Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 11 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i11.49533

PAPER

XAI-PhD: Fortifying Trust of Phishing URL Detection Empowered by Shapley Additive Explanations

Mustafa Al-Fayoumi[1], Bushra Alhijawi[2], Qasem Abu Al-Haija[3](✉), Rakan Armoush[2]

[1]Department of Cybersecurity, Princess Sumaya University for Technology, Amman, Jordan

[2]Department of Data Science, Princess Sumaya University for Technology, Amman, Jordan

[3]Department of Cybersecurity, Faculty of Computer & Information Technology, Jordan University of Science and Technology, Irbid, Jordan

qsabuhaija@just.edu.jo

ABSTRACT

The rapid growth of the Internet has led to an increased demand for online services. However, this surge in online activity has also brought about a new threat: phishing attacks. Phishing is a type of cyberattack that utilizes social engineering techniques and technological manipulations to steal crucial information from unsuspecting individuals. Consequently, there is a rising necessity to create dependable phishing URL detection models that can effectively identify phishing URLs with enhanced accuracy and reduced prediction overhead. This study introduces XAI-PhD, an innovative phishing detection method that utilizes machine learning (ML) and Shapley additive explanation (SHAP) capabilities. Specifically, XAI-PhD utilizes SHAP to thoroughly analyze the significance of each feature in influencing the decision-making process of the classifier. By selectively incorporating input characteristics based on their SHAP values, only the most crucial attributes are assessed, enabling the development of a highly adaptable and generalized model. XAI-PhD utilizes a lightweight gradient boosting machine as its classifier, and a series of rigorous tests are conducted to assess its performance compared to established baseline methods. The empirical findings unequivocally demonstrate the exceptional effectiveness of XAI-PhD, as evidenced by its remarkable accuracy and F1-score of 99.8% and 99%, respectively. Moreover, XAI-PhD exhibits high computational efficiency, requiring only 1.47 milliseconds and 18.5 microseconds per record to generate accurate predictions.

KEYWORDS

explainable artificial intelligence (XAI), phishing, feature engineering, malicious URLs

1 INTRODUCTION

Recently, the number of users using the Internet has expanded considerably. The COVID-19 pandemic lockdown has fueled the proliferation of mobile devices, ad hoc networks, smart sensors, and IoT technologies. As a result, the internet has become an essential part of people's daily lives and activities. According to global statistics, 5.18 billion people worldwide utilize the internet, which accounts for 64.6% of the world's population. Consequently, the number of cybercrime victims is growing

Al-Fayoumi, M., Alhijawi, B., Al-Haija, Q.A., Armoush, R. (2024). XAI-PhD: Fortifying Trust of Phishing URL Detection Empowered by Shapley Additive Explanations. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(11), pp. 80–101. https://doi.org/10.3991/ijoe.v20i11.49533

Article submitted 2024-03-06. Revision uploaded 2024-05-02. Final acceptance 2024-05-07.

sharply [1]. Malicious websites pose a prevalent and ubiquitous hazard to online enterprises. The World Wide Web has significantly increased the probability of cyberattacks. Attackers exploit the internet to conduct malicious activities such as phishing, spamming, and infecting computers with viruses. Phishing is a cybersecurity attack in which the attacker sends fraudulent communications that appear to be from a reputable source, promoting fraud, attacks, and scams. Therefore, identifying malicious websites is crucial to preventing the spread of malware and protecting users from becoming victims. In a simple phishing attack, the attacker sends an email to the victims containing a link to a spoof website to collect their data. Once the attacker obtains the victim's credential information, they can use the credentials to access the legitimate website. Figure 1 illustrates the simplest form of a phishing attack. Phishing can take various forms, including email phishing via the short message service (SMS), voice phishing (vishing), and hijacking a website's page.

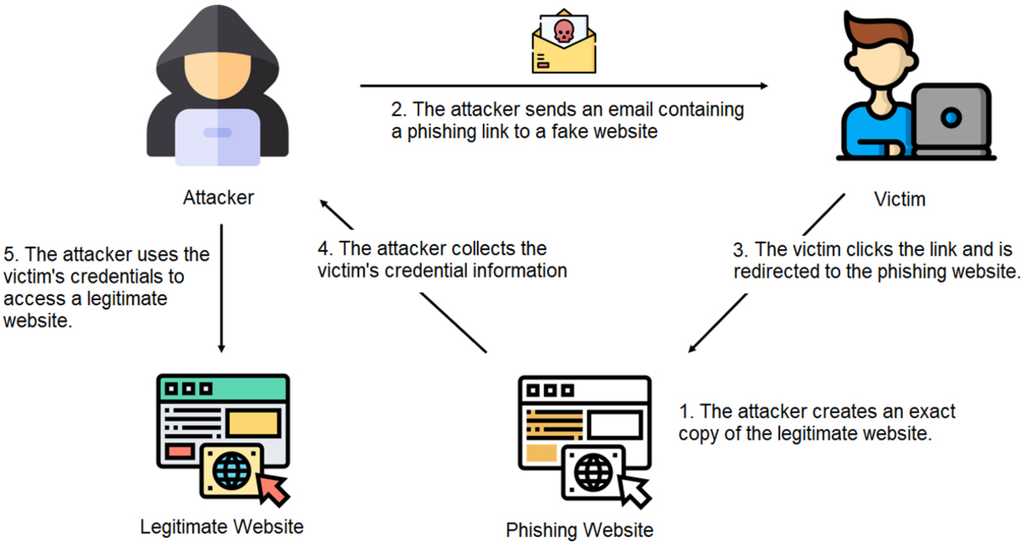

2. The attacker sends an email containing a phishing link to a fake website
Attacker
Victim
5. The attacker uses the victim's credentials to access a legitimate website.
4. The attacker collects the victim's credential information
3. The victim clicks the link and is redirected to the phishing website.
1. The attacker creates an exact copy of the legitimate website.
Legitimate Website
Phishing Website

Fig. 1. Phishing attack scenario

The APWG's phishing activity trends report for the fourth quarter of 2023 revealed that phishing attacks have reached unprecedented levels [2]. The report highlighted 1,077,501 phishing attacks during Q4 2023, culminating in nearly five million attacks, making 2023 the worst year for phishing. Figure 2 shows the trend of phishing websites and phishing emails from Q1 2019 to Q4 2023, with phishing websites experiencing a significant increase in activity during Q1 2023, reaching over 1.6 million attacks. Despite a subsequent decrease in Q2 and Q3 2023, phishing websites remain a persistent threat, with over 1 million attacks recorded in Q4 2023. Phishing emails showed more stability, maintaining relatively consistent levels throughout the observed period. With over 1.13 billion online websites, manually tracking and filtering malicious websites is challenging and time-consuming [1]. As a result, phishing URL detection methods are essential to identify and block phishing websites before users can access them [9]. Several solutions have been developed to detect phishing URLs [3–8], primarily through feature-based and blacklist-based methods. These trends underscore the importance of reliable phishing URL detection methods to safeguard users before they access malicious websites. Feature-based detection and blacklist-based solutions remain essential strategies, with the former relying on automatic analysis of URL features and the latter on expert research and user reports. Identifying trustworthy and robust phishing detection systems remains critical to addressing the rising tide of phishing attacks.

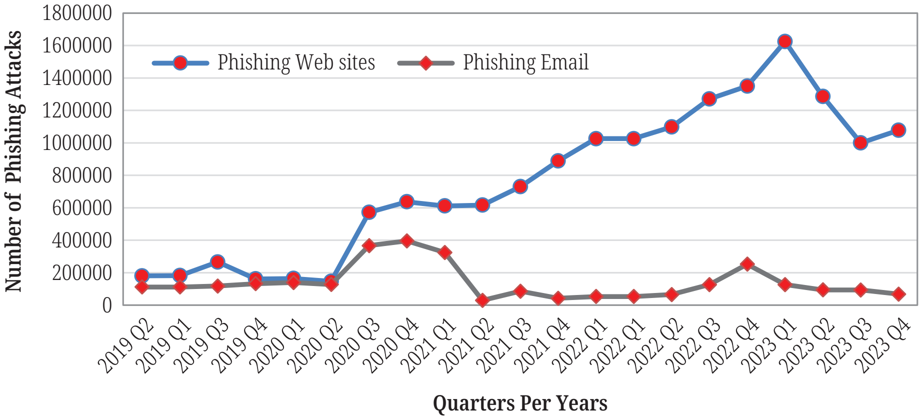


Fig. 2. Summary of APWG Phishing Activity Trend Report from Q1-2019 to Q4-2023

Machine learning (ML) offers opportunities to enhance the accuracy of phishing detection methods. The effectiveness of ML-based approaches depends heavily on the quality of the input data. As a result, the selection of data and the utilization of feature engineering techniques play a crucial role in determining the success of ML-based phishing detection methods. Nevertheless, existing ML methods require greater transparency. Explainable artificial intelligence (XAI) facilitates the creation of more interpretable ML models, allowing individuals to comprehend and have confidence in the output generated by ML systems.

