scientific reports

Check for updates

OPEN

Explainable phishing website detection for secure and sustainable cyber infrastructure

Tanzila Kehkashan[1,4,8], Maha Abdelhaq[2], Ahmad Sami Al-Shamayleh[3], Nazish Huda[4,8], Imran Ashraf Yaseen[5,8]✉, Abdelmuttlib Ibrahim Abdalla Ahmed[6]✉ & Adnan Akhunzada[7]

Phishing is a social engineering attack and a type of cybercrime that is dangerously and constantly on the rise. Phishing attacks can impact various sectors, including governmental, social, financial, and individual businesses. Traditional methods of identifying phishing websites, such as blacklist and heuristic approaches, often fail to provide sufficient protection. Moreover, traditional techniques that combine URLs, webpage content, and external features are time-consuming, require substantial computing power, and are unsuitable for devices with limited resources. Moreover, previous research has often overlooked the critical role of identifying which features are important for detection and their impact on outcomes. Traditional methods might not fully capture the significance of individual features. To overcome this issue, this research applies feature selection techniques, specifically shapley additive explanations, with each model based primarily on the URL to improve the detection process. A dataset with over 11000+ URLs and 30 varied features of the "Phishing Website Detection" was applied from the Kaggle repository. Then, the models, namely support vector machine (SVM), random forest (RF), decision tree (DT), logistic regression(LR), and K-nearest neighbor, were trained and tested. Each model used shapely additive explanations (SHAP) to improve precision and interpretability by highlighting the most important features. It was tested using some key performance metrics such as accuracy, precision, recall, and F1 score. Compared to all the models that were tested, this random forest model indicates 97% accuracy. The proposed system offers an overall and interpretable solution for phishing detection that contributes to a safer digital environment.

Keywords Machine learning, Phishing website detection, RF, SHAP, URL

Phishing is a widespread cyber attack based on social engineering where attackers trick individuals into disclosing sensitive data like usernames, passwords, and financial information. Phishing is a widespread threat across various sectors like government departments, banks, social media websites, and personal users[1]. Due to the fast expansion of internet services like mobile apps and cloud computing, the rate of online transactions and electronic communications has grown significantly[2]. As a result, phishing has become a significant cybersecurity issue, with the attackers constantly evolving their methods in order to bypass current security measures[3]. Phishing sites, which typically depend on misleading URLs to entice victims, are one of the most popular attack vectors for online fraud[4]. Malicious links are typically disseminated through email, SMS, and social media sites, hence turning phishing into an ongoing and ubiquitous threat[5]. It has been reported that over 80% of firms experience a phishing attack every year, which leads to significant financial and operational consequences[6]. The increasing sophistication and magnitude of such attacks emphasize the need for detection methods to be precise, explainable, and cost-effective.

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia. [2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [3]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan. [4]Faculty of Information Technology, University of Lahore, Sargodha 40100, Pakistan. [5]Faculty of Information Technology, University of Lahore, Sargodha 40100, Pakistan. [6]Computer Science Department, Faculty of Computer Science and Information Technology, Omdurman Islamic University, Omdurman, Sudan. [7]College of Computing and Information Technology, Department of Data and Cybersecurity, University of Doha for Science and Technology, Doha 2444, Qatar. [8]Tanzila Kehkashan, Nazish Huda, and Imran Ashraf have contributed equally to this work. ✉email: imranashraf.yaseen@gmail.com; abdelmuttlib@oiu.edu.sd

Traditional detection methods like blacklists and heuristic-based have been widely used[7]. Nevertheless, these approaches are narrow in scope and ineffective for extremely dynamic and new phishing attacks[8]. More sophisticated approaches based on URL examination, web page content, and extrinsic features have been brought up, yet they are computationally costly and unsuitable for low-capacity devices[9]. Feature selection methods, such as Particle Swarm Optimization and Information Gain, have been used to enhance efficiency, but they are still limited by long computation times and partial feature identification[10].

Even with these improvements, phishing detection remains challenging. Attackers use more evasion techniques, making static detection systems useless[11]. Present methods cannot also handle zero-day attacks through dependence on stale or static features[12–14]. In addition, traditional feature selection techniques often do not account for the heterogeneity of phishing URLs, thus compromising model robustness and generalizability[15]. These shortcomings highlight the importance of phishing detection systems that not only work well but also are explainable and sustainable.

To tackle these issues, this research proposes an explainable phishing detection framework that combines Shapley Additive Explanations (SHAP) with supervised machine learning (ML) models. By exploiting URL-based inspection and SHAP-guided feature selection, the suggested solution improves interpretability while being computationally efficient. The novelty of this work is in integrating explainable feature selection with state-of-the-art ML classifiers to achieve both excellent predictive performance and human-interpretable insights, thus making progress in the design of useful and sustainable phishing detection systems.The goals of this work are as follows:

i. To establish a strong method for phishing website detection using URL-based features with enhanced accuracy and efficiency.
ii. To integrate SHAP with ML models like SVM, KNN, RF, DT, and LR for feature interpretation and better model explainability.
iii. To examine the effect of SHAP-based feature selection on the performance of several supervised ML models.

Contributions of this study

This study provides a number of contributions to phishing detection. First, it proposes an innovative explainable detection approach that combines SHAP and supervised machine learning algorithms for phishing website detection. Second, it improves feature interpretation through the identification of the most impactful URL-based features, thus yielding actionable insights for cybersecurity professionals. Third, it presents a benchmark testing by means of systematic experimentation on the Kaggle Phishing Website Detection dataset, with over 11,000 URLs and 30 different features. Fourth, the research shows notable performance improvement, with the Random Forest model augmented with SHAP exhibiting the best accuracy of 97%, surpassing other classifiers when it comes to precision, recall, and F1-score. Lastly, the suggested solution is focused on practical usability in that it is interpretable, cost-effective, and deployable in real-world resource-limited environments.

The rest of the paper is structured as follows. Section 2 discusses a review of related work in phishing detection, including recent developments and lacunae. Section 3 describes the research methodology, including data sources, feature extraction, and ML classifiers. Section 4 discloses the experimental results, comparing the proposed method with current ones and exploring the findings. Section 5 presents the implications of the findings, and Sect. 6 concludes the research by highlighting the greatest contributions and areas for future study.

Literature review

To conduct a literature review, this study follows the framework proposed by Saied et al.[16], which provides guidelines for transparent and reproducible research analysis. Relevant studies on ML, deep learning (DL), and hybrid approaches for phishing detection were selected and evaluated based on dataset characteristics, feature extraction methods, model architecture, and reported performance metrics. This methodology ensures consistent comparisons, highlights gaps in prior work such as limited interpretability, and establishes a solid foundation for proposing the RF model enhanced with SHAP.

Blacklist and whitelist techniques

Traditional phishing detection initially relied on blacklists and whitelists to identify malicious websites[17–19]. Blacklists maintain a repository of known phishing URLs and flag websites accordingly. While straightforward, blacklists are inherently reactive, unable to detect new or zero-day phishing sites. Whitelists, which permit access only to verified legitimate websites, face similar limitations due to attacker evasion strategies. Both methods are further constrained by the dynamic nature of phishing campaigns, as they require continuous updates that are often slow, manual, and resource-intensive[20,21]. These approaches provided a foundation for early phishing detection but proved insufficient against adaptive adversaries. Their reactive design means they fail to anticipate evolving threats, and their reliance on human updating undermines scalability. As phishing campaigns became more automated and sophisticated, the shortcomings of these methods motivated the transition toward data-driven, automated, and more adaptive solutions.

Traditional techniques for phishing detection

Early heuristic-based systems attempted to go beyond blacklists by relying on manually defined rules. Such methods assessed suspicious URL structures, domain names, and the presence of unusual characters or tokens[19–21]. Heuristic rules had the advantage of being lightweight and independent of large datasets, but they frequently produced high false positive rates and lacked the robustness needed for evolving phishing campaigns. For instance, attackers could easily bypass heuristics by slightly altering domain names or obfuscating malicious

indicators. Critically, these approaches offered little to no interpretability. Security analysts could see that a website was flagged as malicious but lacked insight into why. This "black-box flagging" not only reduced trust in automated systems but also limited their adoption in organizational security infrastructures[22,23]. As phishing attacks began to imitate legitimate websites with near-perfect fidelity, heuristic approaches became increasingly inadequate. These limitations directly motivated the shift toward ML and, more recently, explainable AI (XAI), where interpretability is not optional but essential for adoption in high-stakes cybersecurity contexts.

