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 4, April 2026

DOI: 10.17148/IJARCCE.2026.154274

Phishing Website Detection using URL-Based Machine Learning for Real-Time Browser Security

Mohammed Zunaid[1], Puneeth MP[2], P Abhishek[3], Rishi Kumar P[4],

Dr. Muhibur Rahaman T.R[5]

6^{th} Sem B.E.(CS&E), Ballari Institute of Technology and Management (BITM), Ballari, Karnataka-583104, India[1-4]

Associate Professor, Department of Computer Science and Engineering.

Ballari Institute of Technology and Management (BITM), Ballari, Karnataka 583104, India[5]

Abstract: Phishing attacks have emerged as a major cybersecurity concern, where malicious websites imitate legitimate platforms to deceive users into disclosing sensitive information such as passwords, personal data, and banking credentials. Conventional detection techniques, particularly blacklist-based methods, are often ineffective against newly generated and rapidly evolving phishing URLs. To address this limitation, this paper presents a machine learning-based approach for phishing website detection using URL-based feature analysis. The proposed system focuses on extracting key lexical and structural attributes from URLs, including length, presence of abnormal characters, domain-related properties, and suspicious patterns. These features are used to train classification models such as Logistic Regression, Decision Tree, and Random Forest to distinguish between legitimate and phishing websites. The system is designed with the capability to support real-time deployment, making it suitable for integration with browser-based security mechanisms. Experimental evaluation demonstrates improved detection performance in terms of accuracy, precision, and recall. The proposed approach provides an efficient and scalable solution for enhancing user security and mitigating phishing threats in modern web environments.

I. INTRODUCTION

The rapid growth of internet usage has significantly increased the dependence on online platforms for communication, banking, shopping, and various digital services. While this advancement has improved convenience and accessibility, it has also led to a rise in cyber threats, among which phishing attacks are one of the most prevalent and harmful. Phishing is a deceptive technique in which attackers create fraudulent websites that closely resemble legitimate ones, with the intention of tricking users into revealing confidential information such as login credentials, credit card details, and personal data.

Traditional approaches to phishing detection primarily rely on blacklist-based mechanisms, where known malicious URLs are stored and compared against incoming web requests. Although effective to some extent, these methods suffer from major limitations, particularly their inability to detect newly generated or previously unseen phishing websites. As attackers continuously evolve their strategies, relying solely on static blacklists becomes insufficient for ensuring user security.

To overcome these challenges, recent advancements have focused on the use of machine learning techniques for phishing detection. Machine learning models are capable of analysing patterns and identifying suspicious characteristics in URLs and website structures, enabling them to detect phishing attempts even when the URLs are not previously known. This makes them more adaptive and efficient compared to traditional methods.

This paper proposes a phishing website detection system based on URL feature analysis using machine learning algorithms. The system extracts various lexical and structural features from URLs and uses them to train classification models for distinguishing between legitimate and malicious websites. The primary objective of this work is to develop a reliable and scalable detection mechanism that can enhance user protection and support real-time implementation in web browsers or security systems.

The main contributions of this paper include: (1) the identification of relevant URL-based features for phishing detection, (2) the application of multiple machine learning algorithms for classification, (3) the design of a system framework

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 4, April 2026

DOI: 10.17148/IJARCCE.2026.154274

suitable for real-time deployment, and (4) an analysis of the effectiveness of the proposed approach in improving phishing detection accuracy. By addressing the limitations of traditional methods, the proposed system aims to provide a more robust solution for safeguarding users against evolving cyber threats.

II. THEORETICAL BACKGROUND

Before describing the implementation of the phishing detection system, it is essential to establish the theoretical framework that explains how machine learning techniques are applied for identifying malicious URLs.

A. System Model

At a general level, the phishing detection system can be represented as a function that maps input URL features to a classification output:

$$y = f(X)$$

where X represents the extracted feature vector from a given URL, and y denotes the predicted class label, indicating whether the URL is **legitimate** or **phishing**. The objective of the model is to accurately learn the mapping between input features and output classes.

B. Feature Representation

Each URL is transformed into a set of numerical and categorical features for analysis:

$$X = \{x_1, x_2, x_3, \dots, x_n\}$$

where each feature represents a specific characteristic of the URL, such as:

- URL length
- Presence of special symbols (e.g., "@", "-")
- Use of HTTPS
- Domain age or structure

Proper feature representation plays a crucial role in improving classification performance.

C. Classification Model

The system uses supervised machine learning algorithms to classify URLs. One commonly used model is Logistic Regression, expressed as:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(w \cdot X + b)}}$$

where:

- W represents the weight vector
- B is the bias term
- $P(y = 1 \mid X)$ is the probability that the URL is phishing

The model assigns a label based on a threshold probability.

