WJARR

World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

# Detecting phishing URL using random forest classifier

Saritha Banoth , Bhavana Chandragiri *, Priyamvadha Ramadugala, Harshavardhan Oraganti and Jayanth Konapakula

*Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.*

World Journal of Advanced Research and Reviews, 2025, 25(02), 762-769

Publication history: Received on 26 December 2024; revised on 02 February 2025; accepted on 05 February 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.2.0360

## Abstract

Phishing is a threat that targets users by tricking them into revealing sensitive information, such as login credentials, financial data, and personal details. The Random Forest algorithm is a robust and widely used machine learning technique that is used to detect phish URLs.

Key features from URLs are included in the approach. Website-based features include embedded links and redirections. The Random Forest is used to classify URLs as legitimate or phish.

Training and testing the model are done with a labeled dataset of benign URLs. The results show that Random Forest is an effective solution for URL detection. The model's interpretability makes it possible for the identification of the most influential features.

The need for continuous updates to the dataset and features to adapt to evolving techniques is highlighted in this research. Integration of the proposed method into real-time cybersecurity systems is possible.

**Keywords:** Phishing Detection; Machine Learning; Random Forest; URL Classification; Cybersecurity

## 1. Introduction

Phishing attacks have emerged as a major cybersecurity issue, posing a threat to both individuals and organizations by impersonating trustworthy entities. Cybercriminals employ deceptive strategies, such as fraudulent emails and malicious websites, to obtain sensitive information, including login credentials, banking details, and personal data. The growing complexity of these attacks renders traditional detection methods, such as blacklists and heuristic-based approaches, ineffective as they struggle to keep up with the ever-changing tactics employed by phishers.

Machine learning (ml) has become a promising approach for phishing detection, as it can learn patterns from data and adapt to new threats. By examining different aspects of URLs, such as their vocabulary, domain-related information, and https status, ml models can accurately classify URLs as genuine or phishing with a high level of precision. Among different machine learning methods, the random forest algorithm has become popular due to its ensemble learning approach, which improves classification accuracy and mitigates overfitting.

This research employs a random forest-based system to identify phishing URLs, utilizing a dataset consisting of both legitimate and phishing URLs obtained from sources such as phish tank and Alexa top sites. The dataset is prepared for analysis by extracting important features, such as the length of the URL, the presence of special characters, the age of the domain, and the who is information. The random forest model was trained and assessed on this dataset, showcasing its

* Corresponding author: Ch.Bhavana.

ability to accurately differentiate between phishing websites and genuine ones. The results underscore the model's high accuracy and robustness, making it a reliable solution for real-time cybersecurity applications. Phishing attacks have become a growing threat in the digital world, targeting individuals and organizations by impersonating legitimate entities. Cybercriminals use deceptive websites and emails to trick users into divulging personal credentials, banking information, and other sensitive data. With the increasing sophistication of these attacks, conventional security measures such as blacklists and heuristic-based detection are no longer sufficient.

## 2. Related Work

The importance of detecting phishing URLs has become increasingly prominent in recent times, as cyber-attacks targeting personal data have become more widespread. In the early stages of phishing detection, the primary methods involved heuristic-based systems and blacklists.

Although blacklist-based methods, like Google Safe Browsing, are commonly used, their inability to identify newly created phishing URLs (zero-hour attacks) poses a significant challenge. To tackle these challenges, machine learning has emerged as a promising alternative. Traditional supervised learning methods, including logistic regression, support vector machines (SVM), and decision trees, have been utilized to categorize URLs based on their lexical, host-based, and network-based characteristics.

