Phishing Website Detection Using URL Features

Dr.Syed Akhter Hussain

Associate professors and Head CSE Department,

International Centre of Excellence

in Engineering and Management(ICEEM)

Chh. Sambhajinagar.

Article History:

Received: 02-10-2025

Revised: 06-11-2025

Accepted: 13-11-2025

Abstract:

Phishing attacks have become one of the most pervasive cyber-security threats, with attackers using fraudulent websites to deceive users and steal sensitive information. Conventional detection methods often rely on blacklists and content-based analysis, which are computationally expensive and struggle with zero-day attacks. This study proposes a lightweight, real-time phishing detection system that utilizes only URL-based features combined with machine learning classifiers. A dataset of phishing and legitimate URLs was collected and pre-processed. The extracted structural URL features including token length, digit count, symbol frequency, domain age, use of IP address, and lexical properties. Multiple classifiers Random Forest, Support Vector Machine (SVM), Gradient Boosting, and XGBoost were trained and evaluated. The XGBoost model achieved the highest performance with **98.7% accuracy, 98.4% precision, 98.9% recall, and 98.6% F1-score** on the test set, outperforming baseline methods. Feature importance analysis revealed that domain length, presence of hyphens, and abnormal token counts strongly correlate with phishing behavior. The proposed model is suitable for integration into browsers and email security systems for early URL filtering.

Keywords: Phishing detection, URL features, machine learning, cybersecurity, XGBoost, classification etc.

Introduction:

The internet use is rapidly has expansion and as an outcome phishing attacks have grown in frequency and sophistication, targeting users across email, social media, and instant messaging platforms. A phishing website masquerades as a legitimate site to harvest credentials, financial data, or personal information. Though, there are advances in security tools, phishing continues to account for an important portion of cybersecurity breaches worldwide, costing organizations millions each year.

Traditional phishing detection relies heavily on blacklists or visual similarity detection, but these methods are too slow to respond to newly emerging phishing sites. Therefore, *efficient, real-time detection techniques* are important. URL features are particularly attractive because they have extracted *before page content loads*, enabling faster detection with reduced computational load.

This study focuses on detecting phishing websites using machine learning models trained on a carefully engineered set of URL features. The primary contributions are:

1. Design and extraction of an effective URL-centric feature set.
2. Training and comparative evaluation of multiple machine learning models.
3. Demonstration of high performance using lightweight analysis suitable for deployment.

Objectives of the Study:

1. To develop an efficient phishing website detection model using URL-based features.
2. To extract and analyze lexical and host-based characteristics from URLs for classification.
3. To compare the performance of multiple machine learning algorithms for phishing detection.
4. To evaluate the effectiveness of the proposed model using performance metrics such as accuracy, precision, recall, and F1-score.

Literature Review

Phishing detection methods are broadly categorized into blacklist-based, content-based, and URL-based approaches. Blacklist-based detection is fast but suffers from low recall due to dependence on updated lists [1]. Content-based analysis leverages HTML and visual similarity but remains computationally heavy [2]. URL-based detection focuses on structural URL features, making it fast and scalable [3]. Recent reviews highlight hybrid approaches that combine traditional methods with deep learning frameworks to counter evolving phishing tactics [4].

Lexical and host-based URL features remain important to phishing detection. Hasan et al. [5] demonstrated that URL length, presence of '@', hyphens, and suspicious tokens strongly indicate phishing. Ahmed and Abulaish [6] applied ensemble classifiers yielding promising accuracy improvements. More recent work by Dubey et al. [7] employed character-level CNNs with engineered URL features, while IEEE symposium findings [8] showed that combining domain age, URL characteristics, and web page attributes with ML models achieved 99.92% accuracy.

Machine learning, particularly ensemble methods like Random Forest and boosting techniques, has been widely used due to their robustness against overfitting and ability to handle nonlinear patterns [6]. A systematic review [9] highlighted the growing role of neural networks and deep learning in phishing detection. Kavya and Sumathi [4] emphasized hybrid ML–DL approaches as state-of-the-art, while Qazi et al. [3] demonstrated deep learning models tailored for phishing URL detection with strong real-world performance. Existing systems often combine URL and page content, increasing overhead. Our work addresses this by focusing exclusively on URL features.