In that context, this article aims to improve the accuracy of phishing detection methods. The contribution of this paper involves the development of a phishing detection model using ML and Shapley additive explanations (SHAP), called XAI-PhD. XAI-PhD involves two main phases: (1) data and feature engineering, and (2) classifier building. In the data and feature engineering phase, XAI-PhD prepares the input features for training the classifier. XAI-PhD constructs a set of features using the URLs, and then the SHAP values are used to select the most important features for the prediction process. Using SHAP values as a feature selection method reduces the model complexity, thus improving the accuracy and creating a generalized model. This work contributes to the phishing detection domain by initially employing the SHAP values for feature selection. Later, XAI-PhD uses the prepared features to train a light gradient boosting machine (LGBM). Our summarized contributions in this paper can be outlined as follows:

- We present a new and trustworthy XAI-driven phishing URL detection solution that accurately identifies phishing URLs while minimizing prediction overhead.
- We combine LGBM, a ML approach, with SHAP, an XAI-based method. The most crucial features were extracted using SHAP, and LGBM was utilized as an effective supervised learning approach.
- We thoroughly evaluate the performance of our proposed system through a series of tests and benchmark it against research-accepted practices. Detailed findings are presented to enhance understanding of both the issue description and our proposed solution.

The rest of this paper is organized as follows: Section 2 presents and summarizes the works in the literature. Section 3 details the proposed method, XAI-PhD. Section 4 presents the experiments and discusses the gathered results. Finally, Section 5 concludes the paper and presents some future directions.

2 RELATED WORK

Several contributions have been made to phishing detection in the last few decades. This section focuses on reviewing phishing URL-based detection methods. Additionally, a summary of ML techniques, feature selection, and feature reduction methods is presented. Various classification methods are utilized to develop phishing detection methods, including multilayer perceptron (MLP) [10, 11, 14, 21, 22, 23, 32, 33], naive bayes (N.B.) [10, 12, 19, 21, 26, 27], KStar (K*) [10], random forest (RF) [10, 11, 13, 19, 21, 26, 27, 30], k-nearest neighbor (KNN) [11, 19, 21, 26, 27, 30], support vector machine (SVM) [11, 17, 21, 26, 30], decision tree [11, 19, 26, 27, 32, 33], rotation forest (RoF) [11], deep learning [15, 25, 28, 29, 31], Adaboost [19, 30], SMO [19], extreme gradient boosting (XGBoost) [20, 30], logistic regression [26–28], quadratic discriminant analysis (QDA) [27], and logistic model tree (LMT) [33]. The input features significantly influence the performance of the ML classifier. Therefore, several feature selection and reduction methods are used in the literature to choose the relevant features for the prediction task, such as information gain (IG) [21, 22], gain ratio (GR) [21, 22], Relief-F [21, 22], and recursive feature elimination (RFE) [21, 22], principal component analysis (PCA) [21, 22, 33], and correlation analysis [23]. Table 1 summarizes the contributions made to the field of phishing URL-based detection methods.

Ibrahim et al. [10] developed an ML-based phishing detection method using MLP, NB, K*, and RF. Their experimental results demonstrate that R.F. outperformed other classifiers in accuracy. Subas et al. [11] proposed an ML-based approach to classify a website URL as legitimate or phishing. They used different classifiers to implement the proposed method, such as MLP, KNN, SVM, DT, RF, and RoF. The RF-based method achieved better accuracy than other classifiers. Another phishing detection method called SEAHound was developed by Peng et al. [12]. SEAHound integrates semantic analysis and NB to detect phishing emails. Patil et al. [13] suggested a phishing URL detection method that combines blacklist/whitelist, heuristics-based, and visual similarity-based techniques. Their approach monitors the HTTP traffic, compares the URL with a blacklist or whitelist, analyzes the website features, and extracts and compares the CSS of suspicious and legitimate pages. They used D.T., logistic regression, and R.F. to classify a URL as malicious or benign. Ferreira et al. [14] employed an MLP to develop a phishing detection method.

In addition, Ping Yi et al. [15] developed a deep belief network-based detection method based on the original and interaction features of phishing websites. Patil and Patil [16] addressed the challenge as a multiclass classification task, classifying URLs into malicious attacks such as spam, phishing, and malware. Adebowale et al. [17] combined an adaptive neuro-fuzzy inference system and SVM to create a phishing detection method. Their approach analyzes various data types, including text, frames, and images. Yadollahi et al. [18] presented a rule-based learning technique. Sahingoz et al. [19] introduced a real-time anti-phishing system using ML and classifiers such as DT, Adaboost, K*, kNN, RF, SMO, and NB. The R.F. classifier, as in other study findings (e.g., [10, 11]), demonstrated superior accuracy compared to other classifiers. Rao et al. [20] utilized an XGBoost classifier to identify malicious URLs. Zamir et al. devised an ensemble ML method incorporating R.F., MLP, SVM, KNN, and bagging classifiers. Saha et al. [22] introduced an MLP-based approach integrated with various feature selection and reduction techniques. Additionally, Odeh et al. [23] employed MLP to develop a phishing detection method, utilizing correlation analysis for feature selection.

Barlow et al. [24] proposed a TensorFlow-based phishing detection method. Their technique analyzes the binary visualization of the scraped website HTML file, where the HTML file is converted to a 2D image before analysis. Xiao et al. [25] used a hybrid deep learning architecture model combining convolutional neural networks (CNN), long short-term memory, and multi-head self-attention. Yang et al. [29] developed a hybrid detection approach integrating R.F. with an embedding method and a CNN. The embedding method is responsible for representing URLs as fixed-size matrices. CNN used the embedding output to extract features for the R.F. Wang et al. [31] designed a dynamic CNN-based malicious URL detection technique. They also modified the CNN pooling layer to a k-max pooling layer that can be constantly updated based on the URL length. Kumar et al. [26] employed logistic regression, N.B., R.F., D.T., and KNN to classify URLs as benign or phishing. The NB achieved the fewest prediction errors compared to other classifiers. Additionally, Alshirah and Al-Fawa'reh [27] proposed a lexical feature-based phishing detection method. They utilized several classification algorithms to implement their methods, such as R.F., D.T., NB, KNN, logistic regression, SVM, and quadratic discriminant analysis.

Maini et al. [30] developed an ensemble phishing detection model that includes RF, DT, NB, KNN, AdaBoost, SVM, XGBoost, and logistic regression. Shirazi et al. [28] used an adversarial autoencoder (AAE) to enrich existing datasets for building an ML-based phishing detection approach. Al-Haija et al. [32] proposed a phishing detection approach by identifying URL patterns. They utilized a narrow, shallow neural network, a wide, shallow neural network, and the optimizable R.T. Abdulraheem et al. [33] integrated PCA with ML-based methods to detect email phishing. They used MLP, D.T., and LMT as classification algorithms. The PCA+LMT achieved the highest accuracy compared to other classification techniques. Additionally, in [40], the authors introduced PDGAN, a phishing detection model with high performance that utilizes website URLs. PDGAN employs a CNN as a discriminator to differentiate between authentic URLs and phishing URLs and a long short-term memory (LSTM) network as a URL generator. With a detection accuracy of 97.58% and a precision of 98.02%, PDGAN demonstrates impressive results using a dataset of approximately two million URLs collected from PhishTank and DomCop. In the same context, Geng et al. [41] introduced URLGAN, a deep neural network that leverages hierarchical semantic properties to distinguish between malicious and legitimate URLs. Their method enhances its ability to identify various types of harmful URLs by embedding URLs into a hierarchical semantic structure, extracting crucial aspects with BERT, and merging them with generator-generated features.

Furthermore, the authors in [42] proposed a powerful machine learning-based system that uses two classification layers to identify dangerous URLs. The ensemble bagging trees strategy has been shown to be the highest performer, with 99.3% accuracy in binary classification and 97.92% accuracy in multi-classification, exceeding previous solutions. It examines four ensemble learning algorithms using the ISCX-URL2016 dataset. Additionally, Alshingiti et al. [43] presented three distinct deep-learning approaches for phishing website detection in this work. Some techniques include LSTM, CNN, and a hybrid LSTM-CNN approach. Results from experiments show how accurate these methods are, with CNN attaining an impressive 99.2% accuracy rate, LSTM-CNN achieving 97.6%, and LSTM achieving 96.8%. It is important to note that the CNN-based approach performs better regarding phishing detection. Finally, Karim et al. [44] developed and assessed multiple machine-learning models for phishing detection using a dataset of phishing and legal URLs retrieved from over 11,000 websites. These models include naive Bayes, gradient boosting, K-neighbors, support vector, decision tree, linear regression, random forest, and a brand-new hybrid LSD model. The hybrid approach combines soft and hard voting with decision trees, logistic regression, support vector machines, and support vector machines. The suggested hybrid LSD model beats existing models in effectively guarding against phishing attempts, according to a comparative study.