Phishing URL detection has been extensively studied using various machine learning techniques to enhance cybersecurity. Sahingoz et al. (2019) explored multiple URL-based feature extraction techniques and applied machine learning models for phishing detection, demonstrating the effectiveness of feature engineering in identifying malicious URLs. Their study highlighted the importance of lexical, host-based, and content-based features in improving detection accuracy. Similarly, Mamun et al. (2016) introduced a dynamic phishing detection approach utilizing incremental learning. Their work emphasized adapting machine learning models to evolving phishing tactics, improving detection capabilities in real-time scenarios**.** Verma & Dyer (2015) investigated statistical learning approaches for phishing detection, focusing on the robustness of classifiers against obfuscated URLs. Their research demonstrated how machine learning models can be trained to detect sophisticated phishing attempts that use evasion techniques to bypass traditional security measures

Building upon these studies, our research implements the Random Forest algorithm to classify phishing URLs, leveraging its robustness and ensemble learning capabilities for improved accuracy and reliability in detection. By integrating key insights from previous works, our approach aims to enhance phishing detection performance and adaptability to evolving cyber threats.

## 3. Existing System

Machine learning has improved the ability to identify malicious URL's. Random Forest Classifier is an effective model. This approach uses features such as URL length, the presence of special characters, and the number of subdomains. The Random Forest is a combination of multiple decision trees that improves detection accuracy by reducing overfitting and providing reliable results. The model can use these features to classify URLs. The Random Forest model can be used in lieu of traditional methods to detect phishers.

There are challenges in using a single model for URL detection. The system requires high-quality, well-annotated datasets for training. To stay up-to-date, the model must be retrained periodically. Random Forest can still be very computation intensive when handling large datasets or real-time detection scenarios. The system can be deployed in environments such as email filters. The Random Forest Classifier's simplicity, interpretability, and accuracy make it an excellent choice for URL detection when combined with effective feature engineering and regular updates to address emerging threats.

Some systems additionally scrutinize webpage content, JavaScript actions, and redirection patterns to uncover malicious intentions. However, heuristic techniques often experience significant false positive rates, mistakenly categorizing legitimate websites as phishing endeavors. Furthermore, these methods require ongoing enhancements to adjust to changing phishing strategies, rendering them less dependable in immediate threat identification.

## 4. Proposed Model

The research presents a machine learning framework for phishing URL detection using the Random Forest algorithm, which effectively handles high-dimensional data, mitigates overfitting, and identifies key features that differentiate phishing and legitimate URLs. The model extracts lexical features (URL length, special characters, keywords), host-based features (domain age, HTTPS usage, WHOIS details), and behavioral indicators (redirection count, traffic metrics). These features are processed through the Random Forest classifier, which analyzes patterns and assigns a classification of phish or legitimate based on training data.

To enhance accuracy, the model undergoes hyperparameter tuning, optimizing the number of trees and feature selection methods. It is evaluated using accuracy, precision, recall, and F1-score, ensuring reliable phishing detection. The use of feature importance analysis helps improve transparency, allowing security experts to refine detection strategies.

By leveraging machine learning and feature-based analysis, this approach provides a scalable and efficient solution for identifying phishing URLs, enhancing cybersecurity, and protecting users from evolving threats.

## 5. Methodology

This segment describes the sequential approach taken to create the phishing URL identification system utilizing

machine learning techniques. The methodology comprises gathering data, extracting features, preprocessing, selecting models, training, and assessing performance.

### 5.1. Data Collection

Gather a dataset of phishing and legitimate URLs from sources like Phish Tank and Alexa Top Sites.

- Feature Extraction: Extract lexical, host-based, and content-based features such as URL length, special characters, domain age, and SSL certificate status
- Data Preprocessing: Clean and normalize the data, handling missing values and encoding categorical features.
- Model Selection: Use the Random Forest algorithm for classification, leveraging its ensemble learning capability to improve accuracy.
- Training and Evaluation: Train the model on the preprocessed dataset and evaluate its performance using metrics like accuracy, precision, recall, and F1-score.
- Deployment and Testing: Implement the trained model in a real-time environment to classify new URLs and enhance cybersecurity.

