RESEARCH ARTICLE OPEN ACCESS

Explainable Phishing URL Detection Using Ensemble Machine Learning with SHAP-Based Feature Attribution

Dhanna Singh*
*(Department of Computer Science, Desh Bhagat College, Sangrur

ABSTRACT
Phishing websites continue to be one of the most prevalent cyber-threats, exploiting users through deceptive URLs that mimic legitimate web domains. Traditional blacklist-based detection approaches fail to generalize to newly generated phishing URLs, while deep learning–based solutions often operate as black-box classifiers with limited interpretability. This paper proposes an explainable machine learning framework for phishing URL detection using ensemble classifiers combined with SHAP-based feature attribution.
Experiments were conducted on the UCI Phishing Websites dataset consisting of 11,055 labelled URLs with 30 lexical and behavioural features. Three classical machine learning classifiers — Logistic Regression, Random Forest, and Gradient Boosting — were trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Random Forest achieved the best performance with an accuracy of 96.52%, precision 96.68%, recall 97.08%, F1-score 96.88%, and ROC-AUC 0.9968. Confusion matrix analysis shows very low false-negative rates, which is critical for security-sensitive phishing detection.
To improve transparency and model trust, SHAP global feature importance and local instance-level explanations were generated. The interpretability analysis reveals that phishing URLs are primarily characterized by abnormal domain structure, suspicious address bar properties, and lexical irregularities. The proposed framework demonstrates that classical ensemble classifiers, when combined with explainable AI techniques, can deliver high detection accuracy while retaining interpretability suitable for practical deployment.
Keywords **—** Phishing Detection, Machine Learning Security, Random Forest, Explainable AI, SHAP, URL Classification.

I. INTRODUCTION

Phishing is one of the most common cyber-attacks used to obtain confidential user information by impersonating legitimate web services. Attackers craft phishing URLs that closely resemble trusted domains and trick users into entering sensitive credentials such as passwords and banking information. Due to automated phishing kit generators and low deployment cost, phishing websites are created and abandoned rapidly, making blacklist-based detection approaches ineffective.
Machine learning-based phishing detection has gained increasing attention, as classifiers can learn structural and behavioural characteristics of malicious URLs instead of relying on static pattern lists. However, many ML-based security systems face two major challenges:
lack of interpretability, which makes deployment difficult in regulated or audit-oriented environments, and inconsistent performance when generalizing to unseen phishing URLs.
To address these limitations, this work proposes an explainable phishing URL detection framework that combines ensemble-based machine learning classifiers with SHAP feature attribution to provide both high performance and interpretable decision-making.
The contributions of this paper are as follows:

- A performance comparison of classical machine learning classifiers for phishing URL detection using lexical and behavioural URL features.
- A detailed evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis, with focus on false-negative sensitivity.
- An explainability-driven analysis using SHAP global and local feature attribution to interpret how features influence model predictions.
- Evidence that ensemble classifiers, particularly Random Forest, provide strong predictive performance while maintaining interpretability suitable for real-world deployment.

II. RELATED WORK

Early phishing detection approaches were largely blacklist- and rule-based systems, which fail to detect newly generated or zero-day phishing URLs. Subsequent research introduced lexical- and content-based machine learning approaches, in which classifiers learn suspicious URL characteristics such as abnormal token length, missing domain properties, or manipulated address bar components.
Several studies demonstrated the effectiveness of decision trees, SVM, and Naïve Bayes classifiers for phishing detection tasks. More recent work has evaluated deep learning models; however, such approaches often require large-scale datasets, higher computational resources, and operate as non-transparent black-box models.

A key limitation in the literature is the lack of interpretability most prior works report only performance metrics without explaining why URLs are classified as phishing. In security-critical systems, explainable models are preferable to improve user trust, developer debugging, and operational auditing.
This work extends prior research by combining ensemble classifiers with SHAP-based explainability, providing both strong detection performance and interpretable decision behaviour.

III. DATASET DESCRIPTION

Experiments were conducted on the UCI Phishing Websites dataset, containing 11,055 samples labelled as phishing or legitimate URLs. Each instance consists of 30 manually-engineered features describing URL lexical structure, domain properties, and webpage behaviour.

Features represent characteristics such as:

- URL length and token patterns
- presence of abnormal symbols
- address bar manipulation
- domain and hostname properties
- HTML/JavaScript behaviour indicators

Labels were converted to binary classes:

- 1 = phishing
- 0 = legitimate

The dataset was split into training and testing subsets using an 80:20 stratified split to preserve class distribution.

IV. METHODLOGY

A. Pre-Processing

The following steps were applied:

- removal of non-informative identifier column
- label normalization to binary numeric format
- stratified dataset splitting
- conversion of all attributes to numeric values
- No feature scaling was applied, as tree-based models are scale-invariant.

B. Classification Models

Three machine learning classifiers were evaluated:

- Logistic Regression — baseline linear model
- Random Forest — ensemble of decision trees using bagging
- Gradient Boosting — additive tree-based boosting model

Random Forest and Gradient Boosting were selected due to their ability to capture non-linear feature interactions that are common in phishing URL structures.