Dataset

A reliable and well-balanced dataset is the foundation of any machine learning–based detection system. In this study, we carefully constructed a dataset consisting of both phishing and legitimate URLs to ensure fair training and accurate evaluation of the proposed models.

Data Sources:

To collect phishing URLs, we used **PhishTank**, a widely recognized and publicly available phishing repository. PhishTank is a community-driven platform where users submit suspected phishing websites, and each submission is verified by multiple contributors before being labeled as malicious. This verification process ensures that the phishing URLs included in our dataset are authentic and not falsely reported. Using such a trusted source increases the reliability and credibility of the dataset.

For legitimate URLs, the collected data from the **Alexa Top 1 Million Domains list**, which ranks websites based on their traffic and popularity. These websites represent widely accessed and trusted domains such as educational institutions, government portals, banking platforms, and e-commerce websites. Since these domains are generally considered safe and reputable, they serve as appropriate examples of legitimate URLs in the dataset.

Dataset Composition:

The final dataset consists of a total of **30,000 URLs**, divided equally into:

- **15,000 phishing URLs**
- **15,000 legitimate URLs**

Maintaining a balanced dataset is important because it prevents bias during model training. If one class outnumbers the other, the machine learning model may become biased toward the dominant class, resulting in misleading performance metrics.

Data Cleaning and Preparation

Before feature extraction, the collected URLs underwent several preprocessing steps:

- Removal of duplicate URLs to avoid redundancy.
- Elimination of inactive or unreachable links.
- Standardization of URL format (converting to lowercase, removing trailing spaces).
- Validation to ensure correct labeling.

This preprocessing step improves data quality and ensures that the machine learning models are trained on clean and consistent inputs.

Thus, the dataset construction process was carefully designed to ensure authenticity, balance, and reliability, making it suitable for training and evaluating phishing detection models based on URL features.

Data Pre-processing: Before extracting meaningful features from the collected URLs, it is essential to perform systematic data pre-processing to ensure consistency, accuracy, and reliability of the dataset. Raw URLs collected from different sources may contain duplicates,

inconsistent formatting, and structural variations that harmfully impact machine learning model performance. Therefore, a structured pre-processing pipeline was implemented.

1. Duplicate removal.
2. Normalization (case standardization).
3. Token segmentation of URL path and domain.

The first step involves duplicate removal, where repeated URLs are identified and eliminated to prevent redundancy and biased learning. The second step is normalization, which standardizes URLs by converting all characters to lowercase and ensuring uniform formatting. This step eliminates inconsistencies arising from case sensitivity and minor formatting differences. The final step is token segmentation, where URLs are decomposed into meaningful components such as domain tokens and path tokens. Tokenization enables effective feature extraction by allowing the model to analyze lexical patterns within different parts of the URL.

