IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

AI-Driven Phishing Website Detection: An Empirical Evaluation of Machine Learning Algorithms and Feature-Based Classification

Prof. Rana Afreen Sheikh[1], Ms.Vaibhavi Gajanan Awandkar[2], Ms.Saniya Khan[3]
Professor, Department of Master of Computer Applications (MCA)[1]
Students, Department of Master of Computer Applications (MCA)[2,3]
Vidya Bharati Mahavidyalaya, Amravati, Maharashtra, India

Abstract: *Phishing websites are a major cybersecurity threat because they imitate legitimate websites to deceive users into revealing sensitive information, including passwords, financial details, and personal data. Traditional phishing detection techniques, such as blocklists and manually defined rules, may fail to identify newly created or rapidly modified malicious websites. Artificial intelligence (AI) and machine learning (ML) offer a data-driven approach to identifying suspicious websites through the analysis of URL characteristics and other website-related features. This paper presents an empirical evaluation framework for comparing machine learning algorithms for phishing website detection. The proposed study uses the publicly available UCI Phishing Websites dataset, which contains 11,055 instances and 30 predictive features. Four supervised learning algorithms—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are selected for comparative evaluation. The methodology incorporates data preprocessing, stratified training and testing, cross-validation, feature-importance analysis, and multiple performance metrics, including precision, recall, F1-score, and ROC-AUC. Existing research indicates that ensemble learning methods, particularly Random Forest, are promising for this classification task; however, performance depends on dataset characteristics and experimental design. Since no new model execution was conducted for this manuscript, numerical experimental results are not claimed. The proposed framework supports reproducible evaluation and highlights the importance of temporal validation, interpretable predictions, and real-world testing in developing reliable phishing detection systems.*

Keywords: Phishing detection, artificial intelligence, machine learning, Random Forest, cybersecurity, website classification, empirical evaluation

I. INTRODUCTION

The widespread adoption of online banking, e-commerce, cloud services, and digital communication has increased dependence on websites for everyday activities. However, this dependence has also created opportunities for cybercriminals to conduct phishing attacks. Phishing is a form of social engineering in which attackers impersonate trusted organizations, services, or individuals to manipulate users into disclosing confidential information or performing harmful actions.

Phishing websites often resemble legitimate websites in their appearance, domain names, logos, and login interfaces. Users may therefore find it difficult to distinguish fraudulent pages from genuine ones. Successful attacks can result in account compromise, identity theft, financial loss, and unauthorized access to organizational systems.

Conventional detection approaches commonly use blocklists, heuristic rules, and browser warnings. Although these techniques are useful, blocklists may not identify newly registered phishing domains, while manually specified rules

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

can become ineffective as attackers modify their strategies. Machine learning provides an alternative by learning patterns associated with phishing and legitimate websites from labeled examples.

Supervised machine learning formulates phishing website detection as a binary classification problem. Given a set of website features, a trained model predicts whether a website is legitimate or phishing. Relevant features may include URL length, the presence of an IP address in the URL, suspicious symbols, domain characteristics, HTTPS-related indicators, and webpage behavior.

Previous research has investigated decision trees, Random Forest, Support Vector Machines (SVM), neural networks, and boosting algorithms for phishing detection. Nevertheless, reported performance can vary considerably because of differences in datasets, feature extraction methods, class distributions, and validation strategies. A model that performs well on a historical dataset may not generalize to newly emerging phishing websites.

This study therefore focuses on a transparent and reproducible empirical framework for comparing selected machine learning algorithms using a publicly available benchmark dataset.

1.1 RESEARCH PROBLEM

The primary research problem is to determine how effectively different supervised machine learning algorithms distinguish phishing websites from legitimate websites using structured website features. The study also considers whether classification performance alone is sufficient to select a practical detection model.

1.2 RESEARCH OBJECTIVES

The objectives of this study are:

- To examine the application of machine learning to phishing website detection.
- To compare Logistic Regression, Decision Tree, Random Forest, and SVM under a consistent evaluation procedure.
- To evaluate models using precision, recall, F1-score, ROC-AUC, and computational efficiency.
- To identify important predictive features and discuss their implications for interpretable cybersecurity systems.
- To identify limitations affecting the generalization of machine learning models to real-world phishing attacks.

1.3 RESEARCH QUESTIONS

RQ1: Which of the selected machine learning algorithms provides the strongest overall phishing detection performance?

RQ2: How do precision, recall, and F1-score differ across the evaluated classifiers?

RQ3: Which website features contribute most to classification decisions?

RQ4: What methodological limitations affect the practical deployment of these models?

II. LITERATURE REVIEW

Phishing website detection has developed from static list-based approaches toward heuristic, similarity-based, machine learning, and deep learning techniques. Safi and Singh (2023) reviewed phishing detection research and identified machine learning as a frequently used approach. Their review also emphasized the diversity of datasets, algorithms, and evaluation methods used in previous studies.

