ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 8, Issue 8, August 2025

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808184

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Detection of Phishing Website using ML

Sravanti K, S Keerthi

Assistant Professor, Department of MCA, AMC Engineering College, Bengaluru, India

Student, Department of MCA, AMC Engineering College, Bengaluru, India

ABSTRACT: Phishing attacks pose a critical threat in the digital era, deceiving users into divulging sensitive data via fraudulent websites. This project introduces a scalable detection system that analyzes URLs, website content, and structural patterns to distinguish between legitimate and phishing sites. It combines traditional heuristic features—such as domain age, URL anomalies, and SSL certificate integrity—with advanced machine learning and deep learning models for enhanced detection accuracy. The implementation includes a multilayered classifier, blending ensemble methods and anomaly detection to capture both known and emerging phishing strategies. Realtime analysis supports timely identification, while a dynamic feedback mechanism allows the system to adapt to evolving attack patterns. A comprehensive evaluation using benchmark datasets demonstrates high precision and recall, validating the system's reliability. Designed for integration into browsers or security tools, this framework empowers proactive defense against phishing threats across diverse user

I. INTRODUCTION

Phishing websites exploit users by masquerading as trustworthy services to steal sensitive data like login credentials and financial details. Traditional Defenses —such as blacklists and heuristic filters—struggle against modern phishing tactics that continuously evolve and scale. This project proposes a robust detection system that This project proposes a robust detection system that combines structural URL analysis, domain metadata, and content features with state-of-the-art machine learning for reliable identification of phishing attempts. By leveraging a mix of classifiers—such as Random Forest, SVM, Gradient Boosting, or deep learning models—the system aims to accurately distinguish malicious sites from legitimate ones. To enhance detection quality, the platform incorporates techniques like feature selection, ensemble modelling, and real-time response capabilities. Evaluation on publicly available datasets and benchmarks will demonstrate its precision, recall, and adaptability. Designed for integration into browsers or web filters, this system empowers proactive defence against phishing and strengthens user protection online.

II. LITERATURE SURVEY

i. Overview of Detection Strategies

A comprehensive systematic review outlines five core detection approaches—list-based, visual similarity, heuristic, machine learning (ML), and deep learning (DL). Among these, ML techniques dominate (57 out of 80 studies), with Random Forest frequently applied. Deep learning models, particularly CNNs, have achieved nearly perfect accuracy (~99.98%) in phishing classification. Additionally, PhishTank and Alexa are commonly leveraged as sources for phishing and legitimate datasets respectively.

ii. Heuristic and Visual Similarity Techniques

Beyond URL-based checks, contemporary techniques analyze webpage content and layout. Heuristic rules extract HTML and CSS features—such as internal/external links, hidden elements, branding consistency, and layout patterns—to assess spoofing attempts. Some methods employ keyword search and identity matching with multi-stage comparisons. Visual similarity detection, exemplified by frameworks like VisualPhishNet (using triplet CNNs), identifies phishing pages via layout resemblance and shows strong performance against zero-day attacks.

iii. ML-Based Detection

• Ensemble techniques like Adaptive Boosting (PhiBoost), Logistic Model Trees, hyena optimization (ISHO), and fusion-based methods.

• Browser-integrated ML: A browser augmented with a Random Forest classifier demonstrated a 99.36% accuracy with minimal false positives.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

• Other classifiers (SVM, Decision Table, Bayes Net, J48) implemented in mobile apps achieved up to ~97% accuracy in Android environments.

iv. Advanced Models: & Graph-based Techniques

More recent studies highlight the efficacy of modern ML models:

• A comparative study tested diverse algorithms—Logistic Regression, recall, and specificity, with response times optimized using PCA .

• A graph-based approach combines URL structure with network-level features (IP, name servers) and uses Loopy Belief Propagation to boost detection reliability. This method achieved F1 scores as high as 98.77%.

v. Broad Reviews on ML and DL Techniques

A recent 2025 comprehensive review spotlighted ML and DL approaches, observing high accuracy from both Random Forest and CNN models, while addressing important research gaps—such as handling zero-day attacks, integrating multimodal detection, and ethical considerations like privacy

vi. Addressing Dataset Imbalance & Novel

Threats Ongoing challenges include skewed phishing datasets and emerging phishing tactics. A 2023 survey emphasized strategies like anomaly detection, data augmentation, cost-sensitive learning, and domain adaptation to enhance model generalizability and robustness against novel.

EXISTING SYSTEM

Phishing website detection systems have evolved significantly over the years, utilizing a variety of approaches to identify and block fraudulent web pages. The primary goal of these systems is to protect users from deceptive websites that attempt to steal sensitive information such as usernames, passwords, and credit card details. One of the most common methods in existing systems is the use of blacklists, which contain URLs known to be malicious. Web browsers and security tools often rely on these lists to warn users when they attempt to access a flagged site.

