A COMPARISON OF MODEL-BASED AND EXPLANATION-BASED FEATURE IMPORTANCE ESTIMATION FOR URL PHISHING DETECTION

Ammara Zamir *
*Ca' Foscari University of Venice, Italy **
Department of Environmental Sciences, Informatics and Statistics (DAIS)
*Via Torino 155, 30172 Mestre (VE), Italy **

Claudio Lucchese *
*Ca' Foscari University of Venice, Italy **
Department of Environmental Sciences, Informatics and Statistics (DAIS)
*Via Torino 155, 30172 Mestre (VE), Italy **

ABSTRACT

This study compares both model-based and explanation-based feature importance estimation methods across four publicly available phishing datasets. The goal of our experiment is to evaluate whether feature importance estimation methods can consistently and coherently identify the most predictive features across different datasets. We are interested in phishing datasets as phishing strategies may change significantly over time, and detecting predictive features might be very relevant. Experimental results show that model-based and explanation-based feature importance values disagree in the feature ranking produced. Moreover, the most predictive features are different across datasets. Our greatest finding is that explanation-based feature ranking is more effective in identifying the most predictive features. Indeed, we show that model-based feature importance estimation methods are likely to overestimate the predictive power of a few features, while model-based methods prefer to select many features, and allow for building more effective models.

KEYWORDS

Phishing Detection, Feature Importance, Trustworthy AI, Explainable Artificial Intelligence (XAI), Shapley Additive Explanations (SHAP).

1. INTRODUCTION

Phishing is one of the most persistent cyber threats, tricking users into disclosing their sensitive information by cloning legitimate websites. Phishing attacks are often delivered through emails, messages, and website URLs, resulting in a $12 billion financial loss and continuing to grow (Divakaran & Thomas, 2022; Bezerra & Silva, 2025). Attackers now use automated adversarial techniques, and a single click on a malicious URL can expose sensitive data.

URL phishing datasets consist of dozens of raw features, many of which are redundant. Other features, such as those related to the web page targeted by the URL, are difficult to extract, costly, time-dependent, and therefore difficult to manage. Some features may have high predictive power in one dataset but behave differently in another (Mia & Rahman, 2024). Existing techniques have evolved from blacklisting to machine learning and deep learning; although these methods achieve high accuracy, they lack transparency in decision-making (Remya & Kumar, 2025). Phishing detection typically relies on lexical features (length, symbols), content features (HTML structure), technical features (WHOIS, domain age), and ranking information. Each group has strengths and limitations: lexical features are easy to extract but easy to manipulate, content features require page retrieval, and technical or ranking features are often missing or outdated because phishing domains are short-lived. As a result, feature usefulness varies across datasets, making it difficult to determine which

features are consistently reliable. This motivates us to analyze feature importance and examine how consistently different methods identify important features across diverse phishing datasets.

In this work, we investigate feature-importance measures for URL phishing detection, focusing on their reliability, trustworthiness, and generalizability across datasets. We consider two approaches: model-based feature importance, which measures loss reduction during Gradient Boosted Decision Tree (GBDT) training, and explainability-based importance, which estimates the contribution of each feature using Shapley values. The larger the contribution, the more important the feature.

Specifically, we address the following Research Questions (RQ):

- RQ1: Are the most important model-based features the same across datasets?
- RQ2: Are the most important explainability-based features the same across datasets?
- RQ3: Are explainability-based features more powerful than model-based for training?

To address the proposed questions, we analyze four public URL phishing datasets, group features into five categories, and study their importance at the category level. We examine stability by evaluating multiple samples of the same dataset and comparing model-based and explainability-based feature importance estimation methods.

2. RELATED WORK

Literature demonstrates that earlier studies relied on heuristic methods, which use rule-based strategies to identify malicious URLs based on predefined patterns such as IP addresses, "@" symbols, domain length, and missing SSL certificates. These techniques generate a weighted score to decide whether a webpage is phishy and are simple and cost-effective (Da Silva et al., 2020). Several notable heuristic systems were proposed: CANTINA analyzed webpage text and compared it with the domain name using TF-IDF similarity, flagging pages where brand names (e.g., PayPal) appear but do not match the domain (Fu et al., 2006). PhishGuard inspected webpage structure for suspicious hyperlinks, missing forms, or fake SSL information (Joshi et al., 2008). SpoofGuard detected domain mismatches in URLs (Teraguchi et al., 2004). Although these approaches worked well for known patterns, they fail when attackers slightly modify the webpage structure or tokens (Khonji et al., 2013).

