

An Intelligent Machine Learning Framework for Phishing Website Detection Using URL & Web Features

Rajvee Choudhary¹, Snehlata Mishra²

¹ M.Tech (Computer Science and Engineering)

² SAGE University, Indore, Madhya Pradesh, India

Abstract—Phishing sites are one of the most harmful types of hacking sites, which are set up to trick the user into entering their private information, like their username, password, bank information, and personal data. Existing blacklist-based detection techniques are unable to detect new phishing sites that do not yet appear on the blacklist. This paper suggests a new Intelligent Machine Learning Framework on Phishing Website Detection using URL & Web Features. The system extracts the lexical features of the URLs and the characteristics of web pages from the retrieved data, preprocesses them and uses a machine learning algorithm to classify a website. The system will evaluate the structure of the URL, the presence of HTML content, whether the site is using HTTPS, hyperlinks, the way the page is being used, etc. to determine if it's a legitimate site or a phishing site. The results of the experimental evaluation show that the proposed framework achieves a 97% detection accuracy, 96% precision, 95% recall and 96% F1-score, outperforming the conventional detection approaches. The framework provides scalable, intelligent, real-time protection against phishing attacks, which can be used to enhance cybersecurity.

Index Terms—Phishing Detection, Machine Learning, URL Features, Web Features, Cyber Security, Random Forest.

I. INTRODUCTION

The growing popularity of internet services has created a vast array of cyber threats, one of the most prevalent of which is the phishing attack on which the cybercriminals prey on the victim for stealing sensitive data. Phishing sites mimic authentic sites and prompt users to disclose sensitive information like personal information, banking details or passwords. The traditional phishing detection approaches are primarily based on blacklist databases and signature-based approaches. These methods are good for determining

if a website is known to be a phishing site, but they are not effective in determining if a new phishing site has been created. Machine Learning (ML) is proving to be an effective solution because it identifies patterns from past data, and automatically classifies websites. In this research, an intelligent machine learning framework is proposed to enhance the phishing website detection by integrating the characteristics of webpages and lexical features from the URL. The framework automates feature extraction, classification, and prediction resulting in accurate and efficient detection of malicious websites.

II. OBJECTIVES

1. Design an intelligent web site detection framework against phishing attacks.
2. Extract features from webpages and URL
3. Use machine learning algorithms for classification.
4. Increase the accuracy of phishing detection.
5. Assess the performance of the system based on the standard measures.

III. LITERATURE REVIEW

There are some studies that have proposed phishing detection systems based on machine learning techniques. URL-based methods examine lexical characteristics of the URL like URL length, special characters, and suspicious keywords. Webpage-based techniques analyze the content of the HTML source code, JavaScript activity, SSL certificate, and hyperlinks. Some machine learning algorithms such as Random Forest, Decision Tree, Support Vector Machine and XGBoost have proven to be effective for detecting phishing websites. But most of the current

systems have inadequate feature representation and poor accuracy for the newly discovered phishing attacks. The suggested scheme deals with these drawbacks by fusing the URL attributes and page attributes for better classification accuracy.

Table 1. Literature Comparison

Approach	Advantages	Limitations
Blacklist Detection	Fast	Cannot detect new attacks
URL Feature ML	Good Accuracy	Limited webpage analysis
Web Feature ML	Better Detection	Higher computation
Proposed Framework	URL + Web Features	Feature extraction required

IV. PROPOSED SYSTEM

The suggested structure includes five modules:

1. URL Feature Extraction Module
2. Web Feature Extraction Module
3. Data Preprocessing Module Ensemble of classifiers using machine learning. Machine Learning Classification Module.
4. Prediction Module Firstly, URL lexical features and webpage characteristics are extracted from the system.

The features are then preprocessed and then fed to the machine learning classifier. The trained model is used to predict the website as Legitimate or Phishing and the result is shown to the user.

