Intelligent and Converged Networks ISSN 2708-6240
2023, 4(4): 326−341 DOI: 10.23919/ICN.2023.0027

A systematic review: Detecting phishing websites using data mining models

Dina Jibat, Sarah Jamjoom, Qasem Abu Al-Haija*, and Abdallah Qusef

Abstract: As internet technology use is on the rise globally, phishing constitutes a considerable share of the threats that may attack individuals and organizations, leading to significant losses from personal and confidential information to substantial financial losses. Thus, much research has been dedicated in recent years to developing effective and robust mechanisms to enhance the ability to trace illegitimate web pages and to distinguish them from non-phishing sites as accurately as possible. Aiming to conclude whether a universally accepted model can detect phishing attempts with 100% accuracy, we conduct a systematic review of research carried out in 2018–2021 published in well-known journals published by Elsevier, IEEE, Springer, and Emerald. Those researchers studied different Data Mining (DM) algorithms, some of which created a whole new model, while others compared the performance of several algorithms. Some studies combined two or more algorithms to enhance the detection performance. Results reveal that while most algorithms achieve accuracies higher than 90%, only some specific models can achieve 100% accurate results.

Key words: phishing; data mining; machine learning; algorithm; classification

1 Introduction

Due to the great convenience the internet has brought to our modern world, it has become a daily necessity in communication, banking, trading, learning, socializing, and running errands as basic as paying bills. Nonetheless, the internet is associated with inevitable security threats, including but not limited to unsolicited emails, malicious software, viruses, spyware, Denial of Service (DoS) attacks, and phishing. The phishing threat appeared long ago and still exists as phishers find new, creative ways of carrying out their attacks[1]. According to Ref. [1], "phishing is an online identity theft in which an attacker tries to steal a user's personal information, resulting in financial loss of individuals and organizations."

The Anti-Phishing Working Group (APWG) defines phishing as "a criminal mechanism employing both social engineering and technical subterfuge to steal personal identity data and financial account credentials of consumers"[2]. The goal of any phishing attack, regardless of the medium in which the attack is carried out, is to lead the target into following a legitimate-appearing web page but redirect the target to a malicious resource[3]. Differentiating between real and fake web pages is often challenging as their designs are very similar[4]. Phishing is regarded as one of the oldest and easiest ways of personal information theft techniques and is not easy to detect as it does not appear malicious[5]. This is a serious issue with consequences that may affect individuals and firms, from Small and Medium-sized Enterprises (SMEs) to large businesses.

Access to the internet and smartphones is now available for everyone, making this problem more extensive, especially for less educated individuals. Attackers usually target the least knowledgeable groups from old age and young age with less

• Dina Jibat and Sarah Jamjoom are with the Department of Business Intelligence Technology, Princess Sumaya University forTechnology,Amman11941,Jordan.E-mail:din20208043@std.psut.edu.jo; sar20208040@std.psut.edu.jo.
• Qasem Abu Al-Haija is with the Department of Cybersecurity, Princess Sumaya University for Technology, Amman 11941, Jordan. E-mail: q.abualhaija@psut.edu.jo.
• Abdallah Qusef is with the Department of Software Engineering, Princess Sumaya University for Technology, Amman 11941, Jordan. E-mail: a.qusef@psut.edu.jo.
* To whom correspondence should be addressed.
Manuscript received: 2023-03-18; accepted: 2023-09-04

technological background and awareness of similar cases. It was stated by Ref. [6] that there are many ways to detect phishing websites, and this is a vital research topic. Raising users' awareness of fraud, analyzing suspicious characteristics, blacklisting websites, and comparing current attempts to recent attempts that happened are all useful ways to help detect phishing attacks. Different means of detecting phishing attempts should be analyzed to enhance existing tools or to develop new, more effective ones.

This research leads to a systematic review of previous studies regarding detecting phishing websites, URLs, and emails using data mining algorithms to identify the most efficient and most commonly used algorithms for phishing detection. It aims to discover whether a universally accepted model can detect phishing attempts with 100% accuracy.

The remainder of the paper is structured as follows: Section 2 is a review of the state-of-the-art work covering the related work of the studies published in the years 2018–2019 and the related work of the studies published in the years 2020–2021. Section 3 provides the work methodology followed for the review. Section 4 reveals some interesting results. Finally, Section 5 provides the conclusion and remarks.

2 Literature review

There are several types of phishing, including Voice-over Phishing (vishing), Phishing via SMS (smishing), whaling, Mobile Phishing (mishing), social engineering, spear phishing, and clone phishing, among others[3]. Spear phishing collects information and details for the user. Clone phishing is developing an email identical to a legitimate one. Whaling targets senior executives and high profiles whose data was collected in spear phishing[7]. According to Ref. [8], phishing can have significant and undesirable effects on organizations and individuals. Furthermore, as computer users need more anti-phishing knowledge and tend to underestimate the risks and financial losses due to phishing, they have increased vulnerability to phishing attacks. Thus, Ref. [8] emphasizes the important role of education and risk awareness initiatives in raising awareness about security threats, which results in safer user acts.

Ongoing research is conducted to find suitable detection tools and to develop existing solutions. One of the most common phishing web page detection methods was the blacklist approach. However, this method was ineffective as it cannot identify non-blacklisted web pages. Thus, there was a need for more robust detection techniques[9]. Several approaches were developed that can be categorized into three groups: content-based, heuristic-based, and fuzzy rule-based. Content-based system, which, as its name implies, conducts an in-depth analysis of contents. It obtains words from HTML or URLs and specifies their weights. It extracts logos to compare them with the originals and locates web content and URL consistencies[9].

Data Mining (DM) can be used in identifying phishing attempts. One of the most familiar techniques to detect web phishing in DM is classification. Classification aims to build a model by analyzing training data to predict and classify objects whose class labels are unknown in a dataset[10]. In the case of our research topic, it is used to distinguish a web page or an email as phishing or safe. There is a general belief that time consumption and accuracy are major issues in classification when applying DM on high-dimensional datasets. One possible solution is to use feature selection to reduce the dataset dimension. The authors of Ref. [10] performed research to determine whether feature selection can improve accuracy and decrease the computational time of algorithms by comparing four feature selection techniques (information gain, gain ratio, chi-square, and correlation-based feature selection). Results showed that applying the four methods can make the accuracy of Naive Bayes (NB), *k*-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Iterative Dichotomiser 3 (ID3) algorithm decrease. However, it reduces the computation time of KNN, SVM, and ID3 algorithms. Using the four feature selection methods, the most influential attributes are SSL final state, having subdomain, URL of anchor, prefix suffix, Server Form Handler (SFH), domain registration length, links in tags, web traffic, request URL, and Google index. Table 1 presents the advantages and

Table 1 Advantages and disadvantages of most common classifiers.

