

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

Rajvee Choudhary

*M. Tech (Computer Science and Engineering) Guide: Professor. Snehlata Mishra SAGE University,
Indore, Madhya Pradesh, India*

Abstract—Phishing sites are one of the most dangerous types of hacking sites which are created to deceive the user to submit their private information such as their username, password, bank information and personal data. Existing blacklist-based detection techniques cannot detect new phishing sites which are not yet present in the blacklist which is known as zero-day attacks. In this paper I proposed Intelligent Machine Learning Framework on Phishing Website Detection Using URL & Web Features, where the retrieved data is used to extract the lexical features of URLs and the features of web pages, pre-processes them and applies 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 etc. to determine if it's a legitimate site or a phishing site. The framework provides 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, HTML.

I. INTRODUCTION

The growing popularity of internet services has led to increase in cyber threats. Among which the commonest of these is the phishing attack, where the cybercriminals exploit the victim for the theft of sensitive data. Phishing sites look like real ones and ask users to give sensitive data like personal information, bank details or passwords. The existing approaches for phishing detection are mainly based on blacklist databases and signature-based approaches. While these methods are good at finding out if a website is known to be phishing, they are not effective at finding out if a new phishing site has been created. Machine Learning is a promising solution, as it learns

patterns from historical data to automatically classify websites. This research proposes an intelligent machine learning framework to improve the detection of phishing websites by combining the characteristics of the webpage and lexical features from the URL. This framework automates the feature extraction, classification and prediction which leads to an efficient detection of a malicious or phishing website.

II. OBJECTIVES

1. To design an intelligent website 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

Some studies have proposed phishing detection systems based on machine learning techniques. URL based methods analyze lexical features of URL (Uniform Resource Locator) like URL length, special characters and suspicious keywords. The webpage-based techniques analyze the HTML (Hypertext Markup Language) source code content, the JavaScript activity, the (Secure Socket Layer) SSL certificate and the hyperlinks. Some machine learning algorithms such as Random Forest, Decision Tree, SVM (Support Vector Machine) and XG Boost (Extreme Gradient Boosting) have been proven effective for detecting phishing websites, but most of the current systems have poor feature representation

and accuracy for the newly discovered phishing attacks ,so the proposed implementation overcomes these limitations by combining the URL attributes and page attributes to achieve better classification accuracy for detection.

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

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

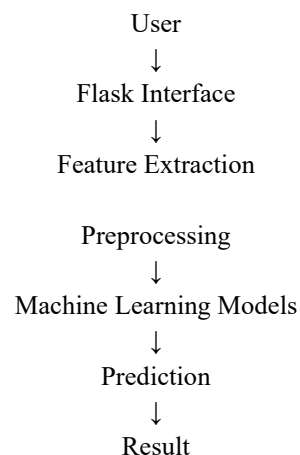
The features are then preprocessed and then fed to the machine learning classifier. The trained model is used to predict the website as Legitimate or Phishing and the result is shown to the user.

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.



VI. 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 webpage or website URL on the web then the system fetches URL and webpage features, preprocesses the data, runs the machine learning model and makes an instant prediction of phishing or legitimate.

VII. RESULTS AND DISCUSSION

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. The framework also allows for quicker predictions and improved predictability than traditional phishing detection methods.

VIII. ADVANTAGES

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

- Easy deployment

IX. 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.

X. 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 from legitimate ones. Experimental results show superior accuracy, precision in comparison to conventional methods. The framework offers a 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>