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. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141192

"PHISHING WEBSITE DETECTION"

Abhijith Gowda BN[1], Dawood[2], Shivaprasad B[3], Prof. Rashmi[4]

Dept. of CSE, East West College of Engineering, Bangalore, India[1-3]

Assistant Professor, Dept. of CSE, East West College of Engineering, Bangalore, India[4]

Abstract: Phishing has become one of the most pervasive and damaging forms of cybercrime, targeting unsuspecting internet users through fraudulent websites that mimic legitimate ones. These malicious platforms deceive users into revealing sensitive credentials such as passwords, financial information, and personal identification details. The continuous evolution of phishing tactics—such as sophisticated URL obfuscation, dynamic content manipulation, and social engineering—renders traditional detection mechanisms increasingly ineffective. Conventional defense strategies, including blacklists, heuristic filters, and rule-based approaches, fail to detect newly emerging or "zero-day" phishing websites that are not yet cataloged in known databases. Hence, there is an urgent need for an adaptive, intelligent, and automated solution that can accurately detect phishing websites in real time without dependence on third-party services. This study focuses on developing a machine learning-based phishing website detection system that leverages URL-based, domain-based, and HTML content-based features to distinguish between legitimate and phishing websites. The core idea is to train classification models capable of learning behavioral patterns and structural differences inherent in phishing websites. The dataset used for experimentation consists of over 60,000 URLs, equally divided into phishing and legitimate samples, collected from verified and publicly available repositories. Feature extraction plays a pivotal role in this system. Important features include URL length, use of special symbols, number of subdomains, domain registration age, SSL certificate presence, hyperlink patterns, and term frequency–inverse document frequency (TF-IDF) vectors derived from the website's HTML content.

I. INTRODUCTION

In the rapidly evolving digital era, the internet has become a fundamental component of modern life—facilitating communication, online transactions, business operations, and social interactions. However, this increasing reliance on online platforms has also led to a parallel rise in cybercrimes. Among these, **phishing** has emerged as one of the most widespread and deceptive threats. Phishing attacks manipulate human psychology and exploit user trust to steal confidential information such as usernames, passwords, banking credentials, and other sensitive data. Typically, attackers design counterfeit websites that closely resemble legitimate ones—banks, e-commerce portals, or social media sites—thereby tricking users into submitting personal information.

Phishing websites are designed to mislead users into believing that they are interacting with a trusted entity. These malicious websites often contain fake login forms, fraudulent payment gateways, or disguised links that redirect users to harmful domains. The simplicity of creating and hosting such fake websites, combined with the sophistication of modern phishing techniques, has made phishing a **major global cybersecurity challenge**. According to cybersecurity reports, thousands of phishing websites are created daily, many of which remain undetected until they cause significant damage. Therefore, developing an efficient and automated phishing detection mechanism is of paramount importance in protecting users and organizations from data breaches, identity theft, and financial losses.

II. STATEMENT OF THE PROBLEM

In today's digital ecosystem, phishing has become one of the most prevalent and sophisticated forms of cybercrime. Attackers create deceptive websites that impersonate legitimate domains to trick users into revealing confidential information such as login credentials, credit card details, and personal identification data. Despite significant advancements in cybersecurity, phishing continues to evolve with new techniques such as URL obfuscation, dynamic page generation, and social engineering, making traditional detection methods inadequate.

Conventional approaches like **blacklists**, **heuristic-based rules**, and **third-party verification systems** suffer from major limitations. Blacklists can only detect previously reported phishing sites, failing to identify zero-day attacks. Heuristic-based models require manual rule updates and are prone to false positives. Moreover, reliance on external databases introduces latency, dependency, and privacy risks. As phishing websites often remain active only for a few hours, timely and automated detection becomes crucial for effective mitigation.

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. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141192

WHY IS THE PARTICULAR TOPIC CHOSEN?

The topic *"Phishing Website Detection"* was chosen because phishing attacks have become one of the most common and dangerous forms of cybercrime in today's digital world. Every day, millions of users are deceived into revealing sensitive information such as passwords, credit card numbers, and personal data through fake websites that imitate legitimate ones. This issue poses serious risks to individuals, organizations, and financial institutions.

