Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

# Detection and Classification of Phishing Websites Using Data Mining and Explainable Artificial Intelligence

**Author:** Jyoti Balwan Sharma
MAEER's MIT Arts, Commerce and Science College, Alandi (D.)

**Corresponding Author:** Mr. Amol Bajirao Kale
MAEER's MIT Arts, Commerce and Science College, Alandi (D.)

**Abstract -** Phishing remains a major cybersecurity threat that exploits user trust to steal sensitive information. Blacklist-based detection approaches are ineffective against zero-day phishing websites due to their reactive nature. This paper proposes a phishing website detection framework using supervised machine learning integrated with Explainable Artificial Intelligence (XAI). A dataset of 10,000 labeled websites was analyzed using URL-based, domain-based, and webpage behavioral features. Five machine learning models were evaluated: Logistic Regression, Support Vector Machine (SVM), Decision Tree, Gradient Boosting, and Random Forest. Ensemble models outperformed standalone classifiers, with Random Forest achieving the highest accuracy (96.8%). SHAP-based explainability highlights influential features driving predictions, improving transparency and analyst trust in practical deployments.

**Keywords:** Phishing Detection, Machine Learning, Explainable AI, Cybersecurity, Data Mining

## 1. INTRODUCTION

The rapid expansion of online services has increased exposure to phishing attacks that trick users into revealing sensitive information such as credentials and financial details. Traditional blacklist databases contain only known malicious URLs and often fail to detect newly created or short-lived phishing sites. Machine learning approaches address this limitation by learning patterns from website features and generalizing to previously unseen (zero-day) attacks. However, many high-performing models lack interpretability; therefore, this work integrates Explainable AI to provide transparent, trustworthy predictions.

*Figure 1. Architecture of the proposed phishing website detection framework*

## 2. RELATED WORK

Prior research has explored neural networks, rule-based systems, and statistical learning techniques for phishing detection. Ensemble methods such as Random Forest and Gradient Boosting generally provide strong performance due to robustness and generalization. Recent work in explainability—particularly SHAP—enables interpretation of model outputs by estimating feature contributions. This study combines high-performing ensemble models with SHAP-based explanations to support practical cybersecurity decision-making.

## 3. PROPOSED METHODOLOGY

The framework includes data collection, feature extraction, model training/evaluation, and explainability analysis. Features are grouped into URL-based, domain-based, and webpage behavioral categories. Preprocessing includes missing-value handling, Min-Max normalization, correlation-based feature selection, an 80:20 train-test split, and five-fold cross-validation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

*Table 1. Categories of features extracted for phishing detection*

| Feature Category | Example Features |
|---|---|
| URL-Based | URL length, special character count, suspicious keywords, IP address usage, subdomain count |
| Domain-Based | Domain age, WHOIS availability, DNS validity, SSL certificate presence, expiration period |
| Behavioral | JavaScript redirection, iframe usage, external resource ratio, form action mismatch, pop-up behavior |

## 4. MACHINE LEARNING MODELS AND METRICS

We evaluated Logistic Regression, SVM (RBF kernel), Decision Tree, Gradient Boosting, and Random Forest. Performance was measured using accuracy, precision, recall, F1-score, and ROC-AUC, which are appropriate for security classification tasks where false negatives are costly.

## 5. EXPERIMENTAL RESULTS

Ensemble-based models achieved superior performance compared to standalone classifiers. Random Forest achieved the best overall results, including 96.8% accuracy and 0.98 ROC-AUC.

*Table 2. Performance evaluation of machine learning classifiers*

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 91.4 | 90.2 | 92.1 | 91.1 | 0.93 |
| SVM | 93.6 | 92.8 | 94.2 | 93.5 | 0.95 |
| Decision Tree | 94.1 | 93.5 | 94.8 | 94.1 | 0.96 |
| Gradient Boosting | 95.7 | 95.0 | 96.1 | 95.5 | 0.97 |
| Random Forest | 96.8 | 96.2 | 97.3 | 96.7 | 0.98 |

## 6. EXPLAINABLE AI (XAI) USING SHAP

To improve transparency, SHAP was applied to the best-performing Random Forest model. Global interpretability ranks the most influential features as domain age, URL length, SSL certificate validity, IP address usage, and subdomain count. Local explanations illustrate how individual features push a given prediction toward phishing or legitimate classification.

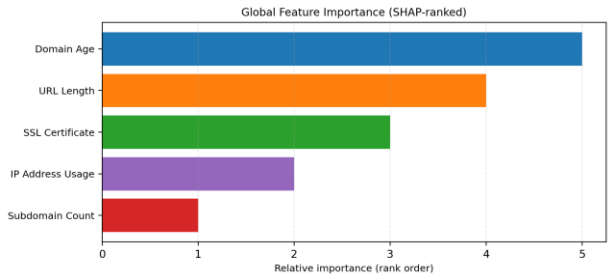


*Figure 2. Global SHAP features importance (ranked) for the Random Forest model*

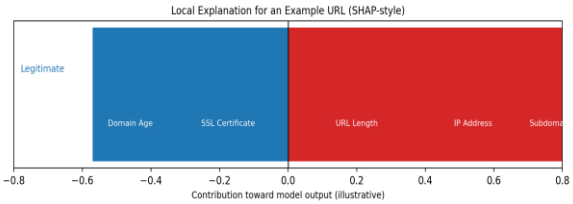


*Figure 3. Local explanation of an example phishing prediction (illustrative SHAP-style plot)*

## 7. DISCUSSION

Results indicate that ensemble learning improves phishing detection by reducing variance and capturing non-linear patterns from heterogeneous features. Unlike blacklist-based mechanisms, the proposed approach can detect zero-day phishing websites by generalizing from learned patterns. Explainability supports analyst validation and trust by clarifying why a URL is flagged. Limitations include potential adversarial manipulation and the need for continuous feature updates as attacker strategies evolve.

## 8. CONCLUSION

This paper presents a phishing website detection framework combining supervised machine learning and Explainable AI. Random Forest achieved the highest accuracy (96.8%) among evaluated models, while SHAP explanations improved interpretability and transparency. The framework is suitable for practical cybersecurity deployments and can be extended toward real-time detection and browser-based protection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

## REFERENCES

[1] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, 2014.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.

[3] A. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, 2014.

[4] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to detect malicious URLs," ACM TIST, 2011.