ML techniques for phishing detection

Machine learning rapidly emerged as a dominant solution to phishing detection, offering adaptability, pattern recognition, and the ability to learn from large datasets. ML methods can analyze URL structures, webpage content, and external metadata to distinguish legitimate sites from phishing attempts[24]. Among these, URL-based approaches remain particularly attractive due to their computational efficiency and reduced dependence on resource-heavy web content parsing[25,26]. Hybrid methods that combine URL features with content and external metadata have achieved higher accuracy, as they integrate diverse sources of information[27,28]. Ensemble learning further strengthens these methods. For example, RF, KNN, and ANN combinations improve resilience and adaptability[29–31]. Similarly, paired classifiers such as SVM-KNN and LR-DT with AdaBoost have demonstrated better generalization to unseen attacks[32,33]. Ahmad et al.[34] reviewed AI-driven phishing detection systems and emphasized that ensembles integrated with real-time threat intelligence significantly improve adaptability against fast-evolving phishing campaigns. However, most ML-based works continue to rely on extensive manual feature engineering, which is resource-intensive and prone to overlooking subtle but critical signals. Moreover, while accuracy rates are frequently reported, interpretability remains underexplored. In practice, decision-makers must understand why a model classifies a site as phishing before acting on it. This gap underlines the importance of XAI-based feature selection methods such as SHAP, which can provide interpretable explanations while maintaining predictive strength.

DL techniques for phishing detection

Deep learning has brought major advances by enabling automatic feature extraction from raw data. CNNs capture spatial and sequential structures in URLs, reducing reliance on manual engineering[35–37]. LSTMs and other RNNs excel at modeling sequential dependencies, particularly valuable in identifying obfuscated URLs that mimic legitimate naming patterns. Incorporating external signals such as SSL certificates, registrar details, and domain age further strengthens detection[38]. Hybrid DL architectures, combining CNN and LSTM layers, have produced state-of-the-art results in detecting increasingly sophisticated phishing techniques[39]. Despite these strengths, DL approaches present significant challenges. They require large labeled datasets and substantial computational resources, which may not be feasible for small organizations or low-resource environments. Moreover, interpretability remains a major barrier: deep models often function as "black boxes," making their decisions difficult to justify in operational or legal contexts[40,41]. Insights from other domains illustrate possible directions. For instance, transformer-based DL architectures have advanced fields such as plant disease detection[42] and micro-expression recognition[43], demonstrating how self-attention mechanisms capture complex dependencies efficiently. In fraud detection, transformer-enabled recruitment fraud detection[44] and AI-driven IoT security frameworks[45] show the utility of advanced DL in high-risk contexts. Similarly, CoAtNet-based medical imaging[46], deep CNN gesture recognition[47], and hybrid ML–DL video captioning studies[48] further highlight the scalability and robustness of attention-driven models. Nevertheless, the trade-off between performance and explainability remains unresolved, reinforcing the need for interpretable and lightweight alternatives.

Quantitative and comparative techniques

Comparative studies benchmark different phishing detection models using quantitative metrics such as accuracy, precision, recall, and F1-score[28,29]. RF frequently outperforms other classifiers due to its ensemble structure and robustness to noisy features, while hybrid models combining classifiers and feature sources achieve stronger generalization[31,32,49]. Yet, a critical observation emerges: performance metrics alone are insufficient. High accuracy may obscure issues such as poor interpretability, long training times, or poor adaptability to unseen zero-day attacks. Several comparative analyses now advocate for a balance of accuracy, efficiency, and explainability. Without this, high-performing models risk remaining academic exercises, rarely adopted in operational cybersecurity infrastructures. This aligns with calls in other domains for explainable solutions, as seen in finance, where XAI frameworks such as SFIX (Scalable Financial-oriented Interpretable eXplanation)[50] and systematic reviews of explainable AI in financial applications[51] highlight the centrality of interpretability for adoption. Such parallels emphasize that phishing detection research must equally embrace XAI if it is to transition successfully into practice.

Feature engineering for phishing detection

Feature engineering remains central to phishing detection, influencing both performance and interpretability. Approaches such as genetic algorithms, permutation importance, and metaheuristics have been applied to optimize features and reduce computational costs[52–56]. More recent work emphasizes interpretable methods like SHAP, which explain predictions by quantifying each feature's contribution[57–60]. Several innovative strategies combine advanced feature selection with resampling and balancing methods. Examples include CatBoost with SMOTE-Tomek balancing, BERT-based embeddings, BMEO-KNN, and document-term matrix feature extraction[61–65]. These approaches show that model transparency and stability improve when robust feature engineering is combined with interpretability tools. In addition, ensemble-based feature engineering methods are emerging as powerful alternatives. For example, ensemble learning for financial data classification has

successfully leveraged encrypted datasets while maintaining interpretability citeKucur2025AuditOpinions, providing a methodological parallel for phishing detection research. Despite progress, persistent gaps remain. Many approaches still depend on handcrafted features, which are vulnerable to adversarial manipulation. Computational costs remain high, limiting practical deployment on low-resource devices. Most importantly, interpretability is often an afterthought rather than a central design criterion. Figure 1 summarizes these limitations, highlighting the need for models that combine efficiency, transparency, and strong generalization. Building on this critical evaluation of existing methods, this study proposes a SHAP-based ML framework that efficiently identifies key URL features, ensures interpretability, and achieves high predictive performance while remaining suitable for deployment in resource-constrained cybersecurity infrastructures.

Methodology

This paper suggests a novel approach to detecting phishing websites using supervised ML models combined with SHAP to incorporate explainability in features and improve detection accuracy. The approach is designed to fill the gap between high-performing detection and feature importance understanding, which is vital in real-world cybersecurity practice. The architecture of the framework consists of five principal components: baseline comparison, model selection, data gathering, preprocessing and model design, and experiment realization.

Baseline method

The baseline for this study is the work of[66], who introduced a ML solution for phishing detection based on URL-based features. The baseline used models like RF and SVM to identify URLs as phishing or not, with the primary aim of detection accuracy. Although the baseline worked well, the baseline was not interpretable and did not offer feature contribution insights. Our suggested methodology builds on this baseline by incorporating SHAP for feature importance analysis while preserving performance in detection without sacrificing interpretability. Figure 2 demonstrates the block diagram of our suggested methodology, indicating the flow of data collection, preprocessing, model learning, SHAP-based feature selection, and performance testing.

Models selection

Our approach utilizes several supervised learning models for high-quality detection and comparative evaluation. RF, SVM, DT, LR, and KNNs are used as the chosen models. RF combines several DTs by majority voting to increase accuracy and avoid overfitting. SVM tackles high-dimensional data efficiently with the help of hyperplanes to classify classes. KNN predicts a data point according to the majority vote of its nearest neighbors, whereas LR estimates the probability of binary classes with a logistic function. DTs divide datasets by feature values to create a tree-like structure, optimizing homogeneity within segments.

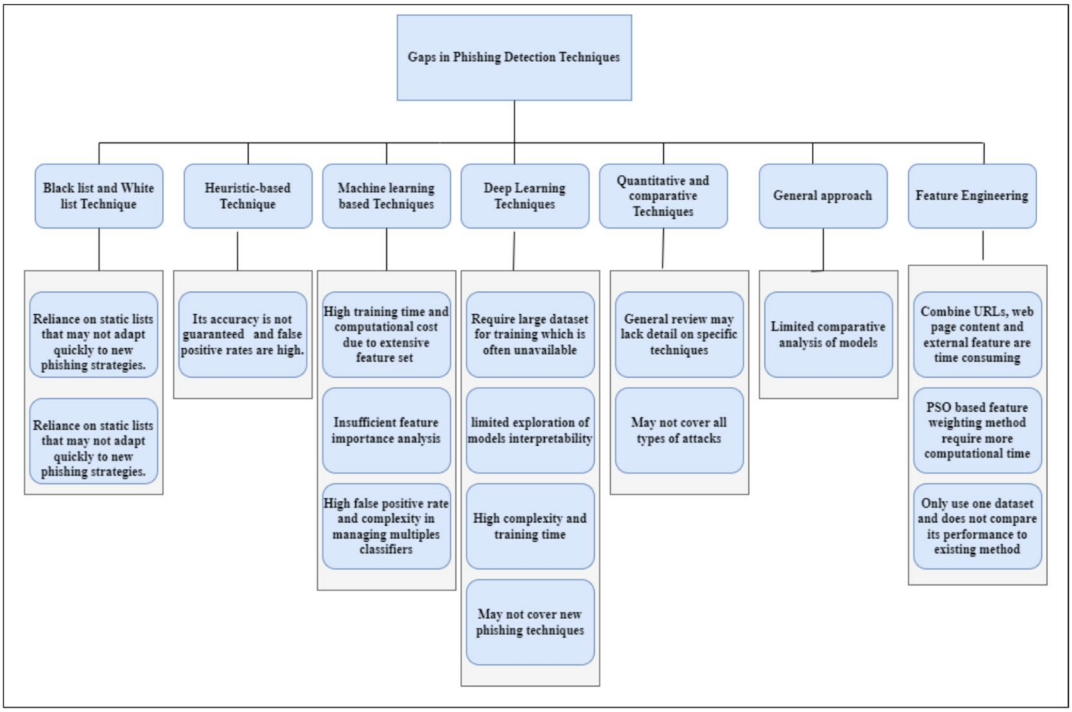

Fig. 1. Gaps in Phishing Detection Techniques: Challenges across blacklists, heuristics, feature engineering, interpretability, and computational constraints.This diagram summarizes the current limitations and unexplored areas in phishing detection methodologies. It highlights key gaps in existing models, such as lack of feature generalization, absence of explainability, and limitations in adapting to real-time phishing campaigns.

