INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
ISSN: 3106-7980 : 3106-7972
Volume 3, Issue 3, 2026

A CROSS-DATASET GENERALIZATION AND EXPLAINABILITY STUDY OF MACHINE LEARNING PHISHING WEBSITE DETECTORS (CAN PHISHING DETECTORS BE TRUSTED IN THE WILD?)

Amoon Shamoon[*1], Khalid Hamid[2], Hafiz Muhammad Rashid[3], Zeeshan Khizar[4], Muhammad Tayyab[5]

[*1,2,3,4,5] *Department of Computer Science and IT, Superior University Lahore, Lahore, Punjab, Pakistan*

[*1]amoonsaab@gmail.com,[2]khalid6140@gmail.com,[3]rashid.liaquat68@gmail.com, [4]zeeshanchaahal@gmail.com, [5]su92-msisw-f25-003@superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.21738657

Keywords
Phishing Detection, Machine Learning, Cross-Dataset Generalization, Explainable AI, SHAP, Cybersecurity.

Article History
Received: 19 May 2026
Accepted: 12 July 2026
Published: 31 July 2026

Corresponding Author: *
Amoon Shamoon

Abstract

It is widely reported that machine learning-based phishing website detectors are able to detect phishing sites with 95–99% accuracy; however, the vast majority of these results are achieved by training and testing a detector using only one phishing website benchmark set. This paper empirically explores whether such models can be applied to unseen phishing data. On two open-source benchmark datasets (the UCI Phishing Websites Dataset and the PhiUSIIL Phishing URL Dataset), we train Random Forest and XGBoost classifiers with five feature-aligned, dataset-agnostic features: IP-address usage, URL length, iframe usage, redirection behaviour, and hyphenated domain names. For both models, the same-dataset (baseline) evaluation results in an accuracy of 76-78%. But if this model is tested on the other dataset, the accuracy drops to 40–48%, and the ROC-AUC is under 0.5, with the Matthews Correlation Coefficient (MCC) becoming negative in both cross-dataset evaluations, meaning that the model fails to outperform random guessing. In addition, SHAP (SHapley Additive exPlanations) analysis indicates that the most influential feature varies across datasets: URL length is most influential on the UCI dataset, while iframe usage is most influential on PhiUSIIL. To validate results, the analysis was performed using a 5-fold cross-validated approach. The results show that high single-dataset accuracy does not necessarily mean high real-world accuracy, as well as that the importance of features is not fixed across data sources. The implications for the design of more generalizable phishing detection systems are discussed, and directions for future work are outlined.

1 INTRODUCTION

Phishing continues to be one of the most common and costliest types of cybercrime, designed using deceptive Web pages and URLs to trick users into providing personal identification and confidential data [1, 23, 33]. Nearly all phishing web detection studies have shown that machine learning (ML) is the biggest and most successful method for automatically detecting phishing websites, and many studies show that classical machine learning algorithms like the random forest, support vector machines (SVM), and gradient boosting

algorithms are extremely accurate, with often just 5% to 10% mistakes, and in some cases, they are almost 100% accurate [34, 35, 36]. This high performance is, however, almost universally reported with a single train–test split taken from a single benchmark dataset, most often the UCI Phishing Websites Dataset. By definition, real-world phishing detection systems have to have encountered new sites and URLs in an ever-changing threat landscape during training. The extent to which models tested on one static benchmark can be used to predict new phishing patterns has been relatively little studied. This study targets that directly. Two research questions are asked: (1) What is the deterioration in classification accuracy when a phishing detection model is used in a different open-source dataset, and (2) are explainable AI techniques able to provide an understanding of the deterioration? In both cases, we provide empirical answers to the two questions by employing two benchmark datasets that are publicly available, two popular classifiers, and feature importance measures extracted from SHAP-based feature analysis [37, 38, 39].

The rest of this paper is organized into two sections. The related works are reviewed in Section 2. In Section 3, the research gap is summarized. The datasets, feature alignment procedure and models are described in Section 4. The experimental setup is described in Section 5. The results of Section 6 are presented and discussed. The paper is concluded and future work is indicated in Section 7. Section 6 presents and discusses the results. Section 7 concludes the paper and outlines future work.

2 Related Work

Most of the phishing detection research uses classical machine learning algorithms and they use the UCI Phishing Websites Dataset as the benchmark for the evaluation of the models, with the majority reporting an accuracy of 95-99% using feature selection and hyperparameter tuning, followed by ensemble stacking methods [26, 27, 9, 17, 11]. There are several publications that specifically consider the class imbalance problem when applying SMOTE and its variants together with boosting classifiers to further enhance the performance reported [24, 13, 14]. These studies show that classical ML models can perform well on the individual benchmark datasets but are almost always tested using a single train–test split from the same dataset [40, 41, 43].

A parallel line of study utilizes deep learning architectures like those based on Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and hybrid CNN–LSTM to learn more complex patterns in the strings of URLs and webpage content [4, 2, 1, 22]. Recently, transformer-based models such as BERT and RoBERTa have been utilized for phishing URL classification [31, 5, 6] and research that investigates the use of large language models (LLMs) using prompt engineering or fine-tuning has emerged as the latest trend in this research stream [20]. A similar structural method is used to find zero-day phishing pages that are similar in appearance to the legitimate brands [1a, 15]. However, these generally need more computational resources and, like their classical counterparts, are usually validated only on a single dataset [42, 44].

Given that high-performing ML models are "black boxes," an increasing number of papers have sought to make them more transparent by adding explainable AI methods, such as SHAP and LIME, which explain how features contribute to model predictions [7, 12, 28, 8]. Most of the studies that focus on XAI do so for a single dataset, but do not consider the consistency of feature importance explainability across datasets.

Another (but related) research direction looks at how robust phishing detectors are to carefully constructed adversarial inputs. Even the most accurate models suffer from low detection rates in the presence of small real-world changes to the URL or webpage features, as shown in studies like Kulkarni et al. [19] and Song et al. [25]. Adversarial robustness is conceptually linked to generalization, but it considers a different threat model (deliberate evasion) than the natural distributional differences between different datasets, and, thus, is not considered here [45].

Recently, several studies, although few in number, have begun to challenge whether reported

accuracy is a true measure of their detection capability, or a specific overfitting of the dataset they are based on. Mia et al. [3] assessed how well eight machine learning models performed on two open phishing datasets, showing that the accuracy of the models [46], when tested on the other dataset, decreased from 92-99% to 51-59%, and that the feature contributions varied in significance between the two datasets. A cross-dataset performance recovery-related ensemble transfer-learning approach is proposed using a feature-space alignment technique [10]. This directly inspires the present study, which expands on this line of research by combining a new dataset pair: UCI Phishing Websites Dataset and PhiUSIIL Phishing URL Dataset [16] (which is larger and more recent), and a feature-transferability analysis based on a common set of features that is explicitly dataset-agnostic.

This gap is also reflected in the wider field, as evidenced by recent research efforts to address cross-dataset generalization and explainability in phishing detection research [18, 31, 28], and both of these remain open challenges in phishing detection research for 2024–2025.

Summary of the Research Gap

Table 1 summarizes the positioning of the present study relative to the three main strands of related work discussed above.

Table 1 Positioning of the present study relative to related work

Study Type	Single-Dataset Accuracy	Cross-Dataset Eval.	SHAP Explainability
Classical ML studies	Yes	No	No
Deep learning / transformer studies	Yes	No	No
XAI-focused studies	Yes	No	Yes (single-dataset)
Cross-dataset studies	Yes	Yes	Partial
This study	Yes	Yes	Yes (comparative)

4 Materials and Methods

4.1 Datasets

Two publicly available, open-source phishing datasets were used:

- 17 Features: lexical and host-based URLs.10,000 labelled URLs with 5000 phishing and 5000 legitimate URLs.
- PhiUSIIL Phishing URL Dataset: 235,795 labelled URLs (134,850 legal, 100,945 phishing) with 54 features extracted from the URL and the source code of the webpage. For computational parity with the UCI dataset, a class-balanced random sample of 10,000 URLs (5,000 phishing and 5,000 legitimate) was sampled.