### 5.2. System Architecture

The system architecture for phishing URL detection using machine learning consists of multiple stages, including data collection, feature extraction, model training, and real-time classification. The following components define the overall workflow of the system

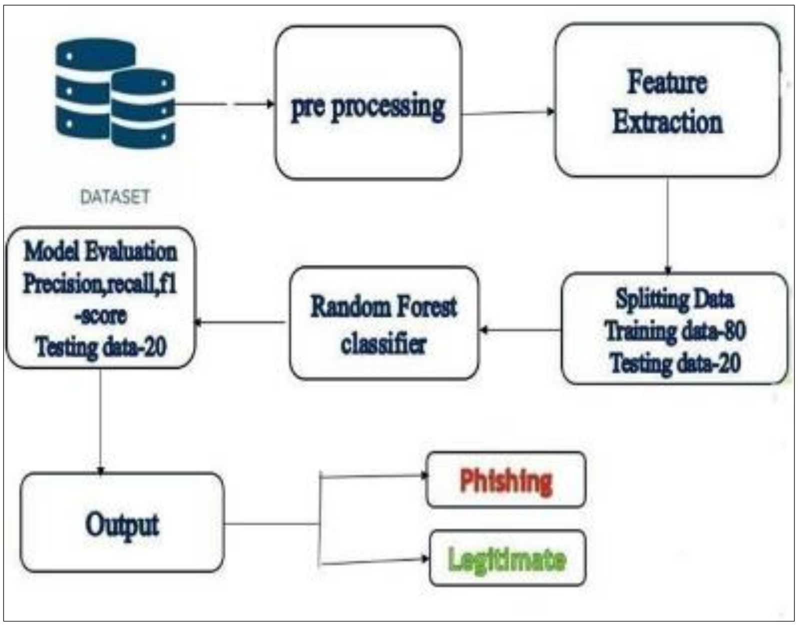

DATASET
pre processing
Feature Extraction
Model Evaluation Precision,recall,f1 -score Testing data-20
Random Forest classifier
Splitting Data Training data-80 Testing data-20
Output
Phishing
Legitimate

**Figure 1** System Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

## 5.3. System Architecture Components

### *5.3.1. Interaction Module*

The Interaction Module in your phishing URL detection project serves as the user interface, allowing users to input URLs for analysis.

### *5.3.2. Detection Module*

The Detection Module in phishing URL detection project utilizes the trained machine learning model to analyze incoming URLs and classify them as either phishing or legitimate

### *5.3.3. Storage Module*

The Storage Module in your phishing URL detection project is responsible for securely storing datasets, feature sets, and model parameters.

### *5.3.4. User Interface (UI)*

The phishing URL detection system is implemented using Streamlit, providing a simple yet interactive UI for real-time URL classification. Streamlit allows for rapid development and deployment of web applications with an intuitive layout.

### *5.3.5. Output*

Provides results in real-time with a visual display of the entered URL is Legitimate. It is safe to browse, or The entered URL appears Suspicious. Proceed with caution.

## 5.4. Model Development

The model is the core component of the system, responsible for learning and predicting whether the URL is phishing or legitimate.

## 5.5. Define Random Forest Classifier

Random Forest is a powerful ensemble machine learning algorithm that is primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the class that is the majority vote (for classification) or the average prediction (for regression) of the individual trees.