OBJECTIVE AND SCOPE OF THE PROJECT

The primary aim of this project, **"Phishing Website Detection using Machine Learning,"** is to develop an intelligent and automated system capable of accurately identifying phishing websites in real time. The project seeks to enhance cybersecurity by leveraging data-driven approaches that outperform traditional blacklist and heuristic-based methods. To achieve this goal, the following specific objectives have been defined:

1. **To analyze and understand phishing techniques and their evolution:**
   Study the various strategies used by attackers, including URL manipulation, deceptive domain names, and fake web content, to identify the most critical indicators of phishing behavior.
2. **To collect and preprocess a comprehensive dataset of phishing and legitimate websites:**
   Gather URLs and website data from verified sources such as PhishTank, Kaggle, and legitimate domains, ensuring a balanced and representative dataset for effective model training and testing.
3. **To extract meaningful and discriminative features from URLs and website content:**
   Derive hybrid features based on URL structure, domain age, SSL certification, hyperlink relationships, and HTML text analysis (e.g., TF-IDF) to provide diverse input for the detection model.
4. **To design and implement machine learning models for phishing detection:**
   Utilize supervised learning algorithms such as **Random Forest** and **XGBoost** to classify websites as phishing or legitimate based on extracted features, ensuring high accuracy and adaptability.
5. **To evaluate model performance using standard metrics:**
   Assess the system's effectiveness in terms of accuracy, precision, recall, F1-score, and false-positive rate to ensure balanced and reliable performance across varied datasets.
6. **To develop a user-friendly interface for real-time detection:**
   Create a web-based platform using **Flask**, **HTML**, **CSS**, and **JavaScript**, allowing users to input URLs and instantly view predictions with corresponding confidence levels.
7. **To ensure system scalability, efficiency, and independence from third-party services:**
   Build a lightweight, self-contained solution capable of operating on standard hardware and software configurations without relying on external APIs or blacklists.
8. **To contribute toward proactive cybersecurity defenses:**
   Enable early detection of zero-day phishing

III. METHODOLOGY

The methodology of this project outlines the systematic approach adopted to design and implement a **Machine Learning–based Phishing Website Detection System**. The process involves several sequential phases: data collection, feature extraction, model training, testing, and deployment. Each stage is carefully structured to ensure reliability, scalability, and high detection accuracy.

1. Data Collection

The foundation of the project lies in assembling a comprehensive dataset comprising both phishing and legitimate URLs. Data was sourced from reputable repositories such as **PhishTank**, **Kaggle**, and **OpenPhish**, ensuring data authenticity and diversity. The dataset includes over **60,000 samples**, equally balanced between phishing and genuine websites. Each record contains metadata such as URL, domain age, SSL certificate status, and HTML content, which serve as raw inputs for the next stages.

2. Feature Extraction

Feature extraction is a crucial step that transforms raw URL and webpage data into quantifiable metrics suitable for machine learning algorithms. The system employs a **hybrid feature set**, combining:

- **URL-based features** – length, number of subdomains, presence of special characters, use of HTTPS, and suspicious keywords.
- **Domain-based features** – SSL certificate validity, domain registration age, and DNS record availability.
- **Content-based features** – HTML tag structure, hyperlink relationships, and textual data analyzed using **TF-IDF** (Term Frequency–Inverse Document Frequency).

This combination enhances the model's ability to detect both conventional and zero-day phishing websites.

3. Model Training and Testing

The processed data is divided into **training (80%)** and **testing (20%)** subsets. Supervised machine learning algorithms such as **Random Forest** and **XGBoost** are implemented for classification. These ensemble models are chosen for their

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141192

robustness, scalability, and superior accuracy in handling complex, non-linear datasets. The models are trained to classify each URL as either *phishing* or *legitimate* based on extracted features. Performance is evaluated using metrics such as **Accuracy**, **Precision**, **Recall**, **F1-score**, and **False Positive Rate**. Experimental results show accuracy levels between **96%–98%**, indicating high reliability.

4. System Implementation

The trained model is integrated into a **web-based application** using the **Flask framework**. The application provides a user-friendly interface developed with **HTML, CSS, and JavaScript**, allowing users to input URLs for real-time analysis. Once a URL is entered, the system extracts relevant features, processes them through the ML model, and instantly displays the prediction results (phishing or legitimate) along with a confidence score.