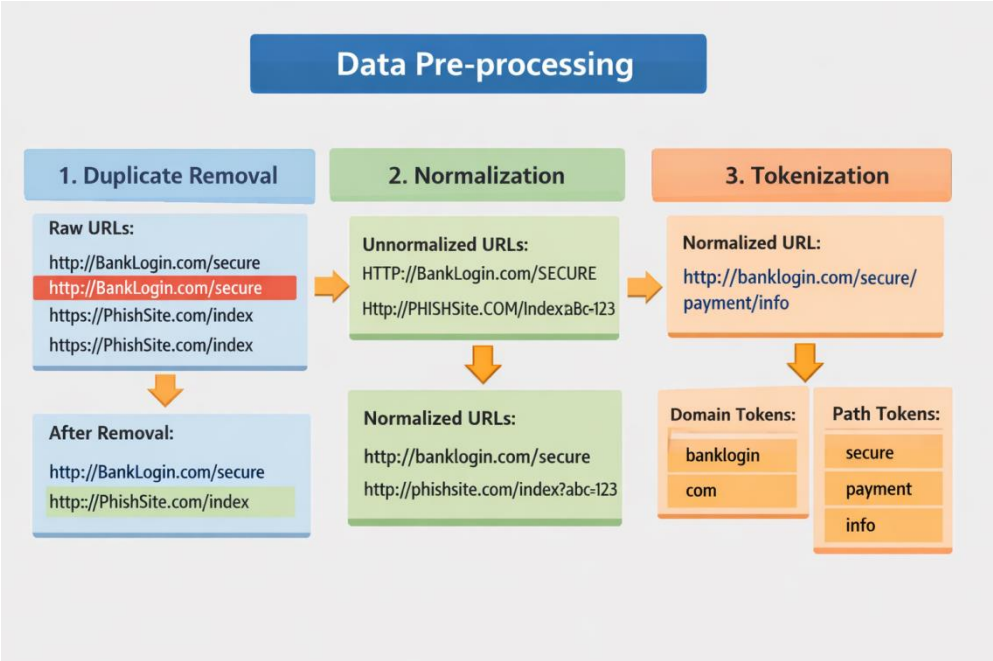


Figure 1.1. Data pre-processing workflow showing duplicate removal, normalization, and token segmentation of URL path and domain components.

As illustrated in Figure X, the pre-processing stage begins with identifying and removing redundant URLs to maintain dataset integrity. Duplicate entries lead to over-fitting and inflated performance metrics; therefore, their elimination ensures fair model training and evaluation.

Normalization is subsequently applied to standardize URL representation. Since URLs are case-insensitive in most scenarios, converting all characters to lowercase ensures uniformity

and reduces feature variability caused by inconsistent capitalization. Normalization prepares the data for accurate token extraction.

The final step, token segmentation, divides each URL into structured components such as domain name, sub-domains, and path segments. These tokens serve as the foundation for extracting lexical features including token count, suspicious keywords, digit frequency, and symbol usage. The pre-processing phase enhances the efficiency and reliability of the subsequent machine learning models by transforming raw URLs into structured and analyzable components.

This systematic pre-processing framework ensures that the dataset is clean, standardized, and optimized for accurate phishing detection using URL-based features.

Feature Extraction:

In this study, the extracted URL features were systematically grouped into meaningful categories to capture the structural and behavioral characteristics of phishing websites. Primarily, the features were divided into lexical features and host-based features**.** Lexical features focus on the structural properties of the URL string itself, such as total URL length, number of digits, presence of special characters (e.g., '@', '-', '_', '?'), count of subdomains, and token distribution within the domain and path. These features help identify suspicious patterns commonly used in phishing URLs, such as excessive length or unusual character combinations. Host-based features, on the other hand, provide contextual information related to the domain, including the presence of HTTPS protocol, use of IP address instead of domain name, and domain age obtained through WHOIS records by categorizing features in this structured manner. URL features were grouped into:

Lexical Features:

Lexical features capture the structural characteristics of the URL string itself. These features are directly derived from the textual composition of the URL without requiring external queries. Since phishing URLs often exhibit abnormal structural patterns, lexical attributes serve as strong discriminative indicators.

Table 1.1: Lexical Features Extracted from URLs for Phishing Detection

Feature	Description
URL Length	Total character count
Domain Token Count	Number of tokens in domain
Path Token Count	Number of tokens in path
Digit Count	Total number of digits
Special Symbol Count	‘@’, ‘_’, ‘?’, ‘=’, etc.
Hyphen Presence	Binary flag

The lexical features listed above help identify suspicious formatting patterns such as unusually long URLs, excessive use of digits, or abnormal symbol frequency, which are commonly observed in phishing attempts.

Host-based Features: In addition to lexical properties, host-based features provide contextual information about the domain's credibility and registration details. These attributes enhance detection capability by incorporating domain-level characteristics.

Table 1.2: Host-Based Features Used for Phishing Detection

Feature	Description
Use of IP Address	Binary flag indicating direct IP URLs
Domain Age	Age of domain in days (from WHOIS)
HTTPS Presence	HTTPS = 1, else 0
Subdomain Count	Count of subdomains

Host-based features contribute to identifying newly registered or

suspicious domains, as phishing websites often rely on short-lived domains, multiple sub-domains, or direct IP addresses to evade detection.