To address these limitations, researchers combined heuristic rules with Machine Learning (ML), allowing models to learn nonlinear relationships that fixed rules cannot capture (Elgharbi et al., 2024). In such hybrid methods, heuristic features serve as inputs to ML models and include lexical features (URL length, token count, subdomains, presence of "@"), technical features (WHOIS, DNS TTL, domain expiration), security features (SSL validity), content features (hidden forms, suspicious elements), and ranking features (PageRank, domain age, traffic). ML models capture interactions among these features, improving accuracy beyond rule-based methods. Several approaches demonstrated this improvement. Gupta et al. showed that combining lexical and hyperlink features, such as internal-to-external link ratios and misleading or empty hyperlinks, enhanced real-time detection and yielded the best performance with XGBoost (Das Guptta et al., 2024). Prince et al. built a hybrid cumulative feature-selection framework using Chi-Square, Mutual Information (MI), Pearson correlation, and Principal Component Analysis (PCA) to generate subsets of strong features. These subsets trained multiple classifiers whose outputs were merged through majority voting, achieving high accuracy with a 32-feature Random Forest model (Prince et al., 2021).

Zamir et al. (2020) proposed a stacked hybrid machine learning approach to evaluate the most important features among lexical, technical, and security-based groups. Their method combined Naïve Bayes (NB), Random Forest (RF), and Support Vector Machines (SVM) in a stacked ensemble to measure the strength of each feature. This helped identify strong feature subsets that contributed to model accuracy and weaker features with limited predictive power. The evolution of phishing detection then shifted toward NLP and deep learning. Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Recurrent Neural Networks (RNNs) automatically learned URL patterns and webpage structure without handcrafted features (Aljofey et al., 2020; Rao et al., 2020). Hybrid deep models such as BGL-PhishNet combined BERT embeddings with LightGBM to capture semantic and structural signals, including URL context and domain metadata—achieving 98.7% accuracy (Remya & Kumar, 2025). Multimodal transformer-based frameworks further integrated lexical signals, HTML DOM graph representations, and visual layout cues, enabling the detection of webpages that mimic legitimate websites through copied design and structure. Despite their

effectiveness, these models' decisions are black boxes, leading to interest in Explainable AI (XAI). Shapley-value-based explanations have been used to interpret model decisions. Fajar et al. (2024) showed that URL length, PageRank, and domain age significantly contribute to boosting model decisions. However, SHAP scores vary across datasets (Mia et al., 2024).

3. RESEARCH METHODOLOGY

3.1 URL phishing datasets

We focus our analysis on four publicly available phishing website datasets, as shown in Table 1. The binary classification task requires discriminating between phishing and legitimate URLs. Each dataset is gathered from PhishTank as the primary source of phishing URLs, while legitimate instances are gathered from Alexa and Yandex. Datasets vary in terms of both the number of cases and the number of features.

Table 1. URL phishing datasets

Dataset	Reference	No. of Instances	Class Distrib. (Benign / Phishy)	Feature Categories	Publication Year	Link
D1	Shafin et al., 2025	11,430	5,715 / 5,715	Lexical, Content, Technical, Security, Ranking	2021 (Kaggle)	https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset/data
D2	Remya & Kumar, 2025	88,647	58,000 / 30,647	Lexical, Technical, Security, Ranking	2020 (Data in Brief)	https://github.com/GregaVrbancic/Phishing-Dataset
D3	Haynes et al., 2021	10,000	5,000 / 5,000	Lexical, Content, Technical, Security, Ranking	2018 (Mendeley)	https://data.mendeley.com/datasets/h3cgnj8hft/
D4	Zamir et al., 2020	11,055	6,157 / 4,898	Lexical, Content, Technical, Security, Ranking	2015 (UCI)	https://archive.ics.uci.edu/dataset/327/phishing+websites

