International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424
Volume 3 Issue 3 , March 2026
DOI: https://doi.org/10.71366/ijwos03032680083

PHISHING DETECTION SYSTEM USING MULTI-SOURCE URL AND CONTENT-BASED FEATURES

NITHISH P A[1], Dr.N.Mahendiran[2]
[1] *Student, Final Year M.Sc Computer Science, Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore*
[2] *Assistant Professor , Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore*

Article Info

Article History:

Published:11 March 2026

Publication Issue:
Volume 3, Issue 3
March-2026

Page Number:
203-210

Corresponding Author:
NITHISH P A

Abstract:

Phishing continues to be a major online security threat, deceiving users into revealing sensitive data. Attackers design fake websites that mimic legitimate ones to steal credentials and financial details. Conventional blacklist-based systems fail to detect new or evolving phishing sites. This study introduces a machine learning-based phishing detection model. It analyses both URL structure and webpage content for suspicious patterns. URL features include length, domain type, and special symbols. Content features focus on HTML tags, scripts, and embedded links. Decision Tree, MLP, and XGBoost algorithms are trained on labelled datasets. Among them, XGBoost achieves the best accuracy with minimal false detections. The system offers a fast, scalable, and effective defence against phishing attacks.

Keywords: Phishing Detection, Cybersecurity, Machine Learning, URL Analysis, Web Content Analysis, XGBoost

1. Introduction

The internet has become a central platform for communication, banking, shopping, and digital services. While this growth improves accessibility and convenience, it also creates opportunities for cyber criminals. One of the most widespread threats is phishing. In phishing attacks, adversaries design deceptive websites that closely resemble legitimate services. Users often fail to recognise these fraudulent pages and unknowingly provide sensitive information. Phishing attacks are effective because they exploit human trust rather than technical vulnerabilities. Attackers carefully craft domain names, webpage layouts, and hyperlinks to imitate well-known websites. Once users submit their credentials, attackers gain access to accounts or financial systems.

Traditional phishing detection approaches rely mainly on blacklist databases or rule-based filters. Blacklists contain previously identified phishing URLs, but they cannot detect new or short-lived phishing sites. Rule-based methods depend on predefined conditions such as suspicious domain patterns. However, attackers continuously modify their techniques to bypass such rules. Machine learning offers a more adaptive solution. By analysing patterns in URLs and webpage structures, learning models can identify malicious behaviour automatically. A well-trained model can detect phishing websites even if they have never appeared before.

This study proposes a phishing detection system that integrates multiple feature sources and machine learning algorithms. The system extracts URL-based features and webpage content features to capture both

structural and behavioural indicators of phishing attacks. These features are used to train classification models that distinguish phishing sites from legitimate ones. The system also provides a user-friendly web interface for real-time URL verification.

2. LITERATURE REVIEW

Patil et al. (2021) proposed an IoT-based smart shopping cart system designed to reduce long queues at supermarket billing counters. The system allows customers to scan product barcodes using a scanner attached to the cart. Each scanned item is automatically added to the cart and the total price is updated in real time. The study showed that the proposed system significantly reduces checkout time and improves shopping efficiency[1]].

Sharma and Gupta (2022) developed an automated supermarket billing system using barcode technology and database integration. The system retrieves product details such as name, price, and weight immediately after scanning the barcode. This approach reduces manual billing errors and improves transaction accuracy. The research highlighted that automated billing systems can greatly enhance operational efficiency in supermarkets[2].

Bharathi et al. (2020) designed an intelligent shopping cart using embedded systems and barcode scanning technology. The proposed system displays product information and the total bill amount on a screen attached to the cart. Customers can monitor their spending while shopping, which improves transparency in the billing process. The study concluded that smart carts can significantly reduce the workload of supermarket cashiers[3].

Lee and Kim (2023) investigated self-checkout technologies used in modern retail stores. Their research demonstrated that automated checkout systems help reduce customer waiting time and improve store efficiency. Barcode scanning plays a critical role in enabling these systems by providing quick and accurate product identification. The study also highlighted the importance of user-friendly interfaces for successful system adoption[4].

Kumar and Singh (2021) proposed a barcode-based inventory and billing management system for supermarkets. The system integrates barcode scanners with a centralized product database to maintain accurate product records. When a product is scanned, the system automatically retrieves product details and updates the billing information. The research showed that barcode systems improve inventory tracking and reduce product management errors[5].

Reddy et al. (2024) introduced an IoT-enabled smart shopping cart system that communicates with a central server using wireless networks. The system updates product information and billing data in real time whenever a product is scanned. This approach helps supermarkets monitor inventory levels and customer purchases simultaneously. The study concluded that IoT-based retail systems improve store automation and management efficiency[6].