In summary, we present an analysis of state-of-the-art phishing detection models in Table 1 below. The table includes a discussion of each phishing detection model in terms of the machine or deep learning models used to develop the detection model, the datasets utilized to evaluate the developed detection models, the number of target classes in each detection model (binary or multi), the primary advantages of the proposed detection model, and the reported performance metrics.

Considerable efforts have been made to develop phishing URL detection methods. The majority of these techniques employed various ML methods. Some techniques were based on cryptographic methods [45]. This paper focuses on techniques that utilize machine and deep learning models. However, only a few studies have focused on preparing and selecting high-quality features to train an accurate generalized model [21–23, 33]. It is widely known that the performance of a data-driven solution is crucial. The existing solutions in the literature involve traditional feature selection and reduction techniques such as I.G., GR, Relief-F, RFE, PCA, and correlation analysis. Moreover, the ML-based approaches lack transparency. This article introduces a novel phishing detection method that combines XAI with LGBM, named XAI-PhD. XAI-PhD utilizes SHAP values to identify the features that impact the model's performance. Consequently, XAI-PhD enhances performance while simplifying the model. This marks the first instance in the phishing detection field where SHAP values are used for feature selection.

Table 1. Summary of contributions made to the phishing detection field

Ref	Year	Model	Datasets	Classes	Advantages	Limitation	Metrics
[10]	2017	R.F.	Phishing Websites Features 2015	2	High Prediction speed High Prediction Accuracy	Small Dataset Shortened and TOR's URLs may not be detected	98.4%
[11]	2017	RF	UCI Phishing Websites 2015	2	High Prediction speed High Prediction Accuracy	Small Dataset Shortened and TOR's URLs may not be detected	97.36%
[12]	2018	NB	Joseph phishing 2014	2	Consider semantic text and lexical features	Small Dataset Shortened and TOR's URLs may not be detected	95%
[13]	2018	R.F.	Alexa.com, rank2traffic.com, siterankdata.com	2	Analyze the visual appearance of the website	High false positive rate Complex and time-consuming	96.58%
[14]	2018	MLP	Phishing Websites	2	High Prediction speed High Prediction Accuracy	Small dataset No realistic testing	98.63%
[15]	2018	DL	Real I.P. flows from ISP	2	Real-time prediction Deeper Inspection	Complex Model High prediction delay	90%
[16]	2018	DT	Alexa Top sites, Malware Domain List, jwSpamSpy	4	Reliable and effective Multiclass classification	Dark URLs may not be classified High prediction delay	98.4%
[17]	2018	SVM	UCI Phishing URL Dataset 2015	3	Reliable and Client-side (browser extension)	Prediction speed No realistic testing	98.3%
[18]	2019	XCS	Private	2	3rd party independent Real-time prediction Language independent	Small dataset No realistic testing	98.3%
[19]	2019	RF	Private Ebbu2017 Phishing	2	Language independent	Shortened and TOR's URLs may not be detected	97.98%

(Continued)

Table 1. Summary of contributions made to the phishing detection field *(Continued)*

Ref	Year	Model	Datasets	Classes	Advantages	Limitation	Metrics
[20]	2019	XGB	Kaggle	2	High Prediction speed High Prediction Accuracy	TOR's URLs may not be detected	96.8%
[21]	2020	KNN-RF-Bagg	Kaggle	2	High Prediction speed High Prediction Accuracy	Small dataset Complex Model	97.4%
[22]	2020	MLP	Kaggle	3	Real-time prediction	Small dataset Complex Model	93%
[23]	2020	MLP	Phish Tank, Miller Smiles, Google	2	High Prediction Accuracy Multi-Source Dataset	Complex Model High prediction overhead Unreliable performance	99.1%
[24]	2020	CNN	Private	2	Reliable + Analyze the visual appearance of the website	Resource consuming Small dataset	94.16%
[25]	2020	CNN	Private	2	High Prediction Accuracy Less feature extraction	Low Prediction speed URL length impacts the model's resilience	98.34%
[26]	2020	NB	Majestic-Million	2	Consider URL Lexical structure analysis	Very high training time Tedious preprocessing is required	98%
[27]	2020	Lex_ Analysis	Alexa, Open Phish	2	Prediction speed Reliable	Small dataset Noncomprehensive results	98%
[28]	2020	AAE	Private	2	High Prediction Accuracy Language independent	No realistic testing High prediction delay	97.45%
[29]	2021	CNN-RF	Alexa, Phish Tank	2	3rd party independent Language independent	Complex Model Low prediction performance	99.26%
[30]	2021	Ensemble Method	Private	2	Reliable High prediction speed	Complex Model Low prediction performance	93.6%
[31]	2021	CNN	Private	2	Less feature extraction efforts	Complex Model Dark web URLs may not be classified	98%
[32]	2021	MLP	Phish Tank, Alexa	2	Prediction speed Pattern Recognition	Complex Model Shortened and TOR's URLs may not be detected	97.4%
[33]	2022	PCA LMT	Phish Tank	2	High Prediction speed	Small dataset Noncomprehensive test	96.92%
[40]	2022	GAN-LSTM-CNN	DomCop and PhishTank	2	Large Scale Dataset Reliable High Accuracy	Low prediction performance Complex Model Very high training time	97.58%
[41]	2022	GAN-BERT	ISCX-URL2016	5	Comprehensive Dataset Thorough Preprocessing Multi-/Binary-class	Only a small portion of the dataset was dedicated to phishing URL samples Low prediction performance	91.61%
[42]	2023	Ensemble Learning	ISCX-URL2016	5	Comprehensive Dataset High-speed multiclass for different URL forms	Only a small portion of the dataset was dedicated to phishing URL samples	93.56%
[43]	2023	CNN, CNN- LSTM, LSTM	Collected from Yahoo and PhishTank	2	High Prediction Accuracy Large Scale Dataset	Huge training time Tedious Feature Engineering Low prediction performance	99.2%, 97.6%, 96.8%
[44]	2023	Hybrid LSD	Retrieved from 11,000 websites	2	New dataset Thorough Preprocessing	Low prediction performance Dark URLs may not be classified	95.2%

3 XAI-PHD: XAI-BASED PHISHING DETECTION METHOD

In this study, we focus on improving the accuracy of phishing detection methods and reducing the time needed to make decisions. This section presents the methodology for building the phishing detection method based on SHAP values. Figure 3 provides an overview of the proposed method. The method consists of two phases: (1) the data preparation phase and (2) the model building phase. The data is cleaned in the data preparation phase, and important features are selected to train the detection system. The key contribution is the use of SHAP values to automate the feature selection process by considering only features that influence the system's decision. The LGBM model is trained and tested using the prepared data in the model-building phase.

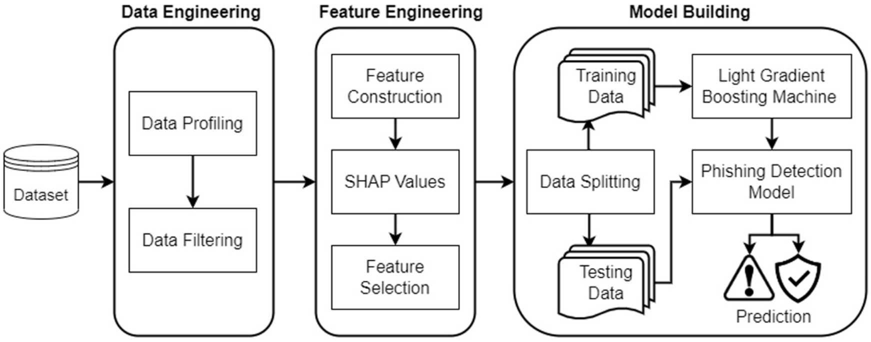
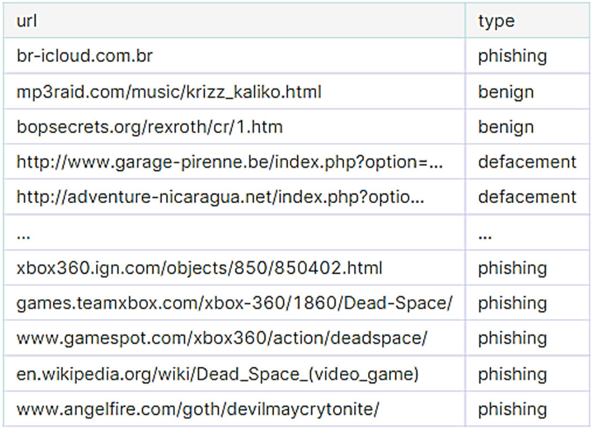


Fig. 3. Architecture of XAI-PhD

3.1 Dataset description

The ISCX-URL2016 dataset will be used to build and evaluate the performance of the proposed phishing detection system. The dataset comprises 114,250 URLs from different sources: the WEBSPAM-UK2007 dataset, the OpenPhish repository, the DNS-BH project, and Alexa. The data involves over 35,300 benign URLs, 12,000 spam URLs, 45,450 defacement URLs, 10,000 phishing URLs, and 11,500 malware URLs. The benign URLs were collected from Alexa Top websites by removing the duplicate and domain-only URLs, whereas malicious URLs were collected from OpenPhish, DNS-BH, Zone-H, and WEBSPAM-UK2007. The dataset includes two features: a URL and a label. Figure 4 shows sample records from the experimental data.