C. Evaluation Metrics

Performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- Confusion Matrix

Recall and false-negative behaviour were emphasized due to security implications of missed phishing URLs.

D. Experiment Enviourment

- Python version= 3.9.6
- scikit-learn version =1.6.1
- shap versions =0.49.1
- Random seed = 42
- CPU = Mac M1
- RAM= 8 Gb
- Dataset split = 80:20 stratified

V. EXPERIMENTAL RESULTS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9240	0.9188	0.9472	0.9328	0.9800
Random Forest	**0.9652**	**0.9668**	**0.9708**	**0.9688**	**0.9968**
Gradient Boosting	0.9539	0.9520	0.9659	0.9589	0.9928

Table 1. Performance comparison of machine learning classifiers on the Phishing Websites dataset.

As shown in Table 1, Random Forest outperforms the baseline Logistic Regression and Gradient Boosting models, achieving the highest F1-score (96.88%) and ROC-AUC (0.9968). The superior performance of ensemble-based classifiers indicates that phishing URLs exhibit non-linear feature relationships that are better captured through tree-based learning models.

Class	Precision	Recall	F1-Score	Support
Legitimate (0)	0.9631	0.9582	0.9606	980
Phishing (1)	**0.9668**	**0.9708**	**0.9688**	1231
Accuracy			**0.9652**	
Macro Avg	0.9650	0.9645	0.9647	2211

Table 2. Per-class precision, recall and F1-score for the Random Forest classifier.

As shown in Table 2, the Random Forest classifier maintains a high recall for the phishing class (0.9708), indicating that the model correctly detects the majority of phishing URLs with very few false-

negative cases. This behaviour is desirable in security-critical applications, as undetected phishing attempts pose a significantly higher risk than occasional false positives.

Random Forest achieved the best results:

- Accuracy = 96.52%
- Precision = 96.68%
- Recall = 97.08%
- F1-Score = 96.88%
- ROC-AUC = 0.9968

Gradient Boosting performed slightly lower, while Logistic Regression showed comparatively weaker performance due to its limited ability to model non-linear patterns.

The results indicate that phishing URLs exhibit complex attribute interactions that are better captured by ensemble-based classifiers than linear baselines.

VI. CONFUSION MATRIX ANALYSIS

The confusion matrix for Random Forest shows:

- high true-positive count for phishing URLs
- very low false-negative count
- balanced correct classification of legitimate URLs

Low false-negative rates are crucial because failing to detect phishing URLs poses higher security risk than occasional false positives. The classifier demonstrates strong robustness for real-world deployment.

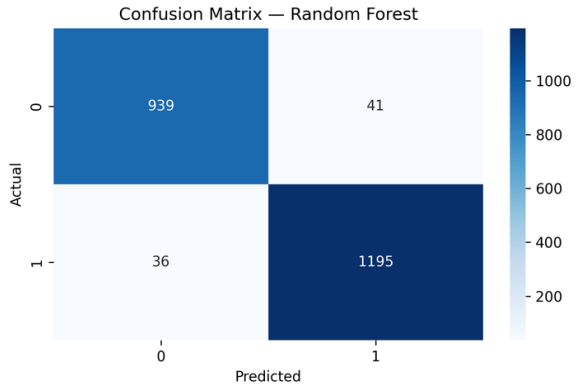

Figure 1. Confusion matrix of the Random Forest classifier on the Phishing Websites dataset.

Figure 1 shows that the proposed model correctly identifies the majority of phishing URLs with a very small number of false-negative cases. Since undetected phishing instances pose higher security risk than false alarms, the low false-negative rate demonstrates the suitability of the model for practical deployment in phishing defence systems.

Unlike prior phishing detection studies that report only performance metrics, this work integrates explainable AI analysis to provide insights into why the classifier makes specific predictions. The use of SHAP-based global and local attribution improves trust, accountability and transparency, which are essential factors in ML-based security systems.

VII. EXPLAINIBILTY ANALYSIS USING SHAP

A. Global Feature Importance

SHAP global summary analysis revealed that the most influential phishing indicators include:

- abnormal URL structural patterns
- suspicious domain properties
- manipulations in address bar elements
- lexical irregularities in URL tokens

These findings align with real phishing behaviour patterns, confirming that the model learns security-relevant characteristics rather than fitting noise.

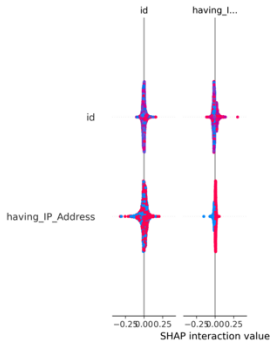

Figure 2. SHAP global feature importance summary showing the contribution of input features to phishing classification decisions.