Algorithm	Advantage	Disadvantage
Logistic regression	Perfect when the dependent variable is collected in 2 classes	Prediction is affected by outliers and repetition
KNN	Efficient and fast, estimation is based on the distance of neighbors	It consumes much memory if the dataset is large
SVM	Easy to implement and works well with many independent variables	Performance is affected by noisy data and is not good with large databases
DT	Easy to explain	Not effective in small datasets or multi-class
NB	Short training time and easy implementation	Lower estimation with fewer data
XGBoost	Boosted decision tree, fast and performs better with each new tree	It takes a long time and can be overfitted
Random forest	Less affected by noise, overfit resistant	It takes a long time for training and needs memory
Artificial neural network	Detects connection between features	Requires high memory

disadvantages of several of the most commonly used classification algorithms.

Classifiers based on Machine Learning (ML) can effectively detect phishing websites[11]. Detecting phishing websites has gained the attention of the ML community. Having that said, training ML models is a huge part of excelling in detecting phishing websites; therefore, the authors of Ref. [12] contributed to creating two benchmark datasets specially built for researchers to train models that aim to detect phishing websites based on the URL, from which the extraction of attributes can be easily done. The attributes are domain properties, URL directory properties, URL file properties, URL parameter properties, URL resolving data, and external metrics. The datasets contain "58 645 and 88 647 websites labeled as legitimate or phishing". These datasets can also be useful for computer security specialists in building malware-detecting systems and firewalls.

Although ML approaches are very common for identifying phishing websites, they are still prone to adversarial learning techniques that aim to decrease the accuracy of trained classifiers. The researchers of Ref. [13] attempted to examine the robustness of ML in detecting phishing attempts facing adversarial learning techniques. After generating adversarial samples and testing them against selected classifiers, it was seen that phishing detection mechanisms are unimmune to adversarial learning techniques. The identification rate of phishing websites was reduced to 70% after changing the value of just one feature. Furthermore, the identification rate dropped to zero when four features were manipulated. By changing at most four feature values, phishing samples that a classifier could have correctly detected can bypass this classifier model. Figure 1 illustrates the significant increase in phishing attacks from 2015 to 2019, according to the APWG.

2.1 Related work in 2018 and 2019

A study[14] was done in 2018 to compare the performance of various classification algorithms on a phishing website dataset that originally consisted of 11 055 records and 31 features. To obtain better performance, the dimension of the dataset was reduced to 27 features by applying feature selection algorithms. After the dataset was reduced, it was used for training and testing. Results showed that Lazy Kstar performed best, with an accuracy of 97.6%. The study also showed BayesNet, Stochastic Gradient Descent (SGD), Lazy Kstar, Random Forest Classifier (RFC), Logistic Model Tree (LMT), and ID3 perform well on the

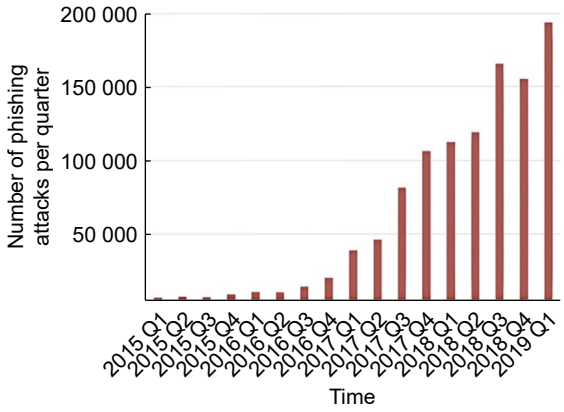

Fig. 1 Phishing activity (2015–2019).

reduced dataset and increase performance. At the same time, multilayer perceptron, Repeated Incremental Pruning (JRip), Partitioning and Regression Trees (PART), J48, Random Forest (RF), and Random Tree (RT) are not useful for the reduced dataset as they decrease in performance.

In Ref. [3], the focus was on detecting phishing websites using URL detection since the easiest manipulation method is creating a malicious URL and then leading the victim to the desired malicious page the phisher wants. They aimed to enhance the efficiency of phishing website detection. In comparing various classification algorithms, Waikato Environment for Knowledge Analysis (WEKA) was used to determine the performance and accuracy of each algorithm. The researchers applied parse to analyze the feature set and minimize it from 31 to 8 due to the significant amount of data to process. Performance metrics showed that RF performed best with an accuracy level of around 95% and was thus chosen for classification. This model used a wide range of metrics, including the True Positives Rate (TPR), True Negatives Rate (TNR), False Negatives Rate (FNR), the F-measure, ROC, precision, and sensitivity for analysis purposes, thus giving a clear view on the performance and accuracy each time the detection takes place.

The researchers of Ref. [15] have also done a study to predict phishing URLs by applying classification mining techniques; they collected a dataset from the University of California Irvine (UCI) Repository Archive composed of 11 055 URLs divided into 7262 valid URLs and 3793 phishing URLs. It contains 31 features obtained from APWG and Phishtank, a website where an individual can confirm whether a suspicious URL is a phishing site through other users' votes. After implementing different algorithms including RT, RF, NB, J48, and LMT, and using Weka to calculate performance metrics such as precision, recall, accuracy, training build data time, specificity, etc., results showed that RT and RF are the best classification algorithms. However, RT algorithm is better since it takes less time to build a model and to test it on training data. Based on this study, tree-based algorithms are the best classifiers.

Based on a Novel Neural Network (NNN) classification method, the researchers of Ref. [16] presented a novel phishing detection model that they have applied to a dataset of 11 055 samples with 30 features already identified as phishing or not. 55.69% of the instances were phishing. The following metrics are used to evaluate the model performance and compare it with other classifiers: accuracy, TPR, FPR, precision, recall, F-measure, and Matthews Correlation Coefficient (MCC). Results conveyed that the proposed model has a high accuracy of 97.71% and a low FPR of 1.7%, indicating its effectiveness in detecting phishing. In addition, in comparison to NB, Logistic Regression (LR), KNN, DT, Linear Support Vector Machine (LSVM), Radial-Basis Support Vector Machine (RSVM), and Linear Discriminant Analysis (LDA), the novel phishing detection model was found to achieve the highest accuracy, TPR, FPR, precision, recall, F-measure, and MCC.

In Ref. [17], specific rules for extracting phishing features were defined and applied to obtain features. Inputs and outputs were determined to classify the Extreme Learning Machine (ELM). ELM achieved the highest accuracy of 95.34% compared to SVM and NB.