2.1 TRADITIONAL DETECTION TECHNIQUES

List-based detection compares a website's URL against previously identified malicious URLs. This method is computationally straightforward and effective for known threats, but it cannot reliably identify every newly created phishing website. Heuristic techniques use predefined indicators, such as suspicious URL structures, unusual domain characteristics, or potentially misleading symbols. Although these indicators are useful, attackers can adapt their techniques to evade fixed rules.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

Similarity-based approaches compare a suspicious website with the appearance or identity of a legitimate website. Such approaches may help identify cloned login pages but can require additional computational resources and access to reference websites.

2.2 MACHINE LEARNING-BASED DETECTION

Machine learning approaches learn classification patterns from labeled examples. Mohammad, Thabtah, and McCluskey developed a widely used set of phishing website features, subsequently made available through the UCI Machine Learning Repository. The benchmark dataset contains 11,055 instances and 30 features, providing a basis for comparing classification methods.

Random Forest combines predictions from multiple decision trees and can capture nonlinear relationships between features. Decision Trees offer comparatively interpretable classification rules, whereas Logistic Regression provides a relatively simple linear baseline. SVM can model complex decision boundaries through appropriate kernels but may require additional computational effort and parameter tuning.

Experimental research by Mohammad and colleagues and subsequent studies has established the value of feature-based machine learning for phishing classification. However, the results of one experiment cannot automatically be generalized to other datasets or future phishing campaigns.

2.3 DEEP LEARNING AND HYBRID APPROACHES

Deep learning models can learn representations from raw URLs, character sequences, webpage content, and screenshots. Adequately trained neural networks may identify patterns that are difficult to express using manually engineered features. Hybrid systems combine URL characteristics with webpage content, domain information, or external threat intelligence.

The literature also identifies challenges associated with data freshness, feature availability, reproducibility, and evaluation on previously unseen websites. A 2021 experimental study on benchmark dataset construction reported that combining different categories of features can improve detection performance, while also highlighting difficulties in reproducing results from older datasets.

2.4 RESEARCH GAP

A key research gap is the limited comparability of reported results when studies use different datasets, feature sets, preprocessing procedures, and train-test strategies. Randomly dividing a historical dataset may also allow closely related records or outdated patterns to appear in both training and testing sets, leading to optimistic performance estimates.

The present study addresses this methodological gap by specifying a consistent comparison procedure, multiple evaluation metrics, and recommendations for independent temporal validation. Its contribution is a reproducible empirical evaluation design rather than a claim that one algorithm is universally superior.

III. RESEARCH METHODOLOGY

3.1 RESEARCH DESIGN

This study adopts a quantitative, comparative experimental design. Four supervised machine learning algorithms are evaluated using the same labeled dataset and a common preprocessing and validation pipeline. The independent variable is the classification algorithm, while the dependent variables are the resulting predictive performance measures and computational requirements.

The experiment is designed to be reproducible using Python and the scikit-learn machine learning library.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

3.2 DATASET DESCRIPTION

The proposed experiment uses the **UCI Phishing Websites dataset**, contributed by Mohammad and McCluskey. It contains 11,055 website instances and 30 predictive features, together with a class label. The dataset includes phishing-related indicators derived from URLs and website characteristics.

The data can be accessed through the UCI Machine Learning Repository:
https://archive.ics.uci.edu/dataset/327/phishing+websites

The dataset is suitable for an initial comparative study because it is publicly accessible, structured, and widely used in phishing classification research. Nevertheless, its historical nature means that it cannot fully represent contemporary phishing techniques.

Before modeling, the dataset's label encoding, feature definitions, missing values, duplicate records, and class distribution should be verified. The target label must be separated from the input features to prevent data leakage.

3.3 DATA PREPROCESSING

Data preprocessing consists of the following steps:

Data inspection: Examine dataset dimensions, feature types, label distribution, missing values, and duplicate records.

Target encoding: Convert the dataset's class labels into a consistent binary representation, documenting which value represents phishing.

Feature-target separation: Define the feature matrix X and target vector y.

Data splitting: Allocate 80% of the records to training and 20% to final testing, using stratified sampling and a fixed random seed.

Feature scaling: Apply standardization for Logistic Regression and SVM where appropriate. Fit scaling parameters only on training data.

Cross-validation: Use stratified five-fold cross-validation on the training set for model selection and hyperparameter tuning.

The final test set must remain untouched until the modeling and tuning decisions have been completed.

3.4 SELECTED MACHINE LEARNING ALGORITHMS

Logistic Regression: Used as a baseline model because it is computationally efficient and provides interpretable coefficients. Its linear decision boundary may limit performance when feature relationships are highly nonlinear.

Decision Tree: Learns classification rules through recursive feature-based splits. It is relatively easy to interpret but can overfit without suitable regularization.

Random Forest: Combines multiple decision trees to improve robustness and capture nonlinear relationships. It is included because of its established use in phishing detection research.