There is no paid access to either dataset and they are distributed under open licenses.

4.2 Exploratory Data Analysis

In order to eliminate class imbalance as a confounding effect in the cross-dataset comparison, both the phishing and legitimate datasets were balanced to 5,000/5,000 (phishing/legitimate) in this study, as can be seen in Figure 1.

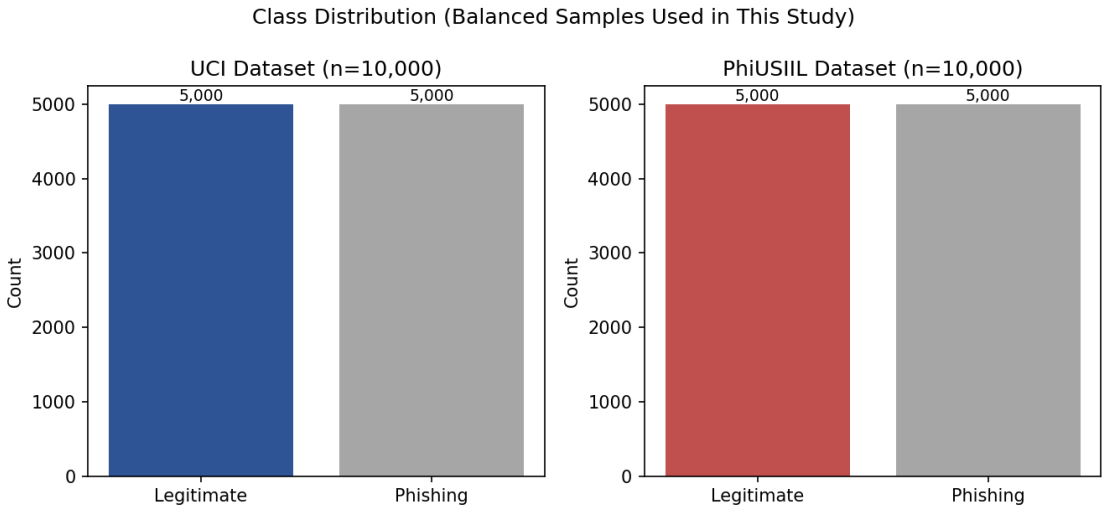

Fig. 1 Class distribution of the balanced UCI and PhiUSIIL samples used in this study.

Figure 2 and Table 2 compare the prevalence of each of the five aligned features between the two datasets — that is, the proportion of URLs for which each binary feature equals 1.

Table 2 Feature prevalence (proportion with feature = 1) by dataset

Feature	UCI	PhiUSIIL
Have_IP	0.0055	0.0035
URL_Length_bin	0.7734	0.5764
iFrame	0.0909	0.2963
Redirects	0.1053	0.1388
Prefix_Suffix	0.0932	0.1970

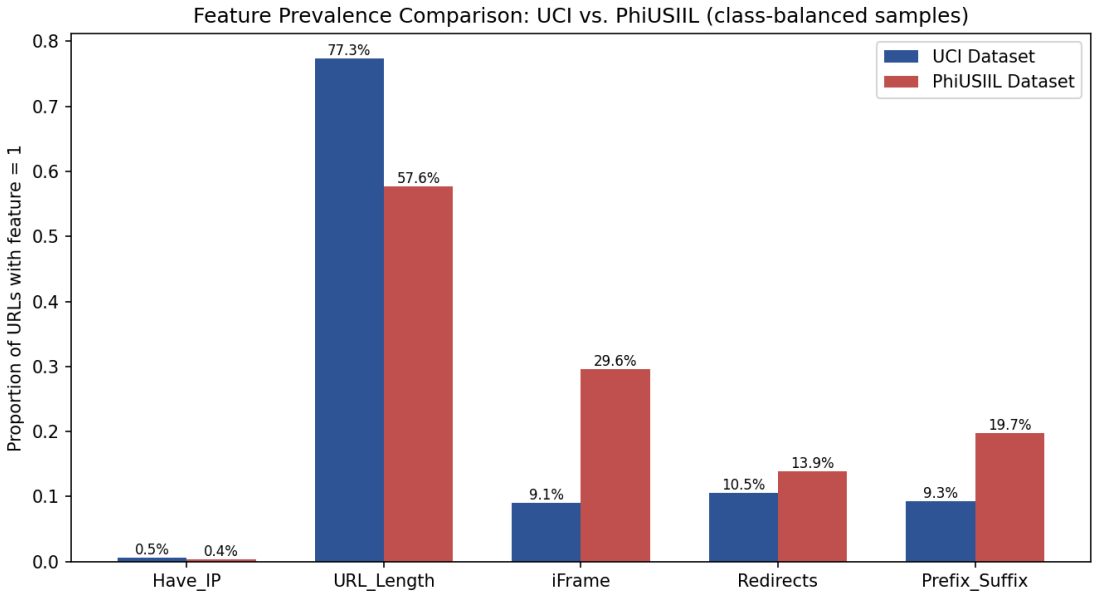

Fig. 2 Prevalence of each aligned feature (proportion of URLs with feature = 1) in the UCI vs. PhiUSIIL datasets.

There are clear differences in the usage of the following features: URLs containing the iframe feature are present in 29.6% of the PhiUSIIL URLs while only 9.1% of the UCI URLs have this feature; URLs with long URLs (URL_Length_bin) are present in 77.3% of the UCI URLs, but only 57.6% of the PhiUSIIL URLs contain this feature; the presence of hyphenated domains (Prefix_Suffix) is nearly twice as high in the PhiUSIIL (19.7%) as in the UCI (9.3%). This shift in feature statistics in the underlying data, even though both features are defined the same, is an early, data-level-understanding of why models trained on one dataset do not generalize to another, as will be seen in the results of explainability in Section 6.8.

4.3 Feature Alignment

Since the two datasets were built separately, the features did not have identical definitions and needed to be aligned manually to find a common, semantically consistent set of features. To both datasets, five conceptually equivalent binary-encoded features were kept:

Table 3: Common feature set used for cross-dataset alignment

Aligned Feature	UCI Source Column	PhiUSIIL Source Column
Have_IP	Have_IP	IsDomainIP (binary)
URL_Length_bin	URL_Length (binary)	URLLength (median-split)
iFrame	iFrame	NoOfiFrame > 0
Redirects	Web_Forwards	NoOfURLRedirect > 0
Prefix_Suffix	Prefix/Suffix	Domain contains '-'

All other features (non-overlapping) have been excluded from this study for a fair comparison

across datasets (an apples-to-apples comparison), and this has been recognized as a limitation in Section 7.

4.4 Models

The classifiers used were two supervised classifiers: the first was Random Forest (200 estimators) and the second was XGBoost (200 estimators, log-loss objective). They are both popular and tree-based ensemble methods for which interpretation can be performed by computing an exact TreeExplainer. For reproducibility, a fixed random seed (42) was used throughout.

4.5 Explainability

To compare feature importance rankings between the UCI-trained and PhiUSIIL-trained models, SHAP values were calculated for the models trained on the full aligned feature set of each of the datasets, using the TreeExplainer method.

4.6 Overall Research Workflow

The overall research process is summarized in Figure 3, which encompasses the entire end-to-end process, from data collection to standardization of data labels, alignment of features, balancing of classes, training the model, evaluating the model with baseline and cross-dataset methods, calculating metrics, and conducting research explainability analysis using SHAP.