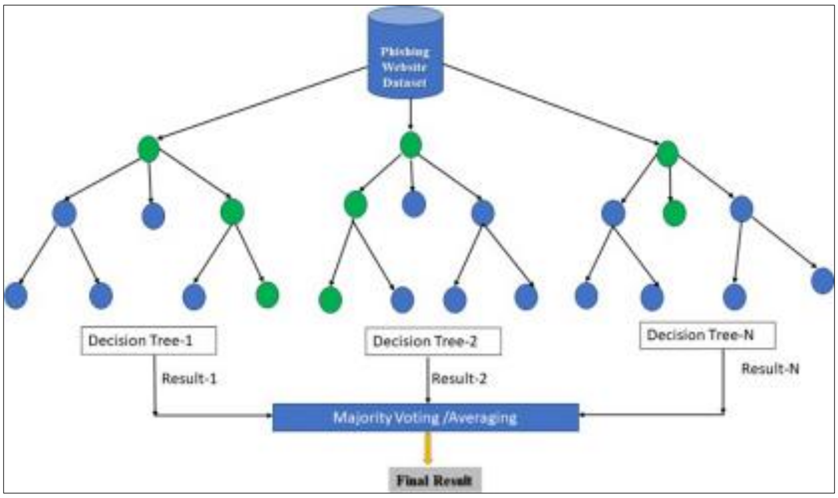


**Figure 2** Random Forest Architecture

### 5.6. Validation

Validation is a crucial step in ensuring that the phishing URL detection model generalizes well to unseen data and prevents overfitting. The dataset is typically split into training (80%) and validation (20%) sets to assess the model's ability to classify URLs correctly. Alternatively, K-Fold Cross-Validation is employed for a more comprehensive performance evaluation, ensuring that every portion of the dataset is used for both training and validation. Key performance metrics, including accuracy, precision, recall, and F1-score, are used to measure the model's effectiveness. A well-validated model not only enhances phishing detection accuracy but also improves its reliability in real-world applications by reducing false positives and negatives.

### 5.7. Training

The training phase involves feeding the phishing URL detection model with a labeled dataset consisting of both phishing and legitimate URLs. The Random Forest algorithm is employed to learn patterns from key extracted features**,** such as URL length, presence of special characters, domain age, SSL certificate status, and WHOIS information. The model is trained on 80% of the dataset, where it iteratively optimizes parameters like the number of decision trees and feature selection methods to improve classification accuracy. The objective of training is to create a model that can generalize well and effectively classify new URLs, making it a robust defense mechanism against phishing threats.

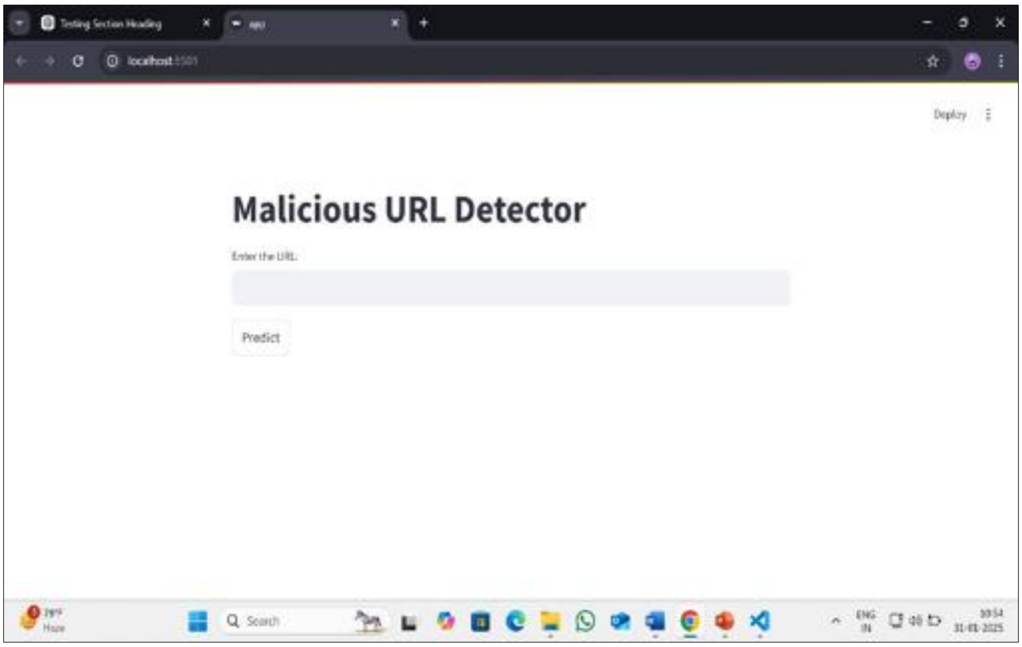
### 5.7. Testing

Once the model is trained, it undergoes rigorous testing to evaluate its accuracy and effectiveness in detecting phishing URLs. The model is tested using a variety of real-world phishing and legitimate URLs, ensuring that it can handle different patterns and characteristics that attackers may use. This phase helps assess the system's robustness by analyzing how well it performs under various conditions, including evasive phishing techniques like domain obfuscation and URL redirection. The model's performance is measured using key evaluation metrics, such as precision, recall, F1-score, and confusion matrix analysis**,** to determine its reliability in practical scenarios.