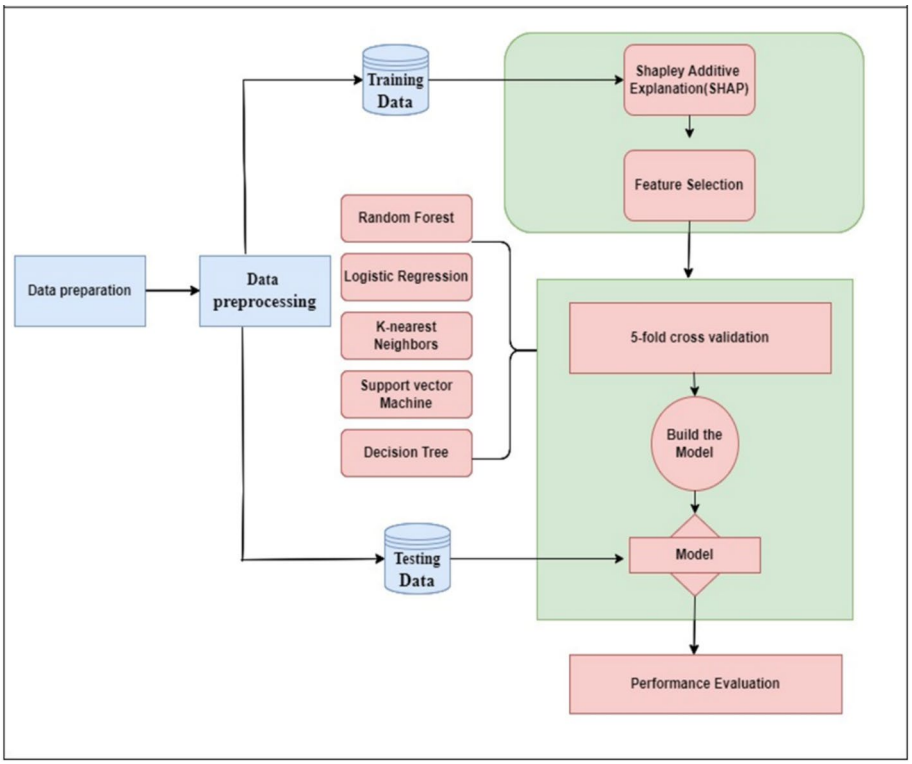

Fig. 2. Proposed Methodology Diagram.This figure presents the end-to-end flow of the proposed phishing detection framework. It includes data collection, preprocessing, feature selection, model training, evaluation, and explainability phases, organized in a sequential and modular fashion.through ensemble averaging.

Each model is paired with SHAP, which computes feature contributions to the end predictions, which increases interpretability and sheds light on which features have the greatest impact on classification decisions. This addresses some of the criticisms of past approaches cited by the reviewers, notably the absence of feature-level interpretability and real-world insight into detecting phishing.

Data collection

The dataset used in this research was obtained from the UCI Machine Learning Repository, originally contributed by Mohammad et al.[63]. It is publicly available at: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites. This benchmark dataset has been widely used in phishing detection research, ensuring reproducibility and comparability of results across studies. The dataset contains 11,055 website records labeled as phishing (1) or legitimate (− 1), with 30 handcrafted features extracted from website URLs and related attributes. Some of the most relevant features include: *UsingIP*, *LongURL*, *ShortURL*, *Symbol@*, *PrefixSuffix-*, *SubDomains*, *HTTPS*, *DomainRegLen*, *Favicon*, *NonStdPort*, *RequestURL*, *AnchorURL*, *IframeRedirection*, *AgeofDomain*, *DNSRecording*, *WebsiteTraffic*, *PageRank*, and *LinksPointingToPage*. The dataset provides a diverse representation of URL patterns and domain-related attributes, allowing models to learn discriminative characteristics of phishing websites compared to legitimate ones. Its capacity is large enough to support vigorous training and validation, and its status as a benchmark dataset increases the certainty of comparative assessment with alternative state-of-the-art approaches.

Preprocessing

Preprocessing is a crucial process to prepare the dataset for effective training of the model and accurate phishing detection. The initial dataset contains over 11,000 URLs with 30 attributes including varied URL features, web content features, and domain features. Data cleaning is initially performed to remove duplicates and unwanted entries in order to preserve the purity of the dataset. Categorical data are encoded numerically, and feature scaling is employed to normalize the range of values over all features to prevent bias towards ranges with high numbers. Pre-processing the dataset into training (80%) and testing (20%) sets prevents weak evaluation while ensuring that there is a sufficient amount of data for model training. URL-specific pre-processing includes tokenization of the URL strings and encoding them numerically in formats suitable for the ML algorithms. Besides, SHAP has been applied to feature selection in order to identify the most influential attributes and eliminate less informative ones. This not only maximizes model performance but also enhances interpretability as it indicates how features contribute to phishing detection results. We empirically chose the top 15 features according to SHAP ranking, since this achieved the best trade-off between interpretability and model accuracy.

Model architecture

The suggested methodology utilizes various supervised learning models, i.e., RF, SVM, DT, LR, and KNNs, with all of them combined with SHAP to offer feature-level interpretability. The input features, like HTTPS usage, Domain Registration Length, URL length, and special symbols, are passed through the models. SHAP values are calculated for every feature to measure their contribution to the predictions made by the model, allowing a transparent assessment of importance of features.

The RF classifier combines the predictions of many DTs to improve robustness and avoid overfitting. Each tree is trained on a random subset of data and, at each node, a random subset of features is used for splitting. The overall prediction is the majority vote across all trees, represented mathematically as:

$$F(z) = \frac{1}{M} \sum_{j=1}^{M} g_j(z) \tag{1}$$

where M is the total number of trees, and $g_j(z)$ denotes the prediction of the j-th tree.

SVMs are applied for classification in high-dimensional feature spaces, separating categories with optimal hyperplanes. The decision function for SVM is given as:

$$D(u) = \alpha \cdot u + \beta \tag{2}$$

where α represents the weight vector, u is the input feature vector, and β is the bias term.

The KNN algorithm determines the class of a sample based on the majority label among its closest neighbors, typically measured by the Euclidean distance:

$$\delta(p, q) = \sqrt{\sum_{k=1}^{d} (p_k - q_k)^2} \tag{3}$$

where p and q are feature vectors, and p_k, q_k denote the values of the k-th feature. In this study, $K = 5$, indicating that the five nearest samples are considered for classification.

DTs partition the dataset recursively by splitting on attribute values, forming a tree where internal nodes correspond to decision criteria, edges represent outcomes, and leaf nodes provide the predicted class. LR estimates the probability of binary outcomes using the logistic function:

$$\Pr(y = 1|u) = \frac{1}{1 + e^{-(\alpha \cdot u + \beta)}} \tag{4}$$

where α are the feature coefficients and β the intercept term.

SHAP values are calculated for all models, enabling interpretation by quantifying the contribution of each feature to the final prediction.

Implementation details

The database has more than 11,000 URLs with binary labels for phishing (1) or legitimate (− 1) web pages. Post-preprocessing, features are domain-specific parameters like URL length, special characters, use of HTTPS, and redirection flags. Categorical features are converted into numerical format by encoding and normalization to gain equal scaling for models. The database is split between training (80%) and testing (20%) sets, and five-fold cross-validation is used to gain reliable evaluation and reduce overfitting. The computation environment consists of Windows 10, 2.4 GHz processor, and Google Colab GPU, which offer adequate resources to train the model. Table 1 gives an overview of components and experimental setup.

SHAP-based feature ranking is performed across all models to identify the most influential features, improving model transparency and enabling interpretable predictions. The methodology ensures that the final

Components	Description
Dataset used	Kaggle / UCI repository
Training and testing split	80/20
Classes	Binary (legitimate/phishing)
No. of samples	11,000 URLs
No. of features	30
System used	Windows 10
GPU	Google colab GPU
Processor	2.4 GHz processor

Table 1. Components Description.This table outlines the dataset specifications and experimental environment used in the study. It details the source of data, dataset size, number of features, classification type, and system configurations such as processor, OS, and computational resources.

prediction is a binary classification of phishing or legitimate websites, combining robust ML performance with feature-level interpretability. The integration of SHAP, careful preprocessing, and ensemble techniques results in a methodology that is accurate, efficient, and applicable in real-world phishing detection scenarios. Algorithm 1 formally summarizes the SHAP-enhanced RF training and evaluation process.