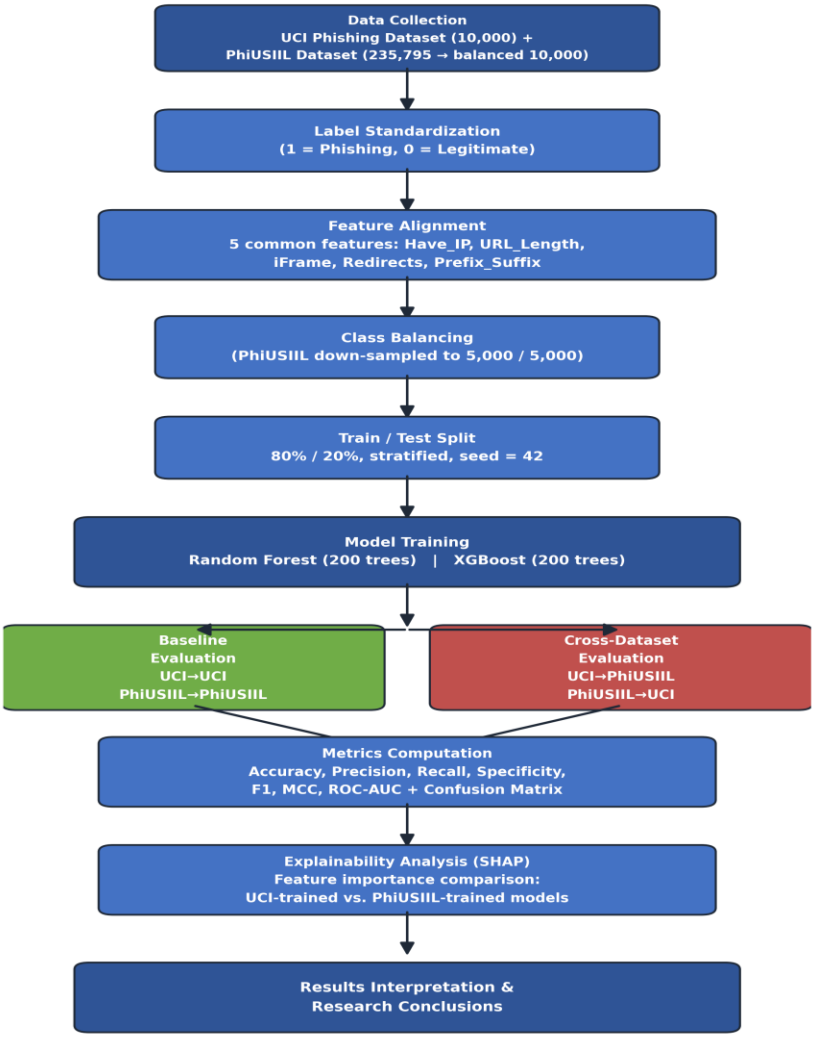

Fig. 3 End-to-end research workflow used in this study.

5 Experimental Setup

All experiments are conducted in Python, with scikit-learn, XGBoost and the SHAP library, without using GPU or paid compute resources, running in a cloud-based Jupyter/Colab environment. An 80/20 stratified train–test split was used for same-dataset (baseline) evaluation. Four experimental scenarios were tested with each of the classifiers:

1. UCI → UCI (baseline): train and test on UCI (80/20 split).
2. PhiUSIIL → PhiUSIIL (baseline): train and test on the balanced PhiUSIIL sample (80/20 split).
3. UCI → PhiUSIIL (cross-dataset): train on the full UCI dataset, test on the full balanced PhiUSIIL sample.
4. PhiUSIIL → UCI (cross-dataset): train on the full balanced PhiUSIIL sample, test on the full UCI dataset.

Evaluated metrics reported are: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, Matthews Correlation Coefficient (MCC), and ROC-AUC. MCC is included because it is a balanced statistic among all four quadrants of the confusion matrix and therefore it would be informative even in the presence of class imbalance, and it ranges from −1 (total

disagreement / inverse prediction) to +1 (perfect prediction), hence, it is particularly suitable to capture the type of generalization failure explored in this paper. Furthermore, to ensure the statistical robustness of the baseline results, 5-fold stratified cross-validation was also conducted with the two classifiers on both datasets separately, for which the mean and the SD across folds were reported for each metric.

6 Results and Discussion

6.1 Baseline (Same-Dataset) Performance

Table 4 reports same-dataset baseline performance. Both classifiers achieved comparable results, given the small (5-feature) common feature space: 78.15% accuracy on UCI and 76.25% accuracy on PhiUSIIL.

Table 4 Baseline (same-dataset) results

Scenario	Accuracy	Precision	Recall	F1	ROC-AUC
RF: UCI → UCI	0.7815	0.9763	0.5770	0.7253	0.7923
RF: PhiUSIIL → PhiUSIIL	0.7625	0.6909	0.9500	0.8000	0.8437
XGB: UCI → UCI	0.7815	0.9763	0.5770	0.7253	0.7923
XGB: PhiUSIIL → PhiUSIIL	0.7625	0.6909	0.9500	0.8000	0.8437

6.2 Cross-Dataset Generalization

The cross-dataset performance is presented in Table 5. At both classifiers, there was a decrease of 30-38 points in accuracy when compared with a baseline trained on the same dataset. What is interesting is that three of the four cross-dataset scenarios had ROC-AUC below 0.5, meaning that the models did worse than random guessing at predicting a new dataset that they were not trained on.

Table 5 Cross-dataset generalization results

Scenario	Accuracy	Precision	Recall	F1	ROC-AUC
RF: UCI → PhiUSIIL	0.4034	0.4305	0.5980	0.5006	0.3860
RF: PhiUSIIL → UCI	0.4776	0.4878	0.8926	0.6308	0.3210
XGB: UCI → PhiUSIIL	0.4032	0.4303	0.5976	0.5003	0.4038
XGB: PhiUSIIL → UCI	0.4774	0.4876	0.8922	0.6306	0.3215

This is an empirical confirmation of the generalization failure reported before by Mia et al. [3] for a new set of data, in a new pairing. It suggests model accuracy does not necessarily translate to real-world reliability, which should be determined by other indicators. This collapse is clearly visible in the visualization in Figure 4: the two baseline scenarios lie well above the 50% random guess line, whereas the two cross-dataset scenarios lie towards or below the 50% random guess line.

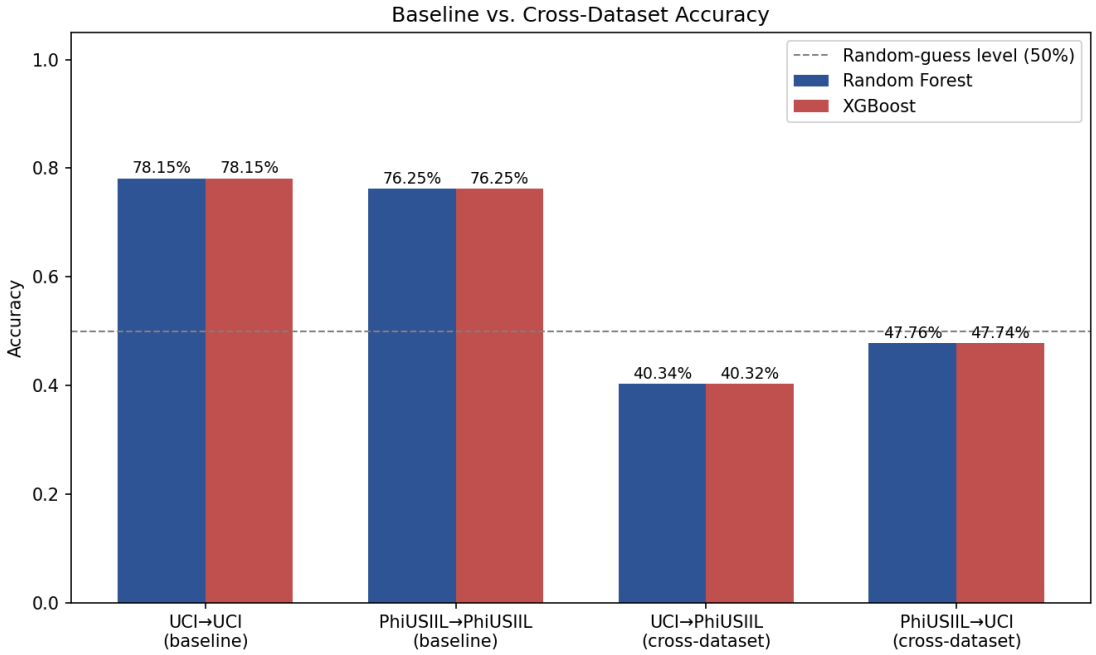

Fig. 4 Accuracy comparison across all four evaluation scenarios for both classifiers, relative to the random-guess baseline (50%).

