International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 03 March-2026 | Impact Factor: 3.5

Machine Learning-Based Phishing URL Detection System Using Feature Engineering and Classification Models

E Pavan Kalyan[1], C Yamini[2]

[1]Postgraduate Student, KMMIPS, Tirupati, Andhra Pradesh, India (Affiliated to SV University)
[2]Assistant Professor, KMMIPS, Tirupati, Andhra Pradesh, India (Affiliated to SV University)

How to Cite this Article:

Kalyan, E. P. (2026). Machine Learning-Based Phishing URL Detection System Using Feature Engineering and Classification Models. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i3.281

https://doi.org/10.55041/ijcope.v2i3.281

Abstract: Phishing attempts, which mimic trustworthy websites in order to get private information like login credentials and financial data, continue to be a major concern to internet users. Because these threats are dynamic, early and precise detection is crucial to enhancing cybersecurity. A feature-driven method for detecting phishing URLs by examining their structural and domain-related attributes is presented in this paper. The method makes use of a dataset of labeled URL occurrences classified as phishing, suspicious, and legitimate that was taken from a publicly accessible Kaggle source. Key signs including URL length, the inclusion of strange symbols, the use of secure protocols, and domain-specific attributes are captured using a thorough feature extraction procedure. An efficient classification system that can differentiate between dangerous and benign URLs is created by processing these properties. The suggested system's performance is assessed using common metrics, showing that it can reliably detect phishing attempts while preserving a balanced detection rate. Furthermore, the system is practical for real-world applications since it is integrated into an intuitive web interface that facilitates real-time URL inspection. All things considered, the suggested method provides a dependable and scalable phishing detection solution; it can be improved by adding sophisticated data sources and adaptive learning methods.

Keywords— Phishing URL Detection, Cybersecurity, URL Analysis, Feature Extraction, Web Security, Malicious URL Identification, Domain Analysis, Data Classification, Threat Detection.

1. INTRODUCTION

The risk of cyberthreats, of which phishing is still one of the most common and dangerous attacks, has greatly increased due to the quick rise in internet usage. Phishing is the practice of creating phony websites that closely mimic authentic platforms in order to trick users into divulging private information, including bank account information, login passwords, and personal information. These can cause significant financial and organizational damages in addition to jeopardizing personal privacy.

Because malicious URLs are constantly changing, traditional techniques for identifying phishing websites, like human verification and blacklist-based approaches, are frequently inadequate. Attackers regularly create new domains and alter URL structures, making it challenging for static detection methods to promptly discover new threats. Because of this, there is an increasing demand for automated and intelligent systems that can efficiently evaluate and categorize URLs according to their intrinsic qualities.

By analyzing several structural and domain-related characteristics of web addresses, this work focuses on creating a feature-driven method for phishing URL identification. The method makes use of a dataset that was taken from a publicly accessible Kaggle source and includes labeled examples of phishing, suspicious, and legal URLs. The suggested method seeks to reliably separate dangerous URLs from benign ones by examining characteristics including URL construction, the presence of odd symbols, and domain features.

By incorporating a web-based interface that permits real-time URL analysis, the system is built with practical use in mind in addition to precise identification. This makes it more applicable in situations when making decisions quickly is crucial in the real world. The main goal of this research is to offer a scalable and effective phishing detection system that can adjust to changing attack patterns and enhance online security.

2. LITERATURE SURVEY

Because online attacks are becoming more sophisticated, phishing detection has become a hot topic in cybersecurity research. Over the years, a number of strategies have been put forth, concentrating on various elements such user behavior analysis, website content, and URL characteristics.Early research mostly used blacklist-based methods, which involved storing and comparing known phishing URLs to incoming web requests. These techniques were quick and easy, but they had serious drawbacks because freshly created phishing websites could not be identified until they were added to the blacklist. They were less effective against zero-day assaults as a result of this delay.