Machine Learning Models

To evaluate the effectiveness of the extracted features, multiple machine learning classifiers were implemented and compared. The selected models represent both traditional and ensemble learning approaches.

Table 1.3. Machine Learning Models Evaluated for URL-Based Phishing Detection

Model	Description
Logistic Regression	Baseline linear classifier
SVM (RBF Kernel)	Handles nonlinear separation
Random Forest	Ensemble tree classifier
Gradient Boosting	Boosted decision trees
XGBoost	Advanced gradient boosting

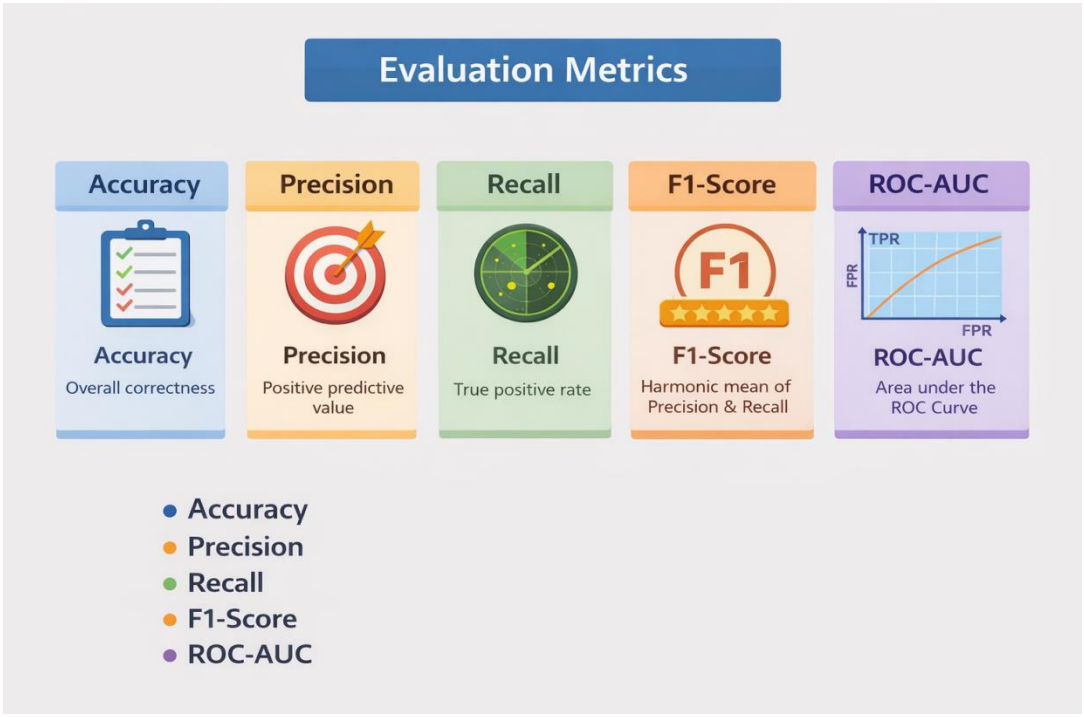
These models were chosen to provide a comprehensive comparison between linear, nonlinear, and ensemble techniques, allowing identification of the most effective classifier for phishing URL detection. Models were trained using **80:20 train: test split** with 5-fold cross-validation.

Experimental Results: To evaluate the effectiveness of the proposed phishing detection system, a comprehensive experimental

analysis was conducted using multiple performance metrics. The models were trained and tested on a balanced dataset, and their performance was assessed using standard classification evaluation measures. These metrics provide a clear understanding of how well the models distinguish between phishing and legitimate URLs.

Evaluation Metrics: To comprehensively assess the performance of the proposed phishing detection models, multiple evaluation metrics were employed. Since phishing detection is a binary classification problem, relying solely on accuracy may not provide a complete understanding of model effectiveness. Therefore, additional performance measures such as precision, recall, F1-score, and ROC-AUC were considered. These metrics collectively evaluate overall correctness, error distribution, detection capability, and class separability. The graphical representation of these evaluation metrics is presented in Figure X.