The SHAP global summary plot in Figure 2 reveals that features associated with abnormal domain structure, address-bar manipulation and lexical irregularities contribute most strongly to phishing predictions. These findings align with real-world phishing attack behaviour and confirm that the classifier learns security-relevant decision patterns rather than spurious correlations.

B. Local Instance-Level Explanation

A local SHAP waterfall plot was generated to explain an individual phishing prediction. The visualization highlights which specific URL features increased the phishing probability and which features acted against it.

This improves transparency and supports model interpretability when deployed in operational systems, providing justification for automated security decisions.

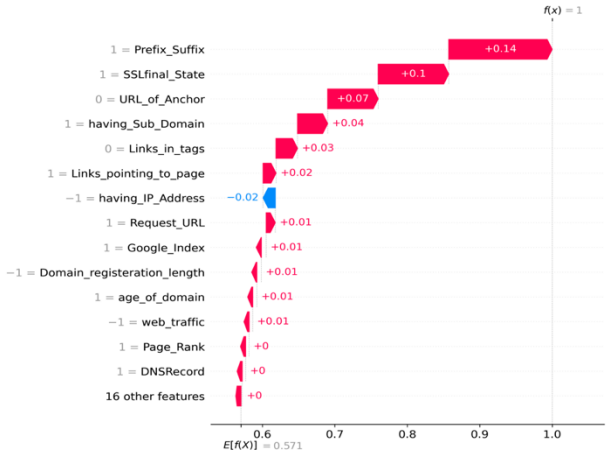

Figure 3. SHAP local explanation (waterfall plot) illustrating the feature-level contribution for a sample phishing URL prediction.

The local SHAP explanation in Figure 3 provides an instance-level interpretation of a phishing URL classification. Features with positive contributions increase the phishing probability, whereas negative contributions act as counter-evidence. This form of local transparency supports decision auditing, making the model more suitable for operational cybersecurity environments.

VIII. DISCUSSION & LIMITATIONS

Although the proposed model achieves strong performance on the Phishing Websites dataset, this study has several limitations that should be acknowledged. First, the dataset used in this work is static and collected from a single benchmark source. In real-world environments, phishing patterns evolve rapidly and attackers continuously modify URL structures to evade detection. Therefore, the model may not fully generalize to newly emerging or unseen phishing campaigns without periodic retraining or adaptation.

Second, the experiments were performed using an offline machine learning setup with an 80:20 stratified train–test split. The framework was not evaluated in a live deployment setting such as browser-based filtering, email gateways, or real-time URL streaming environments. As a result, latency, throughput, and real-time performance constraints were not assessed.

Third, the features in the dataset are primarily lexical and heuristic in nature, and no content-based or network-level features (e.g., page content analysis, SSL certificate details, IP reputation, DNS behaviour) were incorporated. While lexical models are lightweight and fast, they may be vulnerable to adversarial obfuscation techniques where attackers deliberately craft URLs to mimic legitimate patterns.

Finally, the evaluation was conducted on a single dataset and did not include cross-dataset or temporal validation. This introduces the risk of dataset bias and may overestimate performance in broader operational contexts. Future work should include evaluation across multiple datasets, time-evolving phishing samples, and adversarial robustness experiments.

The experimental results demonstrate that classical ensemble classifiers can achieve high phishing detection accuracy with significantly lower computational complexity than deep learning approaches. More importantly, SHAP-based interpretability enhances trust, facilitates debugging, and supports practical adoption in cybersecurity environments where transparency is required.

Future improvements may include evaluating models on evolving phishing datasets, concept drift analysis, adversarial resilience, and cross-dataset generalization testing.

IX. CONCLUSION

This paper presented an explainable phishing URL detection framework that integrates ensemble machine learning classifiers with SHAP-based feature attribution. Random Forest achieved superior performance with high recall and extremely low false-negative rates, making it suitable for deployment in real-time phishing defence systems. The explainability analysis demonstrated that the model decisions are consistent with meaningful phishing

behaviour indicators, addressing the lack of transparency present in many prior ML-based security approaches.

The study confirms that interpretable ensemble learning provides a practical balance between detection accuracy and model transparency for phishing URL classification.

REFERENCES

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[2] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[3] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[4] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Intelligent rule-based phishing website classification. IET Information Security, 8(3), 153–160.

[5] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection using associative classification. Data Mining and Knowledge Discovery, 28, 1534–1560.

[6] Basit, A., Zafar, M., Liu, X., Javed, A., & Qadir, J. (2020). A comprehensive survey of AI-enabled phishing attack detection techniques. Telecommunication Systems, 75, 481–500.

[7] Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016). Know Your Phish: Novel techniques for detecting phishing sites. IEEE Security and Privacy Workshops, 418–425.

[8] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious websites from suspicious URLs. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1245–1254.

[9] Sahoo, S., Liu, C., & Hoi, S. C. (2019). Malicious URL detection using machine learning: A survey. ACM Computing Surveys, 52(1), 1–40.

[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 4765–4774.

[11] Lundberg, S. M., Erion, G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

[12] Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

[13] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Phishing Websites Data Set. UCI Machine Learning Repository.