Email attachments caused two-thirds of all malware in 2016. In addition to the risk of malware installed on a computer, the email recipient may fall to a phishing attempt by responding and giving away money or important, confidential information[18]. Thus, the researchers of Ref. [18] used datasets of phishing websites obtained from UCI and spam emails to train and test ML algorithms optimized through feature selection. The algorithms used are RF, KNN, Artificial Neural Network (ANN), SVM, Light Gradient (LG), and NB. Using WEKA, they concluded that the RF algorithm generated the best spam and phishing detection results with 95.48% and 97.26% accuracy, respectively. Furthermore, after using WEKA's feature selection optimizers to reduce the number of phishing dataset attributes from 31 to 10 and spam dataset attributes from 58 to 16, results showed that the accuracy of the RF algorithm slightly increased. Thus,

both datasets respond well to reducing the number of attributes, which can help minimize the model's cost and complexity.

Another research[11] in 2019 also implemented and compared the performance of several ML classification algorithms, including NB, J48, and Hidden Naive Bayes (HNB), and the performance of an integrated classifier combining HNB and J48. The dataset selected has 2670 instances and 30 attributes. Phishing website detection accuracy was tested based on the manual selection of features and filters with feature selection in the three individual classifiers and the combined classifier. The findings indicated that the address bar based feature group achieved the highest accuracy in detecting phishing websites. The results also showed that combining techniques resulted in a high accuracy rate of 96.3%, proving its efficiency in identifying phishing websites after adding feature selection scenarios. In contrast, the HNB classifier algorithm proved its efficiency in phishing website detection in manual feature selection.

According to the authors of Ref. [19], major setbacks in detecting phishing attacks are low accuracy rates and high FPR. Accordingly, they conducted a comparative analysis of machine classifiers using several algorithms: RF, SysFor, SPAARC, RepTree, RT, LMT, ForestPA, JRip, PART, NNge, OneR, AdaBoostM1, RotationForest, LogitBoost, RseslibKnn, LibSVM, and BayesNet. Using WEKA as a DM tool, the performance of the classifier algorithms was rated using accuracy, which measures the overall rate of correct prediction, precision, which measures the rate of instances correctly detected as phishing concerning all instances detected as phishing, recall, F-measure, root mean squared error, receiver operation characteristics area, root relative squared error, FPR, and TPR. Results showed that RF outperformed other algorithms with an FPR of 1.7% and an accuracy of 98.38%.

Much research has been done on detection systems of web phishing through DM techniques using a single classification algorithm. Therefore, the researchers of Ref. [20] ran a study to add a meta-algorithm to enhance classification performance and develop web phishing detection systems. The meta-algorithms added are bagging, boosting, and stacking. Bagging (bootstrap aggregation) starts with sampling the data population to make a training dataset. The acquired results are then processed, determining the generated predicted value by aggregating the most votes. The meta-algorithms are boosting aims to transform weak learners into strong learners. Stacking manages weak learners by combining different learning models. Results showed that adding meta-algorithms increased accuracy by around 2% as the scenario model using a classifier without adding meta-algorithm is 95.5%. In contrast, after adding boosting, bagging, and stacking meta-algorithms, the classification performance increased to 97.4%, 97.1%, and 97.5%, respectively.

The researcher of Ref. [21] presented another ML approach for detecting phishing by analyzing the hyperlinks in websites' HTML source codes. The hyperlink features are intended to train the algorithm and, in effect, to detect phishing. The presented method is not limited to a specific language and can detect websites of any language. This approach protects user privacy as it is a client-side solution, meaning that it obtains specific features from the browser only and not from third parties like search engines. The dataset used contains 2544 phishing and legitimate websites. Like many other studies, WEKA was used to calculate the evaluation metrics to gauge performance. The performance metrics are TPR (measures the rate of phishing websites classified as phishing out of entire phishing websites), FPR (measures the rate of legitimate websites classified as phishing out of total legitimate websites), TNR (measures the rate of legitimate websites classified as legitimate out of total legitimate websites), FNR (measures the rate of phishing websites classified as legitimate out of total phishing websites), F-measure (the harmonic mean of Precision and Recall), accuracy (measures the overall rate of correct prediction), precision (measures the rate of instances correctly detected as phishing concerning all instances detected as phishing), and recall. After performing experiments on the following classification algorithm: Sequential Minimal Optimization (SMO), NB, RF, SVM, AdaBoost, Neural Networks (NN),

C4.5, and LR, it was observed that LR achieved the best performance with an accuracy of 98.42% and TPR of 98.39%.

Another study that was mainly client-side and extricated different features from URLs and webpages source codes was done by the researchers of Ref. [22], which aimed to produce a consistent detection system that can keep up with fast-moving and changing environments since phishers are always finding new and innovative ways to deceive online users. The detection system was developed using XCS, an adaptive learning classifier system with the advantage of a flexible architecture and thus can be applied in dynamic environments. The dataset was a large collection of 3983 phishing and 4021 non-phishing websites. After experimenting, the performance was compared to several learning algorithms: C4.5 (DT), AdaBoost, Kstart, RF, SMO, and NB. Similar to the previous studies, the performance of these algorithms was evaluated by seven metrics: TPR, FPR, specificity, precision, accuracy, F-measure, and ROC. Based on the results, this method achieved the highest performance, especially in terms of accuracy, F-measure, and ROC, while considering all feature types.

Due to the minimal feature combination of images, frames, and text of phishing and benign sites, current technology must detect innovative phishing attacks correctly. Therefore, the authors of Ref. [23] proposed an integrated approach to considering them all simultaneously. They established an Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm to detect phishing using images, frames, and text features. This approach allows for the automated detection of a cloned website that is extremely similar in content, features, and appearance to a legitimate website instead of being carefully checked for authenticity by the user. Using the F-measure, recall, accuracy, and precision measures to compare its performance with KNN and SVM, results revealed that the proposed solution achieved an overall accuracy of 98.3%, higher than other two algorithms.

2.2 Related work in 2020 and 2021

According to Ref. [24], phishing is a major issue that impacts many fields, such as online marketing, banking, online business, and e-commerce. They described a couple of cases from which a site could be interpreted as a phishing website: when a popup asks for personal information when requesting a URL. Still, objects are loading from a different one, and links within a website redirect to a different domain. All these could be indicators of a phishing website. Their study focused on detecting phishing websites through a stacking model; they proposed two models (RF+NN+bagging) and (KNN+RF+bagging). Working on a dataset with 32 attributes and 11 055 web hits, they focused on selecting the top ten phishing features through different feature selection algorithms such as Recursive Feature Elimination (RFE), relief-F, Information Gain (IG), and Gain Ratio (GR), from which two sets of features were created: N1, which is a combination of all weak features, and N2 which is a combination of all powerful features. Their results reveal that model 1 (RF+NN+bagging), although time-consuming, has achieved an accuracy of 97.4% in detecting phishing websites.