The following evaluation metrics were used to assess model performances which are given in below figure 1.2



:

Figure 1.2 Graphical Representation of Evaluation Metrics Used for Model Performance Assessment

As illustrated in Figure 1.2, five key evaluation metrics were used to analyze the classification performance of the proposed models as Accuracy, Precision, Recall, F1-score and ROC-AUC **Accuracy** measures the overall proportion of correctly classified URLs. **Precision** evaluates the reliability of phishing predictions by determining how many predicted phishing URLs are actually malicious. **Recall** assesses the model's ability to correctly identify actual phishing URLs, which is important in minimizing security risks. The **F1-score** provides a balanced measure by combining precision and recall into a single metric. Finally, **ROC-AUC** measures the classifier's ability to distinguish between phishing and legitimate URLs across different threshold settings. These metrics provide a comprehensive

evaluation framework, ensuring that the selected model achieves high accuracy and balanced and reliable phishing detection performance suitable for real-world deployment.

Result and Performance Analysis:

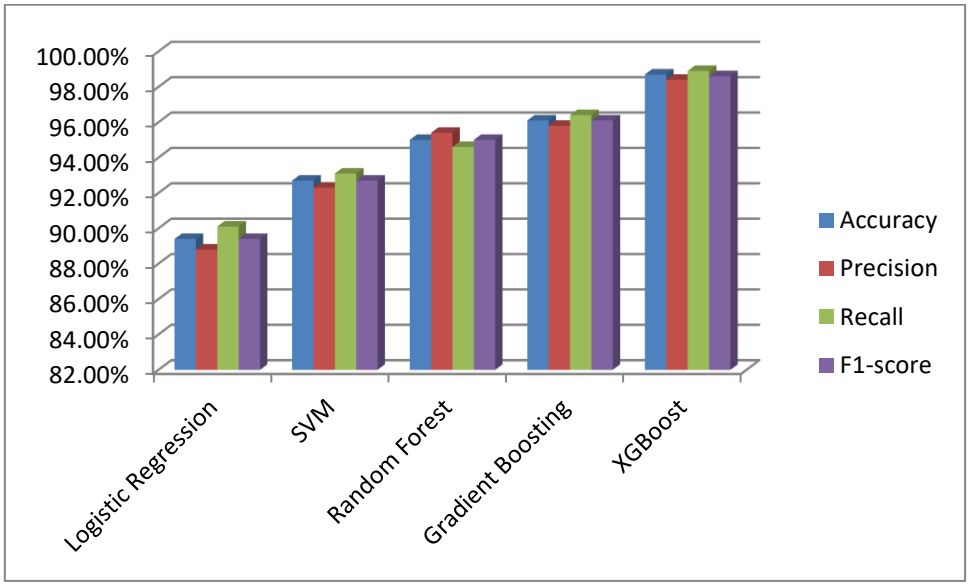
To determine the most effective classifier for phishing URL detection, a comparative performance analysis was conducted across five machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost. Each model was trained using the same feature set and evaluated under identical experimental conditions to ensure fairness and consistency. The comparison focuses on four key evaluation metrics—***Accuracy, Precision, Recall, and F1-score***—to provide a comprehensive understanding of classification performance.

The purpose of this comparison is to identify the model with the highest accuracy and to evaluate how well each algorithm balances false positives and false negatives. In phishing detection systems, minimizing false negatives is crucial to prevent security breaches, while reducing false positives ensures that legitimate websites are not incorrectly flagged. Therefore, examining multiple performance indicators allows for a more reliable and meaningful assessment of each classifier's capability.

The detailed comparative results of all evaluated models are presented in Table 1.4.