url	type
br-icloud.com.br	phishing
mp3raid.com/music/krizz_kaliko.html	benign
bopsecrets.org/rexroth/cr/1.htm	benign
http://www.garage-pirenne.be/index.php?option=...	defacement
http://adventure-nicaragua.net/index.php?optio...	defacement
...	...
xbox360.ign.com/objects/850/850402.html	phishing
games.teamxbox.com/xbox-360/1860/Dead-Space/	phishing
www.gamespot.com/xbox360/action/deadspace/	phishing
en.wikipedia.org/wiki/Dead_Space_(video_game)	phishing
www.angelfire.com/goth/devilmaycrytonite/	phishing

Fig. 4. Sample records of phishing URLs dataset

3.2 Data engineering

The data preparation phase aims to clean the data and select the most important features to make accurate decisions as quickly as possible. This phase mainly involves two primary tasks: (1) data engineering and (2) feature engineering. During data engineering, a data filtering rule is applied to consider only 45,343 benign or phishing URLs. Additionally, an exploratory data analysis is conducted to investigate the dataset's characteristics. Consequently, an imbalanced data challenge is identified. Figure 5 illustrates the data imbalance issue.

3.3 Feature construction

The performance of the phishing detection method depends on the input features. During feature engineering, we generate 18 lexical features based on the URL feature. These include the presence of an IP address in the domain, the use of a shortening service, URL length, subdomain length, top-level domain length, free-level domain length, URL path length, letter count, number count, and punctuation count. Table 2 provides a description of the 18 generated features.

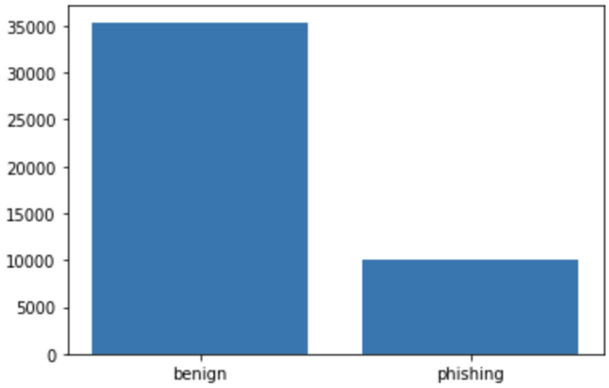


Fig. 5. Data imbalance issue

Table 2. Feature descriptions

Feature	Data Type	Description
is_ip	Boolean	Indicates whether a URL is an I.P. address
contains_shortener	Boolean	Indicates whether a URL used a shortening service
url_len	Number	Length of URL
subdomain_len	Number	Length of subdomain in the URL
tld_len	Number	Length of top-level domain in the URL
fld_len	Number	Length of free-level domain in the URL
url_path_len	Number	Length of the URL's path
url_alphas	Number	Count of letters in the URL
url_digits	Number	Count of numbers in the URL
url_puncs	Number	Count of punctuation in the URL
Count.	Number	Count of "." (dots) in the URL
count@	Number	Count of "@" in the URL

(Continued)

Table 2. Feature descriptions *(Continued)*

Feature	Data Type	Description
count-	Number	Count of "-" in the URL
ount%	Number	Count of "%" in the URL
Count?	Number	Count of "?" in the URL
count=	Number	Count of "=" in the URL
count_dirs	Number	Count of the directories in the URL
first_dir_len	Number	Length of the first directory in the URL

3.4 Feature selection

It is common knowledge that more features result in a more complex ML model that may be overfitted. Thus, we use the SHAP values to select the most important features before employing the proposed phishing detection method. SHAP values are a method based on cooperative game theory used to increase ML models' transparency and interpretability. In this article, we use the SHAP values to measure the impact of the 18 features on the performance of the proposed method. Consequently, the features that contribute to the model are selected as input features. This is the first research paper in the phishing detection domain that applies SHAP value analysis as a feature selection method. Figure 6 presents the results of the SHAP value analysis. The SHAP values analysis shows that ten features contribute to the model, i.e., url_len, subdomain_len, url_path_len, fld_len, count_dirsurl_ digits, count? url_puncs, count., and count_dash. Therefore, eight feature are removed from the input features list.

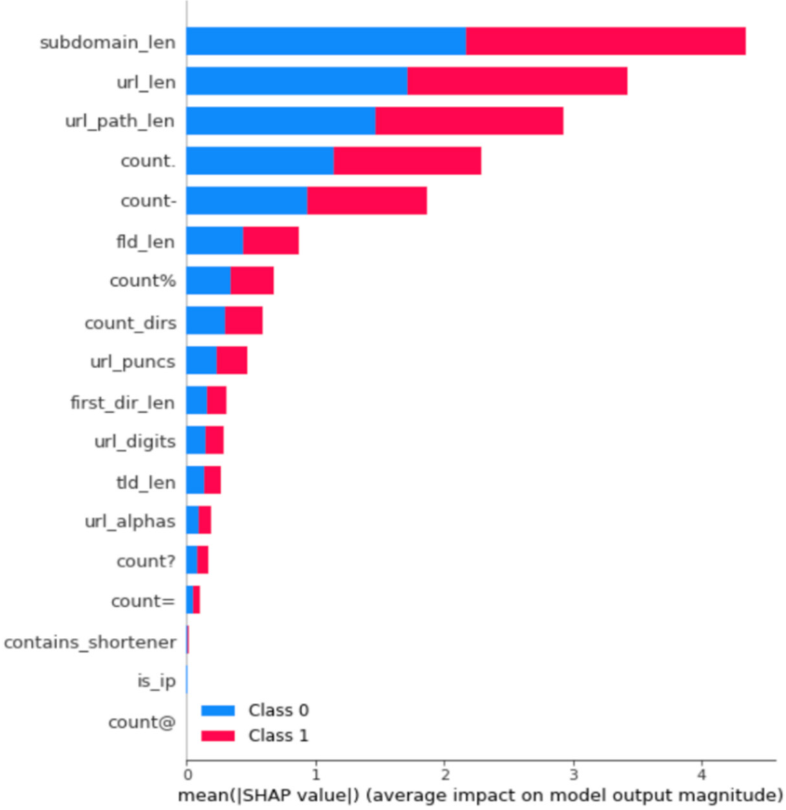


Fig. 6. Ranking the features based on feature importance

3.5 Light gradient boosting machine

The proposed phishing detection method utilizes LGBM as a classifier. LGBM is an open-source distributed gradient-boosting framework for ML [34]. LGBM is built on decision trees (DT) to enhance the model's efficiency and decrease memory usage. With the LGBM, the tree grows leaf-wise, meaning it expands vertically. LGBM selects the leaf that minimizes errors and maximizes efficiency. The LGBM based on a decision tree is illustrated in Figure 7.

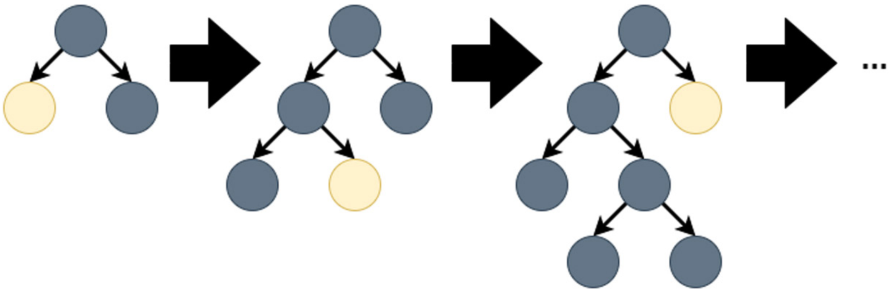


Fig. 7. LGBM using a decision tree

4 EXPERIMENTS AND RESULTS

The evaluation of the SHAP values-based phishing detection approach is presented in this section. A series of experiments has been conducted to assess the effectiveness of using SHAP values as a feature selection technique and the overall effectiveness of the proposed method. The results obtained from the proposed method have been compared with existing approaches in the literature.

4.1 Experiments setup

We conducted two trials to evaluate the effectiveness of XAI-PhD, a proposed phishing detection system based on SHAP values. The first trial analyzed the impact of using SHAP values as a feature selection method. In the second trial, we will compare this approach with traditional ML-based phishing detection methods using the same input attributes. Our study includes three common machine learning techniques: DT, SVM, and KNN. The objectives of these experiments are to test the following hypotheses: (**H1**) Using SHAP values simplifies the model without compromising accuracy; (**H2**) XAI-PhD achieves higher accuracy compared to ML-based methods with the same input features; and (**H3**) XAI-PhD outperforms other phishing detection methods in terms of accuracy.