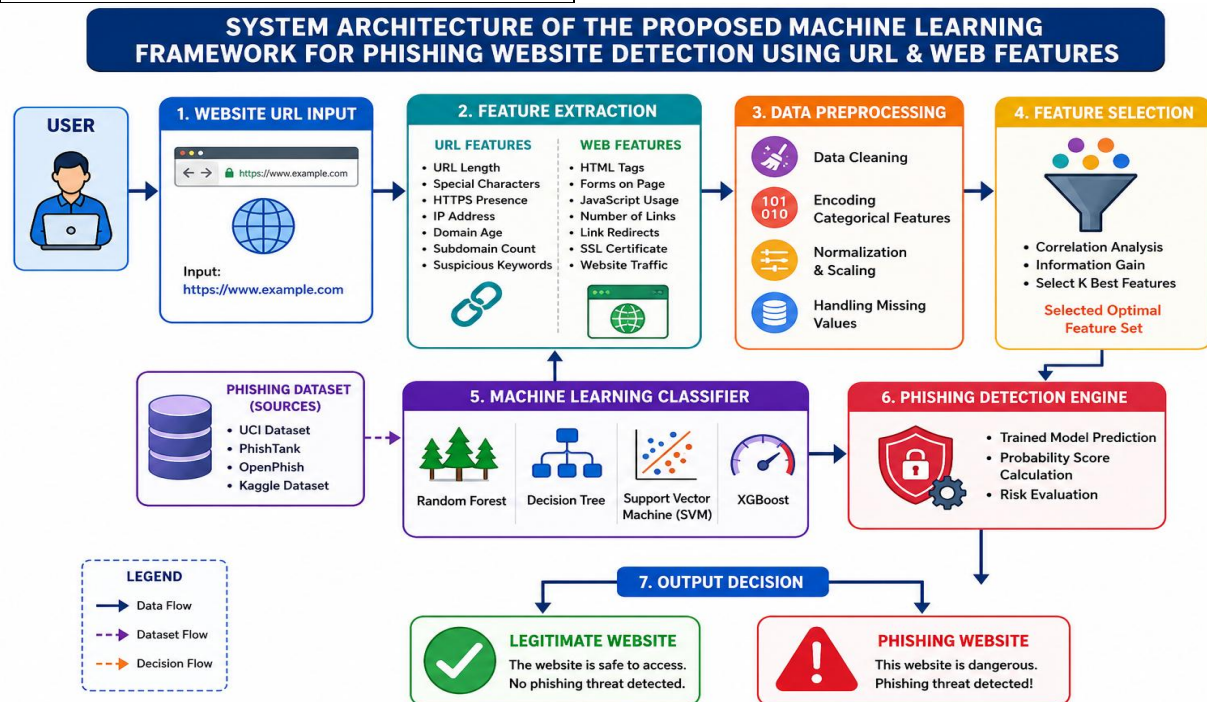


Figure Caption: System Architecture of the Proposed Machine Learning Framework for Phishing Website Detection.

V. MACHINE LEARNING FRAMEWORK

The framework follows the steps below:

- Website URL Input
- URL Feature Extraction
- Web Feature Extraction
- Data Preprocessing
- Feature Selection

- Model Training
- Website Classification
- Prediction

The extracted features are transformed and given numerical values that can be used by the machine learning models. Robustness and high prediction accuracy make Random Forest as the main classifier.