6.3 Confusion Matrix Analysis

The confusion matrices for the XGBoost classifier in all four scenarios are shown in Figure 5. The baseline scenarios have the majority of correct classifications along the diagonal. For the cross-dataset scenarios, misclassifications rise significantly and become more diagonally distributed, especially for the UCI→PhiUSIIL scenario, which provides visual confirmation of a previous finding using the accuracy and ROC-AUC metrics that the system is failing.

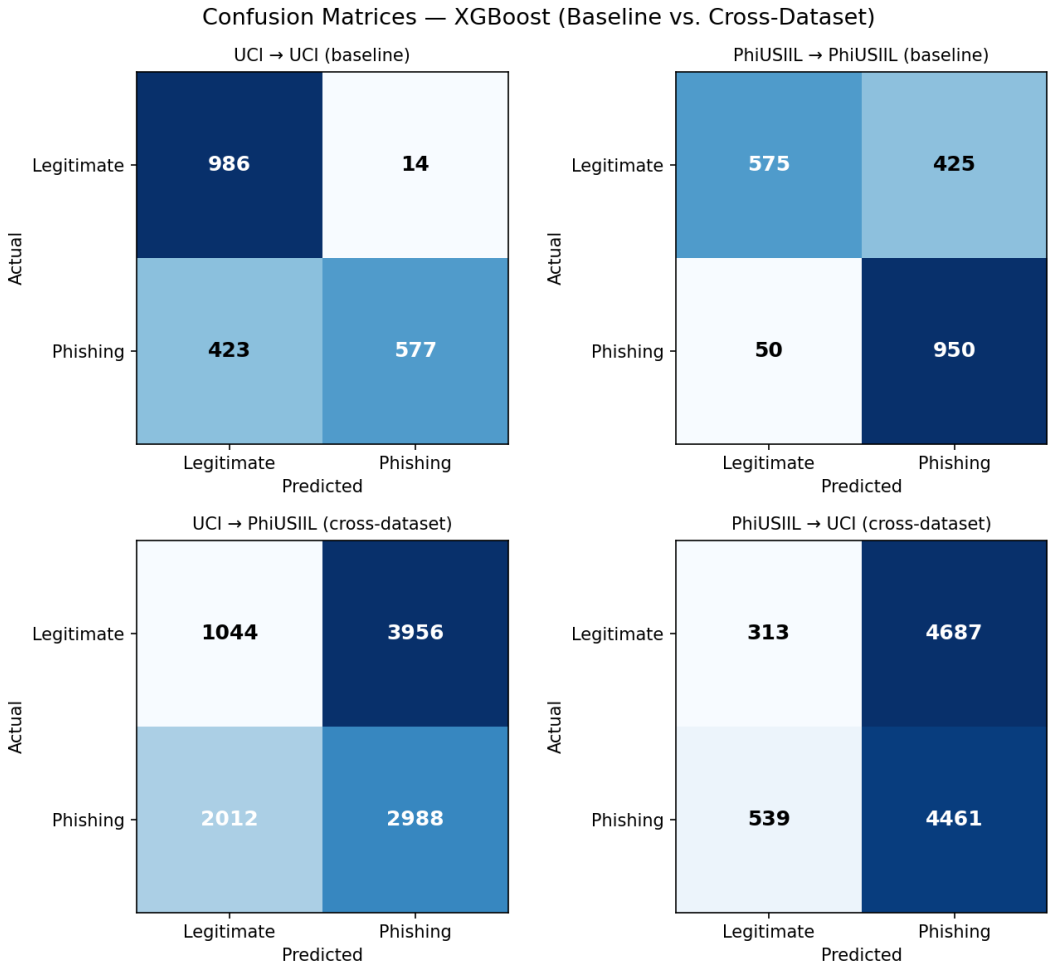

Fig. 5 Confusion matrices for the XGBoost classifier: baseline scenarios (top row) vs. cross-dataset scenarios (bottom row).

6.4 ROC Curve Analysis

All four XGBoost scenarios are plotted over each other in Figure 6. The baseline curves (UCI→UCI, PhiUSIIL→PhiUSIIL) are well above the diagonal line of the random classifier, corresponding to their ROC-AUC values of 0.79 and 0.84, respectively. Conversely, the curves in the cross-dataset plots cluster tightly along or slightly below the diagonal, and visually demonstrate that the cross-dataset classification result does not hold up to the classifiers' ranking ability — not their thresholded accuracy.

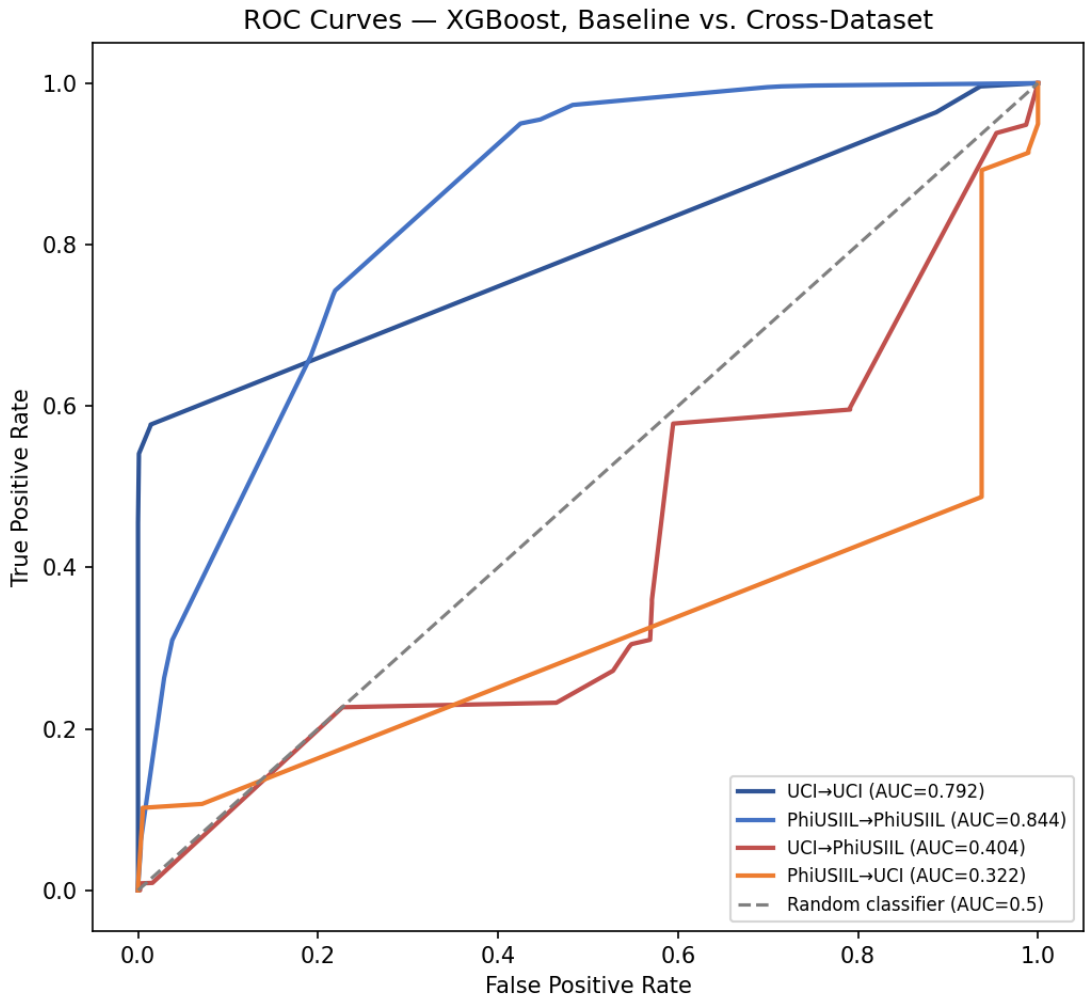

Fig. 6 ROC curves for the XGBoost classifier across baseline and cross-dataset scenarios.

