

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

Phishing Website Detection Using Machine Learning

Ms.Nayudupati Meghana¹, Mrs. V. Munni²

¹B. Tech, Artificial Intelligence and Machine Learning, Andhra Loyola Institute of Engineering and Technology (JNTUK), India

²Assistant Professor, Department of CSE (AI&ML), Andhra Loyola Institute of Engineering and Technology (JNTUK), India

Abstract

Phishing attacks represent a major cybersecurity threat because attackers build fraudulent websites to obtain confidential data including passwords and credit card information. The current detection methods fail to meet the needs of emerging phishing attack methods. This research proposes a machine learning method to effectively identify phishing websites. The system uses features extracted from URLs such as length, presence of special characters, domain age, and HTTPS usage. The researchers trained and assessed different machine learning algorithms which include Logistic Regression Decision Tree and Random Forest. The model demonstrates exceptional accuracy while enabling users to determine website legitimacy through immediate real-time authentication.

Keywords: Phishing Detection, Machine Learning, Cybersecurity, Random Forest, URL Analysis, Classification

1.INTRODUCTION

Phishing attacks have risen sharply because people are using the internet more than ever before. The attackers create counterfeit websites which perfectly replicate real websites to deceive users into providing their private data. The existing methods which use blacklists to detect phishing websites are ineffective at identifying new phishing websites. Machine Learning (ML) provides an efficient way to detect such websites by learning patterns from data. The research project will create a machine learning model which will detect phishing threats through its development. The prototype will extract essential URL-based characteristics which will drive its operations. The model will be implemented through a web-based system which will allow users to access its functions. The team will use ensemble methods to achieve better accuracy results.

2.Literature Review

Several researchers have explored the use of machine learning techniques for phishing website detection. The first methods used basic classification algorithms through their implementation of Logistic Regression and Naïve Bayes, which achieved moderate accuracy but failed to solve complex pattern recognition tasks. Decision Tree models introduced better interpretability, yet they encountered overfitting problems. Support Vector Machines (SVM) have become common because of their capacity to process data with multiple dimensions, yet they need exact parameter adjustments. The current research demonstrates that ensemble techniques, which include Random Forest and Gradient Boosting, are effective because they merge several models to achieve high precision while decreasing overfitting problems. The methods demonstrate better results than their respective single models. The current research shows that feature engineering and real-time deployment of detection systems function as essential components for successful system performance. The project aims to develop systems that achieve both precise accuracy and practical usability, which remains unmet despite current technological progress.

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

3. PROPOSAL METHODOLOGY

A. Data Collection and Feature Selection

LDataset contains:

URL length

Use of HTTPS

Number of dots

Special characters (@, -, etc.)

Domain age

Target:

0 → Legitimate

1 → Phishing

Mathematically:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

Where:

- $T_i(x)$ = prediction from ith decision tree
- N = number of trees

B. Data Preprocessing and Transformation

- Remove missing values
- Convert categorical to numerical
- Normalize data
- Split dataset (Train/Test)

C. Random Forest Model Development

The Logistic Regression

Decision Tree

Random Forest

⇒ Random Forest performs best due to ensemble learning.

D. Prediction and System Integration

· User enters URL

· Features extracted

· Model predicts result

· Output displayed (Safe / Phishing)

E. Performance Evaluation

Metrics used:

Accuracy

Precision

Recall

F1-score

4. SYSTEM ARCHITECTURE

The proposed system consists of the following modules:

Workflow:

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

User Input → Feature Extraction → ML Model → Prediction → Output

Modules:

User Interface (HTML/CSS)