Reference [25] mentioned that setting up a blacklist for phishing websites and URLs is inefficient as new websites are being developed daily. To find a solution for detecting phishing URLs, they proposed a model that can find the relationship between URL characters, resulting in more accurate detection of phishing URLs. The model combines Convolutional Neural Network (CNN) and Multi-Head Self-Attention (MHSA). MHSA checks the URL structure and finds its characters and the relationship between them, while the CNN model does the feature extraction. The dataset used for the experiment contained 43 984 phishing websites and 54 000 legitimate ones. It was seen that CNN-MHSA resulted in an accuracy of 99.84%, which is 6.25% better than CNN-Long Short Term Memory (LSTM) performance on a real environment dataset.

Reference [26] states that analyzing hyperlinks can help detect phishing websites. The researchers proposed Rule Extraction and Integration (REI), which detects phishing websites by performing hyperlink analysis by selecting and integrating hyperlink rules. They validated their approach by comparing the results with 16 other classifiers that are well known for this purpose: KNN, ANN, NB, AdaBoost, SVM, bagging,

RF, Classification Based on Association (CBA), Ripper, C5.0, OneR, Deep Neural Network (DNN), Deep Belief Network (DBN), CNN, Recurrent Neural Network (RNN), and LSTM. The dataset consisted of 1000 phishing and 1000 legitimate websites from phishtank.com and alexa.com. The approach constitutes three steps: data discretization, rule extraction, and integration. Their results showed that a website could be considered suspicious if it has a low value of hyperlink indicators, such as the number of referrals to a website. If it is less than 63.5, the website is considered phishing. Another example was the number of shares this website has since people tend to share legitimate websites, which indicates whether the website is phishing. It is important to mention that this approach outperformed all the classifiers mentioned above in detecting phishing websites by looking at the F-measure, recall, precision, and accuracy with 99.95%, 99.9%, 100%, and 99.95%, respectively.

Artificial Intelligence (AI) techniques have been used to mitigate attacks related to phishing; however, they have some areas for improvement regarding the high rate of false alarms and the inability to detect how phishing websites function. Reference [27] proposes a phishing detecting technique that includes four meta-learner classifiers: Rotation Forest Bagging Extra Tree (RoFBET), Bagging Extra Tree (BET), AdaBoost Extra Tree (ABET), and LogitBoost Extra Tree (LBET). The model was trained on datasets for phishing websites that are comprehensive and recent, containing 11 055 websites and 30 features. The performance was evaluated, showing an accuracy of 97% and an FPR of 2.8% maximum, which is very low. Given the promising results and the proposed model outperformed other ML techniques, this model is recommended for detecting phishing websites.

As previously mentioned, the computers and systems with weaknesses are the attackers' goals by imitating social media, banking, and e-commerce web pages that look legitimate but aim to steal sensitive information. The work to detect these phishing web pages has depended on ML techniques. The researchers of Ref. [28] proposed a system to detect phishing web pages based on ML techniques. This system analyzes the URL features and identifies whether it's legitimate or phishing. They used three datasets to train the model using eight algorithms. The number of URLs in the datasets was 83 857, 82 888, and 1 260 777, respectively. The URL fields such as "domain, subdomain, Top Level Domain (TLD), protocol, directory, file name, path, and query" can differentiate phishing URLs from legitimate ones as they differ in the phishing URL. The benefit of analyzing the webpage URL to detect if it is phishing is that it can be done quickly; other features in the web page, such as content, Cascading Style Sheets (CSS), and layout, take more time. The eight classifiers are ANN, SVM, KNN, NB, DT, RF, LR, and Extreme Gradient Boosting (XGBoost). The experiment showed that the best results were obtained from RF on the three datasets with an accuracy score of 94.59%, 90.50%, and 91.26%, respectively. However, the best training time was obtained from NB.

The authors of Ref. [29] designed a phishing detection model using DL-based techniques by checking a website's content, frames, text, images, and URL. Leveraging the mentioned features, they built a hybrid detection system called Intelligent Phishing Detection System (IPDS) using LSTM and CNN. They used 10 000 images to train their model and a dataset containing 1 million URLs. Results showed that using IPDS with the hybrid features successfully detected Phishing. Also, the model performance was outstanding, with a detection time of 25 s and an accuracy of 93.28%.

The authors of Ref. [30] focused on detecting emails that could be phishing. They designed an optimized algorithm, and their approach included all well-known steps for data classification, preprocessing, feature extraction, and feature selection. The DBN classifier was used, which was trained using the fractional-Earthworm Optimization Algorithm (EWA). Enron (2018) and UCI (2018) datasets were used for analysis. The proposed classifier results were compared with NB, NN, and EWA-DBN, and the accuracy of the proposed model was 85.71%, which was the highest accuracy achieved compared with the other methods mentioned. Also, other performance measures achieved

a high score for the proposed model, such as sensitivity of 81.82% and specificity of 88.00%.

Phishing imposes threats on legitimate websites and internet users, and its risks are complex. The authors of Ref. [31] argued that methods for detecting phishing websites, such as blacklisting and ML models, have somehow failed and can result in inaccurate results. Therefore, they proposed a new way of detecting phishing websites through a Functional Tree (FT) with its variants F1–FT inner nodes and leaves, F2–FT with leaves only, and F3–FT with inner nodes, along with bagging, boosting, and rotation forest meta-learners. Their experiment was conducted on three datasets to test the validation of the proposed models; their results showed that the F2–FT achieved the highest accuracy of 96.07% from NB, SVM, SMO, and Dec Table models. Also, it achieved the highest scores for F-measure, Area Under the ROC Curve (AUC), MCC, TPR, and FPR. Their study concluded that the FT algorithm better detects phishing websites than the other methods listed above. Also, the performance of the FT variant can improve with the use of bagging, boosting, and rotation forest meta-learners. Finally, it was stated that the proposed methods have superiority over previous methods.

Phishing websites are designed in a way that looks similar to legitimate websites. To differentiate, researchers need to look for two main features: similarities in the webpage view, such as logo, copyright, and favicon, and the stealing function implemented to steal the user's data, such as links, redirection, and Domain Name System (DNS). Based on the literature, effective feature extraction significantly affects the model's ability to recognize phishing websites. To detect phishing websites, the researchers of Ref. [32] constructed their dataset, which included phishing and legitimate websites with different content in many languages, given that the dataset quality can indicate if the model can perform as desired in a real environment. The dataset included 3972 legitimate samples and 195 phishing samples; 10 phishing websites were brands such as Apple, Microsoft, Facebook, and others in 12 languages, increasing the difficulty in detection. This study proposed combining "Counterfeiting", "Affiliation", "Stealing", and "Evaluation" (CASE) features with multistage detection, which resulted in more accurate phishing detection since they first focused on excluding legitimate websites. The CASE feature frame identifies if the website is phishing or not based on a set of features: counterfeiting features, affiliation features, stealing features, and evaluation features. The multistage detection included white filtering, fast counterfeit filtering, and accurate recognition, including the CASE framework. For classification, the methodologies used were AdaBoost, SMO, and RF. Results showed that the multistage model with RF (CASE) performed best with an F-measure of 96.59%.