6.5 Full Metric Comparison

Finally, Figure 7 shows a single comprehensive view of the collapse in the performance of the XGBoost classifier across all five evaluation metrics (Accuracy, Precision, Recall, F1, ROC-AUC) side by side for all four scenarios. Remarkably, the Precision (0.49) and sub-random ROC-AUC (0.32) values do not support the model's discriminative power but rather its tendency to over-predict the phishing class, despite the high Recall (0.89) in the PhiUSIIL→UCI cross-dataset scenario.

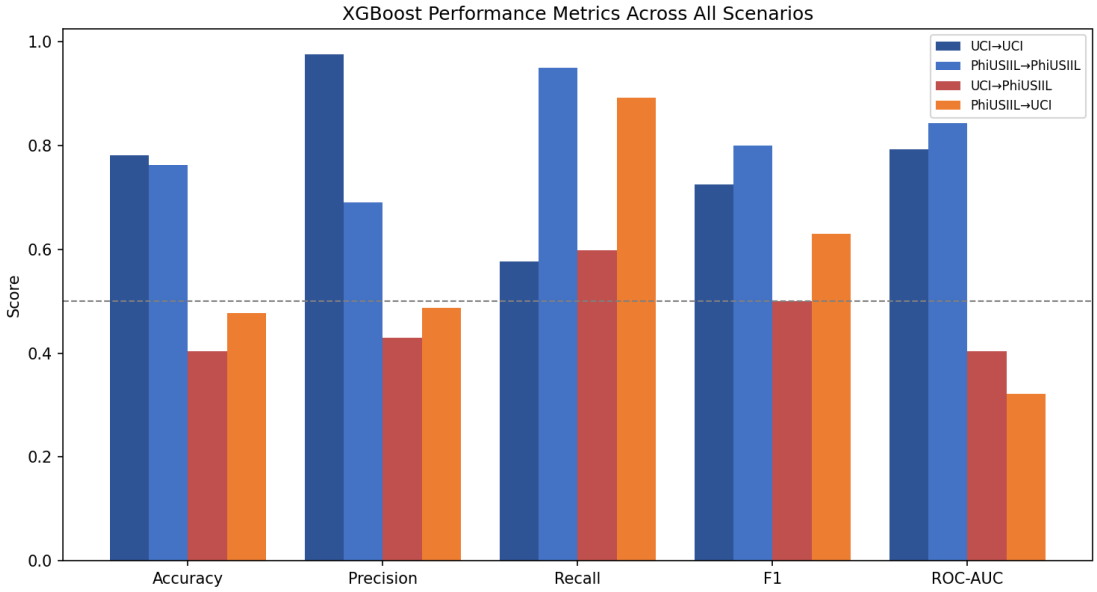

Fig. 7 Full metric comparison (Accuracy, Precision, Recall, F1, ROC-AUC) for the XGBoost classifier across all four scenarios.

6.6 Extended Metrics: Specificity and Matthews Correlation Coefficient

The Specificity and MCC for the XGBoost classifier are given in Table 6 for all four scenarios. The most interesting results from this study are the strong positive results of MCC for both baseline datasets (0.617 and 0.566), but the negative results for both cross-dataset settings (−0.210 for UCI→PhiUSIIL and −0.081 for PhiUSIIL→UCI). The negative MCC makes the prediction of the classifier, on the whole, worse than random, and trends the other way from the truth, which is much more alarming than the sub-random ROC-AUC. Specificity also drops off the floor, from the UCI baseline 98.6% to 6.3% (PhiUSIIL → UCI). This indicates that the model trained on PhiUSIIL fails to recognise 93.7% of the legitimate UCI URLs as legitimate.

Table 6 Extended metrics (Specificity, MCC) for the XGBoost classifier

Scenario	Specificity	MCC	Interpretation
UCI → UCI (baseline)	0.9860	0.6170	Strong agreement
PhiUSIIL → PhiUSIIL (baseline)	0.5750	0.5663	Moderate–strong agreement
UCI → PhiUSIIL (cross-dataset)	0.2088	−0.2101	Worse than random
PhiUSIIL → UCI (cross-dataset)	0.0626	−0.0810	Worse than random

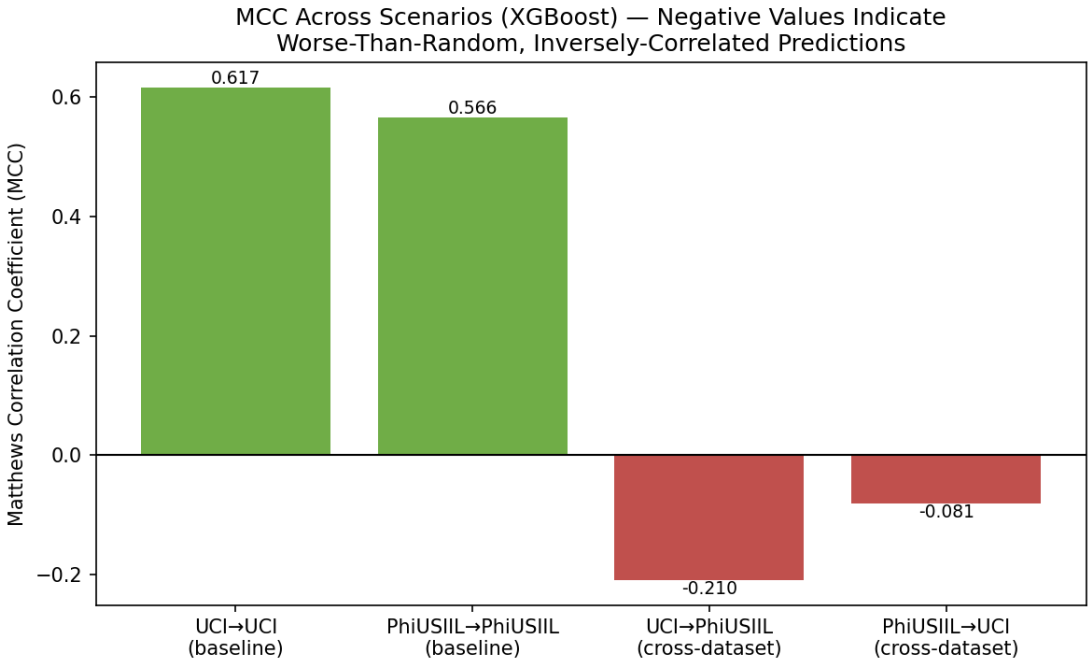

Fig. 8 Matthews Correlation Coefficient (MCC) across all four scenarios (XGBoost). Negative bars indicate predictions that are worse than random.

6.7 Statistical Robustness: 5-Fold Cross-Validation

To ensure the results obtained in Section 6.1 are not a result of a single favourable 80/20 split, the data sets were used separately to perform 5-fold stratified cross-validation. The mean ± SD of the folds is reported in Table 7 and shown in Figure 9. The low standard deviations (all ≤ 0.02) indicate that there is no noticeable variation in performance levels at baseline, and that the performance level is not sensitive to the specific train–test split used, which is important for confidence in the validity of the subsequent cross-dataset comparison.