Require: Dataset D with n samples and m URL-based features
Ensure: Trained Random Forest model with SHAP-based feature interpretability
Split D into training set D_{train} and testing set D_{test}
Preprocess data (normalize features, handle missing values)
Initialize Random Forest classifier RF with tuned hyperparameters
Train RF on D_{train}
Evaluate baseline performance on D_{test}
Compute SHAP values for each feature using trained RF
Rank features by mean absolute SHAP values
Select top-15 most impactful features
Retrain RF on D_{train} using the selected 15 features
Evaluate enhanced model on D_{test} using accuracy, precision, recall, F1-score
Visualize feature importance with SHAP summary plots
return Trained SHAP-enhanced RF model with interpretable feature set

Algorithm 1. SHAP-enhanced random forest for phishing detection

Experiments and results

In this work, we present a systematic evaluation of the new ML–based approach for phishing website detection. Experimental results demonstrate that our approach consistently obtains outstanding classification performance on different evaluation metrics and clearly surpasses some prominent methods reported in recent literature. This section consists of five primary components: model performance, ablation studies, comparison with competing detection models, qualitative explanation, and overall discussion. Each of them has critical observations on the strengths and weaknesses of the system, along with effectiveness and areas of improvement. The results exhibit the performance and reliability of the sophisticated framework to identify genuine and phishing web pages with accuracy. Through integrating high performance with explainability, the study makes a useful contribution to the advancement of realistic and sustainable phishing detection systems.

Performance analysis of proposed method

We first compare the performance of five popular supervised learning classifiers to determine the best phishing website detection classifier. Table 2 reports the comparative analysis in terms of F1-score, precision, recall, and accuracy.

The KNN model achieved an F1-score of 94.9, precision of 94.6, recall of 95.1, and accuracy of 94.3. The SVM achieved an F1-score of 94.2, precision of 92.9, recall of 95.5, and accuracy of 93.5. LR reached an F1-score of 94.1, precision of 92.9, recall of 95.3, and accuracy of 93.3. The DT model obtained an F1-score of 96.4, precision of 96.7, recall of 96.1, and accuracy of 96.0. Finally, the RF model outperformed all others, achieving an F1-score of 97.3, precision of 97.0, recall of 97.6, and accuracy of 97.0. These results provide strong evidence that RF is the most suitable supervised learning algorithm for phishing website detection in our setting.

Analyzing the confusion matrices, in Fig. 3: the RF model correctly identified 937 true negatives and 1206 true positives, with only 39 false positives and 29 false negatives and the SVM achieved 887 true negatives and 1180 true positives, alongside 89 false positives and 55 false negatives. In Fig. 4, LR produced 886 true negatives and 1177 true positives, but slightly higher misclassifications with 90 false positives and 58 false negatives and the DT yielded 936 true negatives and 1187 true positives, with 40 false positives and 48 false negatives. Finally, the KNN (Fig. 5) model achieved 1157 true negatives and 909 true positives, while recording 67 false positives and 60 false negatives. These results confirm that although all models demonstrated competitive performance, the RF classifier consistently achieved the best balance across all metrics (Fig. 6).

To further enhance interpretability, SHAP was applied to evaluate the relative importance of each feature in the classification process. SHAP provides a transparent mechanism to understand feature contributions to model decisions, thereby increasing trust in predictions. Applying SHAP across all models revealed the most influential features that drive accurate phishing detection. The summary plots visualize these contributions by showing the

Algorithms	F1-score	Precision	Recall	Accuracy
KNN	94.9	94.6	95.1	94.3
SVM	94.2	92.9	95.5	93.5
LR	94.1	92.9	95.3	93.3
DT	96.4	96.7	96.1	96.0
RF	97.3	97.0	97.6	97.0

Table 2. Performance Analysis of Proposed Methodology with SHAP (in %)This table presents the performance metrics of various ML algorithms when enhanced with SHAP. It demonstrates the improvements in interpretability and performance across precision, recall, F1-score, and accuracy.

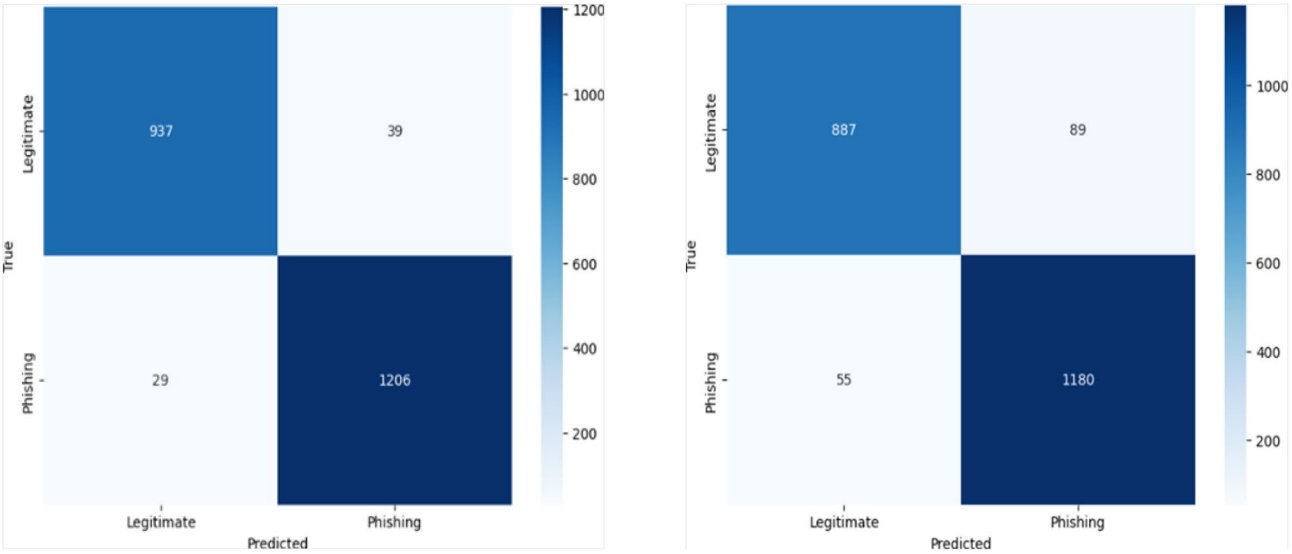

Fig. 3. RF model Confusion Matrix (Right) and SVM model Confusion Matrix (Left).A side-by-side comparison of the confusion matrices for the SVM and RF classifiers. The matrices illustrate the models' classification performance, showing true positives, false positives, true negatives, and false negatives.

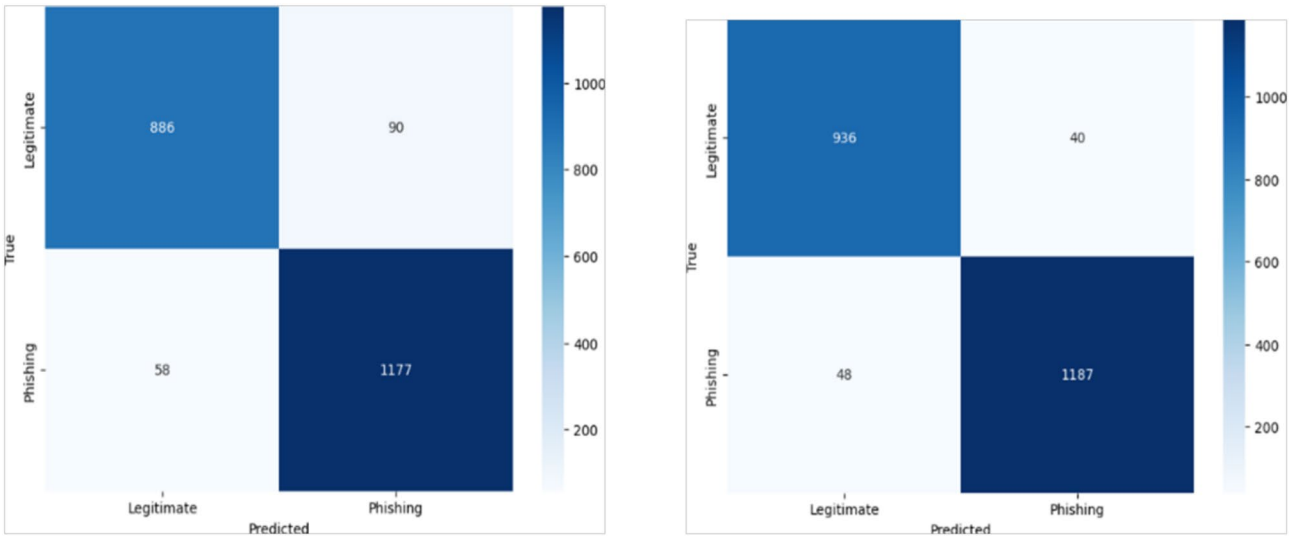

Fig. 4. LR model Confusion Matrix (Right) and DT model Confusion Matrix (Left).This figure presents the confusion matrices for the DT and LR models. It visually compares the predictive accuracy and misclassification rates for both classifiers.

relationship between feature importance (y-axis) and SHAP values (x-axis), clarifying how individual features affect predictions. This interpretability strengthens the transparency and robustness of the detection system.