Reference [33] focuses on detecting phishing URLs using the hostname length, URL, path, first directory, and top-level domain as indicators of a suspicious URL. Also, the count of certain keywords and special characters and the number of directories are all features to consider for checking the legitimacy of a URL. After data collection, a huge, unbalanced dataset was found; the total number of URLs was 450 176, from which 104 438 were suspicious and 345 738 were legitimate. The near-miss technique was used to under-sample the dataset, while the Synthetic Minority Over-sampling Technique (SMOTE) was used to over-sample the dataset. The following methodologies were used on the original dataset, the over-sampled and the under-sampled: KNN, SVM, DT, LR, RF, Gradient Boosting (GB), AdaBoost, XGBoost, and Light Gradient Boosting Machine (LGBM). The results showed that the AdaBoost classifier outperformed the other classifiers for the original and under-sampled datasets with an accuracy of 99.07% and 98.43%, respectively. The SVM classifier performed best for the over-sampled dataset, with an accuracy of 99.50%. It was stated that phishing is a major problem, and the higher the accuracy, the better. That's why over-sampling the dataset is the best way for unbalanced data, given that the accuracy was highest with the over-sampled dataset.

The researchers of Ref. [6] aimed to build a model to classify phishing websites by feature extraction. The results of their study provided an accuracy of 97.63%

for one of the used classifiers. Their datasets consisted of phishing and legitimate URLs with a size of 88 647 and 58 645. They experimented using eight classifiers: XGBoost, SVM, RF, KNN, ANN, LR, DT, and NB. The best results were retrieved from ANN in detecting phishing websites.

3 Research methodology

This systematic research reviewed papers on phishing detection through DM and ML techniques from 2018 to 2021, published in well-known journals' specifically by IEEE, Elsevier, Springer, and Emerald. Specific keywords were used to search for relevant papers, while the search process included three phases: searching, inclusion, exclusion, and extraction. The selected papers were screened to answer the Research Question (RQ) by which this systematic review is directed:

Is there a universally accepted model that can detect phishing websites with 100% accuracy?

3.1 Libraries search

A search was conducted on libraries selected from 2018–2021 with a combination of keywords. Table 2 shows the keywords used for searching.

3.2 Selection execution

Many papers were retrieved early in the search process through Google Scholar. Around 17 000 papers were returned from the first search when applying the targeted keywords from 2018 to 2021; the keywords used at this stage were Detecting Phishing. Since the number returned is impractically large, the authors changed the search to "Detecting Phishing". When applying this, the number of articles returned went down to 1910. To speed up the review process, the authors agreed to split the articles to be reviewed according to a yearly period; articles published in 2018 and 2019 were assigned to the 1st and 3rd authors, while articles published in 2020 and 2021 were assigned to the 2nd and 4th authors of this work. At this stage, the number of papers was 690 for 2018 and 2019 and 1210 for 2020 and 2021. The goal was to reach a reasonable number of studies that can be reviewed manually by skimming through their abstract, which can be included in the review. In order to facilitate the review process, the next phase included exporting the papers retrieved from the previous step. After that, two filtration criteria were followed; the first criterion was based on specific keywords: "Detect" AND/OR "Phishing" AND/OR "Machine Learning" AND "Data Mining". The second criterion was based on the publisher's name; since this systematic review is focused on reviewing studies from well-known journals only, many papers were excluded based on the publisher's name; any paper not published by Elsevier, Springer, Emerald, or IEEE was excluded. At this point, the number of papers remaining was 95, a reasonable number that can be skimmed through. Further filtration of the papers was performed based on their titles not being directly related to the research topic, which resulted in the elimination of 30 papers, reaching 65 papers. The following step included skimming through the paper's abstract, methodology, and results and selecting the studies that serve the review process. So far, The count was 43; however, 10 papers were eliminated since they did not add any new value to the research, with the final count as 33. Below is a list of the inclusion and exclusion criteria that were followed to select the studies that serve the purpose of this research.

Table 2 Keyword combinations.

Keyword	Combination
Phishing	"Phishing" OR/AND "Detect"
Websites	"Websites" AND "Phishing"
Detect	"Detecting" AND "Phishing" AND "Websites"
Data mining	"Phishing" AND "Detection" AND "Data Mining"
URL	"Phishing" AND "URL" OR "Websites" AND "Detect"

Inclusion criteria:

(1) Publications concerned with detecting phishing attempts;

(2) Publications from well-reputed journals;

(3) It helps answer the research question in any way;

(4) Date of publication within the defined research period.

Exclusion criteria:

(1) Publications need clearly defined methods that

serve the review;

(2) Date of publication outside the defined research period;

(3) The journal is not published by IEEE, Elsevier, Emerald, and Springer.

Table 3 shows the number of results retrieved from each library.

Figure 2 summarizes the steps of selecting the studies included in this review and the number of papers included/excluded in each phase.

4 Result and discussion

After reviewing 33 papers concerned with phishing, DM, and ML techniques from 2018 to 2021, it was demonstrated that much effort has been put together to find and create models that can detect and differentiate between phishing and legitimate websites. Some studies focused on detecting phishing websites by finding relationships between the URL characters[25], while others used URL features[22, 25, 28]. References [21, 26] used a different technique to analyze hyperlinks through integration rules from which suspicious ones can be detected. Other experiments studied the content, frames, text, and images, including the URL[29]. At the same time, Ref. [32] stated that favicon, copyright, and logo are important features that can be used for the same goal. The researchers proposed an integrated model that features combinations of images, text, and frames in the detection process. The results show that different algorithms were used, and multiple models were created by combining two to four techniques to get better results[24, 25, 27, 29–32]. Other studies used multiple models and compared their results to find the best approach[6, 11, 28, 34].

The dataset size and whether it is balanced plays a significant role in the model's accuracy and performance. Oversampling and under-sampling techniques are important when verifying the models; SMOTE and near-miss were used by the researchers of Ref. [33], respectively, for that purpose. The authors of Refs. [11, 12] referred to the importance of dimensionality reduction through feature selection in enhancing the accuracy performance of many

Table 3 Extracted papers from each library.