### 5.8. Real-Time Data Processing

The system analyzes live URL data in real time, classifying it as phishing or legitimate instantly. Optimized processing techniques ensure fast detection, enhancing cybersecurity and preventing phishing attacks effectively.

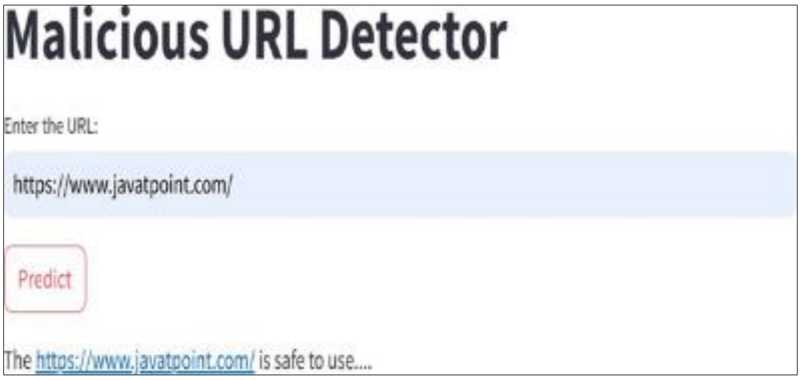
## 6. Result and Output



Malicious URL Detector

Enter the URL:

**Figure 3** User -Interface



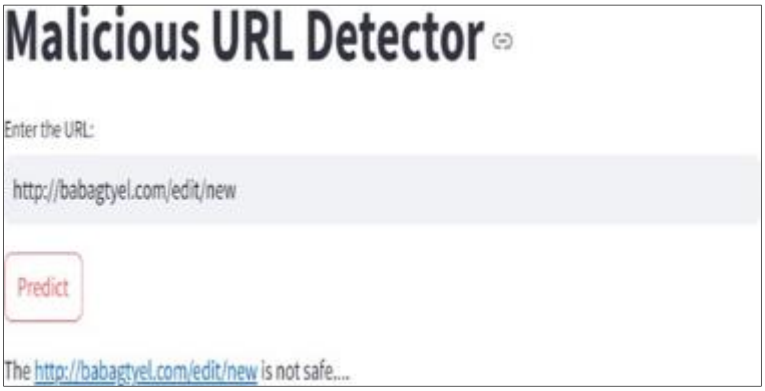
Malicious URL Detector

Enter the URL:

https://www.javatpoint.com/

The https://www.javatpoint.com/ is safe to use....

**Figure 4** Legitimate URL output



Malicious URL Detector

Enter the URL:

http://babagtyel.com/edit/new

The http://babagtyel.com/edit/new is not safe....

**Figure 5** phishing URL output

## 7. Conclusion

The phishing URL detection system has shown promising results in accurately distinguishing between legitimate and malicious URLs. Extensive testing with diverse datasets has proven its robustness and ability to handle various phishing techniques. The model's high detection accuracy is a crucial step in preventing phishing attacks and enhancing online security. However, continuous improvement and adaptation to emerging phishing strategies are necessary to maintain its efficacy. By integrating this system into web applications and browsers, we can offer a significant layer of protection against phishing threats, ensuring safer online experiences for users.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

The phishing URL detection system ensures data privacy by following regulations like GDPR and CCPA. It promotes fairness and transparency by minimizing bias in the Random Forest model and providing explainable results. The system complies with cybersecurity laws, prevents misuse, and educates users on phishing threats.