Table 7: 5-fold stratified cross-validation results (mean ± std. dev.)

Model / Dataset	Accuracy	F1	MCC
Random Forest / UCI	0.7887 ± 0.0087	0.7372 ± 0.0143	0.6274 ± 0.0134
XGBoost / UCI	0.7887 ± 0.0087	0.7372 ± 0.0143	0.6274 ± 0.0134
Random Forest / PhiUSIIL	0.7510 ± 0.0085	0.7937 ± 0.0057	0.5514 ± 0.0142
XGBoost / PhiUSIIL	0.7510 ± 0.0085	0.7937 ± 0.0057	0.5514 ± 0.0142

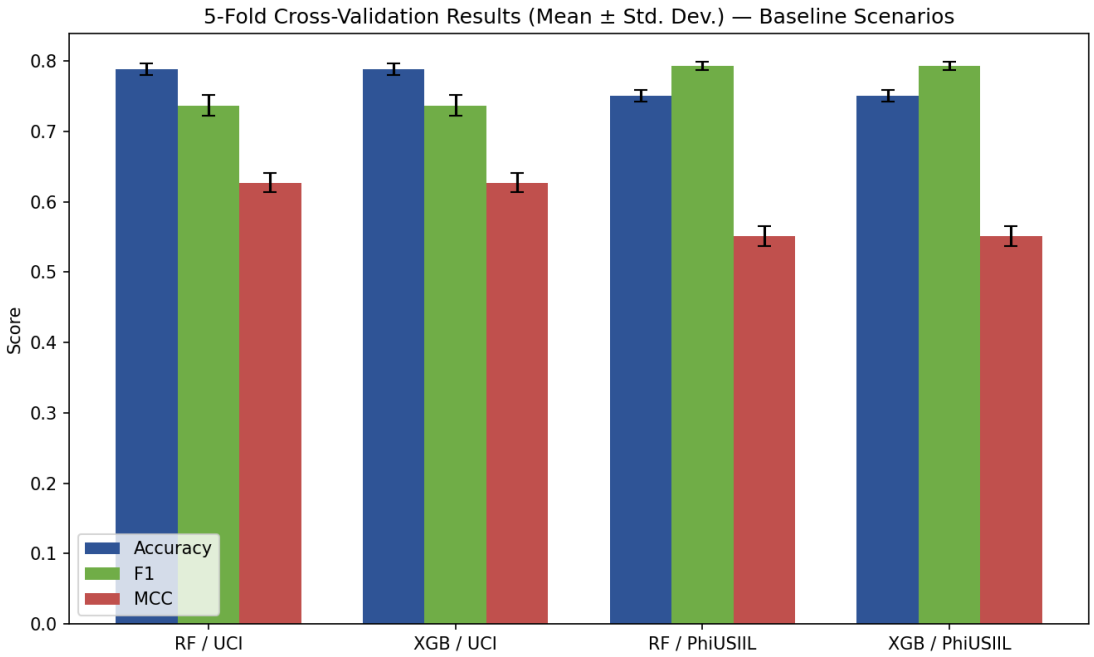

Fig. 9 5-fold cross-validation results (mean ± std. dev.) for Accuracy, F1, and MCC, confirming the stability of baseline performance.

6.8 Explainability Analysis

Figure 10 shows the mean absolute SHAP value importance of the features for the XGBoost models trained on the UCI dataset and the PhiUSIIL dataset. For UCI, the length of the URL is the most important feature. However, the most important feature on the PhiUSIIL set is different: the use of the iframe tag, and not the length of the URL. The corresponding SHAP summary (beeswarm) plots of the two datasets are depicted in Figures 11 and 12, respectively.

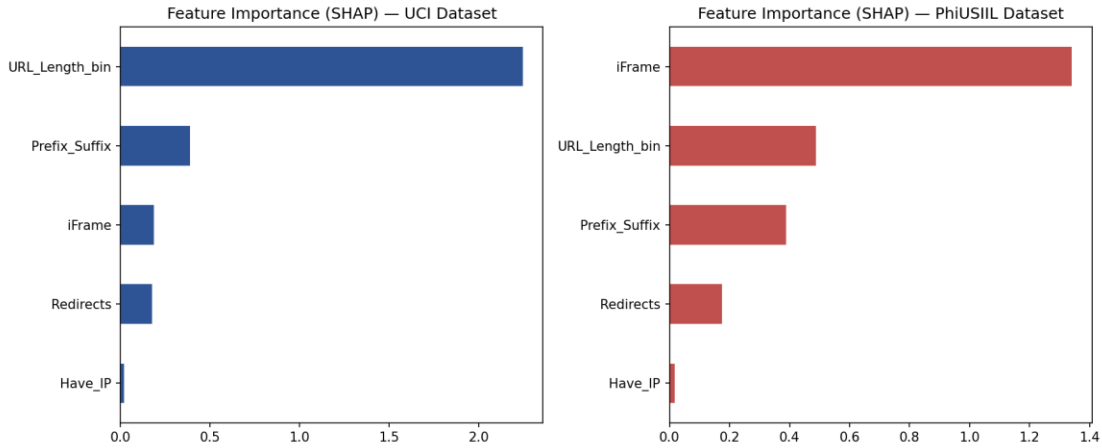

Fig. 10 Mean absolute SHAP feature importance: UCI-trained model (left) vs. PhiUSIIL-trained model (right).

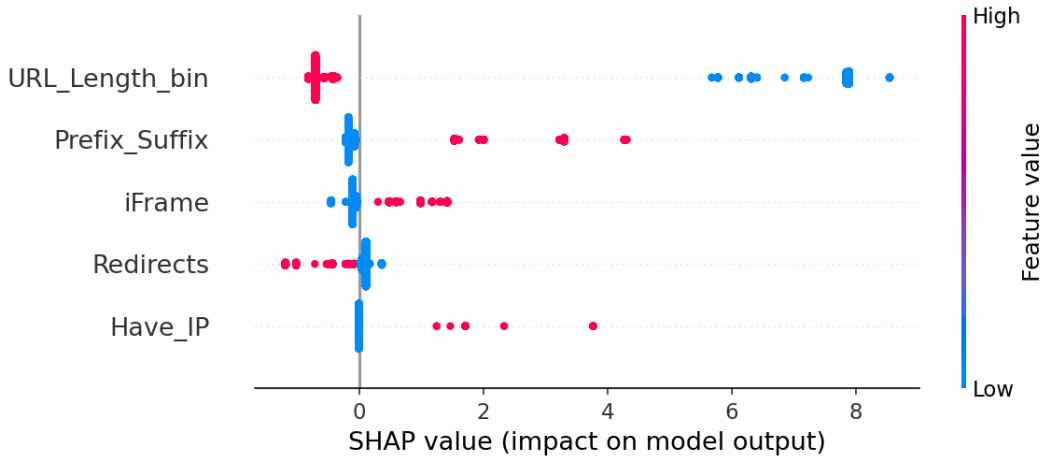

Fig. 11 SHAP summary plot for the UCI-trained model.

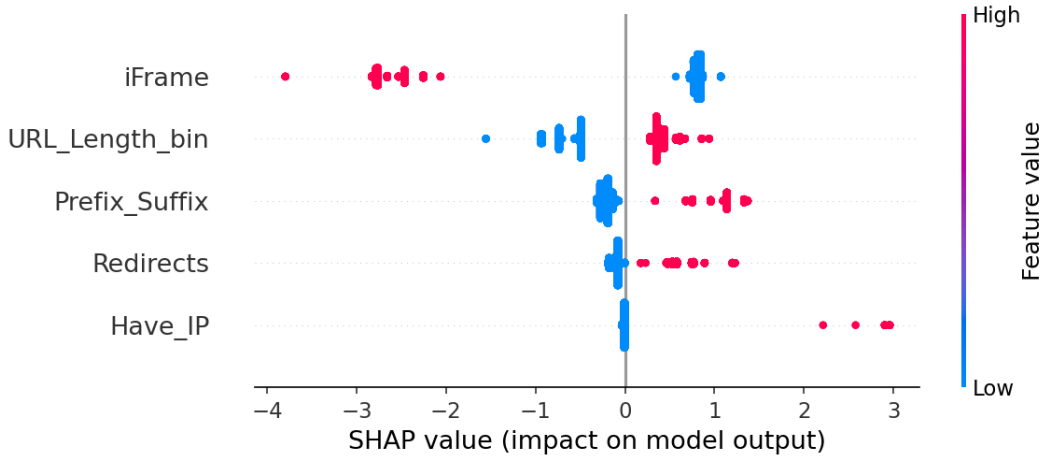

Fig. 12 SHAP summary plot for the PhiUSIIL-trained model.

