

Applying Random Forest Algorithm for Phishing URL Identification

Afthar Kautsar^{1*}, Maghfira Aida², Anita Yulistia³

^{1,2,3} Universitas Islam Negeri Sumatera Utara, Indonesia

* afthar0701212239@uinsu.ac.id

<https://doi.org/10.56427/jcbd.v4i3.782>

ARTICLE INFO

Article History

Received : July 9, 2025
Revised : August 11, 2025
Accepted : September 26, 2025

Keywords

Phishing
URL Detection
Random Forest
Machine Learning
Cybersecurity

ABSTRACT

Phishing attacks continue to be one of the most pervasive cybersecurity threats, particularly through malicious URLs designed to mimic legitimate websites and steal sensitive user information. To address this challenge, this study employs the Random Forest algorithm for automated phishing URL detection using a publicly available dataset from Kaggle. The dataset contains diverse structural, technical, and popularity-based features that capture behavioral and lexical characteristics of each URL. Following data preprocessing and an 80/20 train-test split, the Random Forest classifier achieved strong predictive performance, attaining an accuracy of 94.94%, a precision of 95.19%, and a recall of 96.94%. The model further demonstrated robust classification capability with an F1-score of 96.06% and an ROC AUC value of 0.985, indicating excellent discrimination between phishing and legitimate URLs. Feature importance analysis shows that factors such as the URL's presence in Google's index, page rank metrics, and specific structural patterns significantly influence prediction outcomes. Additionally, performance visualizations including ROC and Precision-Recall curves reinforce the model's reliability and stability. Overall, the findings suggest that Random Forest provides an effective and efficient solution for phishing URL detection, offering promising potential for integration into real-world cybersecurity systems.



JCBD is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

1. Introduction

Phishing continues to be one of the most prevalent and dangerous forms of cyberattacks. While the tactic is relatively straightforward tricking users into providing sensitive information through fake websites the execution has become increasingly deceptive and sophisticated. As a result, conventional detection methods, such as blacklists or rule-based systems, are no longer sufficient to combat modern phishing attempts that evolve on a daily basis. Phishing attacks, particularly through deceptive URLs mimicking legitimate websites to steal sensitive data such as credentials or financial information, have increasingly become a major concern in cybersecurity. Consequently, automatic phishing URL detection has become vital. Among machine learning algorithms, Random Forest (RF) has drawn considerable attention due to its reliability and performance in classification tasks.

For instance, [1] reported that RF achieved a high accuracy of 98.2% in phishing URL detection. [2] also highlighted RF's ability to identify key features that contribute most significantly to the model's performance. [3] found RF outperformed Support Vector Machine (SVM) in a localized dataset, while [4] noted that RF's performance was nearly on par with LightGBM and more stable than Naive Bayes. [5] added further evidence by confirming RF's effectiveness in detecting phishing on global web pages, whereas [6] introduced a combination of RF with bootstrap sampling and LASSO to improve recall.

Efforts to optimize RF include [7] using grid search and RFECV, which significantly improved detection accuracy. [8] incorporated PCA and information gain to reduce feature complexity without sacrificing accuracy. Hybrid approaches such as CNN+RF [9], the ensemble model PhishGuard [10], and the Bi-LSTM + RF combination [11] demonstrated substantial gains in generalization performance. A novel technique is tackled by [12], who integrated graph-based learning and loopy belief propagation with RF to handle complex URL structures. Locally, [13] validated RF's effectiveness on an Indonesian phishing URL dataset. Furthermore, the PhiUSIIL framework by [14] which employs similarity index and incremental learning with RF, achieved detection accuracies exceeding 99%.

Taken together, these studies make clear that Random Forest remains one of the most robust and effective methods for phishing URL detection, whether used standalone or within ensemble architectures. Therefore, this study aims to implement Random Forest to classify phishing URLs using a public Kaggle dataset and evaluate its performance empirically.

2. Research methodology

The research was conducted through several key stages, including data collection, preprocessing, dataset splitting, random forest model, and model evaluation.

a. Data Collection

The dataset used in this study is `dataset_phishing.csv`, which contains a list of URLs along with a variety of features representing each URL's characteristics. These features include structural attributes (such as URL length, number of dots, and presence of "@" symbol), technical attribute (e.g., HTTPS presence, domain age, DNS record), and popularity related attributes (such as web traffic and page rank). The dataset is labeled into two categories: (1) Phishing, (2) Legitimate. The data is provided in CSV format using a semicolon (";") as a separator.

b. Preprocessing

Before training the model, the dataset must be cleaned to ensure it is suitable for processing. The preprocessing steps include:

- 1) Removing irrelevant columns such as Unnamed and url, which contain only full address strings.
- 2) Converting the status label into numeric format: phishing = 1, legitimate = 0.
- 3) Ensuring all feature data types are numeric, including handling non numeric entries.
- 4) Replacing missing values with the median or zero, depending on the context of each feature.
- 5) Dataset Splitting

The cleaned dataset is split into two parts: 80% for training and 20% for testing. This splitting is done using stratified sampling to maintain label distribution across both subsets.

c. Random Forest Model

The Random Forest algorithm is chosen due to its ability to handle heterogeneous features, robustness against overfitting, and its capability to provide feature importance insights. The modeling process involves the following steps:

- 1) Initializing the Random Forest model.
- 2) Tuning model parameters using GridSearchCV with cross validation.
- 3) Training the model using the preprocessed training data.

d. Model Evaluation

The model's performance is evaluated using the test dataset, based on the following metrics:

- 1) Accuracy
- 2) Precision
- 3) Recall
- 4) F1-Score
- 5) ROC AUC

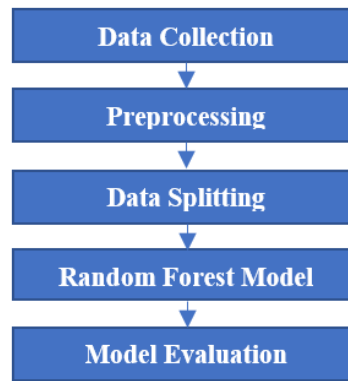


Figure 1. Methodology

3. Results and Discussion

After training and testing the model using the Random Forest algorithm with an 80% training and 20% testing data split, the evaluation results are as follows:

Table 1. Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest	94,94%	95,19%	96,94%	96,06%	0,985

The model demonstrates strong predictive ability. With nearly 95% accuracy, and both precision and recall above 94%, the model is both reliable and balanced.

Table 2. Classification Report

Class	Precision	Recall	F1-score	Support
Legitimate (0)	0,94	0,91	0,93	187
Phishing (1)	0.95	0,97	0,96	327

The model performs especially well in identifying phishing (recall = 0.97). However, some legitimate URLs are incorrectly flagged as phishing (recall = 0.91), indicating a few false positives.

Table 3. Confusion Matrix

	Predicted Legitimate (0)	Predicted Phishing (1)
Legitimate (0)	170	17
Phishing (1)	10	317

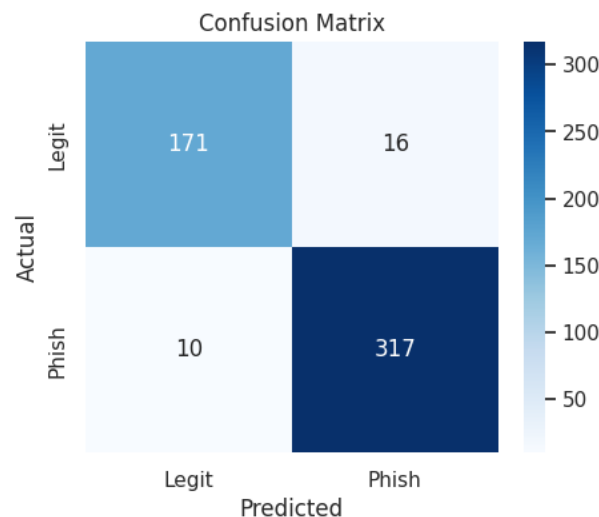


Figure 2. Confusion Matrix

This confusion matrix shows:

- a. 317 phishing URLs were correctly identified.
- b. 17 legitimate URLs were misclassified as phishing (false positives).
- c. Only 10 phishing URLs were missed (false negatives).

The model errs on the side of caution, favoring phishing detection even if it occasionally flags a safe URL.

Table 4. Top 5 Most Important Features

Rank	Feature Name	Contribution	Description
1	google_index	0.184	Phishing pages are rarely indexed by Google.
2	page_rank	0.097	Low-ranked domains tend to be more suspicious.
3	nb_www	0.062	Unusual “www” patterns often appear in deceptive URLs.
4	nb_hyperlinks	0.057	Excessive outbound links may signal untrustworthy content.
5	web_traffic	0.049	Low traffic is common in newly created phishing sites.

Table 4 presents the top five most influential features in the classification process, as identified by the Random Forest algorithm. These features contribute the most to the model’s decision making and are crucial in distinguishing between phishing and legitimate URLs. The google_index feature ranks first with a contribution score of 0.184, suggesting that whether a URL is indexed by Google is a strong indicator phishing URLs are often not indexed. The second feature, page_rank, reflects domain reputation lower ranked sites are more likely to be suspicious. The nb_www feature, which counts occurrences of "www" in the URL, is also important, as phishing links often contain unusual or excessive "www" patterns. Additionally, nb_hyperlinks and web_traffic provide valuable insights about page structure and site popularity, both of which support the model’s classification. By leveraging these features, the model is able to detect phishing patterns with higher precision and confidence.

