**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 15, No. 6, December 2025, pp. 5863~5878
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i6.pp5863-5878 

# Exploring feature engineering and explainable AI for phishing website detection: a systematic literature review

**Norah Alsuqayh, Abdulrahman Mirza, Areej Alhogail**
Information Systems Department, College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia

**Article Info**

***Article history:***

Received Apr 14, 2025
Revised Aug 5, 2025
Accepted Sep 14, 2025

***Keywords:***

Explainable artificial intelligence
Feature engineering
Machine learning
Phishing detection
Phishing websites

**ABSTRACT**

Detecting phishing websites is a rapidly evolving field aimed at identifying and mitigating cyberattacks targeting individuals, organizations, and governments. Ongoing progress in artificial intelligence (AI) has the potential to revolutionize phishing detection by enhancing model accuracy and improving transparency through explainable AI (XAI). However, significant challenges remain, particularly in integrating feature engineering with XAI to address sophisticated phishing strategies including zero-day attacks, that evade traditional detection mechanisms. To overcome these challenges, this examines the impact of feature engineering and XAI in phishing detection, emphasizing their ability to enhance accuracy while providing interpretability. By integrating feature extraction with interpretable models, these techniques improve decision-making transparency and system robustness. This paper presents the first systematic literature review (SLR) focusing on the impact of feature engineering and XAI on state-of-the-art phishing detection approaches. Additionally, it identifies critical research gaps and challenges, including scalability issues, the evolution of phishing techniques, and balancing complexity with interpretability. The findings provide valuable academic insights while offering practical recommendations for developing accurate and interpretable phishing detection systems, aiding organizations in strengthening cybersecurity measures.



***Corresponding Author:***

Norah Alsuqayh
Information Systems Department, College of Computer and Information Science
King Saud University
Riyadh 11543, Saudi Arabia
Email: n.alsuqayh@gmail.com

## 1. INTRODUCTION

Today, the widespread use of technology in many activities conducted by individuals and organizations has greatly simplified life and controlled transactions and therefore has resulted in a simultaneous rise in the sophistication and rate of cyber threats [1]. A cyberattack is defined as the malicious exploitation of computer networks, information systems, and infrastructure [2]. This malicious violation of computing resources is accomplished by using various methods to steal, alter, or destroy financial data, disable systems and networks, and commit identity theft [3].

One cyber threat is phishing, which has emerged a significant concern recently due to its increasing occurrence [1], [2]. Phishing employs social engineering and technical methods to steal personal identity information (PII) and financial credentials. Social engineering deceives victims into trusting the source while directing them to fraudulent websites. A phishing scenario is illustrated in Figure 1, in which a malicious actor fabricates a website that mimics a respectable and well-known company, such as Amazon. Next, using a variety of platforms, including social media and e-mails, the attacker sends the related link to many

possible targets. If a victim falls for the scam, they may access the fake website and provide vital information, and the user's credentials are effectively acquired by the attacker. The attacker then uses the stolen login credentials to gain access to the intended website and commit fraud [4].

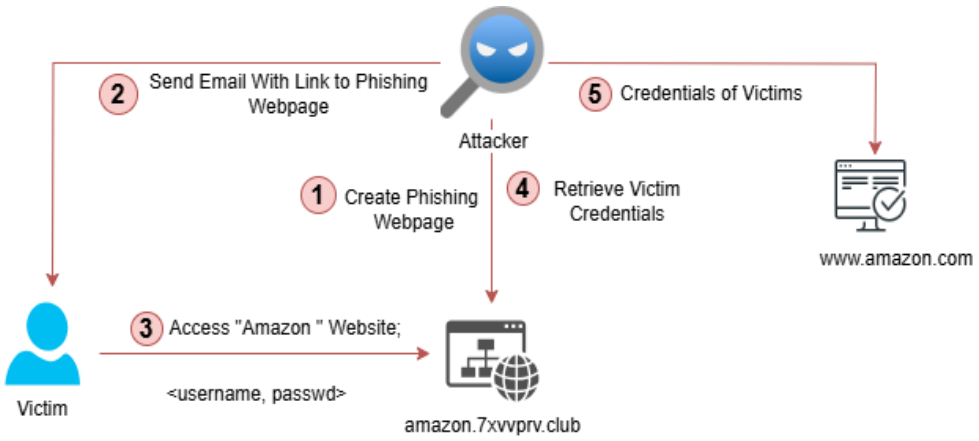



Figure 1. Steps involved in a common phishing scenario [4]

Phishing attacks are not limited to information collection; they are also the most popular way to disseminate ransomware and other malicious software. According to surveys, the financial damage of cybercrime surpassed $6 trillion each year, and is continually increasing. Furthermore, data about the Business Email Compromise (BEC) revealed in 2019 that phishing assaults were responsible for an astounding $26 billion in financial damages [1]. Moreover, a record was set for phishing in 2022 when the anti-phishing working group (APWG) recorded more than 4.7 million attacks, according to its Phishing Activity Trends Report from the fourth quarter of that year [2]. Additionally, the October 2022 sample showing 101,104 phishing e-mail subjects was the largest such sample that APWG had ever seen.

According to recent studies, XAI combined with feature engineering techniques can enhance phishing website detection systems. This is because XAI can provide both accurate predictions and interpretable insights into model behavior. XAI techniques, such as SHAP, enable users to realize the importance of individual features (*e.g.*, uniform resource locator (URL)-based and content-based) to model decisions, increasing confidence and transparency in phishing detection systems. Some of the current approaches have been seen to provide reasonable solutions to the problem of phishing; however, they have some drawbacks including the ability to adapt to zero-day attacks [3], the issue of interpretability [4] and the problems of dealing with imbalanced datasets and the scalability question. To overcome these limitations and since the nature of phishing threats is ever changing, researchers must look for new ways and techniques. Some of the future recommendations for the enhancement of the phishing detection systems include integration of feature engineering with XAI for phishing detection to address these challenges. These approaches improve the efficiency of detection in addition to offering important information regarding the decisions made by the detection systems [5]. Through the incorporation of stable feature selection with the interpretable models, they enhance the performance of the system as well as the trust of the users.

This research connects the fields of cybersecurity, machine learning, and XAI by presenting a systematic review concentrated on both feature engineering and interpretability in phishing detection. The practical significances of this study are significant, contributing to ongoing discussions in the field of secure computing systems. Its results will be valuable to researchers, developers, and policymakers, ensuring its relevance and potential for future citation. The research seeks to understand recent developments in phishing website detection using feature engineering and XAI, analyzing their advantages, drawbacks, and potential paths forward for developing accurate and interpretable detection systems. Multiple studies have been conducted on many categories of phishing detection such as machine learning and deep learning methods; yet, to our knowledge, there has been a shortage of research that focuses on the combination of feature engineering methods with XAI to improve detection accuracy and interpretability. This underscores the importance of conducting deeper investigations to analyze and assess the significant for improving phishing detection systems.

This work represents the first SLR that comprehensively explores the impact of feature engineering and XAI on improving the accuracy and interpretability of phishing website detection systems. The paper

presents novel contributions, including: the classification of phishing detection research utilizing hybrid feature engineering methods, and the evaluation of XAI's role in elucidating model outputs. The review identifies research needs, including scalability, the shifting nature of phishing strategies, and the trade-off between model complexity and interpretability, so providing useful academic insights and a framework for future research. It provides practical direction for developing precise and transparent phishing detection systems, assisting enterprises in enhancing their cybersecurity frameworks and aggressively addressing advanced phishing attacks.

The organization of this review is as follows. In section two, a brief background on feature engineering, XAI and phishing detection methods is presented. Section three views the methodology of the study while section four conducts a systematic literature review of the state-of-the-art works related to phishing detection. In section five, a discussion of the main research studies in this area is provided, along with an exploration of new challenges. Section six presents the suggested directions for future studies based on the findings and finally the conclusion.

## 2. BACKGROUND

Phishing detection is a crucial part of the cybersecurity domain with the goal of identifying and preventing fraudulent attempts at stealing sensitive information. Feature engineering and XAI are important contributors to improving the robustness and reliability of phishing detection approaches. For systems to properly distinguish between legitimate and phishing activities, features should be selected, crafted, and optimized. This is supported by XAI, which ensures that these systems remain understandable and trustworthy through highlighting the decision-making process. Therefore, feature engineering and XAI improve the efficiency, explainability, and user friendliness of phishing detection systems.