Ablation studies

The ablation study evaluates the effect of incorporating SHAP on model performance. Table 3 reports the baseline results without SHAP. Comparing these results with those in Table 2 demonstrates the clear advantage of using SHAP-enhanced feature importance.

For example, the RF classifier achieved 90.0% accuracy without SHAP, which increased to 97.0% after SHAP-based feature refinement. Similar improvements were observed across other models: KNN (Fig. 7) improved from 86.0% to 94.3%, SVM from 86.0% to 93.5%, LR (Fig. 7) from 85.0% to 93.3%, and DT (Fig. 8) from 90% to 93%. This demonstrates that SHAP not only improves interpretability but also enhances predictive accuracy by prioritizing the most relevant features. Overall, the integration of SHAP reduces misclassification rates and strengthens the reliability of phishing detection models.

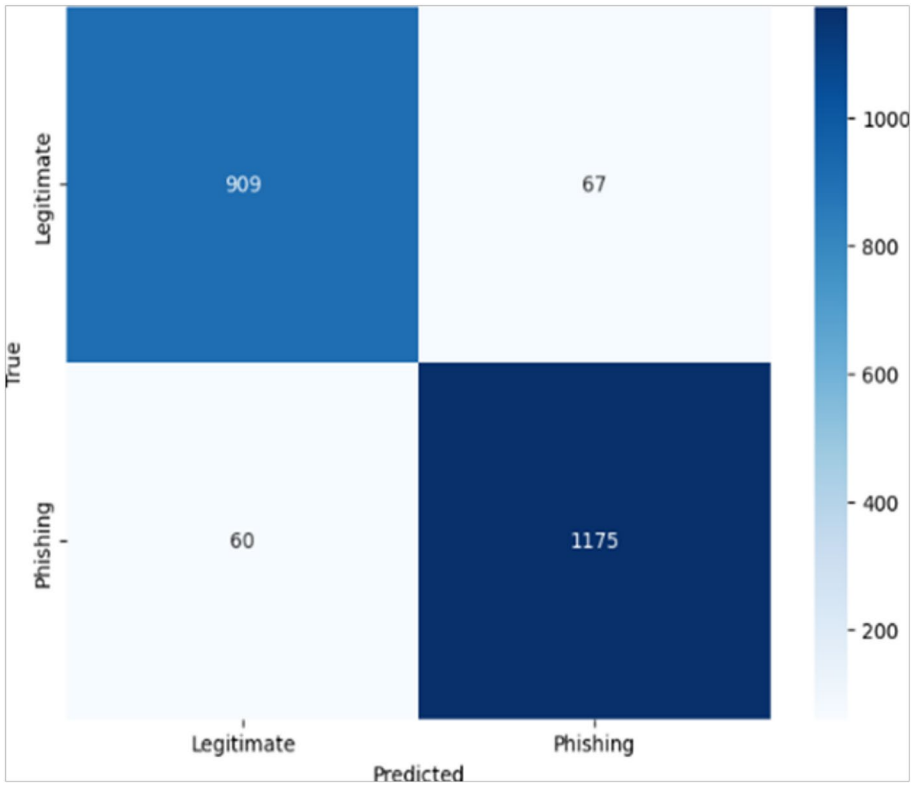

Fig. 5. KNN model Confusion Matrix.The confusion matrix of the KNNs model is shown, highlighting its classification performance in detecting phishing instances based on proximity-based feature matching.

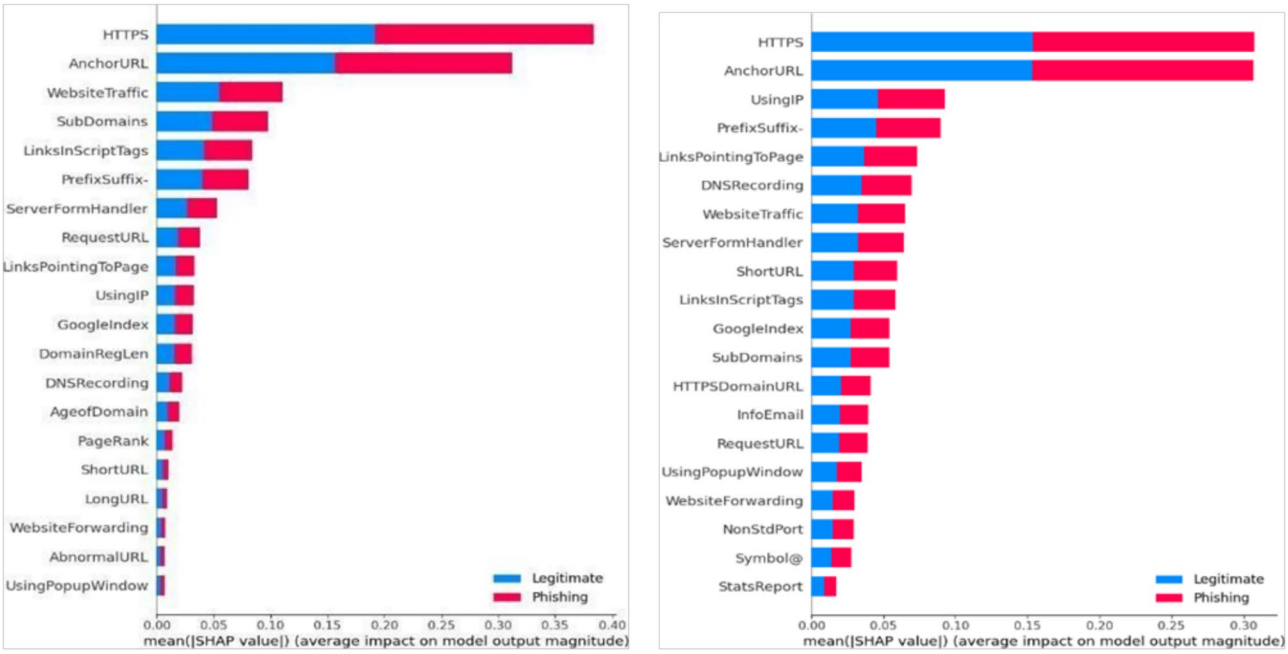

Fig. 6. Feature importance ranked by SHAP for RF (Right) and SVM (Left).SHAP value plots for SVM and RF models, ranking the features based on their contribution to the final prediction. These visualizations enhance interpretability by explaining the impact of each feature on the model's decisions.

Algorithms	F1-score	Precision	Recall	Accuracy
KNN	88.0	87.0	89.0	86.0
SVM	87.0	85.0	88.0	86.0
LR	86.0	84.0	88.0	85.0
DT	91.0	90.0	91.0	90.0
RF	92.0	91.0	93.0	90.0

Table 3. Models Performance Metrics Without SHAP (in %). This table shows the baseline performance of ML models without incorporating SHAP explainability. It serves as a comparison point to highlight the benefits gained through SHAP-based interpretability in the previous table.