Support Vector Machine: Identifies a separating boundary between classes and can model nonlinear relationships using a suitable kernel. Feature scaling and hyperparameter selection are important for effective performance.

The models should be trained on identical training partitions and compared using the same evaluation protocol. Hyperparameters should be selected through cross-validation rather than by repeatedly inspecting the final test set.

3.5 EXPERIMENTAL PROCEDURE

The experiment follows this sequence:

- Download and verify the dataset.
- Inspect and preprocess the features and class labels.
- Create a stratified training-test split.
- Establish baseline model configurations.
- Tune model hyperparameters through cross-validation.
- Train the selected models on the complete training partition.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

- Evaluate each trained model on the held-out test set.
- Calculate performance metrics and confusion matrices.
- Analyze feature importance and computational efficiency.
- Document the software versions, random seed, parameter settings, and final results.

3.6 EVALUATION METRICS

Accuracy alone is insufficient for cybersecurity applications because false negatives can allow malicious websites to reach users.

The following metrics are used:

Accuracy

Accuracy measures the proportion of correctly classified instances.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision

Precision measures the proportion of predicted phishing websites that are actually phishing websites.

$$\text{Precision} = TP / (TP + FP)$$

Recall

Recall measures the proportion of actual phishing websites correctly identified.

$$\text{Recall} = TP / (TP + FN)$$

F1-score

The F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

ROC-AUC

The area under the receiver operating characteristic curve evaluates the model's ability to distinguish the two classes across classification thresholds.

Here, TP represents true positives, TN true negatives, FP false positives, and FN false negatives. The phishing class must be consistently designated as the positive class. Where relevant, precision-recall AUC should also be reported.

3.7 RELIABILITY AND VALIDITY

Reliability is supported through fixed random seeds, documented preprocessing, identical test partitions, and cross-validation. Validity is strengthened by using several complementary metrics rather than relying solely on accuracy.

However, a random split of historical data does not demonstrate generalization to future phishing websites. A stronger follow-up experiment should use temporally separated training and test data, remove duplicated or closely related URLs where appropriate, and evaluate the model on independently collected, recently labeled examples.

IV. LIMITATIONS AND FUTURE SCOPE

The principal limitation is the historical nature of the benchmark dataset. Features that distinguished phishing websites in the past may become less informative as attackers change domain registration practices, use HTTPS, or employ more sophisticated website cloning techniques.

Second, the study focuses on structured features rather than jointly analyzing URL strings, HTML content, screenshots, and domain reputation. Third, results from a single benchmark cannot establish performance across different countries, industries, languages, or attack campaigns.

Future research should evaluate recent phishing feeds and legitimate login websites, perform time-based testing, and compare feature-based models with character-level neural networks and gradient-boosting methods. Explainability methods such as permutation importance and SHAP can help investigators understand individual predictions. A practical deployment study should additionally measure latency, false-alarm rates, calibration, and resilience to adversarial changes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

V. CONCLUSION

This paper presents a reproducible empirical framework for evaluating AI and machine learning techniques for phishing website detection. It describes a comparative study using Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine with the UCI Phishing Websites dataset. The methodology specifies preprocessing, stratified splitting, cross-validation, multiple performance metrics, and feature-importance analysis.

The literature suggests that machine learning can support effective phishing classification and that ensemble methods are promising candidates. Nevertheless, the strongest model must be identified through controlled experimentation rather than assumed in advance. Since the proposed experiment has not been executed for this manuscript, no new accuracy values or winning classifier are claimed.

The central conclusion is that reliable phishing detection requires more than high benchmark accuracy. Reproducible experiments, careful treatment of false negatives and false positives, interpretable predictions, current datasets, and temporal validation are essential for developing systems that remain useful against evolving phishing threats. The proposed framework provides a foundation for conducting and reporting such an empirical investigation.

REFERENCES

1. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012). An assessment of features related to phishing websites using an automated technique. *International Conference for Internet Technology and Secured Transactions (ICITST)*. IEEE. https://doi.org/10.1109/ICITST.2012.6470840
2. Mohammad, R., & McCluskey, L. (2012). *Phishing Websites* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51W2X
3. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University—Computer and Information Sciences, 35*(2), 590–611. https://doi.org/10.1016/j.jksuci.2023.01.004
4. Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2022). Phishing webpage classification via deep learning-based algorithms: An empirical study. *Applied Sciences, 11*(19), 9210. https://doi.org/10.3390/app11199210
5. *Towards benchmark datasets for machine learning based website phishing detection: An experimental study.* (2021). *Engineering Applications of Artificial Intelligence, 104*, 104347. https://doi.org/10.1016/j.engappai.2021.104347
6. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029
7. *Phishing or not phishing? A survey on the detection of phishing websites.* (2023). *IEEE Access, 11*, 18499–18519. https://doi.org/10.1109/ACCESS.2023.3247135