Table 3. Experimental environment hardware specifications

Brand	ThinkPad E560
RAM	16 GB
HD	250 GB SSD
System Processor	Intel(R) Core (TM) i7-6500U CPU © 2.50 GHz 2.60
OS	Windows 10 Enterprise

The experiments are carried out on machines with the hardware specifications shown in Table 3. For all experiments, we followed a ten-fold cross-validation evaluation methodology [35], where the data is randomly split into ten parts. One part of the data is used as testing data for each iteration, and the other parts are used to train the methods. Nine sets were utilized as training sets, while the tenth was kept aside for testing. The performance of XAI-PhD is evaluated in terms of accuracy, precision, recall, and F1-measure. Accuracy (Eq. 1) is the ratio of correct predictions to the total number of predictions. The accuracy measure may need to be a more precise performance measure in the case of imbalanced data. Therefore, other evaluation measures are considered, such as precision, recall, and the F1-measure. Precision (Eq. 2) is the ratio of correct predictions in the phishing class to the total number of predictions. Recall (Eq. 3) is the fraction of correct predictions in the phishing class to the total number of predictions in the phishing class. F1-measure (Eq. 4) is the harmonic mean of precision and recall. The computation of these measures depends on the confusion matrix [36]. Table 4 shows the confusion matrix.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{1}$$

$$Precision = (TP)/(TP + FP) \tag{2}$$

$$Recall = (TP)/(TP + TN) \tag{3}$$

$$F - Measure = (2 * Precision * Recall)/(Precision + Recall) \tag{4}$$

Table 4. Confusion matrix

		Predicted Class	
		Phishing	**Benign**
Actual Class	Phishing	True Positive (TP)	False Negative (FN)
	Benign	False Positive (FP)	True Negative (TN)

4.2 Evaluation of proposed XAI-PhD system

Two experiments have been conducted to evaluate the performance of the proposed method, XAI-PhD. The objective of the first experiment is to examine the impact of SHAP values on the proposed method's performance (i.e., H1). In this experiment, two versions of the proposed phishing detection method are implemented with and without using the SHAP values as a feature selection method. Figure 8 shows the SHAP values for removing eight features from the input list. The results illustrate that removing those features has an insignificant impact on the model's performance. Table 5 presents the gathered results from applying the proposed method with and without SHAP values.

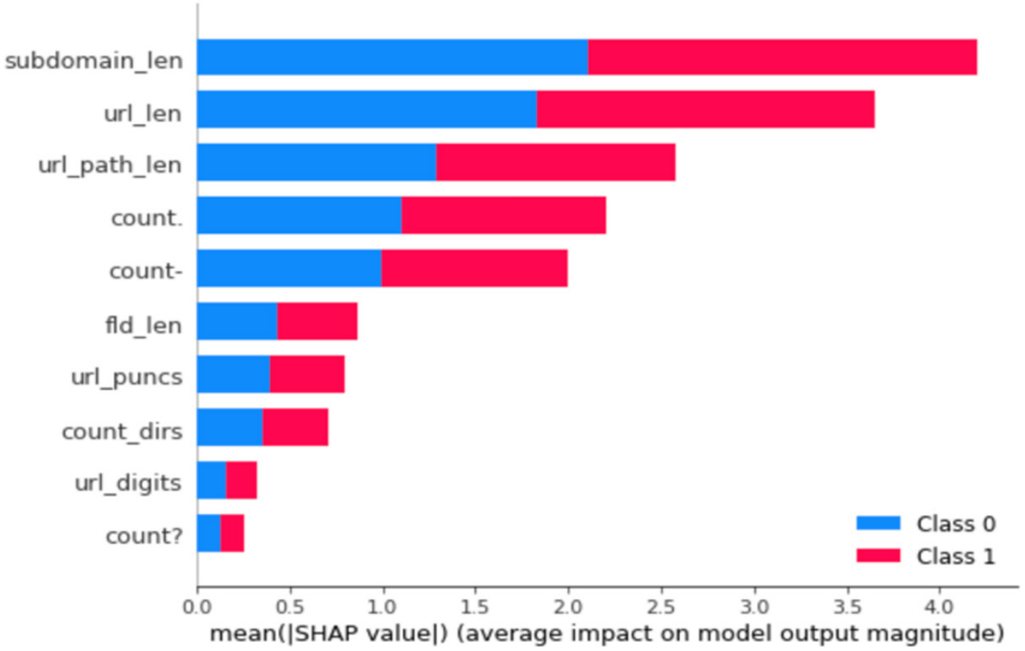


Fig. 8. Results of SHAP values analysis after removing features

The gathered results prove that applying the SHAP values as a feature selection method positively impacts the model's complexity. It is widely known that more input features lead to a more complex model. However, our aim is to reduce complexity without compromising performance or accuracy. Table 5 demonstrates that removing eight features did not affect the accuracy of the proposed method. Moreover, the time required to make predictions using the proposed method with SHAP values is less than that without SHAP values. Utilizing SHAP values as a feature selection method enhanced the prediction time by 40.73%.

Table 5. Accuracy and running time results of XAI-PhD with and without SHAP values

Measure	With SHAP Values	Without SHAP Values
Accuracy	99.8%	99.8%
Precision	1	1
Recall	99%	99%
F1-measure	1	1
Running Prediction Time/Whole Data	48.7 ms ± 473 µs	49.9 ms ± 493 µs
Running Prediction Time/Record	1.47 ms ± 18.5 µs	2.48 ms ± 450 µs

Notes: ms: Millisecond, µs: Microsecond.

Figure 9 illustrates the global explanation of XAI-PhD, which integrates feature importance with feature effect. Features are arranged in descending order from top to bottom based on their importance. The color scheme (red or blue) enables visualization of how variations in a feature's value influence the prediction. For instance, elevated SHAP values for the count feature would suggest a heightened risk of a phishing attack.

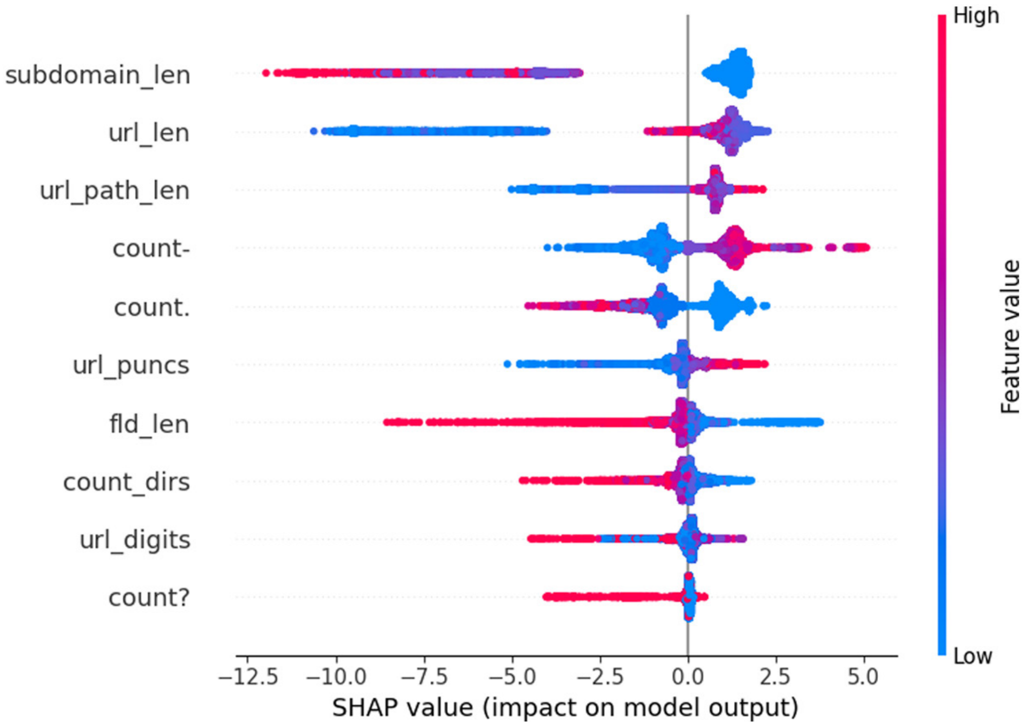


Fig. 9. SHAP summary plot for XAI-PhD input features

The objective of the second experiment is to compare the performance of XAI-PhD with other baseline ML-based techniques (i.e., H2). In this experiment, we consider three popular algorithms: DT, SVM, and KNN. Note that the DT, SVM, and KNN are implemented to use all 18 features. Table 6 shows the results of all methods. XAI-PhD outperformed other ML-based phishing detection methods. The accuracy of the proposed method is 0.31%, 2.1%, and 0.81% better than DT, SVM, and KNN, respectively.