### 2.1. Feature engineering

Phishing detection techniques improved by feature engineering which transforms raw data into related features that enhance the behavior of models. To effectively distinguish between benign and fake websites, phishing detection mechanisms can be based on URL characteristics and content-based indicators that are derived. Advanced techniques, such as Genetic Algorithms and Principal Component Analysis (PCA) can be used to filter the feature space to minimize the dimensionality and at the same time preserve most of the information content [6]. This guarantees that each of the features has its own unique contribution to the predictive ability of the model, by selecting and eliminating redundant features properly. This makes feature engineering a powerful process for enhancing the accuracy and efficiency of phishing detection systems, especially when composite features and domain-specific attributes are created.

### 2.2. Artificial intelligence (AI)

AI is a field that replicates human intelligence to enable systems to do things like predicting future trends in the stock market. It has replaced traditional methods and is comprised of subfields including machine learning and natural language processing that have altered the face of industries like healthcare, finance and autonomous systems. The acceleration in AI development can be attributed to the rapid progress made in research, especially in computer vision (CV) and speech recognition, which highlight the impact of AI on transforming the future of technology and society.

### 2.3. Explainable artificial intelligence (XAI)

In recent years, learning models revolutionized the landscape of automated prediction and decision-making. Artificial neural networks (ANN) and deep learning models have proven highly effective in handling complex tasks and achieving high performance [7]. Despite their performance gains, these models tend to lack transparency and are difficult to interpret. Incorporating interpretability as an additional layer during model development can enhance practical implementation and help identify and address deficiencies for three key reasons [6]:

- It helps ensure integrity in decision-making by enabling the detection and correction of biases present in the training dataset.
- It facilitates model robustness by identifying potential perturbations that may significantly alter the model's predictions.
- It ensures that only meaningful variables contribute to the output, promoting truthful causality and transparency in the model's reasoning process.

Applying XAI techniques in feature engineering for phishing website detection is crucial for several reasons [8], [9]. First, it supplies visible insights into how and why decisions are made behind a model's prediction of a website as phishing or authentic. Second, raises user trust, which is essential for the

acceptance of AI-driven security measures. Third, XAI provides an understanding of how different features such as URL and HyperText markup language (HTML) features influence predictions that result in more precise and trustworthy phishing detection. Lastly, XAI identifies the effect of individual features and their interactions and allows optimized feature engineering by focusing on the most influential factors and discarding redundant or irrelevant data.

### 2.4. Phishing detection methods

Phishing detection methods embrace diverse techniques and strategies to identify and mitigate phishing attacks on websites that are aimed at stealing sensitive information or credentials. Typically, multiple techniques are combined to detect and prevent this attack because phishing is complicated and there is no specific solution to completely prevent this threat. Figure 2 illustrates the phishing detection approaches—user awareness and software-based detection. In the following sections, we focus on discussing the software-based techniques in detail.

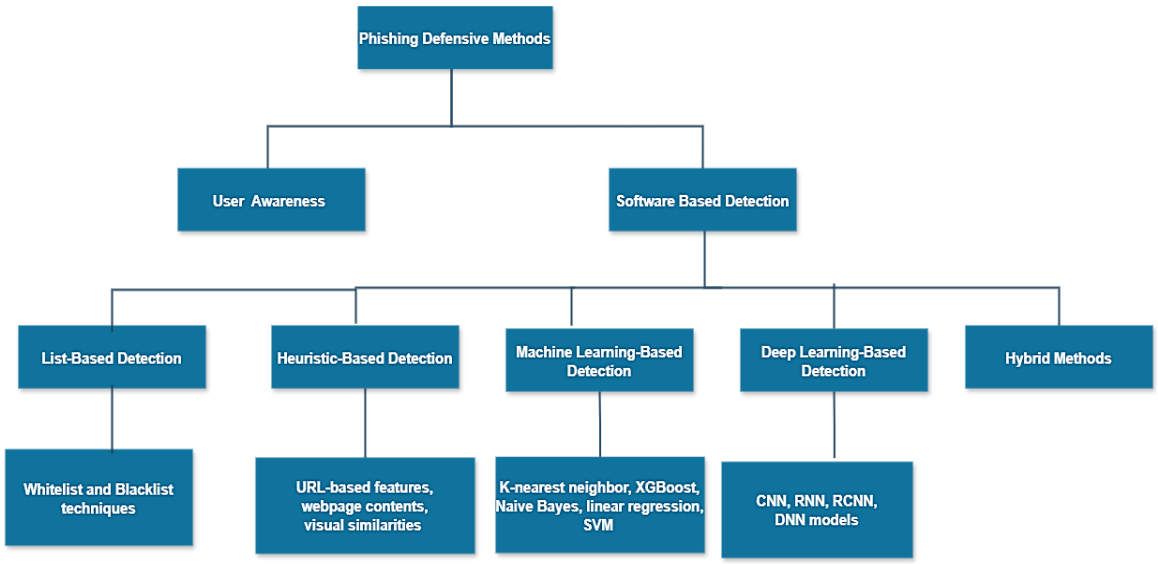



Figure 2. Phishing detection methods [10]–[12]

#### 2.4.1. List-based approach

List-based detection can be implemented in two forms: blacklist detection and whitelist detection [13], [14]. This approach is characterized by ease of implementation and strong operational effectiveness. However, it cannot efficiently identify a phishing attack due to problems with the update mechanisms of these lists [1], which requires a lot of human effort and time to update the lists [10]. The method fails to detect threats from new and unknown URLs, thus making it prone to zero-day attacks [11], [13]. Therefore, the black- and whitelist detection methods are currently less utilized.

#### 2.4.2. Heuristic-based approach

The heuristic approach can identify suspicious content based on indicative cues, thereby enhancing detection efficiency and minimizing phishing-related losses in a timely manner. Unlike the list-based approach, this technique has a high level of performance in detecting threats from new and unknown URLs [11]. However, it often has a relatively higher false positive rate (FPR) and tends to be time-consuming, as it depends on search engines and third-party services such as DNS queries [12]. In addition, the formulation of heuristic strategies is subjective and depends on expert knowledge or observable patterns in phishing attempts. This technique is performed by checking a web page's content, the website URL, or visual similarities.

#### 2.4.3. Machine learning (ML) approach

ML approaches for detecting phishing web pages have previously been extensively discussed [15], [16]. Since phishing detection involves categorizing webpages as either benign or phishing, the models employed are typically binary classifiers [11]. Each data point in the input dataset—such as a URL—is

labeled as either benign or phishing to enable the model to learn the distinguishing features of both classes [17]. Various feature engineering techniques are employed to reduce the number of features and enhance the efficiency and interpretability of dataset visualization [18]. Despite substantial progress in the identification of phishing URLs using ML techniques, several critical challenges remain. One major concern lies in the selection of effective training datasets that accurately represent both phishing and benign websites. Researchers must carefully balance the quantity of URLs used for training with the computational efficiency and scalability of the applied ML algorithms, ensuring both performance and practicality in real-world deployment [19]. Another key obstacle is featuring extraction, as machine learning models typically rely on manual engineering of features to capture relevant patterns [17]. Collecting certain types of features, particularly host-based features, is also time-consuming, which can hinder the efficiency of the phishing detection process [20]. One of the key challenges associated with handcrafted features is their limited generalizability to unseen data. Adversaries, such as phishers, may exploit this by identifying the specific features a model relies on and intentionally crafting URLs or webpages to evade detection.