Indeed, datasets were created independently and at different moments in time, and therefore, they encompass different sets of features. To allow a comparative analysis, we group features into five categories:

- **Lexical:** features that capture lexical and structural characteristics of URLs (URL length, number of slashes, number of dashes, etc.).
- **Content:** features extracted from the actual webpage, such as hidden content, presence of password fields, pop-up windows.
- **Technical:** features capturing domain reputation and lifetime, such as WHOIS age, DNS time-to-live, and traffic.
- **Security:** features describing the use and validity of SSL/TLS/HTTPS protocols.
- **Ranking:** features capturing a URL's positioning in web search engines, such as presence in Google's index and PageRank.

This categorization allows us to discuss feature importance across different datasets. Note that D2 has no content-based features. Before training, datasets undergo the following pre-processing steps. First, non-predictive attributes (e.g., ID or URL columns) are removed because they do not contribute to model prediction.

Since all features in our datasets are numeric, we apply median imputation to handle missing values. Finally, features are normalized by means of standard scaling.

3.2 Machine learning models

To evaluate the goodness of feature selection methods, we consider several machine learning models of different kinds, ranging from simple models such as Logistic Regression (Vajrobol et al., 2024) and Naïve Bayes (Amora et al., 2025), to ensemble methods such as Bagging of decision trees, Random Forest (Rajoju et al., 2024), and GBDT. We use GBDT (Remya & Kumar, 2025) and Scikit-Learn for all other methods. Hyperparameters were selected after 10-fold cross-validation, and the best settings are summarized in Table 2.

Table 2. Hyperparameter settings

Model	Hyperparameter search space	Best settings (from CV)

GBDT	learning rate ∈ {0.05, 0.1}; number of estimators ∈ {150, 250}
	number of leaves = 31; minimum child samples = 20;
	subsample = 0.8; column sample by tree = 0.8
		learning rate = 0.1; No. of leaves = 31;
		number of estimators = 250;
		minimum child samples = 20;
		subsample = 0.8; column sample by tree = 0.8

LR	C ∈ {0.5, 1.0, 2.0}; solver = lbfgs; maximum iterations = 300; StandardScaler (without mean)	C = 1.0; solver = lbfgs
RF	No.of estimators ∈ {150, 300}; maximum depth ∈ {None, 20} minimum samples per leaf ∈ {1, 2}	No. of estimators = 300; maximum depth = None; minimum samples per leaf = 1
NB	variance smoothing ∈ {10^{-9}, 10^{-8}}	variance smoothing = 10^{-9}
NN	epochs = 30; batch size = 128; optimizer = AdamW (learning rate = 10^{-3}); learning rate scheduler = CosineAnnealingLR; hidden layers = [512, 256, 128]; dropout = 0.2	epochs = 30; batch size = 128; optimizer = AdamW; dropout = 0.2

The achieved accuracy is high with all models, with GBDT being the best performing with an accuracy above 0.97% on all datasets. Accuracy is discussed in detail in Section 4.

4. EXPERIMENTAL ANALYSIS

This section discusses the research question previously introduced and answers it by means of an extensive set of experiments across the four datasets. We investigate both model-based and explainability-based feature importance. The comparative analysis sheds light on the effectiveness of the two feature selection methods.

4.1 Model-based feature ranking

Because of its best performance, we use GBDT as a model-based feature selection method. After training a forest of GBDT, we use the recorded feature importance to rank and select features. Recall that the feature importance of a feature is given by the cumulative loss reduction provided by nodes encompassing that feature. We select the smallest set of top-ranked features that covers at least 90% of the total feature importance. To counteract the possible instability of feature importances, we run repeated sampling and feature importance evaluation. Sampling is without replacement, and each sample contains 80% of the original dataset. On each sample, we train a GBDT model and get the feature importance scores. These are accumulated across 50 runs, so that the average feature importance is used to select the top features covering 90% of the total feature importance. This method is dataset independent and allows us to avoid fine-tuning a minimum importance threshold for the feature selection. This adaptive rule reduces the feature set, yielding 18 features out of 88 in D1, 9 out of 111 in D2, 10 out of 49 in D3, and 9 out of 30 in D4. In the remainder, we focus on these restricted sets of features.