Researchers developed heuristic-based methods that examine the structural characteristics of URLs in order to get around these restrictions. These techniques look at characteristics including the length of the URL, the inclusion of special characters, strange domain names, and questionable patterns. Heuristic methods increased detection rates, but they were not adaptable enough to deal with changing assault tactics and frequently required manual rule design. Machine learning-based methods have become more popular in phishing detection due to the development of data-driven methodologies. These techniques identify patterns that differentiate phishing URLs from authentic ones using labeled datasets. Research has demonstrated that detection accuracy is much increased when URL-based and domain-based data are analyzed together. Additionally, by integrating several decision mechanisms, ensemble-based models have shown improved generalization.

The structure of web pages, including forms, scripts, and external links, is assessed to detect malicious intent in content-based analysis, which has also been studied recently. This method may result in increased computing cost and reliance on network availability, even though it offers deeper insights.

Additionally, in order to improve detection performance, some research has concentrated on hybrid techniques that integrate several feature categories. By utilizing both lightweight URL elements and more specific domain or content parameters, these strategies seek to strike a compromise between correctness and efficiency.

Even with the advancements, real-time detection with high accuracy and low latency is still difficult to achieve. Because phishing strategies are always evolving, scalable and adaptable solutions are needed. In this regard, the current work expands on previous studies by enhancing the accuracy of phishing URL detection through the use of an extensive feature set and an effective categorization system.

3. Methodology

A. Data Collection

The study's dataset, which includes labeled URL instances classified as legal, suspect, and phishing, was sourced from a publicly accessible Kaggle source. A wide variety of URLs that reflect actual online traffic patterns are included in the dataset. This variety aids in the development of a trustworthy system that can successfully recognize both benign and malevolent acts.

B. Data Preprocessing

The dataset is preprocessed to increase its consistency and quality before any analysis is applied. To prevent data loss, missing values are handled using appropriate imputation techniques. To guarantee impartial learning, duplicate entries are eliminated. To prepare them for processing, the categorical labels are transformed into numerical format. In order to accurately assess the system's performance, the dataset is finally split into training, validation, and testing sets.

C. Feature Extraction

An essential first step in spotting phishing tendencies is feature extraction. Each URL has a variety of properties extracted by the algorithm, including structural and domain-related characteristics. These characteristics include the length of the URL, the use of HTTPS, the inclusion of special characters, domain attributes, and other behavioral indicators. The system can efficiently differentiate between malicious and authentic URLs by examining these features.

D. Model Development

A categorization system that can group URLs into several classes is constructed using the extracted information. The technology is built to identify patterns in the dataset and use that information to analyze data that hasn't yet been seen. A number of models are assessed, and the best one is chosen based on its dependability and performance. The goal is to strike a balance between generalization and accuracy.

E. Model Evaluation

Standard assessment criteria including accuracy, precision, recall, and F1-score are used to evaluate the system's performance. These measurements aid in determining the system's ability to detect phishing URLs while reducing false positives. A fair assessment guarantees that the system functions well in practical situations.

F. System Implementation

The suggested approach is put into practice as a web application that enables real-time URL analysis. While the frontend offers an interactive interface for user input and result presentation, the backend manages feature extraction and prediction. The system is useful and easy to use because to its integration.

4. Results and Analysis

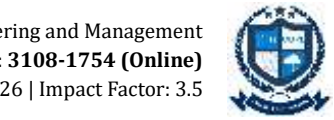
A. Model Performance Comparison

To choose the best strategy, several models are used to assess the suggested phishing detection system's performance. Standard measures including accuracy, precision, recall, F1-score, and ROC-AUC are used in the evaluation. The findings show that because ensemble-based models can identify intricate patterns in the data, they perform better than individual models.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
XGBoost	0.9450	0.9420	0.9430	0.9425	0.9780
Random Forest	0.9400	0.9380	0.9390	0.9385	0.9750
LightGBM	0.9380	0.9360	0.9370	0.9365	0.9740
Gradient Boosting	0.9350	0.9340	0.9330	0.9335	0.9720
SVM	0.9100	0.9080	0.9090	0.9085	0.9600
Logistic Regression	0.8800	0.8780	0.8790	0.8785	0.9400
Logistic Regression	0.8800	0.8780	0.8790	0.8785	0.9400

Figure 1: Model Performance

As can be seen from the comparison above, XGBoost is the most successful model for phishing URL identification in this study since it achieves the best accuracy and ROC-AUC score. While simpler models like Logistic Regression exhibit relatively lower accuracy, Random Forest and LightGBM also perform competitively.