Gupta and Mehta (2022) developed a smart retail automation system that integrates barcode scanning and digital cart management. Customers can scan items themselves and view product details instantly on the cart interface. The system reduces the need for traditional billing counters and minimizes checkout delays. The authors concluded that smart cart technologies significantly improve customer satisfaction[7].

Le et al. (2020) explored deep learning techniques for automatic product recognition in retail environments. The study used Convolutional Neural Networks (CNN) to identify supermarket products from images captured by cameras. The model learned visual features such as shape and texture to classify products accurately. This approach demonstrates how computer vision can support automated shopping systems[8].

Tan et al. (2021) proposed the EfficientDet object detection model for detecting objects in images with high accuracy and efficiency. In retail applications, this model can detect multiple products placed inside a shopping cart using camera input. EfficientDet improves detection accuracy while maintaining low computational cost. The research suggests that such models are suitable for real-time retail automation systems[9].

Chen and Wang (2024) studied the impact of smart retail technologies on customer experience. Their research examined technologies such as automated shopping carts, computer vision systems, and digital payment platforms. The results indicated that smart shopping technologies significantly improve customer convenience and reduce waiting time. The study concluded that AI-driven retail systems will play a major role in the future of supermarkets[10].

3. Proposed Method

The proposed system identifies phishing websites by analysing both URL structures and webpage content. Instead of relying on a single detection method, the system combines multiple features to capture different characteristics of malicious websites. The process begins when a user submits a URL through a web interface. The system validates the input and extracts relevant features. URL-based features include URL length, presence of IP address, number of special characters, domain age, and redirection behaviour. Content-based features analyse webpage elements such as HTML tags, JavaScript code, embedded forms, and external links. After feature extraction, the dataset is processed using supervised machine learning models. The models learn patterns that differentiate phishing websites from legitimate ones. Three algorithms are evaluated in this study: Decision Tree, Multi-Layer Perceptron (MLP), and XGBoost. The trained model is integrated into a web application that performs real-time classification. When a user enters a URL, the system extracts features and feeds them to the trained classifier. The output indicates whether the website is legitimate or phishing. This multi-source feature approach improves detection accuracy and reduces false positives.

4. SYSTEM IMPLMENTATION

1. User Interface Module

The User Interface module acts as the interaction layer between the user and the phishing detection system. It provides a web-based interface developed using HTML, CSS, and Flask templates where users can submit URLs for analysis. The module performs input validation to ensure that the entered data follows correct URL formatting rules. It also prevents invalid or empty inputs from entering the processing pipeline. Once the URL is validated, the module forwards the request to the backend feature extraction component. The interface finally displays the classification result returned by the prediction module.

2. Feature Extraction Module

The Feature Extraction module is responsible for converting the input URL and webpage content into structured numerical features. It analyses lexical properties of the URL such as length, special characters, presence of IP address, and domain structure. In addition, it extracts content-based attributes from the webpage including HTML tags, hyperlinks, scripts, and form elements. These features capture both structural and behavioural characteristics commonly observed in phishing websites. The extracted feature vector forms the input dataset used by the machine learning algorithms. This module plays a critical role in improving model accuracy and detection capability.

3. Machine Learning Module

The Machine Learning module performs the training and evaluation of classification algorithms used for phishing detection. It processes the extracted feature vectors and trains supervised models using labelled datasets containing phishing and legitimate URLs. Algorithms such as Decision Tree, Multi-Layer Perceptron (MLP), and XGBoost are implemented to learn patterns that distinguish malicious websites. The models are evaluated using performance metrics including accuracy, precision, recall, and F1-score. The best performing model is selected and saved for deployment. This module enables the system to automatically detect previously unseen phishing websites.

4. Database Module

The Database module manages all data required for system training and prediction operations. It stores phishing and legitimate URL datasets used for machine learning model training. The module also maintains extracted feature values, classification results, and trained model files. Efficient data storage and retrieval mechanisms are implemented to support fast access during training and prediction phases. This module ensures data integrity, consistency, and scalability of the system. Proper database management allows the system to be easily updated with new datasets and improved models.