RQ1. Are the most important model-based features the same across datasets?

Results are reported in Figure 1, where charts illustrate rankings according to mean feature importance and include the measured standard deviation intervals of feature importance across the 50 runs. Except for dataset D2, all datasets exhibit a similar behavior. One or two features contribute to a significant portion of feature importance, while the remaining ones are less influential and have similar scores. Feature importance values have a small standard deviation across different runs, meaning they are quite stable.

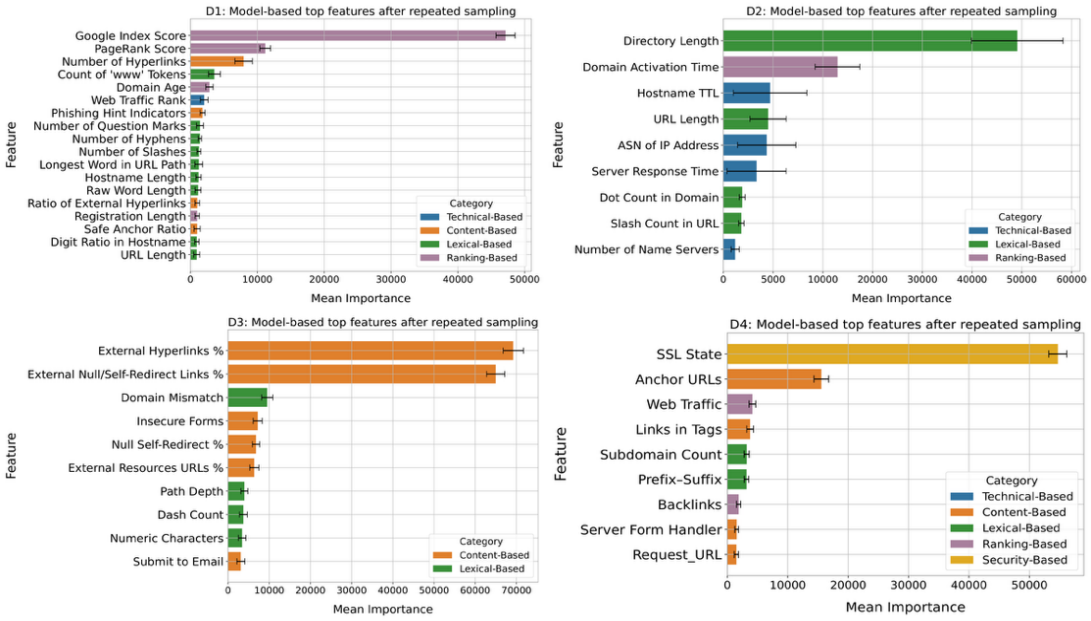


Figure 1. Model-based importance of the selected top features covering 90% of the feature importance after repeated sampling

Therefore, the top two features are stable, while the remaining ones may switch a few places in the ranking, still surviving among the most important ones. Dataset D2 shows instead large standard deviation values, and therefore large variations in the feature importance and rankings observed across the various samples. Note that this is not due to the size of the sample, as D2 is the largest dataset at hand. We note that this dataset does not include content features, and the feature importance distribution is less skewed. We deduce that when highly predictive features are missing, feature importance might not be stable: the model tries to exploit the other features, not necessarily with a consistent order of importance.

To provide a better cross-dataset comparison, we shift the analysis to the category level. For each category, we compute the ratio of the cumulative feature importance of the corresponding features. Figure 2 reports the ratio of feature importance provided by each category on the four datasets, limited to the set of top-selected features. We can see that category distribution is completely different across datasets. In D1, ranking features are dominant, while in D2, lexical features are the most exploited, content features are the most relevant in D3, and finally, D4 sees the most security features.

Answer: No, the most important features are not consistent across datasets. Even feature categories are not consistent. Indeed, each dataset has a dominant feature category, and these differ across the four datasets under analysis. Overall, technical features seem to be the least interesting. In addition, the stability of feature importance depends on feature quality. High-quality features are consistently selected, generating a significant skew in the feature importance distribution. When high-quality features are missing, stability is affected as the model tries to exploit all available features with some degree of randomness. In conclusion, feature importance may be trusted in the presence of a skewed distribution.