It is apparent from this change in feature importance that the model is not just learning a stable, universal concept of "phishing-like" behaviour, but rather a pattern that is specific to the statistical idiosyncrasies of the dataset it was trained on. Other features, like Have_IP and Prefix_Suffix, had a relatively low and stable importance across both datasets and could potentially be more useful and transferable features than the dataset-specific features like URL length or iframe usage.

7 Conclusion and Future Work

In this study, the authors empirically show that, although machine learning-based phishing detectors have 76–78% same-dataset accuracy on a small, dataset-agnostic feature set, they fall to as low as 30–38 percentage points lower in terms of accuracy, with ROC-AUC dropping below chance level and Matthews Correlation Coefficient becoming negative, when tested on an independent, unseen phishing dataset. Perhaps most striking is the negative MCC score, which demonstrates that in the cross-dataset scenario, the classifiers' predictions are not only incorrect, but actually counter to the actual labels. The failure is connected to real distributional shift in feature statistics across datasets and in the features used by the model the most, through explainability analysis using SHAP, with exploratory analysis of

feature prevalence. The baseline (same-dataset) results are statistically stable via 5-fold cross-validation, which helps to rule out the possibility that the observed cross-dataset collapse is a data-specific artifact and further supports the argument that it is a genuine generalization failure. The results support the claim that single-dataset benchmarking grossly exaggerates the true effectiveness of phishing detection systems in the wild and that explainability tools like SHAP are useful not just for interpretability, but for the diagnosis of failures to generalize.

Practical Implications. Practitioners can take these results as guidance not to blindly trust an accuracy claim of a phishing detector in the public benchmark when deciding to deploy it in the wild. When considering third-party or open-source phishing detection tools, organizations should consider requesting or independently measuring cross-dataset / out-of-distribution performance in addition to single-dataset accuracy rates, which they should interpret as an optimistic upper bound for the field. In the worst case, a poorly generalising detector can be counter-productive, misclassifying phishing and legitimate sites more frequently than a naive baseline would do, as in the case of the negative MCC observed in this study.

Limitations and Threats to Validity. The true accuracy of either of the original datasets may be greater than reported by this study, because the datasets were chosen for this study with a deliberately limited common feature set, which excluded additional, non-overlapping features. Limited external validity: only two datasets and two tree-based classifiers evaluated; results may not apply to deep learning or transformer-based classifiers, or other phishing datasets. The UCI dataset is also much older than PhiUSIIL, and part of the reason for the degradation observed may be related to "dataset differences" and "temporal drift." The PhiUSIIL dataset was down-sampled to be computationally comparable with UCI, which was not only class-balanced, but also loses a significant amount of data and may not be representative of the diversity of the original dataset. Last but not least, feature binarization (such as median-split over the length of the URL) adds some information loss compared to using the continuous feature value directly.

Future Work. Future work should (i) expand the common feature set by using more complex features directly from the raw URL and HTML content; (ii) test on other datasets and other deep learning/transformer-based models; (iii) implement some mitigation strategies such as dataset merging, transfer learning, or domain adaptation to recover cross-dataset performance; (iv) investigate adversarial robustness in addition to generalization; and (v) repeat this analysis with all native features of each dataset (not a subset) to estimate the extent to which the observed deterioration is due to feature reduction versus genuine distributional shift.

Abbreviations

The terms used include ML: Machine Learning, XAI: Explainable Artificial Intelligence, SHAP: SHapley Additive exPlanations, LIME: Local Interpretable Model-agnostic Explanations, RF: Random Forest, XGB: XGBoost, MCC: Matthews Correlation Coefficient, ROC-AUC: Receiver Operating Characteristic – Area Under the Curve, CV: Cross-Validation, URL: Uniform Resource Locator, HTML: HyperText Markup Language, TLD: Top-Level Domain, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, BERT: Bidirectional Encoder Representations from Transformers.

Availability of Data and Materials

The UCI Phishing Websites Dataset and the Phishing URL Dataset (PhiUSIIL) employed in the study are open-source, publicly available datasets. Processed/aligned data and analysis codes are available from the corresponding author on request.

Code Availability

The experiments were all conducted in Python with scikit-learn, XGBoost and SHAP. Scripts for the analysis are available on reasonable request from the corresponding author.

Author Contributions

A. Shamoon designed the study, conducted data collection, preprocessing, experimentation and analysis and drafted the manuscript with guidance and supervision of the Information Security department at Superior University.

Funding

This research received no external funding.

Competing Interests

The authors declare no competing interests.

Ethics Approval

This study did not involve human participants, human data, or animal subjects; it relies exclusively on publicly available, anonymized phishing/legitimate URL datasets. Ethics approval was therefore not required.

Acknowledgements

The authors thank the Department of Information Security, Superior University, for supervision and support of this research.

REFERENCES

[1] S. Abdelnabi, K. Krombholz, and M. Fritz, "VisualPhishNet: Zero-day phishing website detection by visual similarity," in Proc. ACM Conf. Comput. Commun. Security (CCS), 2020.

[2] S. Aslam, H. Aslam, A. Manzoor, H. Chen, and A. Rasool, "AntiPhishStack: LSTM-based stacked generalization model for optimized phishing URL detection," Symmetry, vol. 16, no. 2, p. 248, 2024.

[3] M. N. Mia, D. Derakhshan, and M. M. A. Pritom, "Can features for phishing URL detection be trusted across diverse datasets? A case study with explainable AI," in Proc. 11th Int. Conf. Networking, Systems, and Security (NSysS), 2024.

[4] Q. E. U. Haq, M. H. Faheem, and I. Ahmad, "Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks," Applied Sciences, vol. 14, no. 22, p. 10086, 2024.

[5] N. Q. Do, A. Selamat, H. Fujita, and O. Krejcar, "An integrated model based on deep learning classifiers and pre-trained transformer for phishing URL detection," Future Generation Computer Systems, vol. 161, pp. 269–285, 2024.

[6] M. Elsadig et al., "Intelligent deep machine learning cyber phishing URL detection based on BERT features," Electronics, vol. 11, no. 22, p. 3647, 2022.

[7] A. Fajar, S. Yazid, and I. Budi, "Enhancing phishing detection through feature importance analysis and explainable AI: A comparative study of CatBoost, XGBoost, and EBM models," arXiv preprint arXiv:2411.06860, 2024.

[8] S. S. Shafin, "An explainable feature selection framework for web phishing detection with machine learning," Data Science and Management, vol. 8, pp. 127–136, 2025.

[9] T. Kehkashan, M. Abdelhaq, A. S. Al-Shamayleh, N. Huda, I. A. Yaseen, A. A. Ahmed, and A. Akhunzada, "Explainable phishing website detection for secure and sustainable cyber infrastructure," Scientific Reports, vol. 15, 2025.

[10] [Authors to be verified], "Ensemble transfer learning for cross-dataset phishing detection," Franklin Open, Elsevier, 2026, PII: S2773186326000952.

[11] Z. Alamri, A. Alhuzali, B. Alsulami, and D. Alghazzawi, "Enhanced phishing website detection using optimized ensemble stacking models," Int. J. Advanced Computer Science and Applications, vol. 16, no. 8, 2025.