Table 6. Accuracy and running time results of the proposed method with and without SHAP

Measure	XAI-PhD	DT	SVM	KNN
Accuracy	99.8%	99.5%	97.8%	99%
Precision	1	99%	99%	99%
Recall	99%	99%	91%	96%
F1-measure	100%	99%	95%	98%

4.3 Feature analysis and interpretation in the XAI-PhD system

In this section, we delve into feature interpretation and explanation in the XAI-PhD system to demonstrate how individual and paired features contribute to accurate phishing detection predictions. By leveraging the power of SHAP values, we provide comprehensive visualizations that enhance our understanding of the model's predictive behavior and decision-making process. In Figure 10, the heatmap visually represents the correlation between different features used in the XAI-PhD system. Each cell represents the correlation coefficient between two features, with 1 indicating a perfect positive correlation, –1 indicating a perfect negative correlation, and 0 suggesting no correlation. The color intensity reflects the strength of the relationship, with darker colors representing stronger correlations. In the context of phishing detection, the heatmap in Figure 10 underscores the intricate relationships among the features of the URLs. For instance, the strong positive correlation between '**url_len**' and '**url_path_len**' suggests that longer URLs tend to have longer paths, a potential indicator of phishing attempts where attackers may embed malicious parameters within the path. Conversely, a notable negative correlation between '**url_len**' and '**type**' may imply that phishing URLs often diverge from typical benign URL structures. This heatmap not only reaffirms the multifaceted nature of the data but also validates the necessity of considering these relationships during feature selection to bolster the model's predictive accuracy.

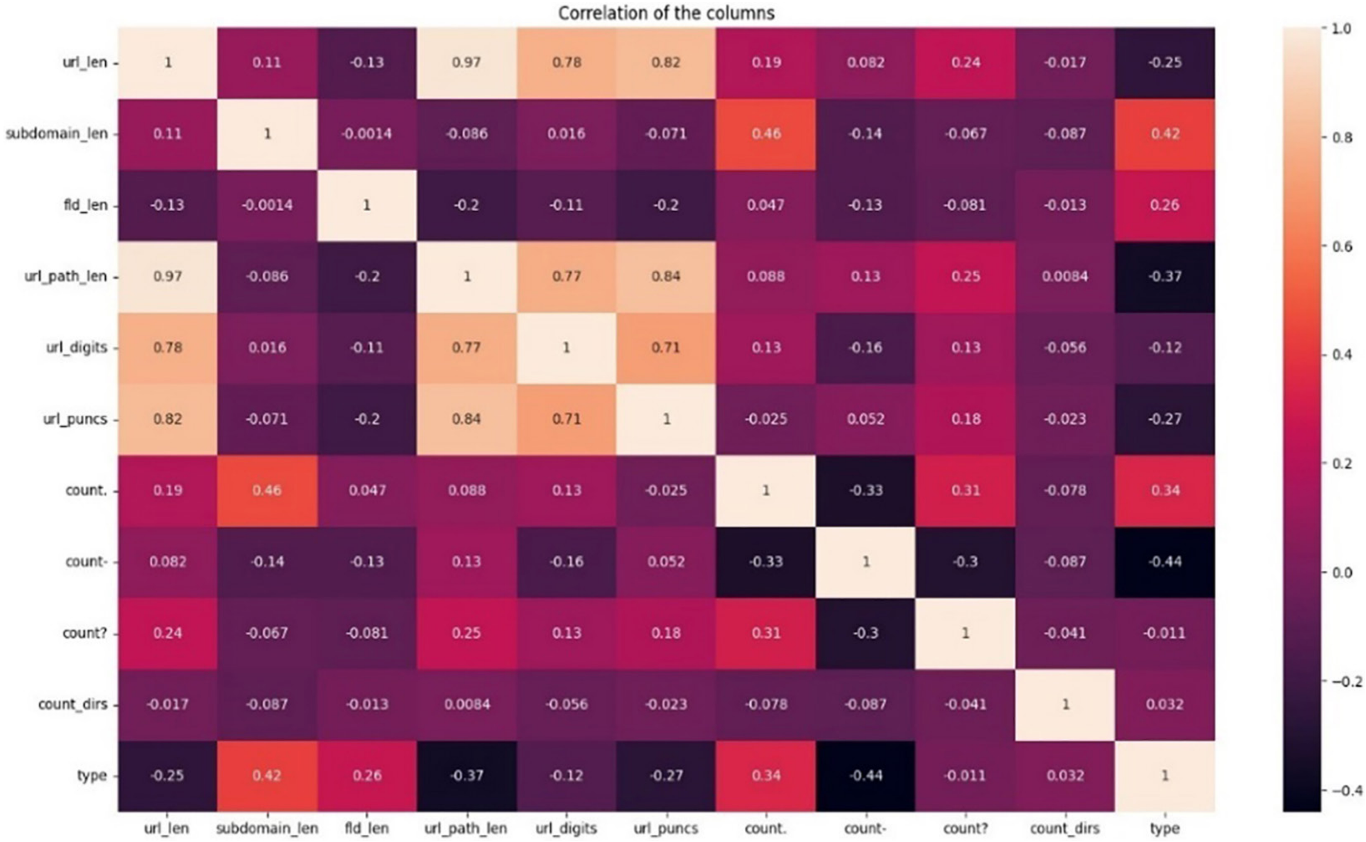


Fig. 10. Correlation heatmap of URL features based on SHAP values in the XAI-PhD

The interaction plots in Figure 11 provide a clear view of how features interact to influence the XAI-PhD system's predictions. For example, the strong positive interaction between the URL length '**url_len**' and the URL path length '**url_path_len**' reveals that longer URLs with extensive paths significantly increase the likelihood of phishing detection. Attackers often use lengthy URLs and paths to hide their intentions, which helps the model recognize this pattern as indicative of phishing. Similarly, the interaction between '**url_len**' and the free-level domain length '**fld_len**' reinforces the complexity of phishing URLs, where longer domain names and overall URLs often conceal malicious behavior. In contrast, the interaction between the dash count '**count-**' and the number of punctuation marks '**url_puncs**' is weaker. Although high values of count slightly reduce phishing predictions, the relationship remains relatively neutral, indicating that special characters alone do not significantly influence predictions. Further insights can be gained by examining the relationships depicted in the interaction values. For instance, '**url_len,**' '**url_path_len,**' and '**fld_len**' features consistently have a positive impact on phishing predictions due to their high SHAP values. These interactions offer valuable insights into how phishing URLs distinguish themselves from benign URLs. The analysis confirms the reliability of the XAI-PhD system, identifying common patterns that inform its predictions: phishing URLs are characterized by longer paths, intricate domain structures, and special characters designed to confuse victims. By interpreting these nuanced interactions clearly, the XAI-PhD system enhances the scientific validity of phishing detection, providing a transparent and credible model for XAI in cybersecurity.

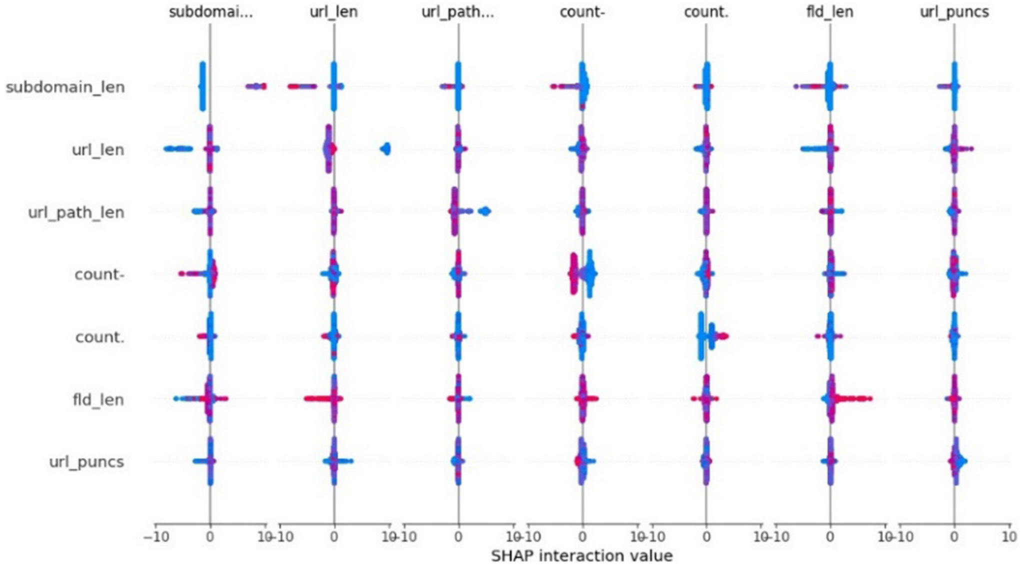


Fig. 11. Interaction plots for feature interactions in the XAI-PhD system

The waterfall plot presented in Figure 12 illustrates how individual features contribute to the XAI-PhD system's prediction for a specific URL. Each step represents a feature's SHAP value and its impact on the final prediction. The feature with the highest absolute magnitude is '**count-**', which negatively contributes –1.26, significantly shifting the prediction to the left and decreasing the likelihood of classifying the URL as phishing. This indicates that a low count of the character '–' (dash) strongly suggests that the URL is benign. On the other hand, the feature with the lowest negative magnitude, '**count?**' has a minimal contribution of –0.03, indicating that the count of question marks has a minor effect on reducing the phishing classification. The feature '**count_dirs**' has the highest positive magnitude of +0.37, pushing the prediction to the right, thereby increasing the probability of classifying this URL as phishing. This suggests that, in this case, a higher directory count is strongly correlated with phishing activity. Other features such as '**fld_len**' (–0.6) and '**url_path_len**' (–0.56) make moderate negative contributions, decreasing the likelihood of phishing.