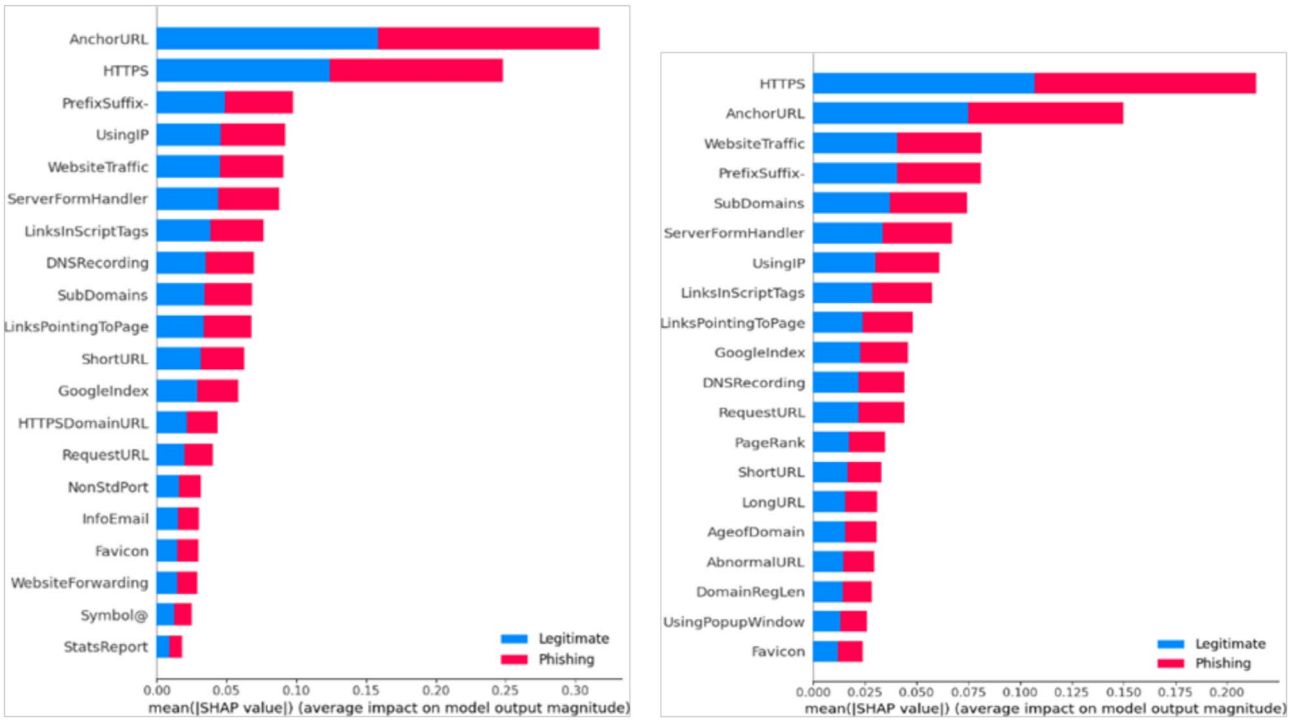

Fig. 7. Feature importance ranked by SHAP for LR (Right) and KNN (Left).This figure displays SHAP feature rankings for the KNN and LR models. It helps interpret which input variables most influenced the prediction in both classifiers.

Comparison with state-of-the-art methods

To critically assess the performance of the proposed method, we conducted a comparative evaluation against several state-of-the-art approaches, including classical ML models, deep learning architectures, hybrid frameworks, and large language model (LLM)-based detectors. Table 4 summarizes this comparison in terms of accuracy, precision, recall, and F1-score. Only studies reporting directly comparable metrics were included.

The proposed RF+SHAP model achieved 97.0% accuracy and an F1-score of 97.3%, surpassing traditional ML classifiers and deep learning-based methods. The baseline RF model performed competitively with 96.25% accuracy, but lacked the interpretability and feature transparency introduced by SHAP. Deep learning models, such as CNN+LSTM and DNN, achieved strong results in the range of 92–93%, but their higher computational costs and limited explainability make them less suited for lightweight, real-world deployments. More recent LLM-based detectors, such as DeepSeek R1 Distill, underperformed significantly with only 75% accuracy, indicating that large models trained for general purposes may not yet be optimized for phishing detection tasks. Overall, these results confirm that the proposed RF+SHAP framework not only achieves state-of-the-art performance but also addresses a critical research gap by providing explainability and resource efficiency. This balance of accuracy, interpretability, and practicality makes the approach particularly relevant for real-world cybersecurity infrastructures.

Discussion

The experimental evaluation of multiple ML models for phishing detection provides clear evidence of the contribution of SHAP in enhancing both predictive performance and interpretability. Across all tested models, the

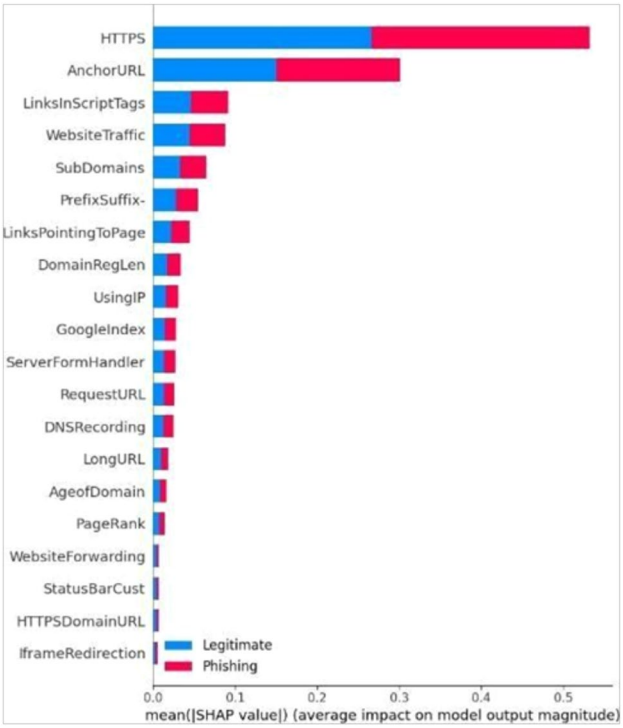

Fig. 8. Feature importance ranked by SHAP for DT.The SHAP plot illustrates the top-ranked features influencing the DT model's predictions. This improves the transparency and trustworthiness of the model by identifying how feature values drive classification.

Model / Paper Name	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Proposed RF + SHAP	97.0	97.3	97.0	97.6
RF (Baseline)[66]	96.25	96.2	97.6	98.3
SVM[67]	94.2	94.8	93.5	96.1
CNN+LSTM (IPDS)[68]	93.28	93.29	93.30	93.27
DNN[69]	92.89	92.21	92.75	93.07
LightGBM[70]	95.1	95.3	95.2	95.5
DeepSeek R1 Distill Qwen 14B Q8[71]	75	76	81	72

Table 4. Comparison of the Proposed RF+SHAP Model with State-of-the-Art Phishing Detection Approaches This table compares the proposed phishing detection model with existing state-of-the-art approaches. It includes models such as RF (baseline), SVM, CNN+LSTM, and DNN, showing the superiority of the proposed model across multiple performance metrics.

integration of SHAP resulted in consistent improvements in accuracy, precision, recall, and F1-score. As shown in Table 4, the proposed RF+SHAP approach achieved 97.0% accuracy, outperforming the baseline RF at 96.25% and surpassing other state-of-the-art methods such as SVM and CNN+LSTM. These findings demonstrate that SHAP facilitates a more effective utilization of features and contributes to a tangible performance advantage. The most significant performance gain was observed in RF (Fig. 6), where accuracy increased from 90.0% without SHAP to 97.0% with SHAP. This improvement reflects SHAP's capability to identify and prioritize the most discriminative features, which was also evident from the SHAP feature importance plots presented in the results. Similarly, the KNN model demonstrated improved balance between precision and recall when SHAP was applied, leading to fewer cases of legitimate websites being misclassified as phishing. This reduction in false positives, confirmed by the confusion matrices, is especially relevant for real-world deployment, where excessive false alarms can reduce user trust. The interpretability provided by SHAP further strengthens its role in phishing detection. Apart from predictive measures, the capacity to interpret model predictions is essential for cybersecurity professionals. SHAP attributes contribution scores to all features, thus demonstrating the features that influence predictions. The experiments indicated that URL-related features and domain indicators made the greatest contribution to classification results, consistent with prior research and validating model outputs for integrity. The findings allow security practitioners to better understand detection mechanisms and maximize defense mechanisms.

Algorithmic variability of performance was observed as well, with ensembling methods such as RF being the highest on average, but other classification methods, such as DT, SVM, and KNN, also showed concrete improvement when applying SHAP. This is proof of SHAP's adaptability across model families, suggesting that ensembles would inherently be more precise, but SHAP's interpretability and incremental gains in accuracy apply across a broad range of algorithms. The experimental outcome, thus, supports the extensive applicability of SHAP to phishing detection tasks. There are, however, certain boundaries to consider. The training and test data used is large and general, but phishing tactics are quickly developing, and new techniques emerging might not be included. This poses a question regarding generalizability to novel patterns, which might influence real-world performance. Moreover, the inclusion of SHAP into the workflow presented computational latency during experimentation. The calculation of SHAP values was observed to increase processing time, particularly for larger folds, which may restrict its applicability in scenarios requiring near real-time responses. Another source of variability lies in the sensitivity to hyperparameters and random initialization. Although five-fold cross-validation reduced this effect, some fluctuations were still observed, indicating that further optimization could stabilize outcomes. While SHAP improved both interpretability and performance, the experiments also revealed that interpretability alone does not guarantee overall model robustness. The models remain susceptible to adaptive phishing strategies and potential adversarial manipulation. Fairness and resilience against evolving threats are not addressed directly by SHAP explanations, meaning that interpretability must be considered alongside robustness and security-oriented evaluation metrics.

