International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Robust and Explainable Machine Learning for Phishing Website Detection

Sankurubhukta Prasanth Kumar
Andhra University College of Engineering
Andhra University, Visakhapatnam, Andhra Pradesh, India

***Abstract* - Phishing website detection research often reports very high classification accuracy on a single benchmark, but benchmark performance can depend strongly on feature engineering and the evaluation protocol. This study evaluates phishing detection using the publicly documented PhiUSIIL dataset. A 50-feature numeric baseline is compared with an 18-feature raw-URL subset that excludes several engineered similarity and probability indicators. Random stratified evaluation is compared with a domain-disjoint split in which no exact domain appears in both training and test sets. Random Forest and HistGradientBoosting models are evaluated using accuracy, precision, recall, F1 score, ROC AUC, PR AUC, and error counts. SHAP ranks URL-level features, and stability of the selected feature set is assessed across five Random Forest seeds. The full 50-feature Random Forest reached 100 percent on the reported test metrics under both evaluation protocols. The 18-feature URL representation produced about 99.7 percent accuracy, while an eight-feature SHAP selected representation retained 99.73 percent accuracy on the domain-disjoint test. The study shows that strong benchmark performance can be retained with a much smaller and more interpretable feature set, while also demonstrating why perfect benchmark scores should not be interpreted as proof of real-world robustness.**

***Keywords* - phishing detection, machine learning, explainable AI, SHAP, URL features, cybersecurity**

I. INTRODUCTION

Phishing websites imitate legitimate online services to persuade users to disclose credentials, payment information, or other sensitive data. Machine learning has become a common approach for detecting phishing because URL structure, domain characteristics, webpage properties, and other observable signals can be represented as classification features. Survey literature reports extensive use of classical machine learning, ensemble methods, and deep learning, but comparisons are complicated by differences in datasets, feature engineering, and evaluation protocols [1], [2].

A central concern is how very high benchmark accuracy should be interpreted. A model may exploit highly informative engineered variables that were derived from webpage characteristics represented in the same benchmark. Such variables can be valuable for detection, but they can also make a benchmark substantially easier than an operational deployment setting. Recent studies have therefore explored explainability and feature selection, including SHAP and LIME-based approaches [3], [4].

This study examines how much predictive performance can be retained when detection is restricted to directly interpretable URL-level features. It also compares a conventional random split with a domain-disjoint split, preventing the same exact domain from appearing in both training and test data. Finally, SHAP is used to identify a compact feature subset and to examine the stability of that subset across different Random Forest seeds.

The contributions are a controlled comparison between full numeric and raw-URL representations, domain-disjoint evaluation, SHAP-based feature reduction, and an explanation stability analysis. The work is presented as a reproducible benchmark study rather than as a claim of a new classification algorithm.

II. DATASET AND DATA PREPARATION

The experiments use the PhiUSIIL Phishing URL Website dataset available through the UCI Machine Learning Repository. UCI documents 235795 instances, including 134850 legitimate URLs and 100945 phishing URLs, with 54 documented features and no missing values [5]. The dataset contains URL and webpage-derived information and is distributed under a Creative Commons Attribution 4.0 license [5].

The uploaded CSV contained 235795 rows and 56 columns because it included the filename field and label in addition to the documented feature fields. The principal numeric baseline excluded filename, label, URL, Domain, TLD, and Title, leaving 50 numeric predictors. A second representation used 18 raw-URL-level variables: URLLength, DomainLength, IsDomainIP,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

TLDLength, NoOfSubDomain, HasObfuscation, NoOfObfuscatedChar, ObfuscationRatio, NoOfLettersInURL, LetterRatioInURL, NoOfDegitsInURL, DegitRatioInURL, NoOfEqualsInURL, NoOfQMarkInURL, NoOfAmpersandInURL, NoOfOtherSpecialCharsInURL, SpacialCharRatioInURL, and IsHTTPS.

The raw-URL representation intentionally excludes engineered similarity and probability indicators and webpage content counts. This separation tests whether strong detection performance can be retained using a smaller feature set whose meaning is easier for a security analyst to inspect.

III. EXPERIMENTAL METHODOLOGY

A. Evaluation Protocols. A stratified random split allocated 75 percent of observations to training and 25 percent to testing using random state 42. A second split used GroupShuffleSplit with Domain as the grouping variable and the same random state. The domain-disjoint test set contained 58572 observations and 55022 unique domains, with zero exact domains shared between training and test.

B. Models and Metrics. The primary classifier was Random Forest with 150 trees, square root feature sampling, balanced subsample class weighting, and random state 42. HistGradientBoosting was additionally evaluated on the 18-feature raw-URL representation using 250 boosting iterations, a learning rate of 0.08, a maximum of 31 leaf nodes, and L2 regularization of 1. Accuracy, precision, recall, F1 score, ROC AUC, PR AUC, false positives, and false negatives were recorded.