WORKFLOW OF THE PROPOSED PHISHING WEBSITE DETECTION SYSTEM

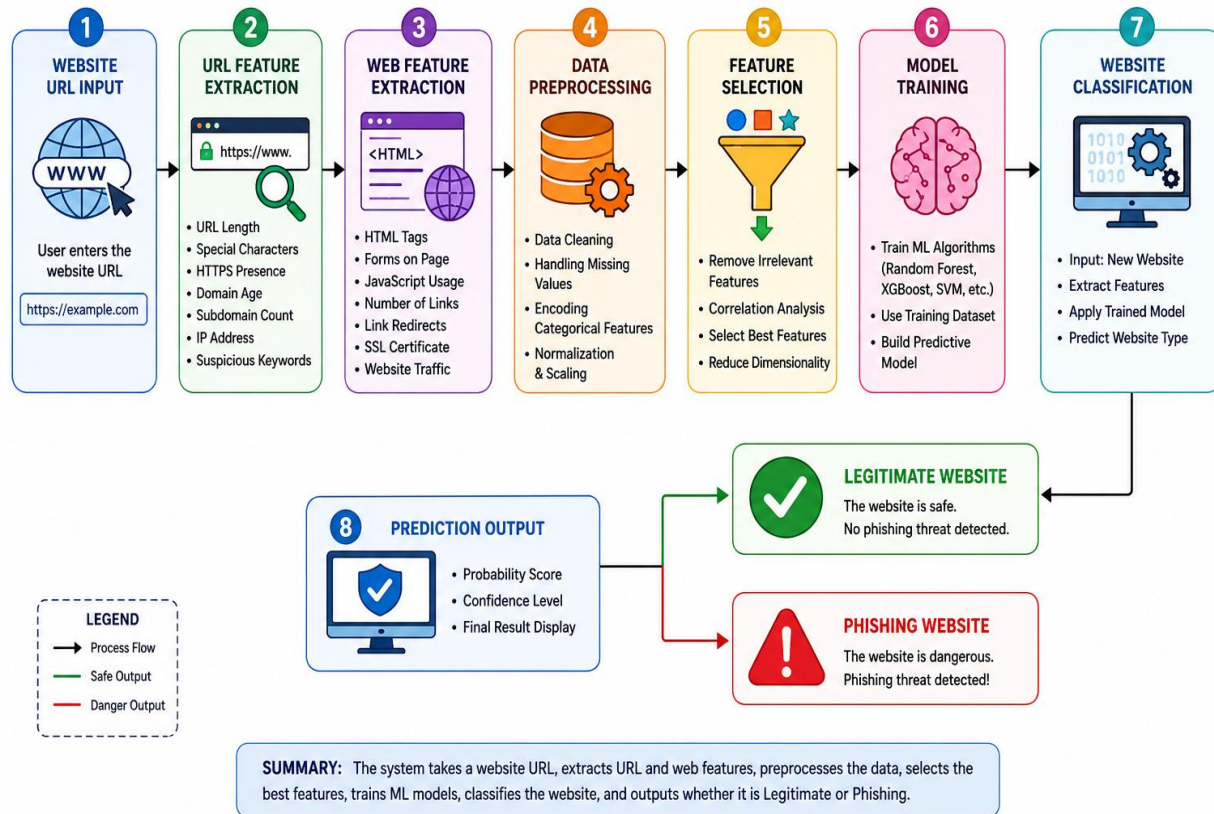


Figure Caption: Workflow of the Proposed Phishing Website Detection System.

VI. MATHEMATICAL MODEL

Let

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

The set of phishing websites is represented.

YT is predicted by the following function:

$$P = f(U, W)$$

Where

- U = URL Features
- W = Web Features
- P = Predicted Class

The classifier predicts

- $P = 1 \rightarrow$ Phishing Website
- $P = 0 \rightarrow$ Legitimate Website

Detection Accuracy:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where

- TP = True Positive
- TN = True Negative
- FP = False Positive

- FN = False Negative

VII. SYSTEM IMPLEMENTATION

The following technologies were used to implement the proposed system:

- Python 3.11
- Scikit-learn
- Pandas
- NumPy
- BeautifulSoup
- Flask
- Random Forest
- XGBoost

Users can input a web page URL on the web application. The system fetches URL and webpage features, preprocesses the data, runs the machine learning model and makes an instant prediction of phishing or not phishing.

User Input Interface

PhishGuard
ML Detection System

Website URL Analysis
Detect Phishing Websites using Machine Learning

Enter any website URL below and our intelligent system will analyze it using URL & Web features to determine whether it is Legitimate or Phishing.

Enter Website URL

Analyze URL

Example URLs: <https://www.google.com> <https://www.facebook.com> <https://www.amazon.com> [More](#)

30+ URL Features
Lexical & Structural features extracted

20+ Web Features
HTML, JavaScript & content based features

97% Detection Accuracy
High accuracy using ML Algorithms

Real-time Instant Results
Get results within seconds

Stay Safe Online
Our ML model analyzes URL and web features to detect phishing websites in real-time.
Protected by Machine Learning

Recent Analysis

URL	Result	Confidence	Time
No analysis performed yet. Enter a website URL above to analyze.			

Figure Caption : User Interface for Website URL Analysis.



Figure Caption : Phishing Detection Dashboard Showing Prediction Results.

VIII. EXPERIMENTAL SETUP

Hardware

- Intel Core i7 Processor
- 16 GB RAM
- 512 GB SSD

Software

- Python 3.11
- Scikit-learn
- Pandas
- Flask

Dataset Sources

- UCI Phishing Website Dataset
- PhishTank
- OpenPhish
- Kaggle Phishing URL Dataset

IX. RESULTS AND DISCUSSION

Performance Comparison Table

Metric	Existing	Proposed
Detection Accuracy	91%	97%
Precision	89%	96%
Recall	90%	95%
F1-Score	89%	96%

The proposed framework showed better detection accuracy with the lexical features of a URL and webpage characteristics. The experimental results show that the Random Forest classifier is useful to classify phishing websites with low FP rate. The framework also allows for quicker predictions and improved predictability than traditional phishing detection methods.