#### 2.4.4. Deep learning (DL) approach

The robustness of DL algorithms has encouraged researchers to explore a range of techniques for website classification, including the extraction of both novel and established features—such as keyword frequency within URLs [21]. In phishing detection, DL techniques offer the potential to develop dynamic feature representations that can adapt to concept drift commonly observed in phishing data [11]. DL algorithms reduce the load of feature extraction and selection. In contrast to ML, DL presents several difficulties in contrast to ML, it necessitates a lengthy training period [22], [23] and excessive computer resources [24]. Furthermore, because these models work as "Blackbox" techniques, it is difficult to explain how the model arrived at a result [25]. Another problem with phishing detection that hasn't been thoroughly discussed yet is real-time detection [25]. DL-based phishing detection models also face the problem of overfitting, in which a model performs well on the training data but fails to generalize to new, unseen data, such as that required to detect phishing websites that were not part of the training [26]. Also, the datasets may contain some duplicate points, and it is challenging to find enough labelled data, and the distribution of real data and the dataset might be different, resulting in the potential requirement for adaptations. Most malicious websites are short-lived and are often offline by the time they are analyzed [27].

#### 2.4.5. Hybrid based approach

Hybrid detection techniques rely on the integration of two or more existing approaches to enhance the performance of phishing site detection [12]. For example, combining heuristics and ML can help form a better system [28]. Another type of hybrid model involves the combination of multiple machine learning algorithms, where the dataset is initially trained using one algorithm, and the resulting output is subsequently fed into a second algorithm for further training [29], [30]. Furthermore, DL methods can be mixed (*e.g.*, creating a convolutional neural network (CNN)–long short-term memory [LSTM] model for phishing detection) [31].

## 3. METHODOLOGY

The study's main goal is to systematically analyze how feature engineering techniques and explainable AI methods might enhance phishing website detection. The methodology comprises a PRISMA-guided systematic review of recent scholarly literature, incorporating quantitative and qualitative evaluation of ML and DL models, XAI frameworks (*e.g.*, SHAP, LIME), and hybrid feature selection techniques. In order to identify common themes, methodological advancements, and current problems, the review synthesizes more than thirty investigations.

This research uses an SLR methodology to discuss the roles of feature engineering and XAI in phishing website detection by investigating the recent techniques for phishing website detection. Moreover, how feature engineering and XAI can enhance the accuracy and interpretability of phishing website detection. Finally, the issues and limitations are associated with phishing website detection. In addition, we identified an appropriate database to deliver relevant results that are limited to a 5-year period between 2019 and 2024 located in ACM Digital, IEEE Explore, Elsevier, Springer, MDPI and Google Scholar. Literature limited to review articles, conference proceedings and researchers' theses. To identify relevant studies and narrow down the number of results included in this review, we followed the systematic review process as illustrated in Figure 3. The review process was divided into three sequential steps: identification, screening and selection.

The following search string was used to retrieve relevant articles: ("Feature engineering") OR ("XAI" OR "explainable AI" OR "explainable artificial intelligence") AND ("phishing detection" OR "phishing website detection"). From the initial search, 102 papers that involved feature engineering for

phishing website detection were obtained. In the screening process, 60 studies that were not in conformity with the requirements were excluded. In the final selection stage, we included 34 papers that met the inclusion criteria for this systematic literature review.

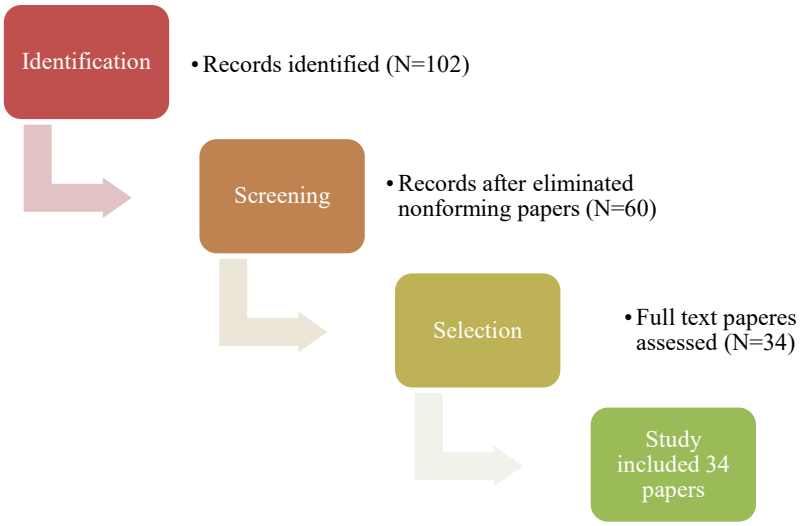



Figure 3. Phases of SLR selection process. N denotes the number of papers at each stage

## 4. RELATED WORKS

Phishing detection approaches, as discussed in section two, can be improved by combining feature engineering and XAI techniques. Feature engineering permits systems to select the most relevant attributes, which raise detection accuracy. Moreover, integrating XAI tools, such as Shapley additive explanations (SHAP) or local interpretable model-agnostic explanations (LIME), supply interpretability to enhance decision-making processes, encourage trust among users and stakeholders. This section is divided into five sub-sections corresponding to the recent phishing detection techniques outlined in section two, combined with feature engineering methods. It also demonstrates how the integration of feature engineering and XAI improves phishing detection models by addressing challenges like adaptability to evolving threats and balancing complexity with interpretability. Finally, this section highlights the limitations and unresolved issues associated with these approaches, paving the way for future research and practical advancements.

### 4.1. List-based approach

Study [16] introduced the automated individual whitelist, a unique anti-phishing strategy based on the Naïve Bayes (NB) classifier. By logging the IP addresses of all well-known login user interfaces (LUIs) that the user has visited, this technique creates a customized whitelist. The system creates a warning about a possible phishing attempt when the user tries to send private information to a LUI that is not on the whitelist. In contrast, Study [32] evaluated their suggested blacklist-based approach using a set of 38 characteristics and random forest (RF) and linear regression (LR) classifiers. The method successfully distinguished between fraudulently registered domains and valid ones with a 97% accuracy rate and a 2.5% FPR.

In order to improve the detection accuracy of phishing attacks, Barraclough *et al.* [33] combined heuristic methods, web content analysis, and blacklists in a machine learning framework that made use of extensive feature sets. The machine learning algorithms that were assessed were J48, JRip, NB, PART, and the adaptive neuro-fuzzy inference system (ANFIS). The best performance was given by PART, which had an execution time of 0.006 seconds and an accuracy of 99.33%. A three-phase attack detection technique using web traffic, web content, and URL features as input was proposed by Nathezhtha *et al.* [34]. According to experimental data, the suggested method detected both phishing and zero-day phishing attempts with an accuracy of 98.9%. Classifying XML-based URLs according to their semantic structural orientation was the subject of a separate study by Murthy *et al.* [35]. An accuracy of 97.36% was attained by their method.

### 4.2. Heuristic-based approach

Study [29] suggested a web phishing detection method that utilized integrated features from a website's text, graphics, and frames. They utilized ANFIS, support vector machine (SVM), and k-nearest

neighbor (K-NN) classifiers, with an accuracy of 98.3%. Feature selection was conducted via Information Gain and Chi-Square methodologies to improve model efficacy. Additionally, Rao *et al.* in [36] introduced a heuristic-based phishing detection method that analyzes the login page and the homepage of a website by utilizing hyperlink and URL-based characteristics. The method employed a Twin SVM classifier to identify intentionally registered phishing websites. Experimental results indicated that the Twin SVM surpassed other variations, attaining a recall of 98.33% and an accuracy of 98.05%.

Furthermore, the study in [37] sought to assess 12 static elements, including keywords and structural patterns, in selected phishing URLs and monitor their prevalence throughout contemporary phishing websites. Alongside this investigation, the researchers performed both quantitative and qualitative evaluations of behavioral patterns. This enabled the identification of critical components, including feature significance, inter-feature correlations, and similarities, which can facilitate the creation of novel heuristic methods or improve existing ones. In order to choose the best classifier, Ramana *et al.* [38] presented an ensemble-based phishing detection model that combines many machine learning methods, such as RF, decision tree (DT), and XGBoost. To improve classification performance, the study also used a number of feature selection strategies, including ANOVA, Information Gain, Fisher Score, Relief-F, and recursive feature elimination. When tested on the Mendeley phishing dataset, the model's accuracy was 98.45%, but it was 97.51% on the UCI phishing dataset. Lastly, Dooremaal *et al.* [39] presented a novel method for detecting phishing attacks by combining textual data from the document object model (DOM) structure with visual features taken from screenshots of webpages. With an overall detection accuracy of 99.66%, this hybrid approach dramatically decreased the phishing misclassification rate by 67%, from 1.02% to 0.34%.