C. Explainability and Feature Selection. The SHAP TreeExplainer was applied to a Random Forest trained on the raw-URL feature set. A 500-instance training sample was used for the final ranking. Mean absolute SHAP values were calculated for the positive class output. The eight highest-ranked features were used as a compact representation and evaluated under both split protocols.

D. Explanation Stability. Five Random Forest models were trained with seeds 11, 22, 33, 44, and 55. SHAP rankings were computed on 400-training observations for each seed. Stability was summarized by the average pairwise Jaccard similarity of the top-eight-feature sets.

IV. RESULTS

The full 50-feature Random Forest produced 100 percent accuracy, precision, recall, F1 score, ROC AUC, and PR AUC on both the random and domain-disjoint test sets, with zero false positives and zero false negatives. This result indicates that the supplied benchmark is highly separable under the selected feature representation; it should not be interpreted as evidence that operational phishing detection is perfect.

When restricted to 18 raw-URL-level variables, Random Forest achieved 99.749 percent accuracy on the random split and 99.732 percent on the domain-disjoint split. HistGradientBoosting achieved 99.752 percent and 99.788 percent accuracy under the same two protocols. The compact eight-feature Random Forest retained 99.730 percent accuracy on the domain-disjoint test.

V. DISCUSSION

The principal result is that most of the benchmark's predictive power can be retained after substantial feature reduction. The 18-feature raw-URL representation achieved more than 99.7 percent accuracy under both evaluation protocols, while the eight-feature SHAP selected representation retained 99.73 percent accuracy on the domain-disjoint test. This suggests that a lightweight detector can be constructed without using every available webpage-derived variable.

The 100 percent result for the full 50-feature representation also demonstrates why benchmark scores should be interpreted in relation to feature construction. The dataset contains engineered similarity and probability variables as well as webpage-derived attributes [5]. These variables can be useful, but they make the classification boundary highly separable. The present study therefore treats the compact raw-URL experiment as the more informative robustness baseline.

The domain-disjoint evaluation prevents exact domain reuse but is not temporal validation. The experiment remains a single dataset study and cannot establish performance against future phishing campaigns or independently collected traffic. SHAP describes model attribution rather than causal importance, so the selected variables should not be interpreted as independent causes of phishing.

The average pairwise Jaccard similarity of 0.867 across five Random Forest seeds indicates substantial overlap in the top-eight SHAP features. In the recorded environment, reducing the numeric feature set from 50 to eight also reduced Random Forest

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

fitting time by roughly 41 to 46 percent, depending on the split. These timing values are environment-specific and are included only as relative computational evidence.

VI. CONCLUSION

This study evaluated robust and explainable machine learning for phishing website detection using the PhiUSIIL benchmark. The full 50-feature Random Forest achieved perfect benchmark performance under both random and domain-disjoint splits, while an 18-feature raw-URL representation achieved about 99.7 percent accuracy. SHAP identified a compact eight-feature representation that retained 99.73 percent accuracy on the domain-disjoint test while reducing the feature count by 84 percent. The selected feature set also showed substantial stability across five model seeds.

The results support a practical conclusion: strong benchmark detection does not require the full feature set, but benchmark accuracy alone should not be interpreted as proof of real-world robustness. Future work should extend the compact feature benchmark to temporally separated and independently sourced datasets, evaluate calibration and adversarial robustness, and test whether explanation rankings remain stable as phishing campaigns evolve.

VII. LIMITATIONS

Only the PhiUSIIL dataset was used for the main experiments. The study therefore does not establish cross-dataset generalization. The domain-disjoint split is not a temporal deployment test. The analysis also focuses on tabular feature-based models and does not compare transformer or character-level neural architectures. Finally, SHAP stability was evaluated using five seeds and limited explanation samples rather than a formal confidence interval framework.

TABLE I. PRINCIPAL CLASSIFICATION RESULTS

Split	Features	Model	Accuracy	Precision	Recall	F1
Random	50	RF	100.000%	100.000%	100.000%	100.000%
Domain disjoint	50	RF	100.000%	100.000%	100.000%	100.000%
Random	18	RF	99.749%	99.619%	99.944%	99.781%
Domain disjoint	18	RF	99.732%	99.647%	99.887%	99.767%
Random	8	RF	99.729%	99.633%	99.893%	99.763%
Domain disjoint	8	RF	99.730%	99.680%	99.851%	99.766%
Domain disjoint	18	HGB	99.788%	99.668%	99.964%	99.816%