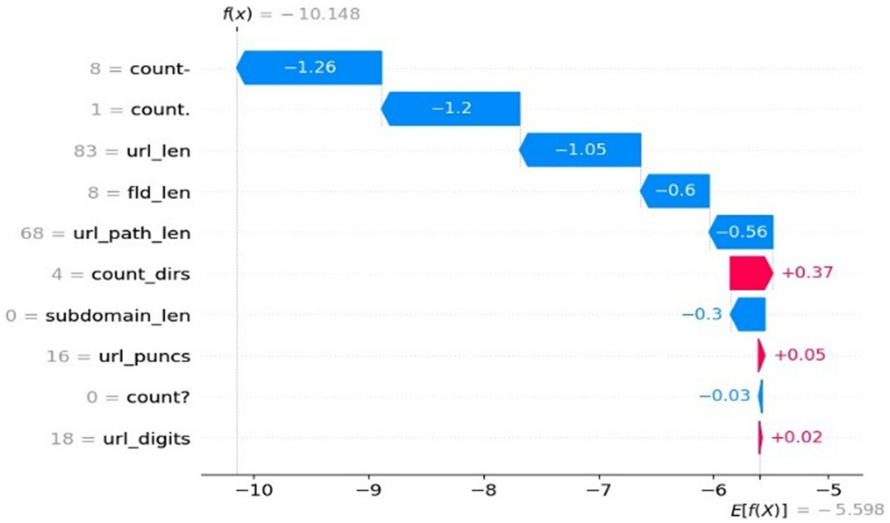


Fig. 12. Waterfall plot compares model output to data distribution based on feature SHAP values

Overall, the combined impact of all features shifts the model's prediction from the base value of –10.148 to the final prediction of –5.598. Blue bars represent features such as '**count-**' and '**url_len**' that decrease the likelihood of phishing, while red bars such as '**count_dirs**' and '**url_puncs**' have positive values that increase it. This visualization offers a clear insight into the decision-making process, emphasizing the significance and relative influence of each feature, thereby improving the interpretability and reliability of the XAI-PhD system.

Figure 13 illustrates the relative impact of features on obtaining a prediction score of 7.77, indicating a benign outcome. The base value starts at –9.305, and features shift this prediction towards either phishing or benign. The threshold value divides the contributions into two categories: red indicates features pushing the prediction toward phishing, while blue indicates features contributing to a benign classification. The features '**fld_len**' = 12, '**url_digits**' = 2, '**count-**' = 1, '**count.**' = 5, '**count_dirs**' = 12, and '**count?**' = 2, shown in red, contribute significantly towards predicting phishing attacks by increasing the final score. In contrast, '**subdomain_len**' = 0, '**url_len**' = 102, and '**url_path_len**' = 82 are displayed in blue, pulling the prediction towards benign. This analysis reveals the importance of key features for the XAI-PhD system's predictions. High directory counts, special characters, and certain URL structures increase the likelihood of phishing, while longer paths and subdomains contribute to benign outcomes. This visual breakdown provides insights into how specific features affect the model's decisions, enabling researchers to identify and leverage the most influential characteristics that enhance the system's predictive performance.

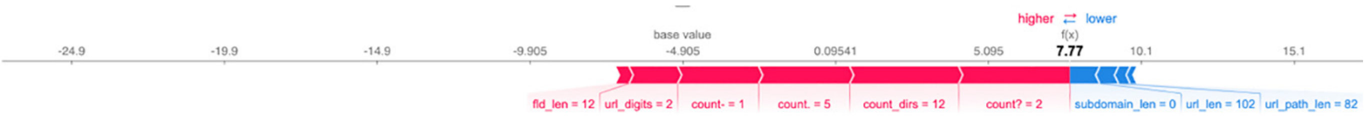


Fig. 13. SHAP force plot used for local explanations

These figures provide a transparent view of the internal workings of the XAI-PhD system, promoting trust and bolstering the model's credibility. They illuminate the intricacies of how particular feature relationships and interactions impact phishing detection, aiding in the identification and refinement of influential characteristics. This in-depth comprehension establishes the foundation for constructing resilient, reliable, and adaptable XAI models that can respond to the ever-changing landscape of phishing threats.

4.4 Comparisons

The proposed performance is compared to other approaches from the literature to examine the third hypothesis (i.e., H3). Table 7 displays the feature selection technique, classification algorithm, and number of features used to achieve the reported results. Generally, the developed methods that utilized less than 40 features demonstrated modest performance compared to those using more than 40. However, XAI-PhD achieved superior accuracy and precision using only ten features. Additionally, the performance of the proposed method is [1.22%–2.66%] better than other methods that employ an XAI technique, such as SHAP values (i.e., [37]).

5 CONCLUSION

A new and robust phishing URL detection technique, XAI-PhD, has been proposed, developed, evaluated, and reported in this paper. XAI-PhD leverages the advantages of integrating explainable AI (XAI) and the classification algorithm (i.e., LGBM). The proposed method involves two main phases: data preparation and classifier building. During the data preparation phase, XAI-PhD extracts features from the URLs and selects appropriate input features using the SHAP values. In XAI-PhD, using the SHAP values improves the model's explainability and generalization. The selected features are used to train the LGBM classifier. A set of experiments is conducted to evaluate and compare the performance of XAI-PhD with other phishing detection methods. The gathered results showed that selecting the features using the SHAP values improves the accuracy and reduces the prediction time of XAI-PhD. Additionally, XAI-PhD outperformed other baseline ML-based methods and similar approaches from the literature. The model may be extended to work with multiclass classification and detect malicious activities.

Table 7. Comparing the proposed model with other studies (where A: accuracy, P: precision)

Study	Feature Selection	Classification Algorithms	No. Features	Best Metric
[18]	–	XCS (a rule-based online learning system)	38	P: 98.39%
[19]	–	R.F.	40	A: 97.98%
[25]	CNN	Deep Learning (CNN)	–	A: 98.34%
[26]	–	Random Forest	26	A: 98.03%
[28]	–	Autoencoder + Gradient Boosting	48	A: 97.45%
[29]	–	Deep Learning	8	A: 89.6%
[30]	–	XGBoost	21	P: 94.6%
[31]	Dynamic CNN	Deep Learning (CNN)	–	P: 99.3%
[37]	Chi-Square	Class Association Rules	12	A: 92.5%
[38]	–	Explainable Boosting Machine	40	P: 97.41%
	–	R.F. + LIME	40	P: 98.8%
	–	SVM – LIME	40	P: 97.90%
[39]	Chi-square, ANOVA	Voting Classifier	47	A: 99.72%
XAI-PhD	**SHAP values**	**LightGBM**	**10**	**A: 99.8%**

6 REFERENCES

[1] "Key Internet Statistics to Know in 2022 (Including Mobile) – BroadbandSearch." [Online]. Available: https://www.broadbandsearch.net/blog/internet-statistics [Accessed: 12-Dec-2022].

[2] Q. Abu Al-Haijaa and M. Al-Fayoumi, "An intelligent identification and classification system for malicious uniform resource locators (URLs)," *Neural Comput. and Applic*., vol. 35, pp. 16995–17011, 2023. https://doi.org/10.1007/s00521-023-08592-z

[3] M. Al-Fayoumi, J. Alwidian, M. Abusaif, and I. M. East, "Intelligent association classification technique for phishing website detection," *International Arab Journal of Information Technology,* vol. 17, no. 4, pp. 488–496, 2020. https://doi.org/10.34028/iajit/17/4/7

[4] K. Haynes, H. Shirazi, and I. Ray, "Lightweight URL-based phishing detection using natural language processing transformers for mobile devices," *Procedia Computer Science,* vol. 191, pp. 127–134, 2021. https://doi.org/10.1016/j.procs.2021.07.040

[5] A. Saleem Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques," in *Materials Today: Proceedings,* 2021, vol. 47, no. 1, pp. 163–166. https://doi.org/10.1016/j.matpr.2021.04.041

[6] Y. A. Yaseen, M. Qasaimeh, R. S. Al-Qassas, and M. Al-Fayoumi, "Email fraud attack detection using hybrid machine learning approach," *Recent Patents on Computer Science,* vol. 12. pp. 1–11, 2019.