### 4.3. ML approach

Study [30] related a multistage phishing detection model and presented an extensive CASE feature architecture, classifying features into four principal categories: Counterfeiting, Affiliation, Stealing, and Evaluation. The suggested method exhibited robust efficacy in practical phishing detection contexts, yielding efficient outcomes with minimized execution durations. A phishing detection algorithm with hybrid cumulative feature selection was proposed in [31]. The methodology utilizes various feature selection approaches, such as Chi-Square, gain ratio, information gain, Pearson correlation coefficient, and PCA, to divide the dataset into n subsets according to the chosen features. A variety of classifiers is employed for each partition, including SVM, NB, C4.5, RF, JRip, PART, and KNN. The RF classifier attained the best accuracy, with 98.24%. Study [40] provided a phishing detection framework utilizing a classifier to facilitate the comparative assessment of detection systems based on 87 distinct features. To mitigate the ephemeral nature of phishing websites, the authors created a dynamic dataset that may adapt over time. Their investigation indicated that webpage content was the least discriminative feature group, but external features—such as domain and hosting attributes—were the most informative. A maximum accuracy of 96.61% was attained by the utilization of hybrid features. Furthermore, applying filter-based ranking method with progressive elimination of less significant features improved the accuracy by 96.83%

In contrast, Gupta *et al.* [41] devised a streamlined phishing detection technique that utilizes merely nine lexical parameters, including URL length, for classification purposes. After assessing the strategy with many machine learning classifiers, the RF algorithm attained the greatest accuracy of 99.57%. Anupam and Kar [42] employed diverse URL-based characteristics—such as the length of the IP address and the validity of the HTTPS request—to categorize websites as phishing or real. A binary SVM classifier was utilized to determine an appropriate hyperplane for classification purposes. Four optimization strategies were employed to improve SVM performance: the bat algorithm, the firefly algorithm, the grey wolf optimizer (GWO), and the whale optimization algorithm. The GWO algorithm surpassed the firefly algorithm regarding detection accuracy.

### 4.4. DL approach

In [12], deep learning-based phishing detection model was proposed using a CNN architecture that relies solely on the website's URL and various feature representations. These include hand-crafted character embeddings, character-level TF-IDF, and character-level count vector features. Notably, the model does not require access to webpage content or any third-party services, nor does it depend on prior knowledge of phishing techniques. Instead, it captures informative and sequential patterns within URL strings for effective detection. The proposed model achieved an accuracy of 95.02% on a custom dataset and recorded accuracies of 98.58%, 95.46%, and 95.22% on three benchmark datasets, outperforming existing phishing URL detection models. In contrast, Vrbančič *et al.* [32] proposed an anti-phishing system that integrates URL-based, natural language processing (NLP)-based, and host-based features to train a range of ML and DL models, including K-NN, LR, SVM, gradient boosting (GB), AdaBoost, RF, and neural network (NN). Among these, the NN model achieved the highest accuracy, reaching 94.89% in phishing URL detection.

In [43], a character-level convolutional autoencoder (CAE) was developed within an anomaly detection framework for phishing detection. Experimental evaluation, conducted using ROC curve analysis and 10-fold cross-validation, demonstrated that the proposed model improved sensitivity by 3.98% compared to the most recent deep learning model. These results confirm the effectiveness of the CAE-based approach in identifying phishing threats. Xiao *et al.* [44] introduced a self-attention-based CNN model that incorporates a generative adversarial network (GAN) to synthesize phishing URLs for training purposes. The proposed architecture consists of four main components: the input block, attention block, feature block, and output block. By combining CNN with multi-head self-attention mechanisms, the model constructs a robust classifier capable of accurately detecting previously unseen phishing URLs. The classifier achieved an accuracy of 95.6%, outperforming baseline models—CNN-LSTM, standalone CNN, and standalone LSTM—by margins of 1.4%, 4.6%, and 2.1%, respectively. AlEroud *et al.* [45] employed GAN to generate URL-based phishing examples capable of evading detection. The synthesized examples were shown to effectively deceive both simple and advanced black-box machine learning-based phishing detection models.

### 4.5. Hybrid approach

Rao and Pais in [46] suggested an ensemble phishing detection model that incorporates extra trees, RF, and XGBoost classifiers. The model assesses the synergistic efficacy of heuristic and blacklist filtering strategies as a cohesive strategy, with an accuracy of 98.72%. Furthermore, Korkmaz *et al.* in [47] created a phishing detection system with a CNN that employs n-gram characteristics derived from URLs. Experimental findings demonstrated that unigrams produced the greatest categorization accuracy. The model attained an accuracy of 88.90% on the URL dataset by utilizing a specific set of 70 characters. Additionally, Orunsolu *et al.* in [48] suggested a phishing detection method that includes a feature selection module to extract pertinent information from URL structure, webpage attributes, and webpage activity using frequency assessment analysis. The methodology was assessed with NB and SVM classifiers. Experimental results indicated an efficient runtime of under 2,000 milliseconds, accompanied by robust performance metrics: 99.96% true positives, 99.96% true negatives, 0.04% false positives, and 0.04% false negatives. Also, Yu *et al.* [49] created a hybrid phishing detection model that combines various deep learning architectures for feature extraction and classification. A multilayer perceptron (MLP) processed custom features, CNN handled image-based features, and a recurrent neural network (RNN) managed text-based feature. The retrieved feature vectors were subsequently integrated using a classification network to get final predictions. The proposed model attained an overall accuracy of 97%.

Furthermore, Ariyadasa *et al.* [50] suggested a phishing detection method that integrates long-term recurrent convolutional networks with graph convolutional networks, employing both URL and HTML characteristics. The approach leverages the sophisticated analytical powers of graph neural networks in the anti-phishing sector. Experimental results indicated a detection accuracy of 96.42% and a false-negative rate of 0.036. Also, study [24] suggested a phishing website detection method that exclusively utilizes the URL, encapsulating its information into a two-dimensional tensor. This tensor is initially processed by a bidirectional long short-term memory (Bi-LSTM) network to extract global contextual information, subsequently followed by CNN to automatically identify the most pertinent components of the URL. The suggested model, PDRCNN, attained a detection accuracy of 97% and an AUC value of 99% in experimental assessments. Study [10] combined CNN and RF by employing character embedding techniques to transform URLs into fixed-size matrices, extracting features at various levels with CNN models, subsequently classifying these features using multiple RF classifiers, and ultimately producing prediction results through a winner-take-all method. A precise rate of 99.26% was attained on the benchmark data. Finally, Study [51] presented HTMLPhish, a phishing detection model that analyzes the HTML content of web pages through CNN to discern semantic relationships within the textual structure, eliminating the need for manual feature engineering. This methodology allows the model to adaptively manage novel features and generalize proficiently to previously unobserved test data. HTMLPhish attained a detection accuracy and true positive rate of 93%, illustrating its efficacy in recognizing phishing websites just through HTML content. Table 1 (in appendix) shows a summation of recent research on phishing detection models.

### 4.6. XAI in phishing website detection

To the best of our knowledge, the application of XAI in phishing detection remains relatively underexplored. The work in [52] explored the interpretability of phishing detection models by applying RF, and SVM in combination with XAI methods, including LIME and explainable boosting machines (EBM). The analysis showed that the most influential URL features, as identified by these techniques, closely matched typical phishing-related attributes. While study [13] explored the application of XAI techniques to enhance the detection of phishing attempts in emails. Their study emphasized the importance of specific words and phrases that significantly influence the classification decisions made by phishing detection models.

Additionally, study [14] proposed a multi-modal hierarchical attention model designed to learn deep phishing indicators from URL, textual, and visual modalities. The model incorporates two levels of attention mechanisms to facilitate the extraction of relevant features and to provide informative interpretability across different modalities. Experimental results demonstrated that the model not only enhances phishing detection performance but also offers hierarchical interpretability, improving transparency in the decision-making process. To improve interpretability, study [53] used a hybrid deep learning-based model that included explainable visual annotations superimposed on screenshots of phishing websites. A two-stage stacked ensemble learning technique was used by study [54], who applied GB and RF classifiers to 21 selected features from a dataset of 651,191 URLs. The accuracy of the suggested model was 97%. The model's decision-making process was then interpreted using XAI approaches, which were also used to examine each feature's contribution to the four-class prediction challenge, which included malware, phishing, defacement, and benign classifications.