Source	Number of papers extracted
IEEE	16
Elsevier	7
Springer	6
Emerald	5

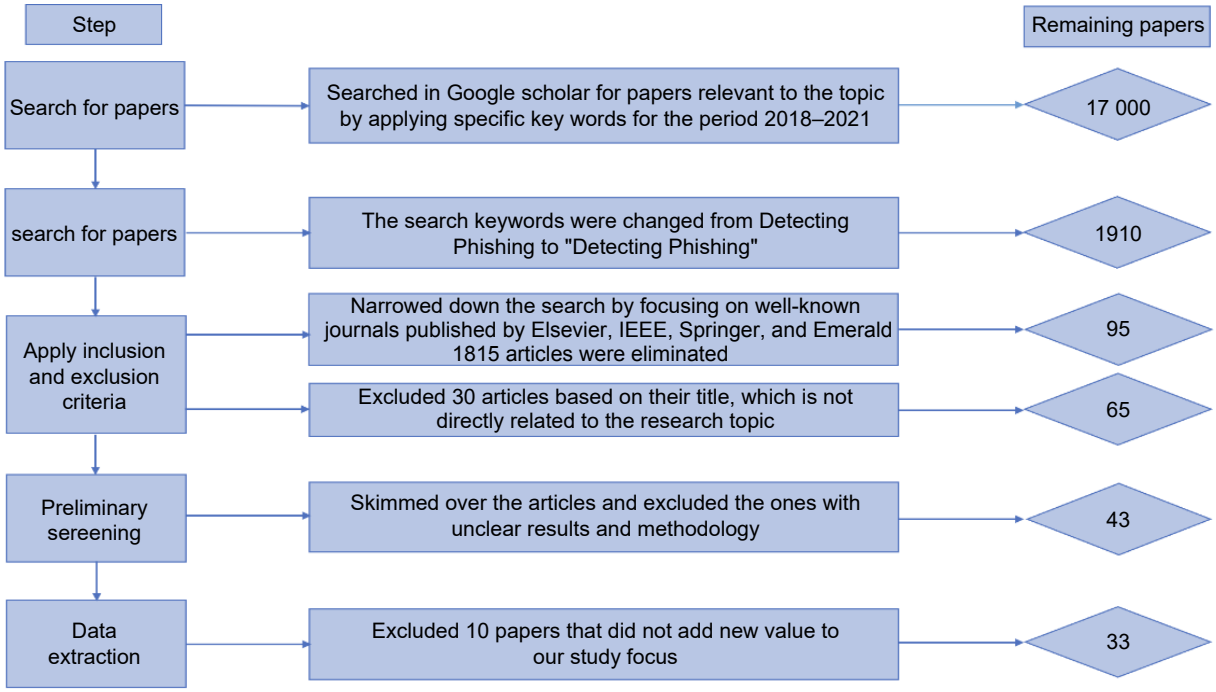

Fig. 2 Selection steps.

algorithms. The authors of Ref. [18] emphasized that email attachments add to the risk of increased susceptibility to phishing attempts and thus used two datasets related to both phishing websites and spam emails to train and test different models to decide on the best-performing algorithm. There are datasets available publicly for researchers to use for their experiments. Some researchers preferred to create their own dataset to test their models, such as the authors of Ref. [6], while the authors of Ref. [12] contributed to creating two benchmark datasets that researchers can use. Feature selection techniques were combined with the algorithms for better model performance; bagging was the most significant.

From the results, most models resulted in accuracies higher than 90%; however, attackers are creating different approaches to overcome those models. Reference [13] indicated that phishing detection mechanisms are, in fact, unimmune to adversarial learning techniques. Based on the above, analyzing and checking the URL features is the most efficient way to detect phishing websites, considering time constraints. Furthermore, until 2021, no universally accepted model can be used to detect phishing websites with 100% accuracy.

Moreover, it was also noticed that some DM techniques were used more than others. RF was the most used technique in detecting phishing websites, used in 13 studies out of the 33 studies reviewed. The reason behind that may be that RF is less affected by noise and overfit resistance. The second most used technique was NB, used in 11 studies. This may be because NB is easy to implement and requires a short time for training. SVM was the most commonly used algorithm after NB; it was used in 9 studies because it is easy to implement and works well with many independent variables. KNN was also one of the most frequently used techniques in detecting phishing attempts, implemented in 8 studies since it is known for its efficiency and speed. The most implemented algorithms were DT and LR, used in 7 and 6 studies, respectively. DT is a simple and easy-to-explain technique, while LR is efficient when the dependent variable is collected from two classes.

The authors of Refs. [11, 16–18, 24] depended on the feature selection method to detect phishing websites and emails. They all commonly used the NB algorithm, among other techniques, which indicates that the NB algorithm is a practical choice for applying feature selection.

Analyzing the content to detect phishing websites was used by three studies[23, 29, 32]. It was noticed that each study relied on a different DM technique; for example, Ref. [23] used SVM, KNN, and ANFIS, where the best model appeared to be ANFIS. CNN and LSTM were used by the authors of Ref. [29], where they created a new model using those mechanisms called IPDS. The authors of Ref. [32] used RF, AdaBoost, and SMO, and they also created a new model for detecting phishing websites called Multistate model + RF (CASE). This reveals that there is no specific agreed-upon algorithm for content analysis.

Analyzing URL features was a commonly used method for detecting phishing attempts. The authors of Refs. [3, 25, 28, 29, 33] followed this approach for detecting phishing. It was observed that the researchers of Refs. [3, 28, 33] commonly used RF, which appeared to be the model with the highest accuracy for Refs. [3, 28], while the highest accuracy algorithm for Ref. [33] was SVM. The authors of Refs. [25, 29] used CNN and LSTM; however, MHSA was also combined with CNN by Ref. [25], which appeared to be the best-performing model than CNN+LSTM.

Three studies applied feature extraction to their datasets to create new features for detecting phishing[6, 14, 22]. They all commonly used RF as part of their study. None of their results showed that RF is the best model.

Two studies depended on analyzing hyperlink features to detect phishing[21, 26]. Their approach was to compare different algorithms' performances and to select the best one. They both commonly used NB and SVM in their comparison, among other techniques. The best-performing technique for Ref. [21] was LR, while REI outperformed all the compared algorithms by Ref. [26].

Finally, Tables 4 and 5 summarize the results of the reviewed studies from 2018–2019 and 2020–2021, respectively.

Table 4 2018–2019 studies summary.