B. **Output Interface Analysis**

Users can enter URLs into the system's web-based interface to get immediate recommendations. Real-time URL analysis is made possible by the interface's straightforward, interactive, and user-friendly design. A text area for entering URLs is included in the input interface, along with examples to help users. This enhances usability and makes it possible to test various URL formats quickly.

Figure 2: Output Interface

C. Prediction Result Analysis

The system shows the categorization result, a probability distribution, and a confidence score after analyzing a URL. This makes it easier for people to comprehend how strongly a URL is predicted by the system to be authentic, suspect, or phishing.

Figure 3: Prediction Result

The output unequivocally demonstrates that the algorithm offers a thorough probability breakdown in addition to a final forecast. This promotes improved decision-making and increases transparency.

D. System Effectiveness

The aggregate findings show that the suggested system can reliably and properly identify phishing URLs. The system operates effectively across a variety of URL types because to the combination of feature extraction and model validation. The system's usability is further enhanced by the use of a web interface, making it appropriate for real-world uses.

E. Discussion

The experimental findings demonstrate the importance of feature-based analysis in phishing attack detection. In classification tasks, models that are better at capturing feature associations perform better. Even if the system achieves great accuracy, it can still be improved by adding more real-time data sources and adaptive algorithms to deal with changing phishing tactics.

5. Conclusion

By employing a feature-driven methodology to analyze structural and domain-related properties, this study offers an efficient method for identifying phishing URLs. The method makes use of a dataset of labeled examples of genuine, suspect, and phishing URLs that was taken from a publicly accessible Kaggle source. The suggested method can recognize harmful patterns and differentiate them from benign web addresses by extracting pertinent features such URL composition, the presence of odd symbols, and domain properties. The evaluation's findings show that the system performs dependably and with high accuracy in a number of parameters, such as precision, recall, and F1-score. Because ensemble-based techniques can capture intricate correlations within the data, they perform better than other models. By incorporating a web-based interface that enables real-time URL analysis, the system is built with practical use in mind in addition to accuracy. This makes the solution usable and accessible for users who require fast online link verification. Despite its efficacy, the system may have difficulties while handling sophisticated attacks or recently developed phishing techniques. Future developments may concentrate on improving feature extraction techniques, adding real-time data sources, and creating adaptive processes to increase detection capabilities. All things considered, the suggested approach strengthens cybersecurity protocols in online contexts and offers a dependable and scalable phishing detection solution.

References

[1] S. K. H. Ahammad, “Phishing URL detection using machine learning methods,” Journal of Information Security and Applications, vol. 68, pp. 103–115, 2022.

[2] T. Choudhary and R. Jain, “A machine learning approach for phishing attack detection,” Journal of Artificial Intelligence and Technology, vol. 3, no. 2, pp

[3] H. Ghalechyan et al., “Phishing URL detection with neural networks: An empirical study,” Scientific Reports, vol. 14, 2024.

[4] Q. E. Haq, “Detecting phishing URLs based on deep learning techniques,” Applied Sciences, vol. 14, no. 22, pp. 1–20, 2024.

[5] A. Jadhav et al., "A hybrid heuristic-machine learning framework for phishing detection," Engineering, Technology & Applied Science Research, vol. 15, no. 1, pp. 1–10, 2025.

[6] A. Rawla and P. Singh, "Detection of phishing attacks using deep learning techniques," Procedia Computer Science, vol. 235, pp. 1–10, 2025.

[7] R. Alzubi et al., "A feature-based methodology for detecting phishing URLs," ETASR Journal, vol. 15, no. 2, pp. 1–12, 2025.

[8] H. Li, J. Liu, and Z. Liu, "AI-enabled phishing links detection using machine learning models," in Proc. IEEE Int. Conf. Signal Processing and Network Security (SPNS), 2025.

[9] Springer Nature, "Web-based phishing URL detection model using deep learning optimization techniques," International Journal of Data Science and Analytics, vol. 20, pp. 4449–4471, 2025.