Study [55] SHAP values were employed to interpret both individual machine learning models and ensemble models—including K-Means, RF, DT, CatBoost, LightGBM, AdaBoost, and a voting classifier—for phishing URL detection classification. Among these, the CatBoost classifier demonstrated superior performance across evaluation metrics. The use of SHAP values played a pivotal role in identifying the most influential features and understanding their effects on the model's outputs, thereby enhancing interpretability and trust in the classification process. Table 2 shows a summary of XAI and feature engineering approaches for phishing websites detection.

Table 2. Summary of XAI and feature engineering approaches for phishing websites detection

| Literature | Type of features | Feature engineering method | XAI technique | Performance metrics |
|---|---|---|---|---|
| [53] | URL | NLP techniques | LIME and EBM. | Precision, recall, F1 score and accuracy |
| [8] | Email | Local feature importance, text highlights as explanations | model-agnostic principles, local feature importance, and search-based explanation generation | False positive rate and classification thresholds |
| [9] | URL, webpage text and webpage image | Shared dictionary learning approach | Hierarchical Attention Mechanism, Attention Score Visualization | Precision, recall, F1 score and accuracy |
| [54] | URL, content and visual features | - | visual comparisons and logo recognition | Identification rate, detection rate, precision and recall |
| [55] | URL | - | ALE (Accumulated Local Effects) | Precision, recall, F1 score and accuracy |
| [56] | URL, content and behavioral features | - | SHAP | Precision, recall, F1 score and accuracy |

## 5. RESULTS AND DISCUSSION

Recently, phishing became a threat in the cybersecurity landscape, targeting users by mimicking legitimate websites to steal sensitive information. This research recognizes the effort on feature engineering and XAI into phishing website detection, with a notable increase in studies since 2019. These models not only enhance the accuracy of detection but also improve the interpretability, which are critical in high-stakes cybersecurity applications. This review enhances existing knowledge through integrating feature selection optimization and model interpretability—two elements frequently examined independently. It offers a comprehensive viewpoint crucial for developing resilient and transparent phishing detection systems. The study prioritizes the explainability of decisions and their reliability in essential security systems, in contrast to previous studies that concentrated exclusively on model accuracy.

A previous review by Safi and Singh [40] divided phishing detection techniques into five approaches; lists based, visual similarity, Heuristic, ML, and DL based techniques and among these, ML techniques have been applied the most. In addition, most studies based on study used ML techniques such as RF while CNN achieved the highest accuracy for detecting phishing websites. Similarly, Catal *et al.* [57], through a comprehensive literature review, recognized deep learning mechanisms for phishing detection. The study demonstrated that all models employed supervised deep learning algorithms and utilized data sources such as URL-and content-related features, third-party metadata about the website, and email data. Among these, DNNs and CNNs emerged as the most widely adopted architecture.

Despite the growing reliance on advanced learning algorithms, it is noteworthy that 72% of the analyzed studies did not implement any form of feature selection during model construction that may compromise both model efficiency and interpretability. Additionally, Subashini *et al.* [58] highlighted several challenges in phishing detection, including imbalanced datasets that can lead to biased classifiers and an

increased risk of false negatives. Moreover, attackers often leverage encrypted traffic to conceal malicious activities. Evasion techniques, such as URL obfuscation and adversarial tactics, further complicate detection by enabling phishing attempts to bypass ML models.

This study focuses on three fundamental aspects of phishing website detection: identifying state-of-the-art techniques for phishing website detection, evaluating the contribution of feature engineering and XAI in improving the performance and interpretability of phishing website detection systems, and delineating key challenges and limitations of applying feature engineering and XAI in phishing websites detection. These focal areas facilitate a comprehensive understanding of the domain and inform future advancements. As you see in section four, state-of-the-art approaches for phishing website detection leverage improvements in ML, DL, and hybrid techniques to resolve the complication of current phishing strategies. ML models, such as SVM [59] and RF [52] employ features like URL length, suspicious keywords for binary classification tasks. Feature engineering techniques such as PCA and RFE raise the accuracy of detection by eliminating redundant or irrelevant attributes. On the other hand, DL approaches, including CNNs [21] and LTSM [60], are capable of learning high-dimensional representations directly from raw inputs, obviating the need for extensive manual feature engineering. XAI tools were recently added into these systems to improve transparency.

Moreover, our findings add nuance by showing how integration of feature engineering and XAI mechanisms improves the accuracy and interpretability of phishing website detection by handling issues in the domain that discussed in section four. Feature engineering enables the selection of relevant features using techniques such as PCA [61] and FSM [48], thereby reducing dimensionality and improving generalization. By focusing on relevant attributes, phishing websites detection reach improved accuracy [61]. XAI methods, such as SHAP [56] and LIME [62] offering interpretability to understand the model decisions and therefore increase trust in the system. Integration treats challenges like zero-day attacks, and improve reliability, leading to more robust, interpretable, and user-friendly phishing websites detection systems [53].

Our review also identifies recurring challenges. One of the main challenges is feature selection, as specifying the effective attributes is complex due to the nature of phishing websites techniques, and irrelevant features can affect the performance of model [61]. Moreover, the complexity of high-dimensional data leads to increased computational costs and reduced model efficiency [58]. Additionally, attackers develop the latest techniques, such as obfuscation attacks, making it difficult for static feature sets to remain effective over time. Moreover, trade-off exists between accuracy and interpretability: while DL models often function as black boxes, making their decisions difficult to explain; conversely simpler models are more interpretable but may lack detection precision [62]. Another issue is scalability and real-time processing, as feature engineering and explainability techniques should work within high-traffic domains without any detection delays [41]. Finally, the absence of standardized datasets and evaluation protocols also hinders reproducibility and consistent benchmarking across studies [63].

This study establishes a robust basis for academic research and organizational application by integrating technological improvements with practical security requirements. From an academic standpoint, the amalgamation of feature engineering with XAI creates opportunities for the advancement of interpretable machine learning models that reconcile performance with transparency—an imperative factor in critical fields such as cybersecurity. The integration of hybrid techniques, classification of phishing detection measures, and examination of XAI methods such as SHAP and LIME enhance comprehension of model behavior and system weaknesses. The findings bring useful insights for cybersecurity experts, emphasizing the significance of developing detection systems that are both reliable and comprehensible, as well as adaptive to emerging threats. These insights enable trust among stakeholders, boost incident response methods, and guarantee compliance with regulations of cybersecurity. The review ultimately recommends for the further development of reliable next-generation AI systems and provides a framework for organizations aiming to enhance their digital defenses against progressively intricate phishing threats.

## 6. FUTURE DIRECTIONS

The combination of feature engineering and XAI in phishing website detection is progressing, with some future directions to enhance accuracy and transparency. First domain is applying automated feature engineering techniques and reducing the dependence of manual feature engineering in response to evolving phishing tactics in employing ML models [1]. Additionally, improving the explainability of these models is also critical; applying advanced XAI methods, such as SHAP and LIME, can provide interpretable classifications without affecting security [64]. Another critical focus is making sure that detection systems operate efficiently in real-time environments with minimum latency which required lightweight feature engineering algorithms and optimized XAI methods. Moreover, the studies will concentrate on improving phishing detection techniques for social networking and mobile platforms by developing sophisticated

detection models based on the characteristics of social media and mobile phishing, including impersonation tactics, and evolving attack vectors [65].

The findings from this systematic study can inform future research in creating adaptive phishing detection systems that enable real-time analysis. The amalgamation of SHAP with automated feature engineering indicates a pathway for developing interpretable deep learning models in cybersecurity. The consequences are crucial for developing AI-driven security solutions that adhere to regulatory and ethical requirements in digital governance.