4.2 Explainability-based feature ranking

To analyze the predictive power and reliability of explainability-based features, we resort to Shapley values of a feature (Shafin, 2025) to measure its contribution to the model prediction. Let us recall that the Shapley value of a feature with respect to a given instance is an additive contribution provided by the feature to the prediction score. This can be more powerful than the tree-based feature importance, which is defined as the loss reduction over training

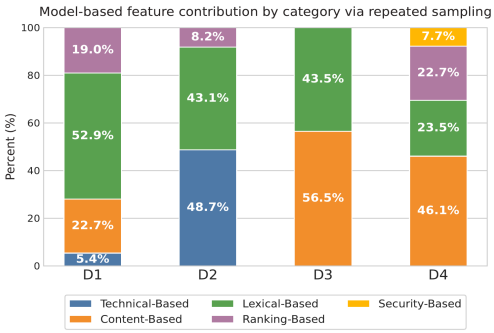


Figure 2. Model-based feature importance contribution by category

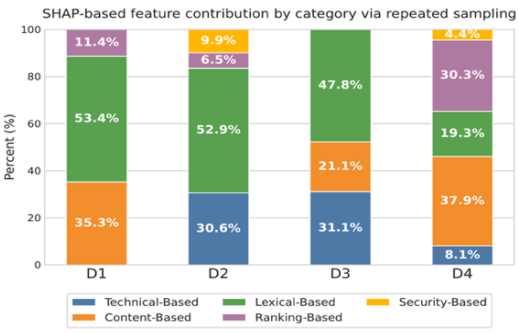


Figure 3. Explanation-based feature importance contribution by category

We consider global Shapley values, which correspond to the average Shapley value of each feature across the full dataset. As in previous analysis, we compute Shapley values on top of a Gradient Boosting Decision Trees model, and we focus on the smallest set of top-ranked features that account for 90% of the Shapley values' cumulative sum. We run the same analysis as in RQ2, by repeatedly sampling from our data and re-evaluating Shapley Values.

RQ2. Are the most important explainability-based features the same across datasets?

Figure 4 shows the Shapley value of the top-ranked features across the four datasets. Shapley values also provide the direction of the contribution of each feature, i.e., whether it pushes the model toward phishing or legitimate classification. Features adopted and their rankings are different across datasets. We highlight that the Shapley value distribution is not very skewed: several features provide a significant contribution.

We observe that standard deviation intervals are large on every dataset. The overlap of standard deviation intervals among different features suggests that the ranking is quite unstable. This is most apparent in D2, but it is true for the other three datasets. This raises some concerns about the explainability power of Shapley values. They describe well the contribution of each feature towards a phishing or legitimate classification. Still, different samples of the same dataset may produce completely different Shapley scores and rankings. This is more apparent in D2, probably suggesting that the features of this dataset have similar prediction power. Figure 3 shows the contribution at the category level. In this case, lexical features are the most important in 3 of the four datasets.

Answer. No, the most important features are not consistent across datasets. A large contribution is given by lexical features, while content-based features are relevant only for D4. Shapley values exhibit significant variation across samples of the same dataset in every experiment we conducted. The feature rankings produced based on Shapley values should be handled with great care, as the ranking is not stable, and a low rank might incorrectly suggest that a feature is not important. However, lexical features are the most relevant in 3 out of four datasets.

RQ3. Are explanation-based features more powerful than model-based for training?

We consider various classifiers using distinct feature configurations: (i) full feature set, (ii) model-based with repeated sampling, and (iii) explanation-based with repeated sampling. Tables 3 and 4 present the results of all classifiers when trained on the above-mentioned feature sets. LGBM on the full feature set outperforms other models in terms of prediction accuracy. This is expected as the model can exploit the information conveyed by the full feature set. When comparing model-based vs. explanation-based, we may observe that explanation-based features provide a small margin of improvement over model-based. On the one hand, this is because a different feature set is selected as discussed above. On the other hand, we may also observe that explanation-

based selection gathers a larger number of features. This is because models tend to overuse a promising feature, limiting the use and inclusion of other important features.