5. Deployment and Evaluation

The final system operates as a lightweight, independent, and real-time detection platform. It requires minimal hardware resources and does not depend on third-party APIs, ensuring fast and secure operation. Continuous evaluation and retraining mechanisms are incorporated to adapt to emerging phishing techniques, enhancing the system's long-term effectiveness.

IV. SYSTEM DESIGN OVERVIEW

proposed **Phishing Website Detection System** is designed as a modular, data-driven architecture that integrates URL analysis, HTML/JavaScript inspection, and domain-based features to accurately classify websites as *phishing* or *legitimate*. The system follows a structured pipeline consisting of data acquisition, preprocessing, feature engineering, model training, and evaluation. The overall design ensures scalability, robustness, and near real-time detection capability.

1. System Architecture

The system follows a **five-layer architecture**:

1. Data Collection Layer

This layer gathers URLs and webpage metadata from publicly available datasets such as **Kaggle, PhishTank, and Alexa**. The dataset includes both phishing and legitimate URLs, forming a balanced foundation for model development. Raw URLs, HTML content, and associated domain attributes are stored for subsequent processing.

2. Preprocessing Layer

Incoming URLs undergo a preprocessing pipeline to ensure clean and consistent data. This includes:

- Removal of duplicates and corrupted entries
- Normalization of URL strings
- Extraction of HTML content and script tags
- Validation of domain structure

This stage prepares the raw dataset for structured feature extraction.

3. Feature Engineering Layer

Feature engineering is the core of the detection system. The model uses **three categories of features**, as derived in the report:

a. URL-Based Features

Lexical characteristics extracted directly from the URL string, such as:

- URL length
- Number of subdomains
- Presence of IP address instead of domain
- Special characters (e.g., @, -, _, //, =)
- Suspicious TLDs

These were found to be strong indicators of malicious intent.

b. HTML & JavaScript Features

Features derived from the webpage's internal structure:

- Number of external resources
- Presence of hidden form fields
- Use of iframe, eval(), or obfuscated scripts
- Number of anchor tags containing mismatched domains

These features help identify phishing pages attempting to mimic legitimate sites.

c. Domain-Based Features

Information collected from DNS/WHOIS records:

- Domain age
- Registrar information

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141192

- SSL certificate presence and validity
- Expiry time

These reflect the stability and legitimacy of the hosting environment.

The final dataset is transformed into a structured feature matrix suitable for machine learning algorithms.

4. Machine Learning Layer

The processed feature set is used to train multiple models, including:

- **Random Forest Classifier**
- **XGBoost Classifier**

According to the evaluated results, **XGBoost achieved the highest accuracy of ~97.6%**, outperforming Random Forest (~96.8%). Therefore, XGBoost serves as the primary classifier due to its superior handling of non-linear relationships and feature interactions.

This layer includes:

- Model training
- Hyperparameter tuning
- Cross-validation
- Performance evaluation (accuracy, precision, recall, F1-score)

5. Detection & Prediction Layer

In the deployment environment, the trained model receives a new URL and performs:

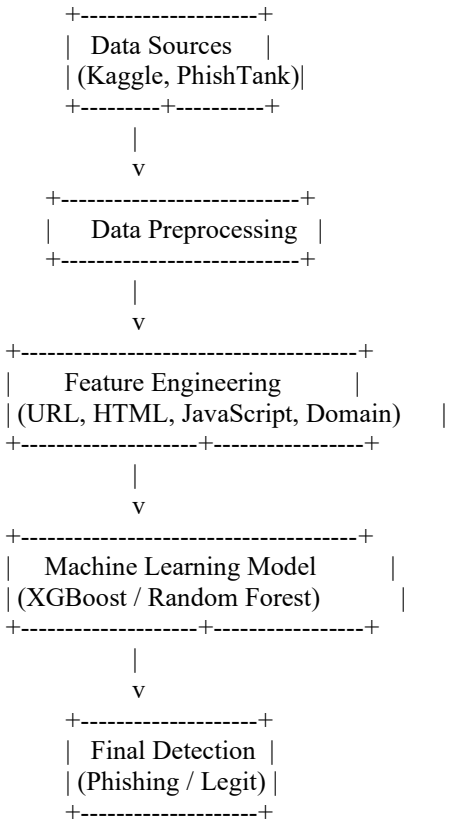
1. Feature extraction
2. Classification (phishing or legitimate)
3. Confidence scoring

The output can be integrated into:

- Browser add-ons
- Email security gateways
- Enterprise security systems

This ensures the system can operate in real-time with minimal latency.