[12] S. Kavya and D. Sumathi, "Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection," Artificial Intelligence Review, vol. 58, no. 2, p. 50, 2024.

[13] A. Kulkarni, V. Balachandran, D. M. Divakaran, and T. Das, "From ML to LLM: Evaluating the robustness of phishing webpage detection models against adversarial attacks," ACM Digital Threats: Research and Practice, 2024.

[14] Y. Lin et al., "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in Proc. USENIX Security Symp., 2021.

[15] A. Awasthi and N. Goel, "Phishing website prediction using base and ensemble classifier techniques with cross-validation," Cybersecurity, vol. 5, no. 1, p. 22, 2022.

[16] M. S. I. Ovi, M. H. Rahman, and M. A. Hossain, "PhishGuard: A multi-layered ensemble model for optimal phishing website detection," arXiv preprint arXiv:2409.19825, 2024.

[17] A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," Computers & Security, vol. 136, p. 103545, 2024.

[18] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? A case study on phishing detection with large language models," Machine Learning and Knowledge Extraction, vol. 6, no. 1, pp. 367–384, 2024.

[19] M. Ramaiah et al., "Hybrid feature selection and class balancing for phishing detection," Applied Mathematics & Information Sciences, vol. 18, no. 6, pp. 1481–1493, 2024.

[20] A. U. Rehman, I. Imtiaz, S. Javaid, and M. Muslih, "Real-time phishing URL detection using machine learning," Engineering Proceedings, vol. 107, no. 1, p. 108, 2025.

[21] K. Sadaf, "Phishing website detection using XGBoost and CatBoost classifiers," in Proc. IEEE Int. Conf. Systems, Computing and Applications (ICSCA), 2023.

[22] K. Omari and A. Oukhatar, "Advanced phishing website detection with SMOTETomek-XGB: Addressing class imbalance for optimal results," Procedia Computer Science, vol. 252, pp. 289–295, 2025.

[23] F. Song et al., "Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers," Int. J. Intelligent Systems, vol. 36, no. 9, pp. 5210–5240, 2021.

[24] J. Sravanthi et al., "Phishing website detection using machine learning," AIP Conf. Proc., vol. 3361, no. 1, p. 050085, 2025.

[25] A. V. Ramana, K. L. Rao, and R. S. Rao, "Stop-Phish: An intelligent phishing detection method using feature selection ensemble," Social Network Analysis and Mining, vol. 11, no. 1, p. 110, 2021.

[26] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 672–694, 2021.

[27] C. Mayo, M. Tchuindjang, S. Brohi, and N. Ersotelos, "A temporally dynamic feature-extraction framework for phishing detection with LIME and SHAP explanations," Future Internet, vol. 18, no. 2, p. 101, 2026.

[28] G. Vrbančič, I. Fister, and V. Podgorelec, "Datasets for phishing websites detection," Data in Brief, vol. 33, p. 106438, 2020.

[29] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or not phishing? A survey on the detection of phishing websites," IEEE Access, vol. 11, pp. 18499–18519, 2023.

[30] P. Maneriker et al., "URLTran: Improving phishing URL detection using transformers," in Proc. IEEE Int. Conf. Big Data, 2021.

[31] O. K. Sahingoz, E. Buber, and E. Kugu, "DEPHIDES: Deep learning based phishing detection system," IEEE Access, vol. 12, pp. 8052–8070, 2024.

[32] K. Hamid, M. W. Iqbal, M. Aqeel, X. Liu, and M. Arif, "Analysis of Techniques for Detection and Removal of Zero-Day Attacks (ZDA)," in Ubiquitous Security, G. Wang, K.-K. R. Choo, J. Wu, and E. Damiani, Eds., Singapore: Springer Nature, 2023, pp. 248–262. doi: 10.1007/978-981-99-0272-9_17..

[33] K. Hamid, F. Naeem, M. Ibrar, A. M. Delshadi, F. Ahmed, M. Aamir, and G. Ahmad, "Behavior and characteristics of ransomware – A survey," Lahore, Pakistan, [Publication details to be verified].

[34] K. Hamid, M. W. Iqbal, M. Aqeel, T. A. Rana, and M. Arif, "Cyber security: Analysis for detection and removal of zero-day attacks (ZDA)," in Artificial Intelligence & Blockchain in Cyber Physical Systems. Boca Raton, FL, USA: CRC Press, 2023.

[35] S. Riaz et al., "Software development empowered and secured by integrating a DevSecOps design," Journal of Computing & Biomedical Informatics, p. 02, Mar. 2025, doi: 10.56979/802/2025.

[36] M. Ibrar et al., "Econnoitering data protection and recovery strategies in the cyber environment: A thematic analysis," Int. J. Electronic Crime Investigation, vol. 8, Dec. 2024, doi: 10.54692/ijeci.2024.0804216.

[37] K. Khaliq, N. Rahim, K. Hamid, M. Ibrar, U. Ahmad, and M. Ullah, "Ransomware attacks: Tools and techniques for detection," 2024, doi: 10.1109/ICCR61006.2024.10532926.

[38] Z. Zafar, K. Hamid, M. Kafayat, M. W. Iqbal, Z. Nazir, and A. Ghani, "AI-based cryptographical framework empowered network security," Journal of Jilin University (Engineering and Technology Edition), vol. 42, pp. 497–510, Apr. 2023, doi: 10.17605/OSF.IO/W69VT.

[39] K. Hamid et al., "Empowering robust security measures in Node.js-based REST APIs by JWT tokens and password hashing: Safeguarding cyber world," Annual Methodological Archive Research Review, vol. 3, May 2025, doi: 10.63075/w2nam443.

[40] M. Delshadi et al., "Empowerment of artificial intelligence (AI) in preventing and detecting ransomware: An analytical review," Spectrum of Engineering Sciences, vol. 3, no. 9, pp. 36–48, Sep. 2025.

[41] M. Delshadi, M. Zubair, K. Hamid, and M. W. Iqbal, "Preventing and detecting malicious activities during phishing attacks using AI-based sandbox bypass," Annual Methodological Archive Research Review, vol. 3, no. 8, pp. 249–261, Aug. 2025, doi: 10.63075/fahtkz35.

[42] M. N. Ahmad, M. N. Amin, K. Hamid, S. M. Rizwan, and S. A. A. Naqvi, "Enhanced IoT network security for network intrusion detection model based on machine learning technique," Annual Methodological Archive Research Review, vol. 3, no. 8, Aug. 2025, doi: 10.63075/0vcg0093.

[43] M. H. Akhtar, T. Jabeen, R. Aziz, M. N. Amin, S. M. Rizwan, and K. Hamid, "Intelligence-based self-healing network design: An automated incident response system for troubleshooting of IoT security breaches," Annual Methodological Archive Research Review, vol. 3, no. 8, Aug. 2025, doi: 10.63075/6dhhj119.

[44] P. Tahir, P. K. Hamid, P. D. A. Kahloon, and P. S. Zubair, "Cyber sovereignty challenges: A strategic framework for national data protection using blockchain authentication," Contemporary Journal of Social Science Review, vol. 3, no. 3, pp. 1314–1328, Aug. 2025, doi: 10.63878/cjssr.v3i3.1100.

[45] W. Aslam, W. Tariq, T. Waheed, K. Hamid, S. M. Rizwan, and A. Ahmed, "Enhancing privacy and security in multi-party computation for data mining in cryptographically protected environments," Annual Methodological Archive Research Review, vol. 3, no. 8, pp. 510–531, Aug. 2025, doi: 10.63075/fxe5gg65.

[46] M. D. Rasheed, K. Hamid, M. W. Iqbal, M. W. Iqbal, M. Ibrar, and M. A. Khan, "Secure evolutionary model empowered smart homes in AI environment: An efficient way of life," Proc. Korean Next Generation Computing Society Conf., pp. 254–257, Dec. 2025.