## 7. CONCLUSION

This paper points out that phishing attacks are still developing and growing in their complexity and that there is a need for the advancement of the detection systems in the sphere of cybersecurity. Some of the current approaches have been seen to provide reasonable solutions to the problem of phishing; however, they have some drawbacks including the ability to adapt to zero-day attacks, the issue of interpretability and the problems of dealing with imbalanced datasets and the scalability question. To overcome these limitations and since the nature of phishing threats is ever changing, researchers must look for new ways and techniques. Some of the future recommendations for the enhancement of the phishing detection systems include integration of feature engineering with XAI for phishing detection to address these challenges. These approaches improve the efficiency of detection in addition to offering important information regarding the decisions made by the detection systems. Through the incorporation of stable feature selection with the interpretable models, they enhance the performance of the system as well as the trust of the users.

## FUNDING INFORMATION

No funding was involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Norah Alsuqayh | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | |
| Abdulrahman Mirza | ✓ | | | ✓ | | ✓ | | | | ✓ | ✓ | ✓ | | |
| Areej Alhogail | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

## REFERENCES

[1] C. Rajeswary and M. Thirumaran, "A comprehensive survey of automated website phishing detection techniques: a perspective of artificial intelligence and human behaviors," in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Mar. 2023, pp. 420–427, doi: 10.1109/ICSCDS56580.2023.10104988.

[2] APWG, "Phishing activity trends report." 2022, [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2022.pdf.

[3] M. C. Calzarossa, P. Giudici, and R. Zieni, "Explainable machine learning for phishing feature detection," *Quality and Reliability Engineering International*, vol. 40, no. 1, pp. 362–373, Feb. 2024, doi: 10.1002/qre.3411.

[4] B. D. Shendkar, P. R. Chandre, S. S. Madachane, N. Kulkarni, and S. Deshmukh, "Enhancing phishing attack detection using explainable AI: Trends and innovations," *ASEAN Journal on Science and Technology for Development*, vol. 42, no. 1, Jan. 2024, doi: 10.61931/2224-9028.1604.

[5] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable artificial intelligence in CyberSecurity: a survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.

[6] A. B. Arrieta and others, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[7] N. Aslam and others, "Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI)," *Sustainability*, vol. 14, no. 12, p. 7375, Jun. 2022, doi: 10.3390/su14127375.

[8] K. Kluge and R. Eckhardt, "Explaining the suspicion: design of an XAI-based user-focused anti-phishing measure," in *IEEE Transactions on Dependable and Secure Computing*, 2021, pp. 247–261.

[9] Y. Chai, Y. Zhou, W. Li, and Y. Jiang, "An explainable multi-modal hierarchical attention model for developing phishing threat intelligence," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2021, doi: 10.1109/TDSC.2021.3119323.

[10] A. Abuzuraiq, M. Alkasassbeh, and M. Almseidin, "Intelligent methods for accurately detecting phishing websites," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Apr. 2020, pp. 85–90, doi: 10.1109/ICICS49469.2020.239509.

[11] R. Purwanto, "Adaptive phishing detection system using machine learning," UNSW Sydney, 2022.

[12] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," *Electronics*, vol. 9, no. 9, Sep. 2020, doi: 10.3390/electronics9091514.

[13] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, Jan. 2012, doi: 10.1145/2063176.2063197.

[14] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *Proceedings of the 4th ACM workshop on Digital identity management*, Oct. 2008, pp. 51–60, doi: 10.1145/1456424.1456434.

[15] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, no. 1, pp. 139–154, Jan. 2021, doi: 10.1007/s11235-020-00733-2.

[16] T. Nagunwa, S. Naqvi, S. Fouad, and H. Shah, "A framework of new hybrid features for intelligent detection of zero hour phishing websites," in *12th International Conference on Computational Intelligence in Security for Information Systems*, 2020, pp. 36–46, doi: 10.1007/978-3-030-20005-3_4.

[17] D. M. Divakaran and A. Oest, "Phishing detection leveraging machine learning and deep learning: a review," *IEEE Security \& Privacy*, vol. 20, no. 5, pp. 86–95, 2022, doi: 10.1109/MSEC.2022.3175225.

[18] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 2015–2028, May 2019, doi: 10.1007/s12652-018-0798-z.

[19] M. Shoaib and M. S. Umar, "URL based phishing detection using machine learning," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mar. 2023, pp. 1–7, doi: 10.1109/ISCON57294.2023.10112184.

[20] N. F. Abedin, R. Bawm, T. Sarwar, M. Saifuddin, M. A. Rahman, and S. Hossain, "Phishing attack detection using machine learning classification techniques," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020, pp. 1125–1130, doi: 10.1109/ICISS49785.2020.9315895.

[21] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN," *Electronics*, vol. 12, no. 1, p. 232, Jan. 2023, doi: 10.3390/electronics12010232.

[22] H.-H. Wang, S.-W. Tian, L. Yu, X.-X. Wang, Q.-S. Qi, and J.-H. Chen, "Bidirectional IndRNN malicious webpages detection algorithm based on convolutional neural network and attention mechanism," *Journal of Intelligent \& Fuzzy Systems*, vol. 38, no. 2, pp. 1929–1941, Feb. 2020, doi: 10.3233/JIFS-190455.

[23] W. Wang, F. Zhang, X. Luo, and S. Zhang, "PDRCNN: precise phishing detection with recurrent convolutional neural networks," *Security and Communication Networks*, vol. 2019, pp. 1–15, Oct. 2019, doi: 10.1155/2019/2595794.

[24] A. S. Bozkir and M. Aydos, "LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition," *Computers \& Security*, vol. 95, p. 101855, Aug. 2020, doi: 10.1016/j.cose.2020.101855.

[25] T. Li, G. Kou, and Y. Peng, "Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods," *Information Systems*, vol. 91, p. 101494, Jul. 2020, doi: 10.1016/j.is.2020.101494.

[26] C. Sur, "Ensemble one-vs-all learning technique with emphatic & rehearsal training for phishing email classification using psychology," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 6, pp. 733–762, Nov. 2018, doi: 10.1080/0952813X.2018.1467496.

[27] E. Zhu, C. Ye, D. Liu, F. Liu, F. Wang, and X. Li, "An effective neural network phishing detection model based on optimal feature selection," in *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainC*, Dec. 2018, pp. 781–787, doi: 10.1109/BDCloud.2018.00117.

[28] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text," *Expert Systems with Applications*, vol. 115, pp. 300–313, Jan. 2019, doi: 10.1016/j.eswa.2018.07.067.

[29] D.-J. Liu, G.-G. Geng, X.-B. Jin, and W. Wang, "An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment," *Computers & Security*, vol. 110, p. 102421, Nov. 2021, doi: 10.1016/j.cose.2021.102421.

[30] M. S. M. Prince, A. Hasan, and F. M. Shah, "A new ensemble model for phishing detection based on hybrid cumulative feature selection," in *2021 IEEE 11th Symposium on Computer Applications \& Industrial Electronics (ISCAIE)*, Apr. 2021, pp. 7–12, doi: 10.1109/ISCAIE51753.2021.9431782.

[31] Y. Peng, S. Tian, L. Yu, Y. Lv, and R. Wang, "A joint approach to detect malicious URL based on attention mechanism," *International Journal of Computational Intelligence and Applications*, vol. 18, no. 03, 2019, doi: 10.1142/S1469026819500214.

[32] G. Vrbančič, I. Fister, and V. Podgorelec, "Swarm intelligence approaches for parameter setting of deep learning neural network," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, Jun. 2018, pp. 1–8, doi: 10.1145/3227609.3227655.

[33] P. A. Barraclough, G. Fehringer, and J. Woodward, "Intelligent cyber-phishing detection for online," *Computers \& Security*, vol. 104, p. 102123, May 2021, doi: 10.1016/j.cose.2020.102123.

[34] T. Nathezhtha, D. Sangeetha, and V. Vaidehi, "WC-PAD: Web crawling based phishing attack detection," in *2019 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2019, pp. 1–6, doi: 10.1109/CCST.2019.8888416.

[35] A. K. Murthy and Suresha, "XML URL classification based on their semantic structure orientation for web mining applications," *Procedia Computer Science*, vol. 46, pp. 143–150, 2015, doi: 10.1016/j.procs.2015.02.005.