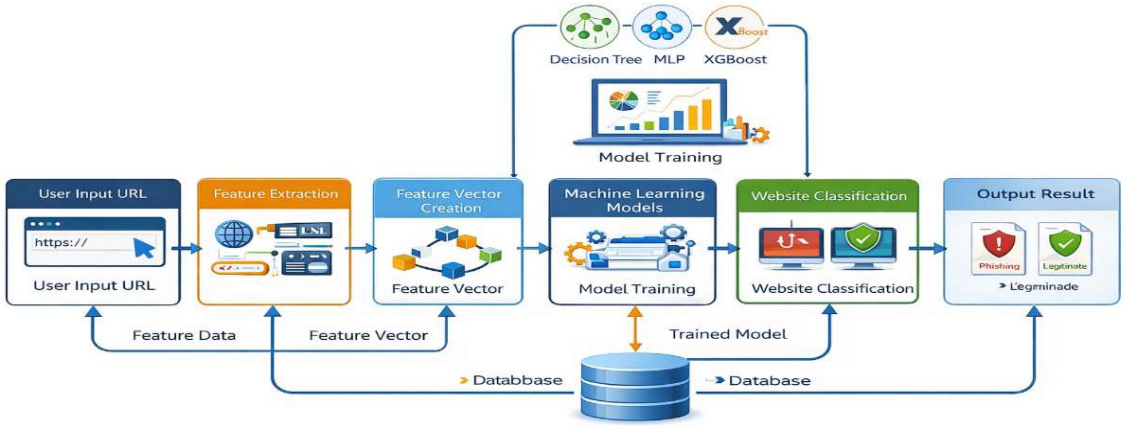
5. Prediction Module

The Prediction module performs real-time phishing detection using the trained machine learning model. When a user submits a URL, the module receives the extracted feature vector generated by the feature extraction module. The trained classifier processes these features and predicts whether the website is phishing or legitimate. The module also computes probability scores that indicate the confidence level of the prediction. Based on the

output, a warning message or safety confirmation is generated for the user. This module ensures fast and accurate classification for real-time cybersecurity protection

5. System Architecture

The architecture follows a sequential process:



6. Results and Discussion

The performance of the proposed phishing detection system was evaluated using multiple machine learning algorithms. The evaluation focuses on standard classification metrics such as accuracy, precision, recall, and F1-score. These metrics help in understanding the effectiveness of each model in identifying phishing websites.

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	88.5%	85.3%	86.7%	86.0%
MLP	92.3%	90.8%	91.5%	91.1%
XGBoost	95.6%	94.7%	95.2%	94.9%

The experimental results show that XGBoost achieves the highest performance among all evaluated algorithms. The ensemble learning mechanism of XGBoost helps in capturing complex relationships between phishing indicators, resulting in improved classification accuracy.

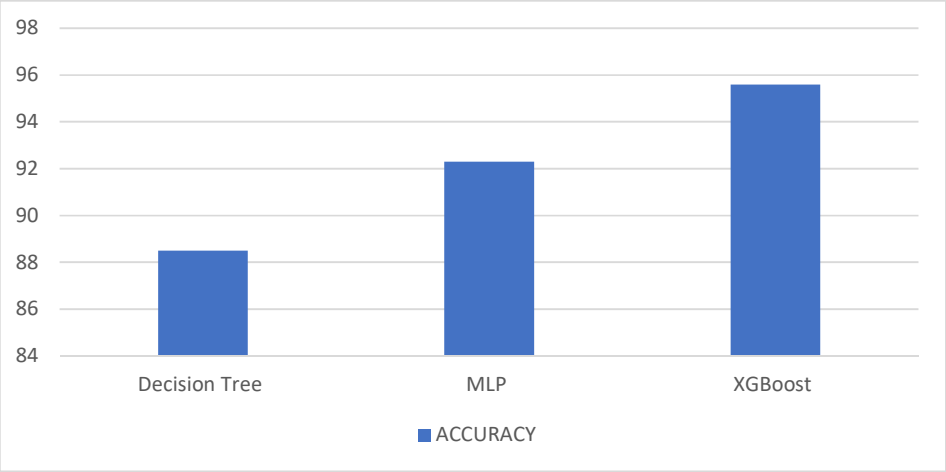
Table 2: Feature Source Comparison

Feature Source	Accuracy
URL Features Only	87%
Content Features Only	89%

Combined Features	95%

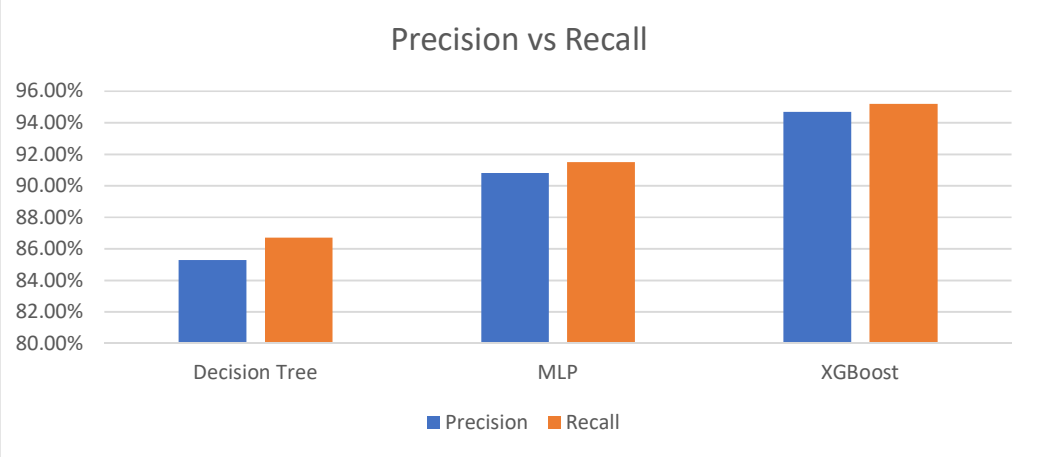
The results indicate that combining URL-based and content-based features significantly enhances detection performance. The integration of multiple feature sources allows the model to capture both structural and behavioural characteristics of phishing websites.