Conclusion and future work

The integration of SHAP with supervised ML models improves both model interpretability and predictive performance in detecting phishing websites. In our experiments, the RF model combined with SHAP outperformed traditional models such as SVM, DT, and standard RF across all key performance metrics, including accuracy, precision, recall, and F1-score. SHAP enables the identification of the most influential features, allowing the model to focus on characteristics that strongly indicate phishing attempts and enhancing the transparency of decision-making processes. Compared to generic feature selection approaches, the proposed method more effectively highlights features with substantial impact on predictions, contributing to a reliable and robust phishing detection system. The five-fold cross-validation and evaluation across multiple supervised algorithms confirm the consistency of these results, showing that models incorporating SHAP not only achieve higher performance but also provide interpretable insights useful for cybersecurity practitioners.

Nevertheless, certain limitations remain. The dynamic and evolving nature of phishing attacks requires periodic retraining or adaptive mechanisms to maintain performance. The computational complexity of SHAP, particularly with very large datasets or deep architectures, can increase training and evaluation time. Furthermore, model generalizability is constrained by the dataset scope; testing primarily on structured URL-based features may not fully capture the diversity of phishing strategies, including those embedded in multimedia or multi-lingual platforms.

Building upon these findings, future research will explore several directions. First, expanding experimentation with larger, more diverse, and real-world datasets including multi-lingual and mobile phishing samples will help validate the scalability of the framework. Second, the development of lightweight SHAP-based variants or approximation techniques can improve computational efficiency, making the approach more suitable for low-resource environments such as IoT and edge devices. Third, integration with advanced deep learning architectures, such as transformer-based models, coupled with explainability tools, offers a promising avenue for combining high accuracy with interpretability. Fourth, adversarial robustness against phishing attempts that actively attempt to evade detection should be systematically studied. Finally, practical deployment and evaluation in real-world cybersecurity infrastructures, such as email gateways, browser plug-ins, or enterprise firewalls, will help assess usability and operational value.

Overall, the proposed approach demonstrates that combining SHAP with RF enhances phishing detection performance while ensuring transparency. By addressing the above limitations and extending research in the proposed directions, future studies can further advance the field toward highly accurate, interpretable, and resource-efficient phishing detection solutions suitable for dynamic and large-scale cybersecurity applications.

Data availability

The dataset used in this research is publicly accessible in the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/327/phishing+websites. This benchmark dataset, provided by Mohammad et al., is commonly employed in phishing detection studies and supports reproducibility and comparability of results across studies.

Received: 29 July 2025; Accepted: 6 November 2025
Published online: 25 November 2025

References

1. Somesha, M., Pais, A. R., Rao, R. S. & Rathour, V. S. Efficient deep learning techniques for the detection of phishing websites. *Sādhanā* **45**, 1–18 (2020).
2. Alkawaz, M. H., Steven, S. J. & Hajamydeen, A. I. Detecting phishing website using machine learning. In *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, 111–114 (OrganizationIEEE, 2020).
3. Jha, A. K., Muthalagu, R. & Pawar, P. M. Intelligent phishing website detection using machine learning. *Multimed. Tools Appl.* **82**, 29431–29456 (2023).
4. Alazaidah, R. et al. Website phishing detection using machine learning techniques. *J. Stat. Appl. Probab.* **13**, 119–129 (2024).

5. Ojewumi, T. O. et al. Performance evaluation of machine learning tools for detection of phishing attacks on web pages. *Sci. Afr.* **16**, e01165 (2022).
6. Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E. & Fujita, H. Deep learning for phishing detection: Taxonomy, current challenges and future directions. *IEEE Access* **10**, 36429–36463 (2022).
7. Ramana, A., Rao, K. L. & Rao, R. S. Stop-phish: An intelligent phishing detection method using feature selection ensemble. *Soc. Netw. Anal. Min.* **11**, 110 (2021).
8. Das Guptta, S., Shahriar, K. T., Alqahtani, H., Alsalman, D. & Sarker, I. H. Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Ann. Data Sci.***11**, 217–242 (2024).
9. Alkhalil, Z., Hewage, C., Nawaf, L. & Khan, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Front. Comput. Sci.* **3**, 563060 (2021).
10. Rani, L. M., Foozy, C. F. M. & Mustafa, S. N. B. Feature selection to enhance phishing website detection based on URL using machine learning techniques. *J. Soft Comput. Data Min.* **4**, 30–41 (2023).
11. Kathrine, G. J. W., Praise, P. M., Rose, A. A. & Kalaivani, E. C. Variants of phishing attacks and their detection techniques. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 255–259 (OrganizationIEEE, 2019).
12. Zamir, A. et al. Phishing web site detection using diverse machine learning algorithms. *Electron. Libr.* **38**, 65–80 (2020).
13. Shafin, S. S. An explainable feature selection framework for web phishing detection with machine learning. *Data Sci. Manage.* **8**, 127–136 (2025).
14. Almheiri, S. J., Shah, A. A., Abbas, S., Ahmad, M. & Khan, M. A. Smart sustainable cyber security: Modelling an interpretable and transparent threat detection with explainable artificial intelligence. *Discov. Sustain.* **6**, 442 (2025).
15. Aljofey, A., Jiang, Q., Qu, Q., Huang, M. & Niyigena, J.-P. An effective phishing detection model based on character level convolutional neural network from URL. *Electronics* **9**, 1514 (2020).
16. Saied, M., Adjogble, F., Guirguis, S., Hemmji, M. & Warschat, J. A framework for systematic scientific research management. In *2023 Portland International Conference on Management of Engineering and Technology (PICMET)*, 1–16, https://doi.org/10.23919/PICMET59654.2023.10216819 (2023).
17. Nti, I. K., Narko-Boateng, O., Adekoya, A. F. & Somanathan, A. R. Stacknet based decision fusion classifier for network intrusion detection. *Int. Arab J. Inf. Technol.* **19**, 478–490 (2022).
18. Bahaghighat, M., Ghasemi, M. & Ozen, F. A high-accuracy phishing website detection method based on machine learning. *J. Inf. Secur. Appl.* **77**, 103553 (2023).
19. Anupam, S. & Kar, A. K. Phishing website detection using support vector machines and nature-inspired optimization algorithms. *Telecommun. Syst.* **76**, 17–32 (2021).
20. Mahajan, R. & Siddavatam, I. Phishing website detection using machine learning algorithms. *Int. J. Comput. Appl.* **181**, 45–47 (2018).
21. Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B. & Joga, S. R. K. Phishing detection system through hybrid machine learning based on URL. *IEEE Access* **11**, 36805–36822 (2023).
22. Jain, A. K. & Gupta, B. B. Towards detection of phishing websites on client-side using machine learning based approach. *Telecommun. Syst.* **68**, 687–700 (2018).
23. Ojewumi, T. O. et al. Performance evaluation of machine learning tools for detection of phishing attacks on web pages. *Sci. Afr.* **16**, e01165 (2022).
24. Sahingoz, O. K., Buber, E., Demir, O. & Diri, B. Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **117**, 345–357 (2019).
25. Basit, A., Zafar, M., Javed, A. R. & Jalil, Z. A novel ensemble machine learning method to detect phishing attack. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 1–5 (OrganizationIEEE, 2020).
26. Altaher, A. Phishing websites classification using hybrid SVM and KNN approach. *Int. J. Adv. Comput. Sci. Appl.***8** (2017).
27. Adeyemo, V. E., Balogun, A. O., Mojeed, H. A., Akande, N. O. & Adewole, K. S. Ensemble-based logistic model trees for website phishing detection. In *Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8-9, 2020, Revised Selected Papers 2*, 627–641 (OrganizationSpringer, 2021).
28. Siddiq, M. A. A., Arifuzzaman, M. & Islam, M. Phishing website detection using deep learning. In *Proceedings of the 2nd International Conference on Computing Advancements*, 83–88 (2022).
29. Roy, S. S., Awad, A. I., Amare, L. A., Erkihun, M. T. & Anas, M. Multimodel phishing URL detection using LSTM, bidirectional LSTM, and GRU models. *Future Internet* **14**, 340 (2022).
30. Almseidin, M., Zuraiq, A. A., Al-Kasassbeh, M. & Alnidami, N. Phishing detection based on machine learning and feature selection methods. *Int. J. Online Biomed. Eng.* **13**, 171–173 (2019).
31. Yang, P., Zhao, G. & Zeng, P. Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access* **7**, 15196–15209 (2019).
32. Alsubaei, F. S. , Almazroi, A. A. & Ayub, N. A novel hybrid deep learning framework for cybercrime forensics. (IEEE Access, Enhancing Phishing Detection, 2024).
33. Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E. & Fujita, H. Deep learning for phishing detection: Taxonomy, current challenges and future directions. *IEEE Access* **10**, 36429–36463 (2022).
34. Ahmad, S. et al. Across the spectrum in-depth review AI-based models for phishing detection. *IEEE Open J. Commun. Soc.* **6**, 2065–2089. https://doi.org/10.1109/OJCOMS.2024.3462503 (2025).
35. Ali, W. & Ahmed, A. A. Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. *IET Inf. Secur.* **13**, 659–669 (2019).
36. Zaimi, R., Hafidi, M. & Mahnane, L. A permutation importance based feature selection method and deep learning model to detect phishing websites. *Res. Square* (2024).
37. Wei, Y. & Sekiya, Y. Feature selection approach for phishing detection based on machine learning. In *International Conference on Applied CyberSecurity*, 61–70 (OrganizationSpringer, 2021).
38. Alenezi, R. & Ludwig, S. A. Explainability of cybersecurity threats data using shap. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–10 (OrganizationIEEE, 2021).
39. Elsadig, M. et al. Intelligent deep machine learning cyber phishing URL detection based on BERT features extraction. *Electronics* **11**, 3647 (2022).
40. Saied Essa, M. & Kamal Guirguis, S. Evaluation of tree-based machine learning algorithms for network intrusion detection in the internet of things. *IT Prof.***25**, 45–56, https://doi.org/10.1109/MITP.2023.3303919 (2023).
41. Saied, M., Guirguis, S. & Madbouly, M. A comparative analysis of using ensemble trees for botnet detection and classification in IoT. *Sci. Rep.* **13**, 21632. https://doi.org/10.1038/s41598-023-48681-6 (2023).
42. Khubaib, M. et al. Data-efficient wheat disease detection using shifted window transformer: Enhancing accuracy, sustainability, and global food security. *IEEE Trans. Consum. Electron.* 1–1, https://doi.org/10.1109/TCE.2025.3582267 (2025).
43. Ashraf, I. et al. Enhancing micro-expression recognition with broadbent attention mechanism: A high-performance approach for emotion detection. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, 189–194, https://doi.org/10.1109/HONET63146.2024.10822916 (2024).
44. Akram, N. et al. Online recruitment fraud (ORF) detection using deep learning approaches. *IEEE Access* **12**, 109388–109408. https://doi.org/10.1109/ACCESS.2024.3435670 (2024).