Performance Comparison Graph

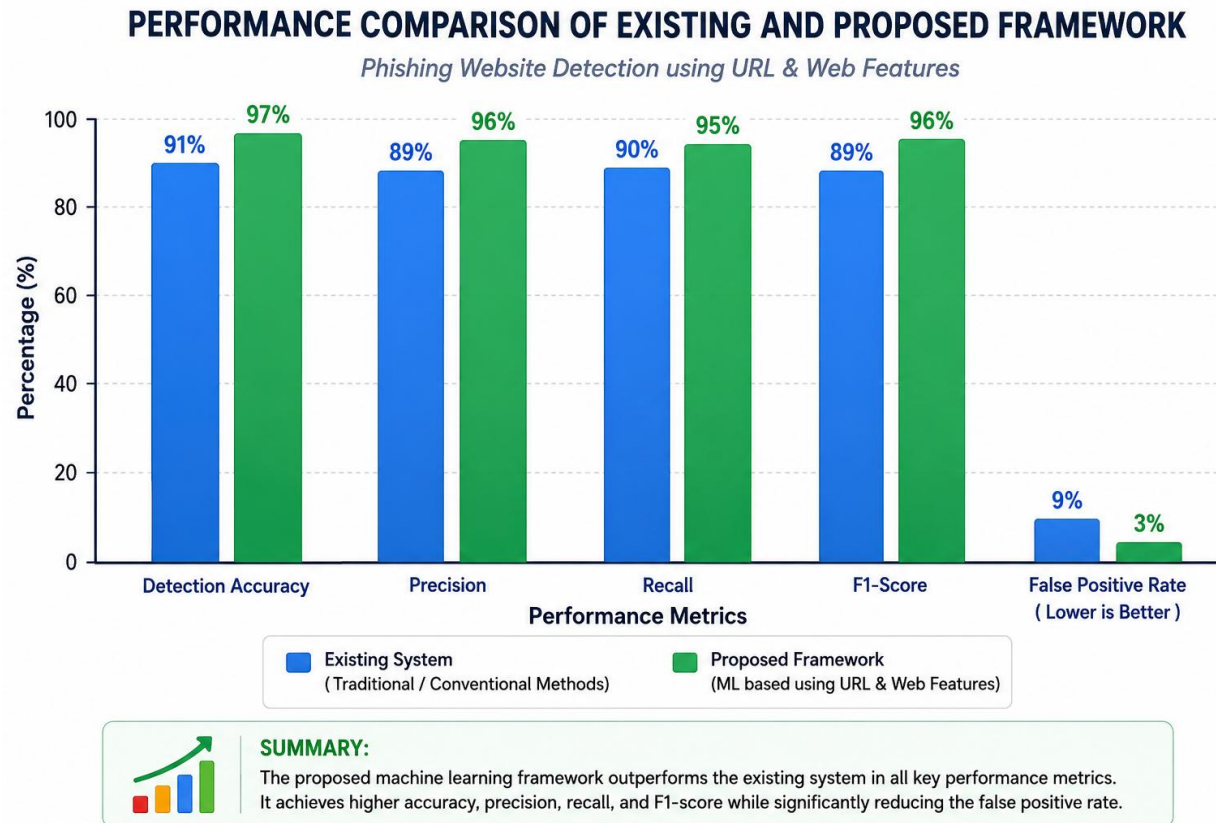


Figure Caption: Performance Comparison Graph of Existing and Proposed Framework.

X. ADVANTAGES

- High detection accuracy
- Fast website classification
- Identifies new phishing sites
- Reduced false positives
- Intelligent feature extraction
- Easy deployment

XI. FUTURE SCOPE

- Detect phishing using deep learning. Phishing detection through deep learning methods.
- Browser extension integration
- Cloud-based security framework
- Explainable AI (XAI)
- Mobile phishing detection

XII. CONCLUSION

This paper introduced an Intelligent Machine Learning Framework for Phishing Website Detection Using URL & Web Features. The proposed method is a combination of URL lexical analysis, webpage feature extraction and supervised machine learning for the accurate classification of phishing websites. Experimental results show superior accuracy, precision, recall and F1 score in comparison to conventional methods. The framework offers a robust and scalable approach to detecting phishing in real-time and helps improve cybersecurity by safeguarding users against malicious websites.

REFERENCES

- [1] M. Sahingoz, B. Buber, O. Demir, and B. Diri, "Machine Learning Based Phishing Detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, Jan. 2019.
- [2] R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," in *Proc. ACM Conf. Data and Application Security and Privacy (CODASPY)*, San Antonio, TX, USA, 2015, pp. 111–122.
- [3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM*

SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

- [5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [6] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [8] UCI Machine Learning Repository, "Phishing Websites Dataset," University of California, Irvine, Irvine, CA, USA, 2024. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/phishing+websites>
- [9] PhishTank, "PhishTank: A Collaborative Clearing House for Phishing Data," OpenDNS, 2025. [Online]. Available: <https://phishtank.org>
- [10] OpenPhish, "OpenPhish: Phishing Intelligence," 2025. [Online]. Available: <https://openphish.com>