Preprocessing Module

ML Model

Output Display

System Architecture

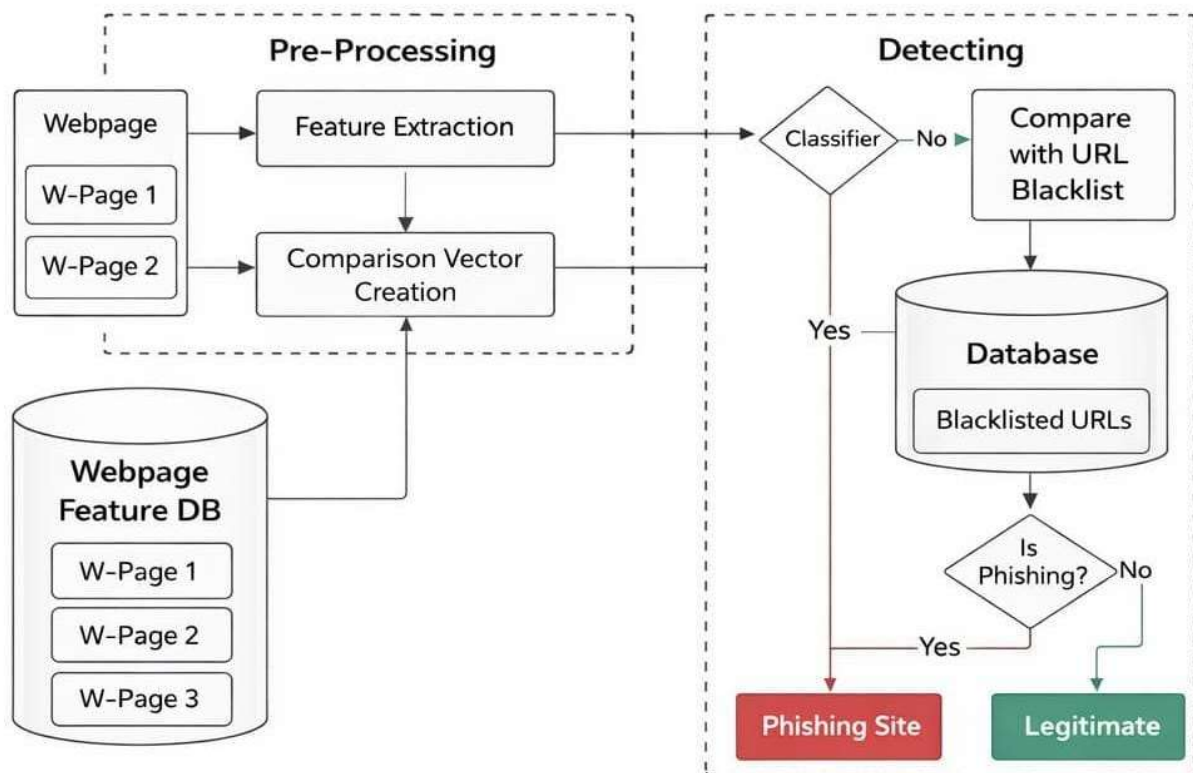


Fig. 1. The above diagram shows how the Phishing Website Detection Using Machine Learning system functions from beginning to end. The system includes two main stages which are Pre-Processing and Detecting. The first phase starts when the system receives input webpages and proceeds to extract their essential features through feature extraction. The extracted features of the webpage include URL structure, special symbol detection, HTTPS implementation, redirection patterns, domain attributes, and additional webpage characteristics. The extracted features create a comparison vector which enables the comparison of extracted webpage features against the existing database of webpage features. This process enables effective organization of input data which gets transformed into a proper format for classification purposes.

5. IMPLEMENTATION DETAILS

A. Development Environment



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

- Python
- Jupyter Notebook
- Libraries: Pandas, NumPy, Scikit-learn

B. Data Handling

The Data handling is a crucial step in the development of the phishing website detection system, as the performance of the machine learning model highly depends on the quality of data used. In this work, a dataset containing both legitimate and phishing website URLs is collected from reliable sources such as public cybersecurity datasets and repositories.

C. Model Training

- Train using dataset
- Test accuracy
- Save model using

D. . Model Saving

The machine learning model for phishing website detection needs its trained model to be saved after the training and evaluation process has been completed. The system achieves efficient operation through model saving because it enables trained models to be reused without requiring retraining for each instance.

E. Web Application Integration

The web application integration serves as a crucial step for the phishing website detection system because it provides users with an easy and effective way to use the machine learning model. The module aims to execute its trained model through a real-time system which allows users to enter URLs and obtain immediate results.

F. Output Generation

The output generation module is the final stage of the phishing website detection system, where the prediction results from the machine learning model are presented to the user in a clear and understandable format. The module functions as an essential component that enables the system to produce results which decision-makers can utilize effectively.

6.EXPLAINABLE INTEGRATION

A. Need for Explainability

The Random Forest machine learning model achieves high accuracy for its ability to identify phishing websites. The model operates as a black box system which prevents users from understanding how URL features affect its final prediction outcome.

B. Feature Importance Analysis

The feature importance scores obtained from the Random Forest model are used to identify the most influential features in detecting phishing websites. The important features of this system include URL length and presence of HTTPS and number of special characters and number of subdomains and domain age.

C. Interpretation of Predictions

This section describes the method used by the system to evaluate how each feature contributes to its final prediction results. The model evaluates different characteristics of the input URL and assigns importance to each feature.

D. Practical Significance

The phishing detection system achieves better usability through Explainable AI because it delivers output results which users find easier to comprehend and believe to be accurate. The



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

system establishes a connection between complex machine learning systems and their end users through its ability to display prediction results.

- The system enables users to understand machine learning predictions better through its transparent design.
- The system boosts user confidence in its ability to detect phishing attempts.
- The system enables users to comprehend the reasons behind a website's classification as either phishing or safe.

7.CONCLUSION

The researchers developed and tested a machine learning system which detects phishing websites in their research study. The system uses URL length and HTTPS presence and special character count and domain attributes as URL-based features to determine whether websites are legitimate or phishing sites. The Random Forest model showed better accuracy and reliability results than all other tested machine learning algorithms. The proposed system functions as a web application which lets users submit URLs to receive instant prediction results. The system provides practical phishing detection capabilities which users can easily operate. The system uses Explainable AI techniques to enhance transparency by showing users the reasons behind model predictions.

8.FUTURE WORK

Future research directions include:

The first direction of research requires real-time threat intelligence system implementation.

The second direction of research needs advanced Explainable AI techniques to improve system interpretation.

The third system deployment direction requires cloud platforms to achieve scalability.

The fourth research direction needs better performance through improved feature extraction methods.

The study requires mobile application integration to enable users from various backgrounds to access the system.

9.REFERENCES

- [1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [2] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd ed., O'Reilly Media, 2019.
- [3] J. Brownlee, "Machine Learning Algorithms for Classification," Machine Learning Mastery, 2020.
- [4] Scikit-learn Documentation, "Classification Algorithms," Available: <https://scikit-learn.org>
- [5] Kaggle Dataset, "Phishing Website Detection Dataset," Available: <https://www.kaggle.com>
- [6] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," in Proceedings of ACM Workshop on Recurring Malcode, 2007.
- [7] R. Verma and R. Das, "What's in a URL: Fast Feature Extraction and Malicious URL Detection," in Proceedings of IEEE Conference, 2017.
- [8] C. Molnar, Interpretable Machine Learning, 2nd ed., 2022.
- [9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas178

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

[10] IEEE Xplore Digital Library, “Research Papers on Phishing Detection Using Machine Learning.”