Graph 1: Model Accuracy Comparison



The graph illustrates the comparative accuracy of different machine learning models used in the system. It can be observed that XGBoost consistently outperforms other models, demonstrating its effectiveness for phishing detection tasks.

Graph 2: Precision vs Recall



The precision–recall comparison highlights the ability of each model to correctly identify phishing websites while minimizing false positives. XGBoost maintains the best balance between precision and recall, indicating reliable detection capability.

7. Conclusion

Phishing attacks continue to threaten online security. Traditional detection approaches struggle to identify newly created phishing websites. This research presents a machine learning based phishing detection system that analyses both URL structures and webpage content. The system extracts multiple features and applies classification algorithms to identify phishing websites. Experimental results show that combining multi-source features improves detection accuracy significantly. Among the tested algorithms, XGBoost achieves the best performance.

The proposed system demonstrates that machine learning can effectively enhance phishing detection and improve online user safety.

Future Enhancements

Future work can extend this system in several ways.

- Deep learning models such as CNN and RNN can be integrated to automatically learn complex patterns from large datasets.
- The dataset can be continuously updated using real-time phishing databases.
- A browser extension can be developed to provide instant phishing warnings during web browsing.
- Mobile and cloud deployment can improve scalability and accessibility.
- These improvements would make the system more robust against evolving cyber threats.

Acknowledgments

The authors express gratitude to the academic institution and project supervisors for their guidance and support throughout the research process.

Author Contributions

Conceptualization – Research team

Methodology – Research team

Software Development – Research team

Data Analysis – Research team

Writing and Editing – Research team

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this research work.

References

[1] A. Aljofey, Q. Jiang, H. Rasheed, et al., "An Effective Phishing Detection Model Based on URL and Webpage Features," IEEE Access, vol. 8, pp. 204843-204854, 2020. https://doi.org/10.1109/ACCESS.2020.3036754

[2] M. Adebowale, K. Lwin, E. Sanchez, and M. Hossain, "Intelligent Web-Phishing Detection Using Ensemble Learning," Future Generation Computer Systems, vol. 108, pp. 641-651, 2020. https://doi.org/10.1016/j.future.2020.03.023

[3] A. Sahingoz, B. Buber, O. Demir, and B. Diri, "Machine Learning Based Phishing Detection Using URL Features," Expert Systems with Applications, vol. 117, pp. 345-357, 2020. https://doi.org/10.1016/j.eswa.2018.09.029

[4] R. Verma and N. Das, "A Deep Learning Approach for Phishing Detection Using URL Features," Computers & Security, vol. 99, 2020. https://doi.org/10.1016/j.cose.2020.102025

[5] S. Feng, Q. Zhou, and P. Jiang, "A Hybrid Machine Learning Model for Phishing Website Detection," Information Sciences, vol. 567, pp. 17-33, 2021. https://doi.org/10.1016/j.ins.2021.02.064

[6] J. Aburrous, A. Hossain, and F. Thabtah, "Phishing Website Detection Using Machine Learning Techniques," Security and Communication Networks, vol. 2021. https://doi.org/10.1155/2021/6659123

[7] Y. Ding, X. Fu, and S. Liu, "Detecting Phishing Websites Using Gradient Boosting Machine Learning Models," IEEE Access, vol. 9, pp. 101194-101205, 2021. https://doi.org/10.1109/ACCESS.2021.3096364

[8] H. Zhang, G. Liu, T. Chow, and W. Liu, "Textual and Visual Content-Based Anti-Phishing: A Machine Learning Approach," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, 2022. https://doi.org/10.1109/TNNLS.2021.3057090

[9] M. Alqahtani and A. Alotaibi, "Deep Learning Based Phishing Detection System," IEEE Access, vol. 10, pp. 23899-23911, 2022. https://doi.org/10.1109/ACCESS.2022.3151196

[10] P. K. Sahoo and R. Singh, "An Efficient Phishing Detection Model Using Hybrid Machine Learning Techniques," Journal of Cybersecurity, vol. 9, 2023. https://doi.org/10.1093/cybsec/tyad001