Warning notifications are generated when suspicious websites are identified.

System Limitations: The performance of the proposed system depends on the quality and availability of the dataset. Newly created phishing websites and advanced attack techniques may sometimes be difficult to identify. Changes in website structure and missing domain information can also affect prediction accuracy.

Future Improvements: The system can be further enhanced by integrating browser extensions, deep learning techniques and automatic dataset updates. Advanced webpage content analysis and threat intelligence integration can improve detection capabilities. These improvements can make the system more reliable, scalable and effective against emerging phishing attacks.

ADVANTAGES OF THE PROPOSED SYSTEM

High Detection Accuracy: The system uses Random Forest and XGBoost algorithms to improve phishing website detection and distinguish between legitimate and malicious URLs.
Real-Time Detection: It allows users to check website URLs in real time and receive quick detection results without manually analysing websites.
Multiple Feature Analysis: The system analyses various features, including URL length, domain age, HTTPS status, DNS records, SSL certificates and webpage content, for better identification.
Instant Warning Notifications: It generates warning messages whenever a phishing website is detected, helping users avoid suspicious links.
Reduced False Predictions: By combining multiple website features and machine learning techniques, the system aims to reduce false-positive and false-negative predictions.
User-Friendly Interface: The system provides a simple and easy-to-use interface that allows users to enter URLs and understand the detection results.
Improved Online Security: It helps protect users from fraudulent websites, credential theft and other phishing-related cyberattacks.
Scalable and Flexible: The system can be enhanced with new datasets, additional detection features and advanced machine learning techniques to identify emerging phishing attacks.

CONCLUSION

The proposed Phishing Website Detection Using Machine Learning system aims to identify and prevent phishing attacks by analysing website URLs and their important features. It uses machine learning algorithms such as Random Forest and XGBoost to classify websites as phishing or legitimate. By considering features such as URL characteristics, domain age, HTTPS status, DNS records, SSL certificates and webpage content, the system aims to improve detection performance.

The proposed system provides real-time URL checking, quick prediction results and warning notifications for suspicious websites. It helps users identify potentially harmful websites and improves online security. Overall, this system offers an effective approach to phishing website detection and can be further enhanced using advanced machine learning techniques, browser extensions and automatic dataset updates.

FUTURE SCOPE

Advanced Machine Learning Techniques: Future versions can integrate deep learning algorithms and advanced machine learning models to improve phishing website detection accuracy.

Browser Extension Integration: The system can be developed as a browser extension to automatically detect phishing websites while users browse the internet.

Real-Time Database Updates: Phishing URL databases can be updated automatically to identify newly created phishing websites and emerging cyber threats.

Advanced Website Analysis: Future improvements can include webpage content analysis, HTML structure inspection, image analysis and suspicious login form detection.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151114

Mobile Application Development: The system can be extended into an Android or iOS application to help users check suspicious URLs directly from their mobile devices.

Explainable AI Integration: Explainable AI techniques can be introduced to provide reasons for classifying a website as phishing or legitimate.

Email and SMS Phishing Detection: The system can be expanded to identify phishing links received through emails, SMS messages and social media platforms.

Cloud-Based Deployment: The proposed system can be deployed on cloud platforms to support scalability, faster processing and access from multiple devices.

ACKNOWLEDGMENT

The authors thank nayanashree D and Lavanya N L for their guidance and continuous support throughout this project, and the Department of Computer Science and Engineering, East West College of Engineering, for providing the resources necessary to carry out this work.

DECLARATIONS

- **Funding Source:** No funding was received for this work.
- **Acknowledgement:** The authors thank East West College of Engineering and the Department of Computer Science and Engineering for their support.
- **Data Availability:** Data is available from public sources or upon reasonable request.
- **Authors' Contribution:** All authors contributed to the research, implementation, analysis, and manuscript preparation.
- **Use of AI and AI-Assisted Technologies:** AI tools were used for language improvement and grammar checking. The authors verified the content.
- **Conflict of Interest:** The authors declare no conflict of interest.
- **Copyright Permissions:** Copyright permissions were obtained for all figures and tables used.

REFERENCES

[1]. S. A. S. et al., "URL Phishing Detection using Machine Learning Techniques Based on URLs Lexical Analysis," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021. DOI: 10.1109/ICICS52457.2021.9464539.
[2]. "A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information," 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), 2022. DOI: 10.1109/PRAI55851.2022.9904055. IEEE Xplore
[3]. "Integrated Machine Learning Approach to Phishing Detection: Comparing SVM, Random Forest, and XGBoost Models," 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), 2024. DOI: 10.1109/ICTACS62700.2024.10840493.
[4]. "XGBoost-Based URL Phishing Detection Method With Cross-Dataset Validation," IEEE Access, vol. 14, pp. 43200–43212, 2026. DOI: 10.1109/ACCESS.2026.3672690.
[5]. PhishTank, "Developer Information and Phishing URL Data Feed." Available: PhishTank Developer Information.
[6]. Scikit-learn, "Machine Learning in Python." Available: Scikit-learn Official Documentation.
[7]. XGBoost, "Extreme Gradient Boosting Documentation." Available: XGBoost Official Documentation.

 This work is licensed under a Creative Commons Attribution 4.0 International License