Answer. Yes. Overall, findings confirm that explanatory features derived from SHAP are more powerful and reliable for model training.

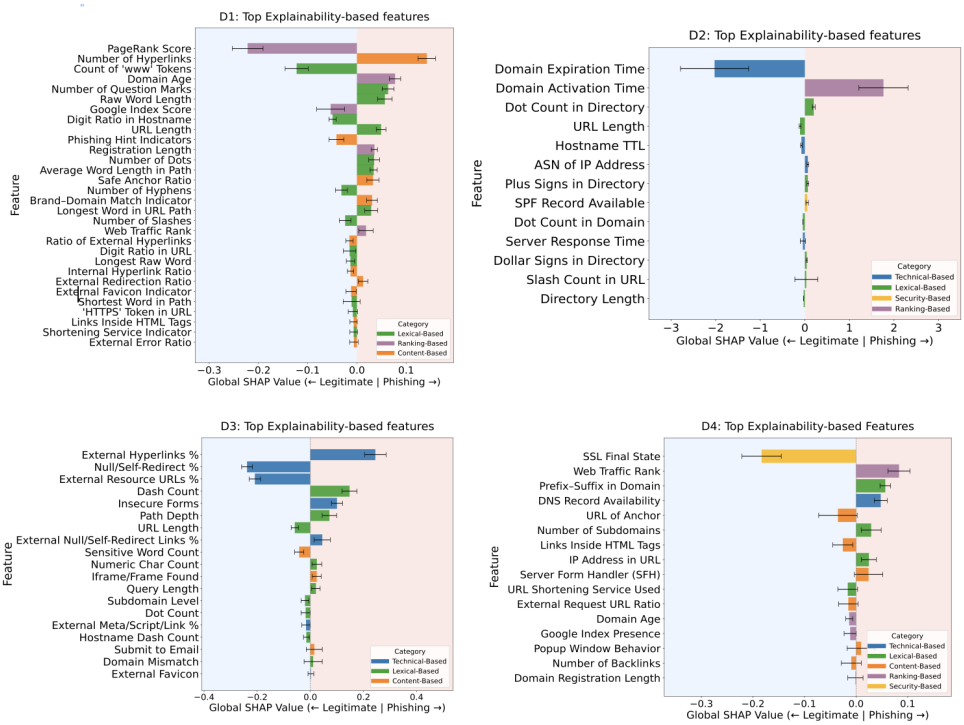


Figure 4. Explanation-based top features after repeated sampling

Table 3. Accuracy (All Features / Model-Based / Explanation-Based)

Model	D1 (88 / 18 / 30)	D2 (112 / 9 / 13)	D3 (49 / 10 / 19)	D4 (30 / 9 / 17)
GBDT	0.9705 / 0.9689 / 0.9703	0.9720 / 0.9682 / 0.9682	0.9883 / 0.9787 / 0.9864	0.9729 / 0.9478 / 0.9682
RF	0.9667 / 0.9638 / 0.9654	0.9710 / 0.9670 / 0.9669	0.9849 / 0.9769 / 0.9819	0.9726 / 0.9475 / 0.9677
NN	0.9661 / 0.9591 / 0.9658	0.9670 / 0.9552 / 0.9580	0.9802 / 0.9758 / 0.9615	0.9721 / 0.9458 / 0.9650
LR	0.9458 / 0.9325 / 0.9355	0.9330 / 0.9134 / 0.9160	0.9452 / 0.9126 / 0.9328	0.9272 / 0.9203 / 0.9238
NB	0.6943 / 0.8479 / 0.8365	0.7980 / 0.8705 / 0.8899	0.8426 / 0.8204 / 0.8437	0.6205 / 0.5989 / 0.6097

Table 4. Average accuracy across all datasets

Model	All	Model-Based	Explanation-Based
GBDT	**0.9759**	**0.9659**	**0.9733**
RF	0.9738	0.9638	0.9705
NN	0.9714	0.9590	0.9626
LR	0.9378	0.9197	0.9270
NB	0.7389	0.7844	0.7949