45. Akhunzada, A., Al-Shamayleh, A. S., Zeadally, S., Almogren, A. & Abu-Shareha, A. A. Design and performance of an ai-enabled threat intelligence framework for IoT-enabled autonomous vehicles. *Comput. Electr. Eng.* **119**, 109609. https://doi.org/10.1016/j.compeleceng.2024.109609 (2024).
46. Khan, M. A. et al. 7th International Congress on Human-Computer Interaction. *Optimiz. Robot. Appl. (ICHORA)*1–7, 2025. https://doi.org/10.1109/ICHORA65333.2025.11017101 (2025).
47. Rehman, A. *et al.* Enhanced sign language detection with deep CNN: Achieving accuracy in hand gesture recognition. In *2024 5th International Conference on Innovative Computing (ICIC)*, 1–6, https://doi.org/10.1109/ICIC63915.2024.11116573 (2024).
48. Kehkashan, T., Alsaeedi, A., Yafooz, W. M. S., Ismail, N. A. & Al-Dhaqm, A. Combinatorial analysis of deep learning and machine learning video captioning studies: A systematic literature review. *IEEE Access* **12**, 35048–35080. https://doi.org/10.1109/ACCESS.2024.3357980 (2024).
49. Saied, M., Guirguis, S. & Madbouly, M. Review of artificial intelligence for enhancing intrusion detection in the internet of things. *Eng. Appl. Artif. Intell.* **127**, 107231. https://doi.org/10.1016/j.engappai.2023.107231 (2024).
50. Cil, A. E. & Yildiz, K. SFIX:scalable financial-oriented interpretable explanation. *Internet Things* **33**, 101713. https://doi.org/10.1016/j.iot.2025.101713 (2025).
51. Cil, A. E. & Yildiz, K. A systematic literature review on applications of explainable artificial intelligence in the financial sector. *Internet Things* **33**, 101696. https://doi.org/10.1016/j.iot.2025.101696 (2025).
52. Minocha, S. & Singh, B. A novel phishing detection system using binary modified equilibrium optimizer for feature selection. *Comput. Electr. Eng.* **98**, 107689 (2022).
53. Hussain, S. et al. A novel feature engineered-catboost-based supervised machine learning framework for electricity theft detection. *Energy Rep.* **7**, 4425–4436 (2021).
54. Gualberto, E. S., De Sousa, R. T., Thiago, P. D. B., Da Costa, J. P. C. & Duque, C. G. From feature engineering and topics models to enhanced prediction rates in phishing detection. *IEEE Access* **8**, 76368–76385 (2020).
55. Wahyudi, R. et al. Algorithm evaluation for classification "phishing website" using several classification algorithms. In *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, 265–270 (OrganizationIEEE, 2018).
56. Odufisan, O. I., Abhulimen, O. V. & Ogunti, E. O. Harnessing artificial intelligence and machine learning for fraud detection and prevention in Nigeria. *J. Econ. Criminol.* **7**, 100127 (2025).
57. Puri, N., Saggar, P., Kaur, A. & Garg, P. Application of ensemble machine learning models for phishing detection on web networks. In *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 296–303 (OrganizationIEEE, 2022).
58. Alhaji, U. M., Adewumi, S. E. & Yemi-peters, V. I. Classification of phishing attacks using machine learning algorithms: A systematic literature review. *J. Adv. Math. Comput. Sci.* **40**, 26–44 (2025).
59. Makhdoomi, P. M. S. *et al.* Network-based intrusion detection: A comparative analysis of machine learning approaches for improved security. *J. Cyber Secur. Technol.* 1–28 (2025).
60. Fletcher, G. & Shi, T. Financial fraud detection with self-attention mechanism: A comparative study. *J. Comput. Sci. Softw. Appl.* **5**, 10–18 (2025).
61. Bacanin, N. et al. Addressing feature selection and extreme learning machine tuning by diversity-oriented social network search: An application for phishing websites detection. *Complex Intell. Syst.* **9**, 7269–7304 (2023).
62. Sawe, L., Gikandi, J., Kamau, J. & Njuguna, D. Sentence level analysis model for phishing detection using KNN. *J. Cybersecur (2579-0072).* **6**, 25 (2024).
63. Mohammad, R. & McCluskey, L. Phishing websites. How Dpublished UCI machine learning repository (2012). noteDOI: https://doi.org/10.24432/C51W2X
64. Altwaijry, N., Al-Turaiki, I., Alotaibi, R. & Alakeel, F. Advancing phishing email detection: A comparative study of deep learning models. *Sensors* **24**, 2077 (2024).
65. Ahmad, S. et al. Across the spectrum in-depth review ai-based models for phishing detection. *IEEE Open J. Commun. Soc.* 1–1, https://doi.org/10.1109/OJCOMS.2024.3462503 (2024).
66. Jalil, S., Usman, M. & Fong, A. Highly accurate phishing URL detection based on machine learning. *J. Ambient. Intell. Humaniz. Comput.* **14**, 9233–9251 (2023).
67. Krishna Reddy, V., Sai, Y. N., Keerthi, T. & Reddy, K. A. Detection of phishing website using support vector machine and light gradient boosting machine learning algorithms. In *International Conference on Innovative Computing And Communication*, 297–308 (OrganizationSpringer, 2023).
68. Adebowale, M. A., Lwin, K. T. & Hossain, M. A. Intelligent phishing detection scheme using deep learning algorithms. *J. Enterp. Inf. Manag.* **36**, 747–766 (2023).
69. Gopal, S. et al. Autoencoder-based architecture for identification and mitigating phishing URL attack in IoT using DNN. *J. Inst. Eng. India Ser. B* **104**, 1227–1240 (2023).
70. Foozy, C. F. M., Anuar, M. A. I., Maslan, A., Adam, H. A. M. & Mahdin, H. Phishing URLs detection using Naives Baiyes, random forest and Lightgbm algorithms. *Int. J. Data Sci.* **5**, 56–63, https://doi.org/10.18517/ijods.5.1.56-63.2024 (2024).
71. Thapa, J., Chahal, G., Gabreanu, S. V. & Otoum, Y. Phishing detection in the gen-ai era: Quantized llms vs. classical models. **2507**, 07406 (2025).

Acknowledgements

Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2025R97), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. We also acknowledge the technical insights and collaborative support provided by the VLCMatrix Lab during this research.

Author contributions

T. Kehkashan conceptualized the study, designed the methodology, and drafted the main manuscript. M. Abdelhaq and A. S. Al-Shamayleh contributed to the development of the predictive model and the interpretation of results. A. Akhunzada provided technical oversight, contributed to refining the manuscript, and supervised the overall research direction. A. I. A. A. Ahmed contributed to the literature review and data curation. N. Huda, and I. Ashraf supported data preprocessing, experiment setup, and results visualization under supervision. All authors reviewed and approved the final version of the manuscript.

Funding

This research was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2025R97), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to I.A.Y. or A.I.A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.