[36] R. S. Rao, A. R. Pais, and P. Anand, "A heuristic technique to detect phishing websites using TWSVM classifier," *Neural Computing and Applications*, vol. 33, no. 11, pp. 5733–5752, Jun. 2021, doi: 10.1007/s00521-020-05354-z.
[37] C. M. R. da Silva, E. L. Feitosa, and V. C. Garcia, "Heuristic-based strategy for Phishing prediction: A survey of URL-based approach," *Computers & Security*, vol. 88, p. 101613, Jan. 2020, doi: 10.1016/j.cose.2019.101613.
[38] 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, Dec. 2021, doi: 10.1007/s13278-021-00829-w.
[39] B. van Dooremaal, P. Burda, L. Allodi, and N. Zannone, "Combining text and visual features to improve the identification of cloned webpages for early phishing detection," in *Proceedings of the 16th International Conference on Availability, Reliability and Security*, Aug. 2021, pp. 1–10, doi: 10.1145/3465481.3470112.
[40] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.004.
[41] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.
[42] S. Anupam and A. K. Kar, "Phishing website detection using support vector machines and nature-inspired optimization algorithms," *Telecommunication Systems*, vol. 76, no. 1, pp. 17–32, Jan. 2021, doi: 10.1007/s11235-020-00739-w.
[43] S.-J. Bu and S.-B. Cho, "Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection," *Electronics*, vol. 10, no. 12, p. 1492, Jun. 2021, doi: 10.3390/electronics10121492.
[44] X. Xiao and others, "Phishing websites detection via CNN and multi-head self-attention on imbalanced datasets," *Computers & Security*, vol. 108, p. 102372, Sep. 2021, doi: 10.1016/j.cose.2021.102372.
[45] A. AlEroud and G. Karabatis, "Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks," in *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, Mar. 2020, pp. 53–60, doi: 10.1145/3375708.3380315.
[46] R. S. Rao and A. R. Pais, "Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 9, pp. 3853–3872, 2020, doi: 10.1007/s12652-019-01637-z.
[47] M. Korkmaz, E. Kocyigit, O. K. Sahingoz, and B. Diri, "Phishing web page detection using N-gram features extracted from URLs," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Jun. 2021, pp. 1–6, doi: 10.1109/HORA52670.2021.9461378.
[48] A. A. Orunsolu, A. S. Sodiya, and A. T. Akinwale, "A predictive model for phishing detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 2, pp. 232–247, Feb. 2022, doi: 10.1016/j.jksuci.2019.12.005.
[49] S. Yu, C. An, T. Yu, Z. Zhao, T. Li, and J. Wang, "Phishing detection based on multi-feature neural network," in *2022 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Nov. 2022, pp. 73–79, doi: 10.1109/IPCCC55026.2022.9894337.
[50] S. Ariyadasa, S. Fernando, and S. Fernando, "Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML," *IEEE Access*, vol. 10, pp. 82355–82375, 2022, doi: 10.1109/ACCESS.2022.3196018.
[51] R. Yang, K. Zheng, B. Wu, C. Wu, and X. Wang, "Phishing website detection based on deep convolutional neural network and random forest ensemble learning," *Sensors*, vol. 21, no. 24, p. 8281, Dec. 2021, doi: 10.3390/s21248281.
[52] M. A. Al Ahasan, M. Hu, and N. Shahriar, "OFMCDM/IRF: a phishing website detection model based on optimized fuzzy multi-criteria decision-making and improved random forest," in *2023 Silicon Valley Cybersecurity Conference (SVCC)*, May 2023, pp. 1–8, doi: 10.1109/SVCC56964.2023.10165344.
[53] P. R. G. Hernandes, C. P. Floret, K. F. C. De Almeida, V. C. Da Silva, J. P. Papa, and K. A. P. Da Costa, "Phishing detection using URL-based XAI techniques," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2021, pp. 1–6, doi: 10.1109/SSCI50451.2021.9659981.
[54] Y. Lin and others, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 3793–3810.
[55] S. Poddar, D. Chowdhury, A. D. Dwivedi, and R. R. Mukkamala, "Data driven based malicious URL detection using explainable AI," in *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Dec. 2022, pp. 1266–1272, doi: 10.1109/TrustCom56396.2022.00176.
[56] N. Puri, P. Saggar, A. Kaur, and P. Garg, "Application of ensemble machine learning models for phishing detection on web networks," in *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Jul. 2022, pp. 296–303, doi: 10.1109/CCiCT56684.2022.00062.
[57] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: a systematic literature review," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1457–1500, 2022, doi: 10.1007/s10115-022-01672-x.
[58] K. Subashini and V. Narmatha, "Detecting phishing websites using recent techniques: A systematic literature review," in *ITM Web of Conferences*, Nov. 2023, vol. 57, p. 1008, doi: 10.1051/itmconf/20235701008.
[59] A. Altaher, "Phishing websites classification using hybrid SVM and KNN approach," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017, doi: 10.14569/IJACSA.2017.080611.
[60] Q. Li, M. Cheng, J. Wang, and B. Sun, "LSTM based phishing detection for big email data," *IEEE Transactions on Big Data*, vol. 8, no. 1, pp. 278–288, Feb. 2022, doi: 10.1109/TBDATA.2020.2978915.
[61] E. Kocyigit, M. Korkmaz, O. K. Sahingoz, and B. Diri, "Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection," *Applied Sciences*, vol. 14, no. 14, p. 6081, Jul. 2024, doi: 10.3390/app14146081.
[62] Z. Fan, W. Li, K. B. Laskey, and K.-C. Chang, "Investigation of phishing susceptibility with explainable artificial intelligence," *Future Internet*, vol. 16, no. 1, p. 31, Jan. 2024, doi: 10.3390/fi16010031.
[63] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020, doi: 10.1109/ACCESS.2020.2969780.
[64] R. Alenezi and S. A. Ludwig, "Explainability of cybersecurity threats data using SHAP," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2021, pp. 1–10, doi: 10.1109/SSCI50451.2021.9659888.
[65] M. 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 *Proceedings of the 11th International Conference on Networking, Systems, and Security*, Dec. 2024, pp. 137–145, doi: 10.1145/3704522.3704532.
[66] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104347, Sep. 2021, doi: 10.1016/j.engappai.2021.104347.

[67] C. Opara, B. Wei, and Y. Chen, "HTMLPhish: enabling phishing web page detection by applying deep learning techniques on HTML analysis," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207707.

## APPENDIX

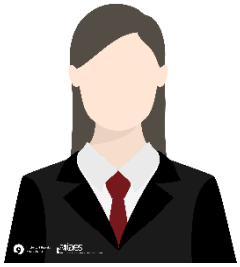
Table 1. Summary of state-of-the-art phishing detection approaches (*continue*)