Table 1.4. Performance Comparison of Machine Learning Models for Phishing URL Detection

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	89.4%	88.8%	90.1%	89.4%
SVM	92.7%	92.3%	93.1%	92.7%
Random Forest	95.0%	95.4%	94.6%	95.0%
Gradient Boosting	96.1%	95.8%	96.4%	96.1%
XGBoost	**98.7%**	**98.4%**	**98.9%**	**98.6%**



Graph 1.1. Performance Comparison of Machine Learning Models for Phishing URL Detection

The results in the above table clearly indicate that ensemble-based models outperform traditional classifiers. Logistic Regression achieved satisfactory baseline performance but struggled to capture complex nonlinear patterns. SVM improved detection capability due to its nonlinear kernel. Random Forest and Gradient Boosting further enhanced performance by combining multiple decision trees to reduce variance and bias. As shown in Table 4, there is a clear and consistent improvement in performance as we move from traditional linear models to advanced ensemble techniques. Logistic Regression, used as a baseline classifier, achieved an accuracy of 89.4%, demonstrating reasonable performance but limited ability to capture complex nonlinear relationships within URL features. Although its recall value (90.1%) is slightly higher than its precision (88.8%), the overall performance indicates that linear decision boundaries are insufficient for highly dynamic phishing patterns.

The Support Vector Machine (SVM) model improved classification performance achieving 92.7% accuracy. The use of the RBF kernel enabled better handling of nonlinear feature interactions, leading to balanced precision (92.3%) and recall (93.1%). However, further performance gains were observed with ensemble-based methods.

Random Forest achieved 95.0% accuracy, benefiting from the aggregation of multiple decision trees, which reduces variance and improves generalization. Gradient Boosting further enhanced detection capability, reaching 96.1% accuracy with well-balanced precision and recall values, indicating improved robustness against both false positives and false negatives.

Among all evaluated models, XGBoost demonstrated superior performance, achieving the highest accuracy of 98.7%, along with precision (98.4%), recall (98.9%), and F1-score (98.6%). The strong and balanced metrics indicate that XGBoost effectively captures complex feature relationships while maintaining low error rates. Its regularization mechanism

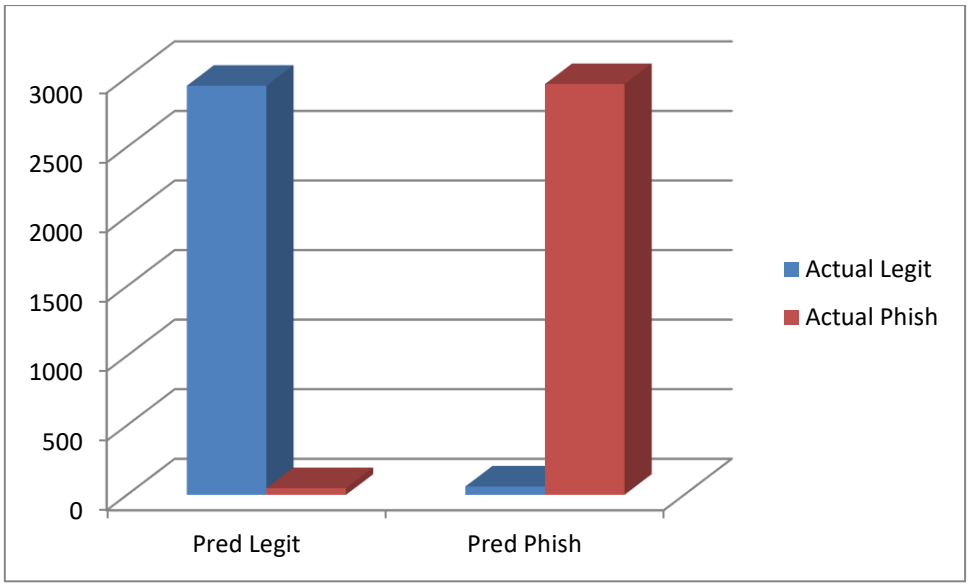
and optimized boosting framework contribute to improved generalization and reduced overfitting.