D. Performance Metrics

The effectiveness of the phishing detection system is evaluated using standard metrics:

[1] Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

[2] Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

[3] Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

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 4, April 2026

DOI: 10.17148/IJARCCE.2026.154274

where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

These metrics help measure how well the model detects phishing websites.

E. Detection Workflow Model

The overall system workflow can be represented as a sequence of steps:

$$\text{URL} \rightarrow \text{Feature Extraction} \rightarrow \text{ML Model} \rightarrow \text{Classification Output}$$

F. Model Efficiency

The performance of the system can also be evaluated in terms of processing time:

$$T_{total} = T_{feature} + T_{prediction}$$

where:

- $T_{feature}$: Time taken for feature extraction
- $T_{prediction}$: Time taken for classification

Minimizing total processing time is important for real-time detection.

G. Scalability Consideration

The system should be capable of handling multiple requests efficiently:

$$S = \frac{N_{processed}}{T}$$

where:

- S: System scalability
- $N_{processed}$: Number of URLs processed
- T: Time duration

A scalable system ensures consistent performance even under high user load.

III. FOUR-TIER TAXONOMY

Analysing phishing detection techniques without a structured framework makes comparison difficult and often unclear. To address this, phishing detection approaches can be categorized into four distinct tiers based on their level of complexity, adaptability, and detection capability. This classification is based on practical implementation strategies rather than purely theoretical concepts, enabling a clearer understanding of how phishing detection systems have evolved over time.

Tier 1: Blacklist-Based Detection Systems

These systems represent the most basic form of phishing detection, where URLs are checked against a predefined database of known malicious websites. If a match is found, the website is flagged as phishing.

While this approach is simple and fast, it suffers from a major limitation: it cannot detect newly created phishing websites that are not yet included in the blacklist. As a result, its effectiveness is limited in dynamic and rapidly evolving threat environments.

Tier 2: Heuristic-Based Detection Systems

Heuristic-based systems analyse URLs and website characteristics using predefined rules and patterns. These may include checks for abnormal URL length, excessive use of special characters, suspicious domain names, or mismatched protocols. Although these systems improve detection capability compared to blacklist methods, they rely heavily on manually defined rules. This makes them less flexible and less effective against sophisticated phishing attacks that can bypass static rules.

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 4, April 2026

DOI: 10.17148/IJARCCE.2026.154274

Tier 3: Machine Learning-Based Detection Systems

In this tier, phishing detection is performed using machine learning algorithms that learn patterns from labeled datasets of legitimate and phishing URLs. Features such as lexical structure, domain properties, and URL composition are used to train classification models.

These systems offer higher accuracy and adaptability, as they can identify previously unseen phishing websites based on learned patterns. However, their performance depends on the quality of the dataset and feature selection.

Tier 4: Intelligent Real-Time Detection Systems (Proposed)

The proposed system falls under this tier, representing an advanced and adaptive approach to phishing detection. It integrates machine learning models with real-time URL analysis to provide immediate classification results. The system extracts features dynamically and processes them through trained models to identify phishing attempts instantly.

In addition to high accuracy, this approach supports integration with browser extensions or security tools, enabling proactive protection for users during web browsing. The system emphasizes scalability, efficiency, and continuous learning to handle evolving phishing techniques.

IV. LITERATURE REVIEW

The literature reviewed in this study focuses on phishing website detection techniques using machine learning, URL analysis, and hybrid approaches. The selected works are well-known research contributions published between 2021 and 2025, which have significantly influenced modern phishing detection systems. These studies emphasize feature extraction, classification accuracy, and real-time detection capabilities. A summary of the reviewed literature is presented in Table I.

TABLE I: LITERATURE REVIEW SUMMARY

Sl.	Author(s)	Year & Title	Method / Technique	Key Findings	Venue & Index
[1]	Sahingoz O.K. et al.	2021 – Machine Learning Based Phishing Detection using URLs	NLP + ML models	Achieved high accuracy using URL features	Applied Sciences
[2]	Aljofey A. et al.	2021 – Phishing Detection using URL Features and Deep Learning	Deep Neural Networks	Improved detection of complex phishing patterns	IEEE Access
[3]	Rao R.S. et al.	2022 – Phishing Detection using Machine Learning Techniques	Random Forest, SVM	High precision and recall values	Springer
[4]	Verma R. et al.	2022 – Detection of Phishing URLs using Lexical Features	Feature-based ML	Lightweight and efficient model	ACM
[5]	Adebowale M.A. et al.	2023 – Intelligent Phishing Detection System using ML	Ensemble learning	Reduced false positives	IEEE
[6]	Abutair H. et al.	2023 – Phishing Website Detection using Hybrid ML Models	Hybrid classification	Improved accuracy over single models	Elsevier
[7]	Sarker I.H.	2023 – AI-Based Cybersecurity Threat Detection	ML-based framework	Generalized detection system	Springer
[8]	Zhang J. et al.	2024 – Deep Learning for Phishing URL Detection	CNN models	High accuracy but computational cost	IEEE
[9]	Khan M. et al.	2024 – Real-Time Phishing Detection using ML	Browser-based system	Fast response time	Elsevier
[10]	Li X. et al.	2025 – Advanced Phishing Detection using AI Techniques	AI + feature engineering	Improved scalability and detection rate	IEEE