Model Visualization

To further illustrate the model's performance, three key visualizations were generated the ROC curve, the Precision-Recall curve, and the top 20 feature importance chart.

- a. ROC Curve (Receiver Operating Characteristic)

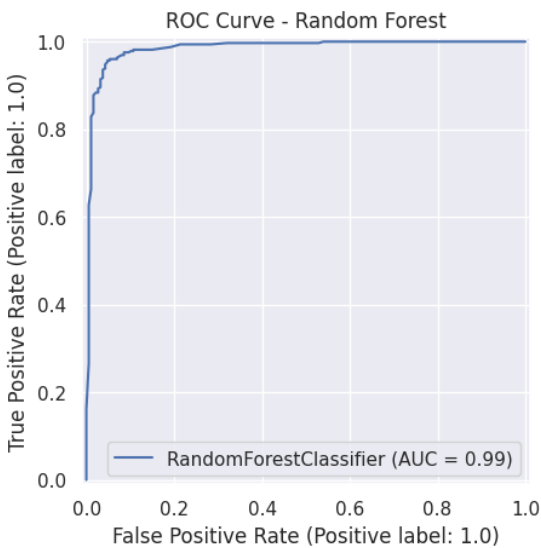


Figure 3. ROC Curve

The ROC curve evaluates how well the model distinguishes between the two classes phishing and legitimate. The curve appears close to the top left corner, indicating excellent performance. This visual aligns with the ROC AUC score of 0.985, reflecting high separability.

b. Precision-Recall Curve

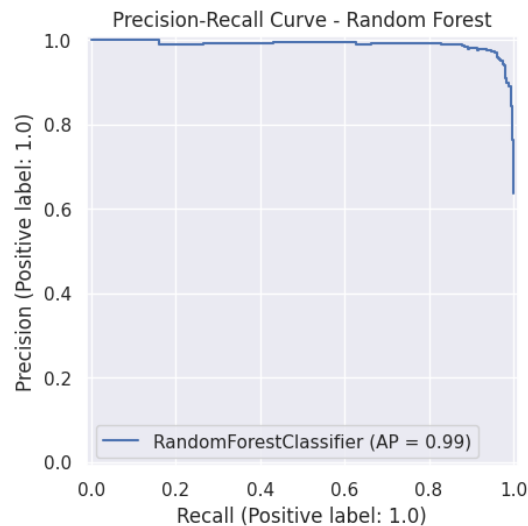


Figure 4. Precision-Recall Curve

The Precision-Recall curve demonstrates how the model balances identifying phishing (recall) and predicting them accurately (precision). The curve maintains a high precision even as recall increases, which is ideal in phishing detection scenarios where false negatives can be dangerous.

c. Top 20 Feature Importance

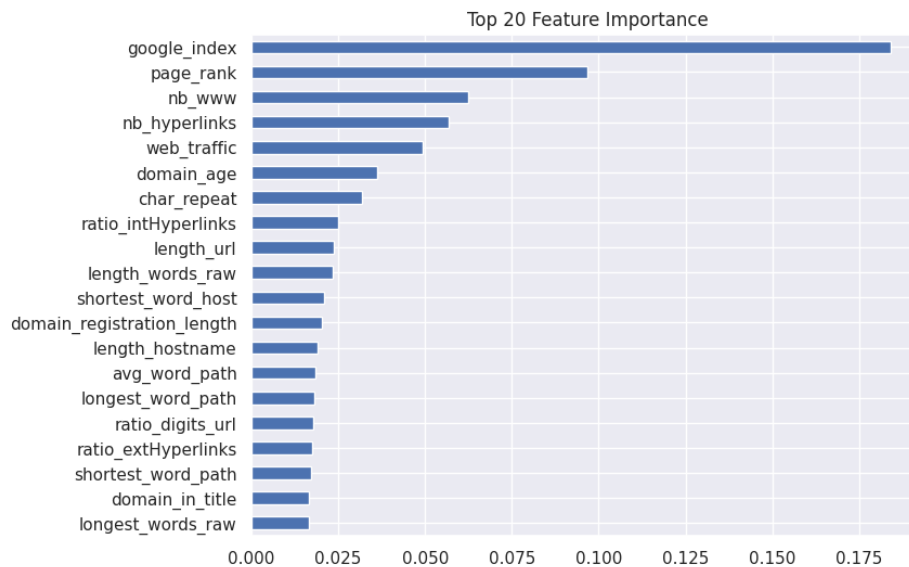


Figure 5. Top 20 Feature Importance

To understand which features most influence the classification, a bar chart displays the top 20 features ranked by their importance. Features like `google_index`, `page_rank`, and `nb_www` are among the highest, indicating that being absent from Google's index or having low page rank are strong signals of phishing behavior.