5. CONCLUSION

This study compares model-based versus explanation-based features importance estimation methods. We use repeated sampling to consider the possible variability of the estimated feature importance. Our experiments, conducted on four public datasets and on a variety of machine learning models, show that explanation-based methods are the best performing. Moreover, explanation-based feature selection puts lexical features as the most relevant in three out of four datasets. We motivate this better performance of explanation-based feature selection based on (a) the greedy nature of model-based features, where the model excessively re-uses those features that may reduce the loss, resulting in a limited number of features, (b) the most intersecting approach of explanation-based methods that focus on the predictions of a method rather than on the loss function. The category level analysis also provides a more stable view on the phishing phenomenon, suggesting that lexical features are set of features the most adopted by malicious users. Based on our findings, we conclude that explanation methods are best recommended for feature importance analysis.

REFERENCES

Aljofey, A. et al, 2020. An effective phishing detection model based on character-level convolutional neural network from URL. *Electronics*, 9(9), p.1514.

Amora, E.N.O., 2025. Phishing URL detection using Naive Bayes classification: A robust method for accurate detection. *Proceedings of the 2025 International Conference on Visual Analytics and Data Visualization (ICVADV),* pp.212–217.

Bezerra, A. et al, 2025. A case study on phishing detection with a machine learning net. *International Journal of Data Science and Analytics,* 20(3), pp.2001–2020.

Da Silva, C.M.R. et al, 2020. Heuristic-based strategy for phishing prediction: A survey of URL-based approach. *Computers & Security*, 88, p.101613.

Das Guptta, S. et al, 2024. Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Annals of Data Science*, 11(1), pp.217–242.

Divakaran, D.M. and Oest, A., 2022. Phishing detection leveraging machine learning and deep learning: A review. *IEEE Security & Privacy*, 20(5), pp.86–95.

Elgharbi, S.E. et al, 2024. Online phishing detection: A heuristic-based machine learning framework. *Proceedings of the 2024 Mediterranean Conference on Embedded Computing (MECO)*, pp.1–4.

Fajar, A. et al, 2024. Enhancing phishing detection through feature importance analysis and explainable AI: A comparative study of CatBoost, XGBoost, and EBM models. *arXiv preprint, arXiv*:2411.06860.

Fu, A.Y. et al, 2006. Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD). *IEEE Transactions on Dependable and Secure Computing*, 3(4), pp.301–311.

Joshi, Y. et al, 2008. PhishGuard: A browser plug-in for protection from phishing. *Proceedings of the 2008 International Conference on Internet Multimedia Services Architecture and Applications*, pp.1–6.

Khonji, M. et al, 2013. Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, 15(4), pp.2091–2121.

Mia, M. et al, 2024. Can features for phishing URL detection be trusted across diverse datasets? A case study with explainable AI. *Proceedings of the 11th International Conference on Networking, Systems, and Security,* pp.137–145.

Prince, M.S.M. et al, 2021. A new ensemble model for phishing detection based on hybrid cumulative feature selection. *Proceedings of the 2021 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp.7–12.

Rajoju, R. et al, 2024. Text phishing detection system using Random Forest algorithm. *Proceedings of the 2024 International Conference on Applied Artificial Intelligence and Computing (ICAAIC),* pp.1332–1339.

Rao, R.S. et al, 2020. CatchPhish: Detection of phishing websites by inspecting URLs. *Journal of Ambient Intelligence and Humanized Computing*, 11(2), pp.813–825.

Remya, S. et al, 2025. BGL-PhishNet: Phishing website detection using hybrid model-BERT, GNN, and LightGBM. *IEEE Access*, 13, pp.47552–47569.

Shafin, S.S., 2025. An explainable feature selection framework for web phishing detection with machine learning. *Data Science and Management,* 8(2), pp.127–136.

Teraguchi, N.C. and Mitchell, J.C., 2004. Client-side defense against web-based identity theft. *Stanford University Technical Report*.

Vajrobol, V. et al, 2024. Mutual information-based logistic regression for phishing URL detection. *Cyber Security and Applications*, 2, p.100044.

Zamir, A. et al, 2020. Phishing website detection using diverse machine learning algorithms. *The Electronic Library*, 38(1), pp.65–80.