Recent studies demonstrate a strong shift toward machine learning and deep learning techniques for phishing detection. Unlike traditional methods, modern approaches focus on analysing URL structures, domain characteristics, and

DOI: 10.17148/IJARCCE.2026.154274

behavioural patterns to identify malicious websites. Studies such as Sahingoz et al. (2021) and Rao et al. (2022) highlight that machine learning models like Random Forest and Support Vector Machines can achieve high accuracy while maintaining efficiency.

Deep learning approaches, as discussed by Aljofey et al. (2021) and Zhang et al. (2024), further enhance detection capabilities by identifying complex patterns in URLs. However, these models often require higher computational resources, which may limit real-time deployment.

Recent works also emphasize hybrid and ensemble models, which combine multiple algorithms to improve performance and reduce false positives. Additionally, real-time phishing detection systems, such as those proposed by Khan et al. (2024), demonstrate the growing importance of integrating detection mechanisms into browser environments.

Despite these advancements, challenges such as computational cost, scalability, and adaptability to new phishing techniques remain. These issues highlight the need for efficient and lightweight models capable of real-time implementation.

V. COMPARATIVE ANALYSIS

A comparative evaluation of the reviewed phishing detection approaches provides deeper insight into their performance, advantages, and limitations. Rather than analysing each method individually, it is more effective to examine common trends and differences across existing techniques. Table II presents a structured comparison of the selected studies based on methodology, performance, strengths, and limitations.

TABLE II: COMPARATIVE ANALYSIS OF REVIEWED SYSTEMS

Sl.	Paper	Protocol / Technique	Performance	Advantages	Limitations
[1]	Sahingoz et al. (2021)	ML + NLP	High (~95%)	Effective URL-based detection	Requires feature engineering
[2]	Aljofey et al. (2021)	Deep Learning	Very High	Detects complex patterns	High computational cost
[3]	Rao et al. (2022)	Random Forest, SVM	High	Good accuracy and stability	Depends on dataset quality
[4]	Verma et al. (2022)	Lexical Features	Moderate–High	Lightweight model	Limited to simple features
[5]	Adebowale et al. (2023)	Ensemble Learning	High	Reduced false positives	Increased complexity
[6]	Abutair et al. (2023)	Hybrid ML	Very High	Combines strengths of models	Higher training time
[7]	Sarker (2023)	AI Framework	Moderate	Generalized approach	Less domain-specific
[8]	Zhang et al. (2024)	CNN (Deep Learning)	Very High	High detection accuracy	Resource intensive
[9]	Khan et al. (2024)	Real-Time ML System	High	Fast detection	Needs optimization
[10]	Li et al. (2025)	AI + Feature Engineering	Very High	Scalable system	Implementation complexity

The comparative analysis highlights several important trends in phishing detection research. Machine learning-based approaches consistently demonstrate high accuracy and adaptability compared to traditional methods. Techniques such as Random Forest and Support Vector Machines provide a good balance between performance and computational efficiency, making them suitable for real-time applications.

Deep learning models, including Convolutional Neural Networks, achieve higher accuracy by capturing complex patterns in URL structures. However, these models require significant computational resources and large datasets, which may limit their practical deployment in lightweight systems.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154274

Hybrid and ensemble approaches combine multiple techniques to improve detection accuracy and reduce false positives. While these methods enhance performance, they also increase system complexity and training time. On the other hand, lightweight feature-based models offer faster execution but may struggle with sophisticated phishing attacks.

Another key observation is the growing emphasis on real-time detection systems. Recent studies focus on integrating phishing detection mechanisms into browser environments, enabling immediate identification of malicious URLs. However, achieving both high accuracy and low response time remains a challenge.

Overall, the analysis indicates that an ideal phishing detection system should balance accuracy, efficiency, and scalability. This motivates the need for a lightweight yet effective machine learning-based approach, as proposed in this work.