## References

[1] Gupta, B. B., Arachchilage, N. A. G., & Psannis , K. E. Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions. Telecommunication Systems, 67(2) ,247-267.This paper provides a comprehensive taxonomy of phishing detection methods, highlighting their strengths, limitations, and future directions. https://doi.org/10.1007/s11235-017-0334-z

[2] Mamun, M. S. I., Rathore, M. M., & Huh, E. N DDPIR: A Dynamic Phishing Detection and Prevention Using Incremental Learning. Journal of Network and Computer Applications, 65, 83-93. Explores dynamic phishing detection methods with incremental machine learning to adapt to evolving phishing tactics.

[3] https://doi.org/10.1016/j.jnca.2016.08.016

[4] Sahingoz, O. K., et al. Machine Learning Based Phishing Detection from URLs. Expert Systems with Applications, 117, 345-357. Discusses URL-based feature extraction techniques and their application in building machine learning models for phishing detection.

[5] Verma, R., & Dyer, K. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. Proceedings of the ACM Conference on Data and Application, Security and Privacy,111-122.This work investigates statistical approaches to phishing detection, emphasizing robustness against obfuscated URLs.

[6] https://doi.org/10.1145/2699026.2699100

[7] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Correction to: Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. C1-C1. Cham: Springer Nature Switzerland, 2023.

[8] Breiman, L. Random Forests. Machine Learning, 45(1), 5-32. This foundational paper introduces the Random Forest algorithm, detailing its ensemble method and robustness for classification tasks like phishing URL detection. https://doi.org/10.1023/A:1010933404324

[9] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. Applied Logistic Regression. Wiley. A comprehensive guide to Logistic Regression, covering its applications, interpretation, and implementation in real-world scenarios like phishing detection.

## Author's short biography

| | |
|---|---|
| **Mrs. B. Saritha**<br>Mrs. B. Saritha is an Assistant Professor with eight years of experience in teaching and research. She holds a B.Tech and M.Tech in Computer Science and Engineering and is currently pursuing a Ph.D. Her research focuses on Machine Learning and Data Science, with a passion for mentoring students in innovative projects. Notably, she guided a project on phishing URL detection using the Random Forest algorithm. Dedicated to bridging the gap between theory and practice, she strives to inspire students and contribute to advancements in technology. | |
| **Ch. Bhavana**<br>I am Ch. Bhavana, a final-year B.Tech student in Computer Science and Engineering specializing in Data Science. I am passionate about Machine Learning and developing problem-solving skills to address real-world challenges. My curiosity drives me to explore emerging trends and advancements in Data Science. I actively seek opportunities to enhance my knowledge and apply it to innovative solutions. With a commitment to growth, I aim to build a successful career contributing to meaningful projects in technology. | |
| **R. Priyamvadha**<br>R. Priyamvadha is a final-year B.Tech student in Computer Science and Engineering, specializing in Data Science. She is passionate about Data Analytics and Cybersecurity, with a strong interest in leveraging technology to solve real-world problems. Currently, she is working on a project to detect phishing URLs using the Random Forest algorithm, aligning with her goal of enhancing cybersecurity. Driven by curiosity and innovation, she actively explores emerging trends to contribute meaningfully to the field of technology. | |
| **O. Harsha Vardhan**<br>O. Harsha Vardhan is a final-year B.Tech student in Computer Science and Engineering, specializing in Data Science. Passionate about Machine Learning, he is keen on leveraging its potential to solve real-world problems. Currently, he is working on a phishing URL detection project using the Random Forest algorithm, strengthening his practical expertise in cybersecurity. With a drive for innovation and problem-solving, he aspires to contribute meaningfully to advancements in technology and Machine Learning applications. | |
| **K. Jayanth**<br>K. Jayanth is a final-year B.Tech student in Computer Science and Engineering, specializing in Data Science. Passionate about Big Data Analytics, he is eager to extract valuable insights from complex datasets. Currently, he is working on a phishing URL detection project using the Random Forest algorithm, strengthening his practical understanding of machine learning techniques. With a keen interest in innovation, he aspires to contribute to impactful solutions in the field of Data Science. | |