Attribute		Ref. [14]	Ref. [3]	Ref. [15]	Ref. [16]	Ref. [17]	Ref. [18]	Ref. [11]	Ref. [19]	Ref. [21]	Ref. [22]	Ref. [23]
General information	Year	2018	2018	2018	2018	2018	2018	2019	2019	2019	2019	2019
	New model	No	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes
	Dataset source	UCI	Phishtank	UCI	-	UC Irvine	UCI&Spam Emails	UCI	UCI	-	-	California Irvine & Huddersfield
	Dataset Size	11 055	-	11 055	11 055	11 000	11 055	2670	11 055	2544	8044	13 000
Data mining technique	SVM					√	√			√		√
	LR				√		√			√		
	NB			√	√	√	√	√		√	√	
	DT				√					√	√	
	ANN						√					
	NN									√		
	RF	√	√	√			√		√	√	√	
	KNN				√		√					√
	AdaBoost								√	√	√	
	SMO			√						√	√	
	Lazy Kstar	√										
	BayesNet	√							√			
	SGD	√										
	LMT	√		√					√			
	ID3	√										
	Multilayer perceptron	√										
	JRip	√							√			
	PART	√							√			
	J48	√		√				√				
	RandTree	√		√					√			
	LSVM				√							
	RSVM				√							
	LDA				√							
	ELM					√						
	HNB							√				
	HNB & J48							√				
	NNN				√							
	SysFor								√			
	RseslibKnn								√			
	LibSVM								√			
	SPAARC								√			
	RepTree								√			
	ForestPA								√			
	NNge								√			
	OneR								√			
	RotatForest								√			
	LogitBoost								√			
	Kstart										√	
	XCS										√	
	ANFIS											√
	RFC	√										
Best performance	Model	Lazy Kstar	RF	RT+RF	NNN	ELM	RF	HNB&J48	RF	LR	XCS	ANFIS
	Accuracy (%)	97.6	95	99	97.71	95.34	97.26	96.3	98.4	98.42	98.3	98.3

Table 5 2020–2021 studies summary.

Attribute		Ref. [24]	Ref. [25]	Ref. [26]	Ref. [27]	Ref. [28]	Ref. [29]	Ref. [30]	Ref. [31]	Ref. [32]	Ref. [33]	Ref. [6]
General information	Year	2020	2020	2020	2020	2020	2020	2020	2021	2021	2021	2021
	New model	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No
	Dataset source	-	-	Phishtank & Alexa	UCI & Kaggle	-	-	Enron & UCI	-	-	-	-
	Dataset size	11 055	88 984	2000	11 055	83 857 82 888 1 260 777	1 million URLs & 10 000 Image	-	11 055 10 000 1353	4167	450 176	147 292
Data mining technique	SVM			√		√			√		√	√
	LR					√					√	√
	NB			√		√			√			√
	DT			√		√					√	√
	FT								√			
	ANN			√		√						√
	NN	√										
	RF	√		√		√				√	√	√
	KNN	√		√		√					√	√
	RoFBET				√							
	BET				√							
	ABET				√							
	LBET				√							
	CNN		√	√			√					
	LSTM		√	√			√					
	MHSA		√									
	XGBoost					√					√	√
	DBN			√				√				
	AdaBoost			√						√	√	
	SMO								√	√		
	GB										√	
	LGBM										√	
	GBA			√								
	Ripper			√								
	DNN			√								
	OneR			√								
	RNN			√								
	Bagging			√					√			
	Boosting								√			
	Rotation forest								√			
Best performance	Model	RF+NN+bagging	CNN+ MHSA	REI	RoFBT+ BET+ ABE+ BET	RF	IPDS	WDA+ DBN	FT+ bagging+ boosting+ rotation forest	Multistate model + RF (CASE)	SVM	ANN
	Accuracy (%)	97.4	99.84	0.995	97	94.95	93.28	0.857	97.86	-	99.50	97.63

5 Conclusion

Web applications and internet technology have become essential in our day-to-day activities, from searching for a cooking recipe to paying bills to performing banking transactions online. With the immense

dependency on cyberspace, cybersecurity problems have arisen with threats to steal the end-user's personal and credit card information, known as phishing. Phishing is a serious issue with many severe consequences. It may affect individuals as well as large firms and businesses. There are many ways to detect phishing websites, making this a vital research topic in recent years. This paper reviewed 33 research journal papers for detecting phishing websites using DM techniques published between 2018 and 2021 by the top four publishes: Elsevier, Emerald, Springer, and IEEE. Based on the review results, it was shown that analyzing and checking the URL features is the most efficient way to detect phishing websites, considering time constraints. Also, many DM techniques and models achieved 90% or higher accuracy in detecting phishing websites. Still, only some universally accepted models of phishing websites with 100% accuracy are accepted. Future work should include more studies from different journals, as this review exclusively relied on four well-known journal publishers and thus excluded papers that were relevant to the research topic.

References

[1] D. Goel and A. K. Jain, Mobile phishing attacks and defence mechanisms: State of art and open research challenges, *Comput. Secur.*, vol. 73, pp. 519–544, 2018.

[2] Q. Abu Al-Haija and M. Al-Fayoumi, An intelligent identification and classification system for malicious uniform resource locators (URLs), *Neural Comput. Appl.*, pp. 1–17, 2023.

[3] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, A new method for detection of phishing websites: URL detection, in *Proc. 2018 Second Int. Conf. Inventive Communication and Computational Technologies* (*ICICCT*), Coimbatore, India, 2018, pp. 949–952.

[4] M. Baykara and Z. Z. Gürel, Detection of phishing attacks, in *Proc. 2018 6th Int. Symp. Digital Forensic and Security* (*ISDFS*), Antalya, Turkey, 2018, pp. 1–5.

[5] I. Vayansky and S. Kumar, Phishing - challenges and solutions, *Comput. Fraud Secur.*, vol. 2018, no. 1, pp. 15–20, 2018.

[6] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, Phishing website detection from URLs using classical machine learning ANN model, in *Proc. 17th EAI Int. Conf. Security and Privacy in Communication Systems*, virtual, 2021, pp. 509–523.

[7] Q. Abu Al-Haija, M. Alohaly, and A. Odeh, A lightweight double-stage scheme to identify malicious DNS over HTTPS traffic using a hybrid learning approach, *Sensors*, vol. 23, no. 7, pp. 3489, 2023.

[8] R. Butler and M. Butler, Assessing the information quality of phishing-related content on financial institutions' websites, *Inf. Comput. Secur.*, vol. 26, no. 5, pp. 514–532, 2018.

[9] A. A. Zuraiq and M. Alkasassbeh, Review: Phishing detection approaches, in *Proc. 2019 2nd Int. Conf. New Trends in Computing Sciences* (*ICTCS*), Amman, Jordan, 2019, pp. 1–6.

[10] S. Adi, Y. Pristyanto, and A. Sunyoto, The best features selection method and relevance variable for web phishing classification, in *Proc. 2019 Int. Conf. Information and Communications Technology* (*ICOIACT*), Yogyakarta, Indonesia, 2019, pp. 578–583.

[11] S. Zaman, S. M. Uddin Deep, Z. Kawsar, M. Ashaduzzaman, and A. I. Pritom, Phishing website detection using effective classifiers and feature selection techniques, in *Proc. 2019 2nd Int. Conf. Innovation in Engineering and Technology* (*ICIET*), Dhaka, Bangladesh, 2020, pp. 1–6.