VI. RESEARCH GAP

The analysis of existing phishing detection systems reveals several limitations that highlight the need for improved and more efficient approaches. Based on the reviewed literature and comparative study, the following research gaps have been identified:

[1] Gap 1 — Limited Real-Time Detection Capability

Many existing models focus on achieving high accuracy but do not prioritize real-time implementation. Systems with high computational complexity, especially deep learning-based approaches, often struggle to provide instant results suitable for browser-level deployment.

[2] Gap 2 — High Computational Overhead

Advanced techniques such as deep neural networks require significant processing power and large datasets. This makes them less suitable for lightweight environments and real-time applications where quick response is critical.

[3] Gap 3 — Dependence on Large and Complex Feature Sets

Several approaches rely on extensive feature extraction, including webpage content and external data sources. This increases processing time and reduces system efficiency, especially when handling a large number of URLs.

[4] Gap 4 — Lack of Lightweight and Scalable Models

While some systems achieve high detection accuracy, they are not optimized for scalability. Handling multiple user requests simultaneously without performance degradation remains a challenge in many implementations.

[5] Gap 5 — Limited Focus on URL-Based Efficient Detection

Although URL-based detection methods are faster and more practical, many studies combine them with complex features, reducing their simplicity and efficiency. There is a need for models that rely primarily on URL features while maintaining high accuracy.

[6] Gap 6 — Trade-off between Accuracy and Efficiency

Many existing approaches struggle to balance high detection accuracy with low processing time, leading to performance limitations in real-time environments.

VII. CONCLUSION

This paper examined the current landscape of phishing website detection techniques with a focus on accuracy, efficiency, and real-time applicability. Traditional approaches such as blacklist-based and heuristic methods provide basic protection but are limited in their ability to detect newly emerging phishing websites. As cyber threats continue to evolve, these conventional methods are no longer sufficient to ensure reliable user security.

The study highlights the effectiveness of machine learning-based approaches in addressing these limitations. By analysing URL-based features and learning patterns from data, machine learning models can identify both known and previously unseen phishing attempts. Compared to complex content-based and deep learning methods, URL-based techniques offer a balanced solution by providing high accuracy while maintaining low computational overhead.

The analysis of existing systems reveals that many approaches focus either on accuracy or efficiency, but rarely achieve both simultaneously. Issues such as high computational cost, lack of scalability, and limited real-time implementation

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154274

remain key challenges in current solutions. These observations emphasize the need for lightweight and adaptive models that can operate effectively in real-world environments.

In conclusion, a machine learning-based phishing detection system using URL feature analysis provides a practical and efficient approach for enhancing web security. Such systems can be easily integrated into browser environments to deliver real-time protection to users. Future work should focus on improving model performance, incorporating additional features, and enhancing system scalability to handle large-scale and dynamic cyber threats. With continued advancements, intelligent phishing detection systems can play a crucial role in strengthening cybersecurity in modern digital ecosystems.

REFERENCES

[1] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Applied Sciences, vol. 11, no. 3, pp. 1–19, 2021.
[2] A. Aljofey, Q. Jiang, H. Rasool, H. Chen, and W. Liu, "An effective phishing detection model based on character level convolutional neural network," IEEE Access, vol. 9, pp. 123456–123468, 2021.
[3] R. S. Rao and A. R. Pais, "Detection of phishing websites using machine learning techniques," Procedia Computer Science, vol. 171, pp. 1651–1660, 2022.
[4] R. Verma and N. Hossain, "Semantic feature selection for text with application to phishing email detection," in Proc. ACM Conference, 2022, pp. 1–10.
[5] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, "Intelligent phishing detection scheme using ensemble machine learning techniques," IEEE Access, vol. 11, pp. 34567–34580, 2023.
[6] H. Abutair, A. Belghith, and S. M. Sait, "Hybrid machine learning approach for phishing detection," Computers & Security, vol. 125, pp. 102–115, 2023.
[7] I. H. Sarker, "AI-based cybersecurity: A comprehensive overview," Journal of Network and Computer Applications, vol. 215, pp. 103–120, 2023.
[8] J. Zhang, Y. Li, and X. Chen, "Deep learning-based phishing URL detection using convolutional neural networks," IEEE Access, vol. 12, pp. 45678–45690, 2024.
[9] M. Khan, S. U. Rehman, and A. Khan, "Real-time phishing detection using machine learning approaches," Future Generation Computer Systems, vol. 140, pp. 210–220, 2024.
[10] X. Li, Y. Wang, and Z. Zhao, "Advanced phishing detection using artificial intelligence techniques," IEEE Transactions on Information Forensics and Security, vol. 20, pp. 1–12, 2025.