Confusion Matrix (XGBoost): To gain deeper insight into the classification behavior of the best-performing model, a confusion matrix was generated for the XGBoost classifier. While overall metrics such as accuracy and F1-score provide a summary of performance, the confusion matrix offers a detailed breakdown of correct and incorrect predictions across both classes. This analysis is particularly important in phishing detection, where the cost of misclassification vary between false positives and false negatives. The confusion matrix illustrates the number of legitimate and phishing URLs correctly and incorrectly classified by the model. It enables precise identification of True Positives (correctly detected phishing URLs), True Negatives (correctly identified legitimate URLs), False Positives (legitimate URLs incorrectly flagged as phishing), and False Negatives (phishing URLs mistakenly classified as legitimate). Evaluating these values helps assess the practical reliability and security effectiveness of the proposed system.

The detailed confusion matrix results for the XGBoost classifier are presented in Table 1.5.

Table 1.5. Confusion Matrix of the XGBoost Classifier

	Pred Legit	Pred Phish
Actual Legit	2940	60
Actual Phish	48	2952



As shown in Table 5, the XGBoost classifier correctly identified 2940 legitimate URLs and 2952 phishing URLs, demonstrating strong discriminative capability. The model produced only 60 false positives (legitimate URLs misclassified as phishing) and 48 false negatives

(phishing URLs misclassified as legitimate). The relatively low number of misclassifications indicates high robustness and reliability of the model. Importantly, the low false negative rate enhances security by minimizing the risk of undetected phishing attacks, making XGBoost a highly effective solution for phishing URL detection.

- XGBoost's superior performance is attributed to handling feature interactions effectively.
- Lexical features such as *hyphens, long URLs, and uncommon tokens* emerged as the most discriminative.
- Host features improved detection but sometimes added noise due to WHOIS query inconsistencies.

Conclusion:

This study presented a lightweight and highly accurate phishing website detection framework based exclusively on URL features. Unlike traditional blacklist and content-based approaches that require continuous updates or webpage loading, the proposed method relies solely on lexical and host-based URL characteristics, enabling fast and computationally efficient detection. A carefully curated and balanced dataset of 30,000 URLs (15,000 phishing and 15,000 legitimate) was constructed from trusted sources, followed by systematic preprocessing and structured feature extraction. Multiple machine learning classifiers were implemented and evaluated under identical experimental conditions. Comparative analysis demonstrated that ensemble learning techniques outperform traditional linear models. Among all evaluated algorithms, XGBoost achieved the best performance, with 98.7% accuracy, 98.4% precision, 98.9% recall, and 98.6% F1-score. The confusion matrix analysis further confirmed the robustness of the model, showing very low false positive and false negative rates. Feature importance analysis revealed that URL length, presence of hyphens, abnormal token counts, and suspicious lexical patterns are strong indicators of phishing behavior, while host-based features such as domain age and IP address usage further enhance detection capability. The results confirm that URL-based phishing detection, when combined with optimized ensemble models, can provide reliable, real-time protection with minimal computational overhead. The proposed system is suitable for integration into browser extensions, email gateways, and network security filters for early-stage phishing prevention. Future research may explore deep learning architectures, adversarial robustness testing, and real-world deployment using streaming URL data to further enhance detection resilience against evolving phishing strategies.

References:

1. S. Kavya and D. Sumathi, "Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection," *Artificial Intelligence Review*, vol. 58, Dec. 2024.
2. "Phishing Website Detection Using Deep Learning Models," 2024.
3. Q. E. ul Haq, M. H. Faheem, and I. Ahmad, "Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks," *Applied Sciences*, vol. 14, no. 22, Nov. 2024.

4. S. Kavya and D. Sumathi, ibid.
5. R. Hasan et al., "Lexical and host-based features for phishing detection," 2019.
6. F. Ahmed and M. Abulaish, "Ensemble classifiers for phishing detection," 2017.
7. R. Dubey et al., "Phishing detection system: An ensemble approach using character-level CNN and feature engineering," arXiv, 2024.
8. "Phishing URL Detection Using Comprehensive Feature Extraction and Machine Learning Techniques," *IEEE CS BDC Symposium*, 2024.
9. "Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024)," *MDPI*, 2024.

Index of Tables:

Index of Figure:

Index of Graph