[12] G. Vrbančič, I. Fister Jr, and V. Podgorelec, Datasets for phishing websites detection, *Data Brief*, vol. 33, p. 106438, 2020.

[13] H. Shirazi, B. Bezawada, I. Ray, and C. Anderson, Adversarial sampling attacks against phishing detection, in *Proc. 33rd Annual IFIP Conf. Data and Applications Security and Privacy*, Charleston, SC, USA, 2019, pp. 83–101.

[14] M. Karabatak and T. Mustafa, Performance comparison of classifiers on reduced phishing website dataset, in *Proc. 2018 6th Int. Symp. on Digital Forensic and Security* (*ISDFS*), Antalya, Turkey, 2018, pp. 1–5.

[15] S. Gupta and A. Singhal, Dynamic classification mining techniques for predicting phishing URL, in *Soft Computing*: *Theories and Applications*, M. Pant, K. Ray, T. K. Sharma, S. Rawat, A. Bandyopadhyay Eds. Singapore: Springer, 2018: 537-546.

[16] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Wang, The application of a novel neural network in the detection of phishing websites, *J. Ambient Intell. Humaniz. Comput*., pp. 1–15, 2018.

[17] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avcı, Phishing web sites features classification based on extreme learning machine, in *Proc. 2018 6th Int. Symp. on Digital Forensic and Security* (*ISDFS*), Antalya, Turkey, 2018, pp. 1–5.

[18] I. Salihovic, H. Serdarevic, and J. Kevric, The role of feature selection in machine learning for detection of

Spam and phishing attacks, in *Proc. Int. Symp. Innovative and Interdisciplinary Applications of Advanced Technologies* (*IAT*), Jahorina, Bosnia and Herzegovina, 2018, pp. 476–483.

[19] N. N. Gana and S. M. Abdulhamid, Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis, in *Proc. 2019 2nd Int. Conf. IEEE Nigeria Computer Chapter* (*NigeriaComputConf*), Zaria, Nigeria, 2020, pp. 1–8.

[20] A. F. Nugraha and L. Rahman, Meta-algorithms for improving classification performance in the web-phishing detection process, in *Proc. 2019 4th Int. Conf. Information Technology, Information Systems and Electrical Engineering* (*ICITISEE*), Yogyakarta, Indonesia, 2020, pp. 271–275.

[21] A. K. Jain and B. B. Gupta, A machine learning based approach for phishing detection using hyperlinks information, *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 2015–2028, 2019.

[22] 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 *Proc. 2019 5th Int. Conf. Web Research* (*ICWR*), Tehran, Iran, 2019, pp. 281–286.

[23] M. A Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, Intelligent web-phishing detection and protection scheme using integrated images, frames, and text features, *Expert Syst. Appl.*, vol. 115, pp. 300–313, 2019.

[24] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum, and M. Hamdani, Phishing web site detection using diverse machine learning algorithms, *Electron. Libr.*, vol. 38, no. 1, pp. 65–80, 2020.

[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 Netw.*, vol. 125, pp. 303–312, 2020.

[26] C. Wang, Z. Hu, R. Chiong, Y. Bao, and J. Wu, Identification of phishing websites through hyperlink analysis and rule extraction, *Electron. Libr.*, vol. 38, nos. 5/6, pp. 1073–1093, 2020.

[27] Y. Ahmad Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, AI meta-learners and extra-trees algorithm for the detection of phishing websites, *IEEE Access*, vol. 8, pp. 142532–142542, 2020.

[28] M. Korkmaz, O. K. Sahingoz, and B. Diri, Detection of phishing websites by using machine learning-based URL analysis, in *Proc. 2020 11th Int. Conf. Computing, Communication and Networking Technologies* (*ICCCNT*), Kharagpur, India, 2020, pp. 1–7.

[29] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, Intelligent phishing detection scheme using deep learning algorithms, *J. Enterp. Inf. Manag.*, vol. 36, no. 3, pp. 747–766, 2023.

[30] M. Arshey and K. S. Angel Viji, An optimization-based deep belief network for the detection of phishing e-mails, *Data Technol. Appl.*, vol. 54, no. 4, pp. 529–549, 2020.

[31] A. O. Balogun, K. S. Adewole, M. O. Raheem, O. N. Akande, F. E. Usman-Hamza, M. A. Mabayoje, A. G. Akintola, A. W. Asaju-Gbolagade, M. K. Jimoh, R. G. Jimoh, et al., Improving the phishing website detection using empirical analysis of Function Tree and its variants, *Heliyon*, vol. 7, no. 7, p. e07437, 2021.

[32] 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, *Comput. Secur.*, vol. 110, p. 102421, 2021.

[33] A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra, and A. Mani Jha, Machine learning approach based on hybrid features for detection of phishing URLs, in *Proc. 2021 11th Int. Conf. Cloud Computing, Data Science & Engineering* (*Confluence*), Noida, India, 2021, pp. 954–959.

[34] Q. Abu Al-Haija and A. Al Badawi, URL-based phishing websites detection via machine learning, in *Proc. 2021 Int. Conf. Data Analytics for Business and Industry* (*ICDABI*), Sakheer, Bahrain, 2021, pp. 644–649.

Dina Jibat received the master degree in business analytics from Princess Sumaya University for Technology, Jordan in 2023. She is a senior software quality analyst team lead at Aspire Company in Jordan. Her research interest includes software quality, information security, and communication topics.

Sarah Jamjoom received the master degree in business analytics from Princess Sumaya University for Technology, Jordan in 2023. Her research interests include big data analytics, internet mobile banking systems, and information security.

Abdallah Qusef received the PhD degree in software engineering from the University of Salerno, Italy in 2012. He has over ten years of industry experience as a senior systems analyst and project manager. Currently, he is an associate professor at Princess Sumaya University for Technology, Jordan, a senior member of IEEE and ACM, and a member of the program committee of many international and local conferences and journals. In many organizations and ministries, he presents consultation services related to software quality assurance, project management, and e-business topics. His research interests include software engineering, project management, information security, and e-business topics.

Qasem Abu Al-Haija received the PhD degree from Tennessee State University (TSU), USA, in 2020. He is an assistant professor with Department of Cybersecurity, Princess Sumaya University for Technology, Jordan and also with Department of Cybersecurity, Faculty of Computer & Information Technology, Jordan University of Science and Technology, Jordan. He authorized more than 100 scientific research papers and book chapters. His research interests include Artificial Intelligence (AI), cybersecurity and cryptography, Internet of Things (IoT), Cyber-Physical Systems (CPS), Time Series Analysis (TSA), and computer arithmetic. Recently, he was listed as one of the world's top 2% of scientists list released publicly by Stanford University and Elsevier.