4. Conclusion

The results of this study clearly demonstrate that the Random Forest algorithm is a highly effective approach for phishing URL detection. With an accuracy of 94.94% and a strong F1-score of 96.06%, the model proves capable of identifying malicious URLs with a high level of reliability. Particularly notable is its high recall value of 96.94%, indicating that nearly all phishing attempts in the dataset were correctly identified. Although a small number of legitimate URLs were misclassified as phishing, the model consistently prioritized minimizing false negatives a favorable trait in cybersecurity applications, where failing to detect phishing can be far more damaging than occasional false alarms. The top contributing features such as `google_index`,

page_rank, and nb_www align well with known phishing characteristics, reaffirming the model's interpretability and trustworthiness.

While the current results are promising, future work may explore the integration of dynamic features (e.g., WHOIS data or time based patterns) to improve detection under evolving attack strategies. Nevertheless, this study confirms that Random Forest stands as a robust and practical solution for phishing URL detection in real world settings.

References (11 pt)

- [1] M. Fahri, "Implementation of the Random Forest Algorithm for Phishing Detection on Websites," *JITSI J. Ilm. Teknol. Sist. Inf.*, vol. 6, no. 2, pp. 186–194, 2025.
- [2] E. I. E. Asker, N. S., & Essa, "An empirical study on feature importance and model performance for phishing website detection using a random forest classifier," *Int. J. Commun. Inf. Technol.*, vol. 5, no. 2, pp. 1–10, 2024.
- [3] D. Tampinongkol, F. F., Kamila, A. R., Wardhana, A. C., Kusuma, A. W. C., & Revaldo, "Implementation of Random Forest Classification and Support Vector Machine Algorithms for Phishing Link Detection," *INISTA J. ISS Softw. Eng.*, vol. 7, no. 1, pp. 127–137, 2024.
- [4] H. Foozy, C. F. M., Anuar, M. A. I., Maslan, A., Adam, H. A. M., & Mahdin, "Phishing URLs Detection Using Naïve Bayes, Random Forest and LightGBM Algorithms," *Int. J. Data Sci.*, vol. 5, no. 1, pp. 56–63, 2024.
- [5] R. Gürfidan, "Intelligent methods in cyber defence: machine learning based phishing attack detection on web pages," *J. Eng. Sci. Des.*, vol. 12, no. 2, pp. 416–429, 2024.
- [6] D. R. I. M. Sarasjati, W., Rustad, S., Purwanto, H. A. S., & Setiadi, "Phishing detection using Random Forest-based weighted bootstrap sampling and LASSO+ feature selection," *IETA/IIETA IJSSE*, vol. Dec 2024, 2024.
- [7] S. Pangemanan, C. V., & Nurafni, "Optimalisasi akurasi deteksi URL phishing dengan hyperparameter -tuning RFECV dan grid search pada algoritma Random Forest," *Data Enthus. J.*, vol. 1, no. 1, pp. 17–24, 2024.
- [8] K.-K. Daniel, M. A., Chong, S.-C., Chong, L.-Y., & Wee, "Optimising phishing detection: A comparative analysis of machine learning methods with feature selection," *J. Informatics Web Eng.*, vol. 4, no. 1, pp. 200–212, 2025.
- [9] X. Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, "Phishing website detection based on deep convolutional neural network and Random Forest ensemble learning," *Sensors*, vol. 21, no. 24, p. 8281, 2021.
- [10] M. A. Islam Ovi, M. S., Rahman, M. H., & Hossain, "PhishGuard: a multi-layered ensemble model for optimal phishing website detection (Random Forest, XGBoost, CatBoost). *arXiv preprint*," *arXiv Prepr.*, 2024.
- [11] S. Baskota, "Phishing URL detection using Bi-LSTM.," *arXiv preprin*, 2025.
- [12] R. Q. Guo, W., Wang, Q., Yue, H., Sun, H., & Hu, "Efficient phishing URL detection using graph-based ML and Loopy Belief Propagation," *arXiv Prepr.*, 2025.
- [13] H. Kresna Kencana, A., Ananda, F. D., Hartanto, A. D., & Hartatik, "Implementasi metode Random Forest klasifikasi untuk phishing link detection," *Intechno J. IT*, vol. 4(2), Dec, 2022.
- [14] S. Prasad, A., & Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," vol. 136, 10354, 2024.