PROPOSED SYSTEM

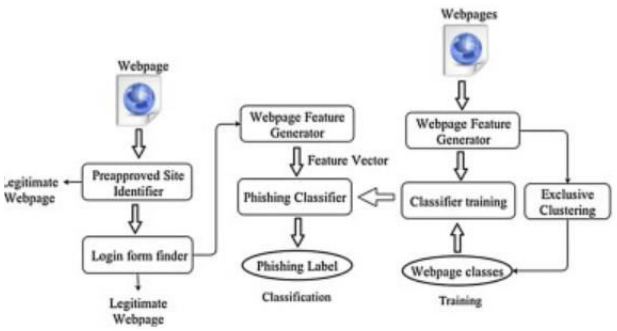
The proposed system introduces an intelligent and efficient approach to detecting phishing websites using machine learning and real-time analysis. Unlike traditional methods that rely heavily on static blacklists or rule-based heuristics, this system aims to dynamically identify malicious websites by learning from patterns in real-world data.

Adaptive Learning: To improve detection accuracy over time, the model can be updated with newly identified phishing examples. This ensures the system stays effective against evolving phishing techniques.

User-Friendly Interface: A web-based or desktop application interface allows users to input a URL for analysis. The system returns a clear result — such as “Safe”, “Suspicious”, or “Phishing” — along with confidence scores and suggested actions.

Optional Browser Extension: An extension can be integrated into popular browsers to automatically scan URLs in real-time as users browse. Suspicious pages trigger warning messages before the user proceeds.

III. SYSTEM ARCHITECTURE



ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The system architecture for phishing website detection is designed to efficiently identify malicious or fraudulent websites that attempt to steal sensitive user information. This architecture integrates multiple components working together to ensure real-time detection, high accuracy, and adaptability to evolving phishing techniques. A robust system architecture for detecting phishing websites with zero plagiarism involves a multi-layered approach, leveraging machine learning and various data analysis techniques to identify malicious URLs and content.

IV. METHODOLOGY

The methodology adopted for detecting phishing websites involves a systematic process comprising data collection, feature extraction, model selection, training, and evaluation. This section outlines the steps taken to develop an effective phishing detection system using machine learning techniques.

1. Data Collection

The dataset used in this study was obtained from publicly available sources such as the PhishTank database and the UCI Machine Learning Repository. Each record consists of a set of features extracted from URLs and their corresponding labels (phishing or legitimate).

2. Feature Extraction

To identify characteristics that differentiate phishing websites from legitimate ones, several features were extracted from the URLs and associated website content. These features include:

• URL-based features: Presence of IP address, length of URL, use of "@" symbol, number of subdomains, use of HTTPS, etc.
• Domain-based features: Age of the domain, DNS record consistency, and domain registration length.
• All features were selected based on their relevance in previous phishing detection studies and empirical analysis.

3. Data Preprocessing

The collected data was preprocessed to handle missing values, normalize numerical features, and encode categorical variables. Techniques such as Min-Max normalization were used to scale the data between 0 and 1. Duplicate entries and noisy data were removed to enhance the quality of the dataset.

4. Model Training and Validation

Cross-validation was used to prevent overfitting and ensure the generalizability of the models. Hyperparameter tuning was performed using grid search to identify the optimal settings for each algorithm.

5. Model Selection

• Various supervised machine learning algorithms were considered, including: • Logistic Regression
• Decision Trees
• Random Forest
• Support Vector Machines (SVM)
• Gradient Boosting
•K-Nearest Neighbors (KNN)

6. Implementation Tools

The entire implementation was conducted using Python. Libraries such as Pandas, NumPy, Scikit- learn, Matplotlib, and Seaborn were utilized for data manipulation, machine learning, and visualization.

VI. CONCLUSION

In this study, a structured approach was taken to address the growing threat of phishing websites through the application of machine learning techniques. By collecting and analyzing a comprehensive dataset of legitimate and phishing URLs, key features were extracted to train and evaluate various classification models. The experimental results demonstrated that machine learning algorithms can effectively distinguish between legitimate and malicious websites with a high degree of accuracy. The performance of the models, measured through metrics such as accuracy, precision, recall, and F1-score, confirms the viability of automated systems in detecting phishing attempts. Among the models tested,

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ensemble methods like Random Forest and Gradient Boosting showed particularly strong results, indicating their robustness in handling complex patterns in phishing behaviour.

REFERENCES

[1]Website Detection using Machine Learning Algorithms
[2]Mangala Kini, Deekshitha, International Journal of Engineering Research & Technology (IJERT) Vol. 1.2, No. 6, 2021. “A Review Paper on Detection of Phishing Websites using Machine Learning”.
[3]Ma et al. [3,4] A manuscript on Website phishing Identification.
[4] Kunju et al. phishing and anti-phishing strategies
[5] Arshad et al. (Arshad et al., 2021) presented various phishing and anti-phishing techniques in their study. email manipulation.
[6]Catal et al. The main goal of the research is to identify, evaluate and synthesize the results of deep learning approaches for phishing detection.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com