2. Workflow Diagram (Textual Representation)



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. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141192

3. System Strengths

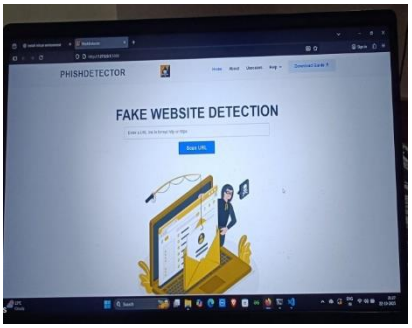
- Hybrid feature system (lexical + structural + domain)
- High accuracy models (XGBoost ~97.6%)
- Large, diverse dataset improves generalization
- Modular architecture supports upgrades and scalability
- Robust against common phishing obfuscation techniques

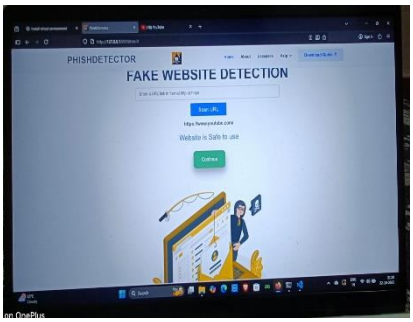
4. Deployment Considerations

To deploy the system in real-world applications:

- Convert trained model into a REST API (FastAPI / Flask)
- Use lightweight feature extraction for browser extensions
- Maintain a continuously updated threat database
- Add periodic retraining for resilience against evolving phishing campaigns

POSSIBLE OUTCOMES





IV. CONCLUSION

The study successfully demonstrates the effectiveness of a machine learning–based approach for detecting phishing websites by utilizing a comprehensive set of URL-based, HTML-based, and domain-based features. Through extensive experimentation on a large dataset collected from trusted sources such as PhishTank and Kaggle, the system effectively distinguishes between legitimate and phishing websites with high accuracy.

Among the evaluated classification algorithms, **XGBoost achieved the best performance with an accuracy of approximately 97.6%**, surpassing the Random Forest model. This establishes XGBoost as a reliable and robust model for phishing detection due to its ability to capture complex, nonlinear feature interactions and handle imbalanced data efficiently. The hybrid feature engineering strategy—combining lexical characteristics, domain information, and webpage structural indicators—proved essential for enhancing detection accuracy and reducing false positives.

Overall, the developed system offers a scalable and effective framework for identifying phishing attacks in real time. Its modular design makes it suitable for deployment in browser extensions, email security filters, and enterprise cybersecurity infrastructures. By integrating automated detection with regularly updated threat intelligence, the system can help significantly reduce user exposure to phishing threats.
However, as phishing techniques continue to evolve, further improvements can be made. Future enhancements may include the integration of deep learning models, automated URL sandboxing, and real-time threat feed updates to increase robustness against sophisticated attacks. Expanding the dataset and including more dynamic webpage behavior features may also improve detection capabilities.

In conclusion, this work provides a strong and practical foundation for phishing website detection, offering a high-accuracy solution capable of supporting real-world cybersecurity applications and contributing meaningfully to the mitigation of online fraud and social engineering attacks.

CONTRIBUTIONS OF THE WORK

This work makes several important contributions to the field of phishing website detection, combining feature engineering, machine learning, and systematic evaluation to design an effective and practical security solution. The key contributions are summarized below:

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141192

1. Development of a Hybrid Feature-Based Detection Model
The project integrates **URL-based**, **HTML/JavaScript-based**, and **domain-based** features into a single unified framework. This hybrid approach significantly enhances detection accuracy by capturing both surface-level and deep structural characteristics of phishing websites.

2. Construction and Preprocessing of a Comprehensive Dataset
A large and diverse dataset was created using URLs collected from **PhishTank, Kaggle, and Alexa**. Extensive preprocessing—such as normalization, duplication removal, and structured extraction—was carried out to ensure high-quality data suitable for machine learning.