| Literature | Type of features | Feature engineering method | Main finding | Limitations |
|---|---|---|---|---|
| | | | **List-based Approach** | |
| [14] | XML | - | 100% true positive rate and 0% FPR in identifying a successful login process. | If attackers control the whitelist, the whole application will lose its efficacy. |
| [32] | URL and content | - | Achieves 97% accuracy with a 2.5% FPR. | The model relies on third-party-based features, which introduce latency and consequently slow down the detection process. |
| [33] | URL and content | - | Among the evaluated classifiers, PART delivered the highest performance, achieving 99.33% accuracy and a processing speed of 0.006 seconds. | A comparatively high error rate of 0.6%. |
| [34] | Web traffic, web content and URL | Web crawlers | The method attained an accuracy of 98.9% in identifying both conventional phishing and zero-day phishing assaults. | It uses third-party based feature (search engine) that slows the process. |
| [35] | XML | - | The framework achieved an accuracy level of 97.36%. | May not be effective at detecting zero-day attacks because it relies on known patterns and structures. |
| | | | **Heuristic-based approach** | |
| [28] | Text, Frame and Image | Chi-Square and Information Gain | The classifier achieved 98.3% accuracy. | Accurately identifying phishing sites may prove challenging, since phishers can utilize techniques to conceal visuals and frames to avoid detection. |
| [36] | URL, similarity and hyperlink | - | Among the evaluated models, the Twin SVM achieved superior performance, with a recall of 98.33% and an overall accuracy of 98.05%. | The model is unable to identify low-content webpages, such as single sign-on pages, due to its reliance on webpage DOM for feature extraction. As a result, if a phisher attempts to obtain sensitive information using low-targeted content, the model may misclassify the website. |
| [37] | URL | - | Perform quantitative and qualitative assessments of behavioral patterns to ascertain critical components—such as feature significance, interconnections, and resemblances—that might facilitate the creation of novel heuristic methods or improve current ones. | Some features are not present in the proposed study, for example, the Google page rank feature. |
| [38] | URL and content | ANOVA, information gain, fisher score, relief-f, Recursive feature elimination. | Using the UCI dataset (Dataset 1), the experiment's detection accuracy was 97.51%; using the Mendeley phishing dataset for ML, the experiment's accuracy was 98.45%. | The approach relies on third-party features, which introduces latency and slows down the overall detection process. |
| [39] | DOM of web page and its screenshot of it. | Region filtering | By lowering the misclassification rate from 1.02% to 0.34%, the method achieved a 67% reduction in phishing misclassification and an accuracy of 99.66%. | Because it uses search engine-based filtering, the method may eventually yield various results for the same query. |
| | | | **ML Approach** | |
| [29] | Counterfeiting, Affiliation, Stealing and evaluation | CASE | A multistage phishing detection framework that integrates rapid filtering with precise identification. | There is potential to enhance both the model's accuracy and the efficiency of its training process [1]. |
| [30] | - | Chi-Square, Gain Ratio, Information Gain, Pearson Correlation Coefficient, PCA | Using the RF classifier, the model demonstrated strong performance with an accuracy of 98.24%. | The feature reduction method was not used in the study to eliminate overlapping characteristics. [41] |

Table 1. Summary of state-of-the-art phishing detection approaches *(continue)*

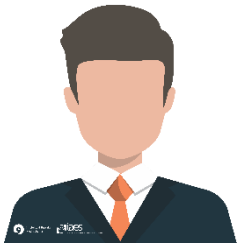
| Literature | Type of features | Feature engineering method | Main finding | Limitations |
|---|---|---|---|---|
| [66] | URL and content | - | An accuracy of 96.61% was attained by combining hybrid features with the RF classifier, representing the best performance among the evaluated classifiers. | The study did not specify the percentage used for the train-test dataset split. |
| [41] | URL | Feature correlation and K best method | RF produced the best accuracy of 99.57% | It requires additional training time when applied to large datasets and demonstrates reduced efficiency [1]. |
| [42] | URL | - | The system's highest accuracy, 90.38%, was attained using GWO, surpassing the performance of all other tested algorithms. | The accuracy achieved in this trial was comparatively lower than that reported in other studies utilizing the same dataset and classification model. |
| | | | **DL Approach** | |
| [12] | URL | - | The proposed model achieved an accuracy of 95.02% on a bespoke dataset, and 98.58%, 95.46%, and 95.22% on established benchmark datasets. | The experimental results yielded lower accuracy levels compared to other studies that utilized the same dataset and classifier. |
| [32] | URL and content | - | An accuracy of 94.89% was attained by NN model in phishing URL detection, outperforming other classifiers in the study. | The utilization of third-party features, such as WHOIS data, adversely affects processing performance and diminishes the efficacy of the detection system. |
| [43] | URL | CAE | Utilizing ROC curve analysis and 10-fold cross-validation, the suggested model exhibited a 3.98% enhancement in sensitivity compared to the most recent deep learning method. | Focusing only on character-level features and not considering the structure and content of the web address might decrease the accuracy of phishing URL detection |
| [44] | URL and content | - | The integration of a self-attention mechanism with CNN significantly enhanced performance, achieving an accuracy rate of 95.6%. | The proposed approach does not incorporate features derived from HTML content, which can play a significant role in detecting phishing websites. |
| [45] | GAN | - | The approach employs GAN to generate URL-based phishing examples capable of deceiving complex black-box ML phishing detection models. | GAN-based approaches have not yet been evaluated within the context of graph-based phishing detection techniques. |
| | | | **Hybrid Approach** | |
| [46] | URL and content | Simhash algorithm | The ensemble model integrated extra trees, RF, and XGBoost classifiers to evaluate the combined effectiveness of heuristic-based and blacklist-based filters as a unified approach, achieving an accuracy of 98.72%. | The dataset used in this study contains a limited number of instances, and the system exhibits a relatively high response time overall. |
| [47] | URL and content | N-gram | Using *n*-gram features derived from URLs, the CNN-based model attained an accuracy of 88.90% on the URL dataset. | Compared to similar studies, the accuracy achieved in this work is relatively lower. |
| [48] | URL features, web document properties, and web behavior attributes | Feature selection module (FSM) | Features were derived from the URL, webpage attributes, and behavioral characteristics. The method achieved high accuracy, with 99.96% true positive and true negative rates, and only 0.04% false positives and false negatives. | The dataset is small (5000 instances) and without any new feature. |
| [49] | URL and content | - | The model's accuracy achieves 97%. | The dataset is small (6000 instances) and using powerful processor (GPU) |
| [50] | URL and content | - | A false-negative rate of 0.036 and a detection accuracy of 96.42% were recorded. | The model's accuracy is low and required powerful processor (Xeon with 4 cores) |
| [23] | URL and content | Cumulative distribution function gradient (CDF-g) algorithm | PDRCNN yielded a detection accuracy of 97% and an AUC score of 99%, according to experimental evaluations. | If a phishing URL lacks relevant semantic information, the model may fail to classify it accurately, as it does not account for whether the corresponding website is active or accessible. |
| [51] | URL | - | The model attained an accuracy of 99.26% when evaluated on benchmark data. | The model is unable to determine whether a URL is active. Additionally, URLs that do not mimic legitimate websites may go undetected. |
| [67] | URL | - | The model enabled the integration of new features and ensured smooth generalization to test data, achieving 93% accuracy and TPR of 93%." | The model requires a longer training time and is unable to determine whether a URL is currently active. |

## BIOGRAPHIES OF AUTHORS



**Norah Alsuqayh** Ph.D. candidate in Information Systems at King Saud University, Saudi Arabia, and holds a master's degree in information systems from the same university. She currently works as a Cybersecurity Specilalist and has previous experience as a Teaching assistant at Princess Nora Bint Abdulrahman University. Her research interests include cybersecurity, XAI, and data analytics. She can be contacted at email: n.alsuqayh@gmail.com.

**Abdulrahman Mirza** received the Ph.D. degree in computer science from Illinois Institute of Technology. He is currently a professor with the Information Systems Department, King Saud University (KSU). He is also a Consultant with the Deputyship of Planning and Information, Ministry of Education, and the acting Director of the National Center for Research on Educational Policies. Some of his previous leadership positions include the Vice Dean of academic affairs with the College of Computer and Information Sciences, KSU; and a General Supervisor of the General Directorate of Teachers Affairs, Ministry of Education. He was also a Senior Advisor to the Minister of Education and the Minister of Higher Education. He also held other positions, such as the Director of quality and accreditation with Saudi Electronic University, the Deputy Director of the Center of Excellence in Information Assurance, the Chairperson of the Information Systems Department, and the CIO with the King Abdullah Foundation for Developmental Housing. His research interests include software engineering, e-commerce, and information security. He can be contacted at email: amirza@ksu.edu.sa.



**Areej Alhogail** received the B.Sc. degree in computing and multimedia systems from Leeds Beckett University, U.K., the M.Sc. degree in information systems management from De Montfort University, and the Ph.D. degree from King Saud University, Saudi Arabia, specializing in information security. She is currently an assistant professor with the Department of Information Systems and the Vice-Dean of electronic transactions and communication with Saudi Arabia. In addition, she has numerous research papers in related international journals and conferences on the themes of information security management, the human aspect of information security, information security change management, IoT security, and blockchain. She has also been involved in workshops, competitions arbitration, and delivered many seminars, and events regarding information security to society. She can be contacted at email: aalhogail@ksu.edu.sa.