[7] S. R. Zahra, M. A. Chishti, A. I. Baba, and F. Wu, "Detecting Covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining-based intelligence system," *Egyptian Informatics Journal,* vol. 23, no. 2, pp. 197–214, 2022. https://doi.org/10.1016/j.eij.2021.12.003

[8] M. Al Fayoumi, A. Odeh, I. Keshta, A. Aboshgifa, T. Alhajahjeh, and R. Abdulraheem, "Email phishing detection based on naïve Bayes, Random Forests, and SVM classifications: A comparative study," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* 2022, pp. 7–11. https://doi.org/10.1109/CCWC54503.2022.9720757

[9] Z. Luo, "A study of accuracy and reliability of Cbir-based phishing filter," Purdue University, 2013.

[10] D. R. Ibrahim and A. H. Hadi, "Phishing websites prediction using classification techniques," in *Proceedings – 2017 International Conference on New Trends in Computing Sciences (ICTCS),* 2017, pp. 133–137. https://doi.org/10.1109/ICTCS.2017.38

[11] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," in *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA),* 2017, pp. 1–5. https://doi.org/10.1109/ICECTA.2017.8252051

[12] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *Proceedings – 12th IEEE International Conference on Semantic Computing (ICSC),* 2018, pp. 300–301. https://doi.org/10.1109/ICSC.2018.00056

[13] V. Patil, P. Thakkar, C. Shah, T. Bhat, and S. P. Godse, "Detection and prevention of phishing websites using machine learning approach," in *Proceedings – 2018 4th International Conference on Computing, Communication Control and Automation (ICCUBEA),* 2018, pp. 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697412

[14] R. P. Ferreira *et al.*, "Artificial neural network for websites classification with phishing characteristics," *Social Networking,* vol. 7, no. 2, pp. 97–109, 2018. https://doi.org/10.4236/sn.2018.72008

[15] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," *Wireless Communications and Mobile Computing,* vol. 2018, no. 1, p. 4678746, 2018. https://doi.org/10.1155/2018/4678746

[16] D. Patil and J. Patil, "Feature-based malicious URL and attack type detection using multiclass classification," *The ISC International Journal of Information Security,* vol. 10, no. 2, pp. 141–162, 2018.

[17] 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, 2019. https://doi.org/10.1016/j.eswa.2018.07.067

[18] M. M. Yadollahi, F. Shoeleh, E. Serkani, A. Madani, and H. Gharaee, "An adaptive machine learning based approach for phishing detection using hybrid features," in *2019 5th International Conference on Web Research (ICWR),* 2019, pp. 281–286. https://doi.org/10.1109/ICWR.2019.8765265

[19] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications,* vol. 117, pp. 345–357, 2019. https://doi.org/10.1016/j.eswa.2018.09.029

[20] P. Varaprasada Rao, S. Govinda Rao, P. Chandrasekhar Reddy, B. S. Anil Kumar, and G. Anil Kumar, "Detection of malicious uniform resource locator," *International Journal of Recent Technology and Engineering,* vol. 8, no. 2, pp. 41–47, 2019.

[21] A. Zamir *et al.*, "Phishing website detection using diverse machine learning algorithms," *Electronic Library,* vol. 38, no. 1, pp. 65–80, 2020. https://doi.org/10.1108/EL-05-2019-0118

[22] I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana, and S. Hossain, "Phishing attacks detection using deep learning approach," in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT),* 2020, pp. 1180–1185. https://doi.org/10.1109/ICSSIT48917.2020.9214132

[23] A. J. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron," *International Journal of Interactive Mobile Technologies,* vol. 14, no. 11, pp. 22–31, 2020. https://doi.org/10.3991/ijim.v14i11.13903

[24] L. Barlow, G. Bendiab, S. Shiaeles, and N. Savage, "A novel approach to detect phishing attacks using binary visualisation and machine learning," in *Proceedings – 2020 IEEE World Congress on Services (SERVICES),* 2020, pp. 177–182. https://doi.org/10.1109/SERVICES48979.2020.00046

[25] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN–MHSA: A convolutional neural network and multi-head self-attention combined approach for detecting phishing websites," *Neural Networks,* vol. 125, pp. 303–312, 2020. https://doi.org/10.1016/j.neunet.2020.02.013

[26] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, "Phishing website classification and detection using machine learning," in *2020 International Conference on Computer Communication and Informatics (ICCCI),* 2020, pp. 1–6. https://doi.org/10.1109/ICCCI48352.2020.9104161

[27] M. Alshira'H, "Detecting phishing URLs using machine learning lexical feature-based analysis," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 9, no. 4, pp. 5828–5837, 2020. https://doi.org/10.30534/ijatcse/2020/242942020

[28] H. Shirazi, S. R. Muramudalige, I. Ray, and A. P. Jayasumana, "Improved phishing detection algorithms using adversarial autoencoder synthesized data," in *Proceedings – Conference on Local Computer Networks (LCN),* 2020, pp. 24–32. https://doi.org/10.1109/LCN48667.2020.9314775

[29] 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, 2021. https://doi.org/10.3390/s21248281

[30] A. Maini, N. Kakwani, B. Ranjitha, M. K. Shreya, and R. Bharathi, "Improving the performance of semantic-based phishing detection system through ensemble learning method," in *2021 IEEE Mysore Sub Section International Conference (MysuruCon),* 2021, pp. 463–469. https://doi.org/10.1109/MysuruCon52639.2021.9641614

[31] Z. Wang, X. Ren, S. Li, B. Wang, J. Zhang, and T. Yang, "A malicious URL detection model based on convolutional neural network," *Security and Communication Networks*, vol. 2021, no. 1, p. 5518528, 2021. https://doi.org/10.1155/2021/5518528

[32] Q. A. Al-Haija and A. Al Badawi, "URL-based phishing websites detection via machine learning," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI),* 2021, pp. 644–649. https://doi.org/10.1109/ICDABI53623.2021.9655851

[33] R. Abdulraheem, A. Odeh, M. Al Fayoumi, and I. Keshta, "Efficient Email phishing detection using Machine learning," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* 2022, pp. 354–358. https://doi.org/10.1109/CCWC54503.2022.9720818

[34] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems,* vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[35] M. Al-Fayoumi and Q. A. Al-Haija, "Capturing low-rate DDoS attack based on MQTT protocol in software Defined-IoT environment," *Array,* vol. 19, p. 100316, 2023. https://doi.org/10.1016/j.array.2023.100316

[36] M. Al-Fawa'reh, M. Al-Fayoumi, S. Nashwan, and S. Fraihat, "Cyber threat intelligence using PCA-DNN model to detect abnormal network behavior," *Egyptian Informatics Journal,* vol. 23, no. 2, pp. 173–185, 2022. https://doi.org/10.1016/j.eij.2021.12.001

[37] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," *Applied Soft Computing,* vol. 48, pp. 729–734, 2016. https://doi.org/10.1016/j.asoc.2016.08.005

[38] 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), 2021 – Proceedings,* 2021, pp. 4–9..

[39] H. M. J. Khan, Q. Niyaz, V. K. Devabhaktuni, S. Guo, and U. Shaikh, "Identifying generic features for malicious URL detection system," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON),* 2019, pp. 0347–0352.

[40] S. Al-Ahmadi, A. Alotaibi, and O. Alsaleh, "PDGAN: Phishing detection with generative adversarial networks," *IEEE Access,* vol. 10, pp. 42459–42468, 2022. https://doi.org/10.1109/ACCESS.2022.3168235

[41] J. Geng, S. Li, Z. Liu, Z. Cheng, and L. Fan, "Effective malicious URL detection by using generative adversarial networks," in *Web Engineering. ICWE 2022.* in Lecture Notes in Computer Science, T. Di Noia, I. Y. Ko, M. Schedl, and C. Ardito, Eds., vol. 13362, Springer, Cham, 2022, pp. 341–356. https://doi.org/10.1007/978-3-031-09917-5_23

[42] Q. Abu Al-Haija and M. Al-Fayoumi, "An intelligent identification and classification system for malicious uniform resource locators (URLs)," *Neural Comput. and Applic.,* vol. 35, pp. 16995–17011, 2023. https://doi.org/10.1007/s00521-023-08592-z

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

[44] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing detection system through hybrid machine learning based on URL," *IEEE Access,* vol. 11, pp. 36805–36822, 2023. https://doi.org/10.1109/ACCESS.2023.3252366

[45] M. A. Snober, A. Droos, and Q. A. Al-Haija, "Prevention of phishing website attacks in online banking systems using visual cryptography," in *6th Smart Cities Symposium (SCS 2022)*, 2022, pp. 168–173. https://doi.org/10.1049/icp.2023.0391

7 AUTHORS

Mustafa Al-Fayoumi is with the Department of Cybersecurity, Princess Sumaya University for Technology, Amman, Jordan.

Bushra Alhijawi is with the Department of Data Science, Princess Sumaya University for Technology, Amman, Jordan.

Qasem Abu Al-Haija is with the Department of Cybersecurity, Faculty of Computer & Information Technology, Jordan University of Science and Technology, PO Box 3030, Irbid 22110, Jordan (E-mail: qsabuhaija@just.edu.jo).

Rakan Armoush is with the Department of Data Science, Princess Sumaya University for Technology, Amman, Jordan.