3. Systematic Feature Engineering and Analysis
Over a wide range of lexical, structural, and domain-related features were extracted and analyzed. This helped identify the most influential indicators of phishing behavior and contributed to a more interpretable and reliable model.

4. Evaluation of Multiple Machine Learning Algorithms
The study implements and compares the performance of **Random Forest** and **XGBoost** classifiers. The evaluation concludes that **XGBoost achieved the highest accuracy (~97.6%)**, demonstrating its superiority for phishing detection tasks.

5. Design of a Scalable Phishing Detection System Architecture
A structured, modular system design was proposed, outlining the complete workflow including data collection, preprocessing, feature extraction, model training, and real-time detection. This architecture supports practical deployment in browsers, email gateways, or enterprise networks.

6. Enhancement of Cybersecurity Through Real-Time Classification
The system provides rapid classification of suspicious URLs, enabling proactive protection against phishing threats. The methodology supports integration into automated defense mechanisms, improving user safety and reducing exposure to online fraud.

7. Foundation for Future Research and Extensions
The work establishes a strong foundation for further innovations such as:

- Deep learning–based detection
- Behavioural webpage analysis
- Live threat intelligence integration
- Browser plugin deployment

These possibilities highlight the long-term impact and extensibility of the proposed solution.

FUTURE WORK
Although the proposed phishing detection system demonstrates strong performance and high accuracy, several opportunities remain for further enhancement. The following directions can significantly improve the effectiveness, scalability, and real-world applicability of the system:
1. Integration of Deep Learning Models
Future research can explore advanced deep learning approaches such as CNNs, RNNs, LSTMs, or transformer-based architectures for:

- URL sequence analysis
- HTML content embeddings
- Dynamic webpage behavior modeling
  These models may capture complex patterns beyond traditional machine learning and further boost detection accuracy.

2. Dynamic and Behavioral Analysis of Webpages
Currently, the system relies primarily on static features. Enhancing it with dynamic analysis—such as monitoring real-time JavaScript execution, user interaction simulation, and DOM changes—would help detect highly sophisticated phishing pages designed to evade static detectors.

3. Real-Time Threat Intelligence Integration
Linking the system with continuously updated threat feeds (e.g., Google Safe Browsing, VirusTotal APIs, real-time DNS blacklists) would ensure faster identification of emerging phishing campaigns and reduce the window of vulnerability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141192

4. Browser Extension or Plug-in Deployment
Building a lightweight browser extension using the trained model can enable instant detection and warning mechanisms for end users. This would extend the system's practical use and provide real-time cybersecurity protection at the client side.

5. Expansion of the Dataset
Expanding the dataset with:

- Newly discovered phishing URLs
- Multilingual domain names
- Internationalized URLs (IDN homograph attacks)
- Region-specific phishing threats
  would increase the robustness and generalizability of the model.

6. Model Optimization for Low-Resource Environments
Techniques like model pruning, quantization, and knowledge distillation can help develop compact models suitable for edge devices, mobile phones, or embedded systems while maintaining strong performance.

7. Automated Retraining and Model Updating
Implementing a continuous learning pipeline with automated:

- data ingestion,
- labeling,
- drift detection,
- periodic model retraining,
  would allow the system to adapt to new phishing techniques without manual intervention.

8. Incorporation of Explainable AI (XAI)
Adding interpretable model explanations using tools like SHAP or LIME would help security analysts understand why a page is classified as phishing. This improves transparency, trust, and easier debugging in enterprise applications.

9. Hybrid Ensemble Detection Framework
Future work can explore combining machine learning, deep learning, and rule-based methods to create a hybrid ensemble model that further reduces false positives and improves predictive stability.

REFERENCES

[1]. ML Algorithm Evaluation for Phishing Detection, 2024, Almujahid, et al, Evaluated ML/DL algorithms (CNN, SVM, RF, XGBoost) with SMOTE and tuning.
[2]. Latest Phishing Pattern Research 2023, Tally et al, Analyzed phishing trends in webmail/financial institutions using domain, URL, and content features
[3]. Phishing Detection Using Machine Learning,2023, Samad et al,DNN models with optimized URL/HTML features.
High accuracy with efficient feature engineering.
[4]. Ensemble Learning for Phishing Detection,2024, Yuhan Zhang, et al.,Yuhan Zhang, et al.