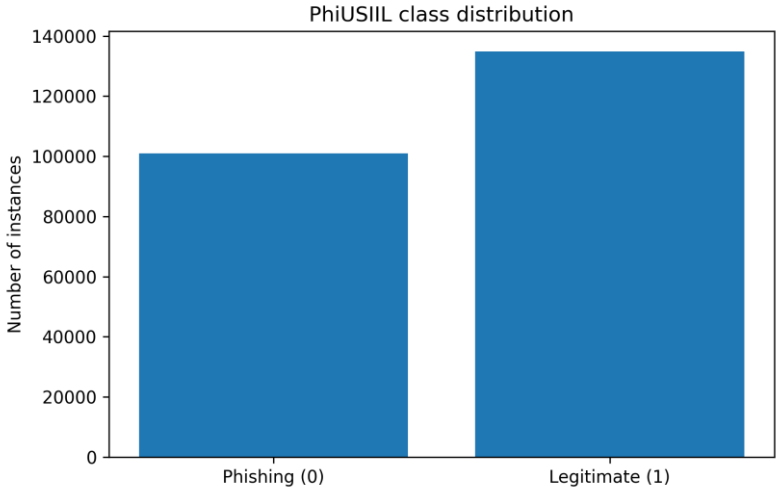

Fig. 1. Class distribution in the PhiUSIIL dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

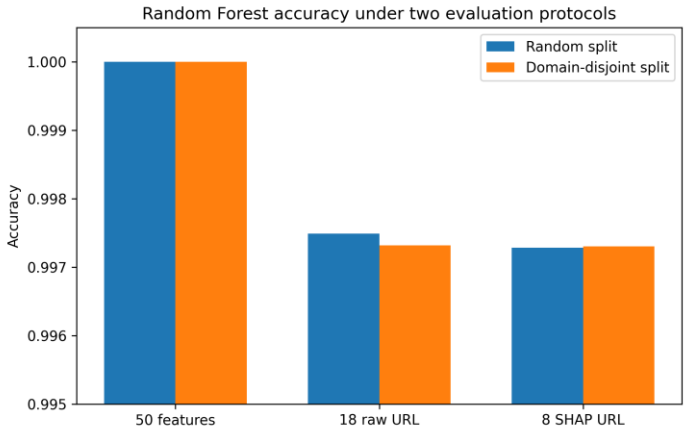

Fig. 2. Random Forest accuracy for the evaluated feature representations.

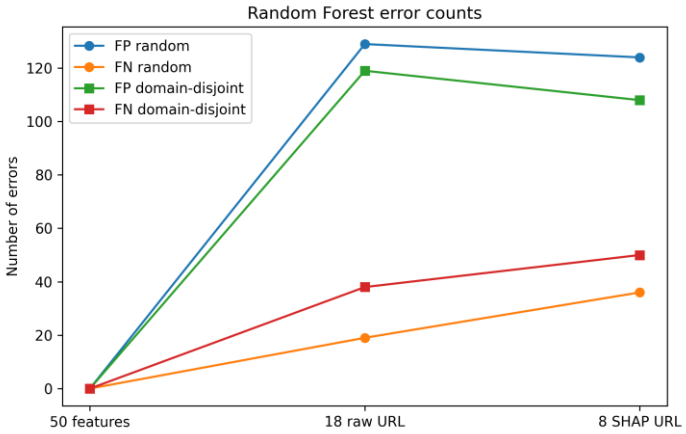

Fig. 3. False positive and false negative counts under the two evaluation protocols.

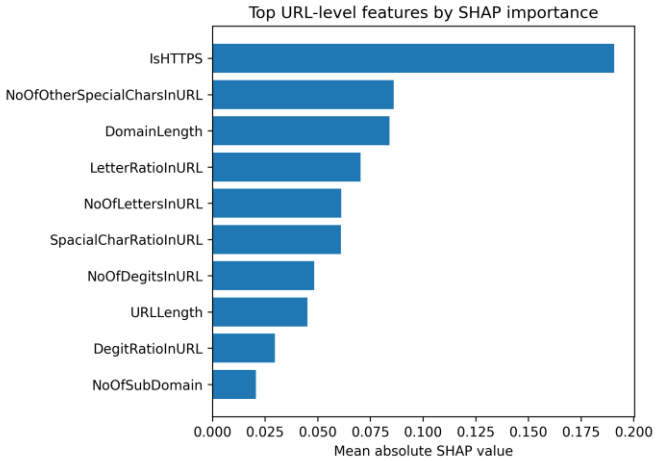

Fig. 4. Mean absolute SHAP importance for the most influential raw-URL features.

REFERENCES

[1] M. C. Calzarossa, P. Giudici, and R. Zieni, Phishing or not phishing A survey on the detection of phishing websites, IEEE Access, vol. 11, pp. 18499 to 18519, 2023, doi: 10.1109/ACCESS.2023.3247135.

[2] A. Aljofey et al., A systematic literature review on phishing website detection techniques, Journal of King Saud University Computer and Information Sciences, vol. 35, no. 2, pp. 590 to 611, 2023, doi: 10.1016/j.jksuci.2023.01.004.

[3] An explainable feature selection framework for web phishing detection with machine learning, Data Science and Management, vol. 8, no. 2, pp. 127 to 136, 2025, doi: 10.1016/j.dsm.2024.08.004.

[4] Hybridizing Explainable AI for intelligent feature extraction in phishing website detection, Electronics, vol. 15, no. 2, 350, 2026.

[5] A. Prasad and S. Chandra, PhiUSIIL Phishing URL Website, UCI Machine Learning Repository, 2024, doi: 10.1016/j.cose.2023.103545.