Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Phishing Attack Recognition Mechanism Based on Improved Decision Tree Algorithm

Haoran Yang[1,✉], Yi Li[2], Chang Liu[3], Yichuan Zhou[4]

[1] *Beijing Troy Cloud Data Technology Co., Ltd., Beijing, 100071, China*
[2] *School of Computer Science and Technology, Jilin University, Beijing, 100010, China*
[3] *Department of Hospitality and Business Management, The Technological and Higher Education Institute of Hong Kong, 999077, Hong Kong*
[4] *Shanghai Shiyun Information Technology Co., Ltd., Shanghai, 200120, China*

ABSTRACT
Phishing has become an increasing threat on online networks with evolving Web, mobile device and social networking technologies. Therefore, there is an urgent need for effective methods and techniques used to detect and prevent phishing attacks. In this paper, a phishing detection model based on decision tree and optimal feature selection is proposed. An optimal feature selection algorithm based on a newly defined feature evaluation metric (f_Value), decision tree and local search is designed to prune out negative and useless features. The overfitting problem in the process of training neural network classifiers is mitigated. The optimal set of sensitive features for feature selection and the optimal structure for training the neural network classifier are constructed by tuning the parameters. Experiments on CART-based phishing detection system and comparative experiments based on different phishing detection models are also conducted. The experimental results show that the model precision, accuracy, and recall of the improved decision tree-based algorithm proposed in the article are 92.7%, 96.5%, and 88.3%, respectively, on the dataset of phishtank, and the three metrics are 98.3%, 99.1%, and 99.5%, respectively, on the datasets of Vrban ˇ ci ˇ c-small and show that the proposed CART has a higher performance than the many existing method models.

Keywords: decision tree, optimal feature selection, phishing detection model, neural network

1. Introduction

In today's digital era, the network has become an indispensable part of our life and work. However, along with the popularization and development of the network, a variety of network security threats are emerging, and phishing attacks are one of the common and harmful threats. Phishing attacks are designed to bring huge losses to users by tricking them into obtaining their sensitive information, such

✉ Corresponding author.
E-mail address: admin@adysec.com (H. Yang).
Received 15 April 2024; Revised 05 July 2024; Accepted 25 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-156
© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

as usernames, passwords, credit card numbers, etc. [1]. With the development of information technology, phishing attacks are also evolving, from the beginning of contact baiting attacks to hidden attacks in the network [2]. At present, common phishing attacks include fake emails, fake websites, instant messenger scams, phone scams, and malware [3]. Among them, fake emails are one of the most common means used by phishing attackers. Attackers will disguise themselves as legitimate organizations or companies, such as banks, e-commerce platforms, etc., and send users seemingly formal emails, which usually contain urgent prompts, such as "there is an abnormality in your account, please click the link to verify immediately", and attach a link to a fake website [4]. Once the user clicks on the link, he or she may be directed to enter sensitive personal information. And fake websites are fake websites that attackers will create that look almost exactly like the legitimate website, except that the URL may be slightly different [5]. When users enter relevant keywords in a search engine, they may mistakenly click into these fake websites. Unknowingly, the information entered by the user on the fake website will be stolen by the attacker. In addition, attackers usually disguise themselves as websites of credible entities or organizations and hide their attacks in seemingly normal mediums such as links, images, and files, which are difficult for users to detect, in order to gain users' trust and control their behaviors, which are highly disguised and hidden, and therefore have a long latent time [6-9]. In addition, through WeChat, QQ and other instant messaging tools, the attacker will contact the user as a friend, colleague or acquaintance, send false information or links to induce the user to click or provide sensitive information, for this type of social platform, some scholars have also proposed to put a phishing attack, so that the user is protected from its harm [10-11]. The most contact with the life of the telephone voice fraud, the performance of the attacker may pose as a bank customer service, government staff, etc., through the phone to contact the user, with a variety of reasons to require the user to provide personal information or transfer operations [12]. Malware is a way of attacking the user's device and obtaining information by implanting malicious codes or programs. Attackers may induce users to install applications with malware through emails, social media, downloading software, etc. Once the user installs malware, the attacker can monitor all the user's operations, steal sensitive information, and may cause irreparable damage to the user's device [13]. Against this background, phishing attack identification has become a research priority. In the identification stage, it is first necessary to understand the distinctive features of phishing attacks, i.e., disguising authenticity, creating urgency, requesting personal sensitive information, and inducing malicious operations [14]. Phishing attackers often take advantage of the user's fear or curiosity to create urgency and induce the user to take quick action by informing the user that the account may be stolen or there is an unfinished transaction, etc., so as to make the user lose their vigilance, which also indicates side by side that the user's susceptibility to this type of phishing attack is low [15-17]. Phishing attackers will guide users to carry out malicious operations, such as downloading virus software, transferring money to unidentified accounts, clicking on malicious links, etc., by means of forged platform pages or email links, etc., which will lead to serious consequences such as data leakage and loss of funds [18-19]. Although designers have set up a number of anti-phishing attack techniques for websites, emails and other research, identification remains the focus of phishing attacks [20]. Therefore, it is crucial to understand the recognition mechanism of phishing attacks for better network security.

There are many identification and detection methods for phishing attacks, and literature [21] analyzes the characteristics of phishing objects with rough sets and formal concepts to create an identification model for phishing attacks. The study starts with the attack object and performs a limited analysis over a large number of data samples. Literature [22] created a privacy-aware framework to identify and mitigate phishing attacks on unsuspecting web users and IoT devices and the effectiveness of the framework is not limited to a single type of phishing attack. Literature [23] proposed a spear phishing attack detection system for enterprise credentials by focusing on user-registered domain names, which strongly combats phishing attacks on fake emails.

Identifying phishing attacks can be regarded as a classification problem, so there are a number of scholars between classification algorithms applied to it. Literature [24] concludes that random forest is the best classification algorithm for phishing attack identification in websites by comparatively studying each technique in machine learning. And literature [25] jointly random forest and support vector machine to identify the classification phishing attacks, the accuracy of identification in the hybrid identification model is 94%, random forest is about 93% and support vector machine is 90%. Literature [26] proposed a new feature dataset and identified phishing attacks on emails with the support of Support Vector Machines, Simple Bayes and Long Short Term Memory algorithms and the accuracy of these algorithms were 99.62%, 97% and 98% each.

In addition, literature [27] builds an identification model for phishing attack detection with K-nearest neighbor algorithm support combined with Uniform Resource Locator (URL) and the accuracy of this identification method is at 85%. Literature [28] modeled the recognition of phishing by character-level convolutional neural network, which is used to obtain the information in the URL and its sequential patterns to classify them, obtaining an average accuracy of 96%. Whereas, literature [29] utilizes the consistency features of URLs and web pages to identify phishing attacks in communication networks and the accuracy of this method is 99.1%. Literature [30] firstly identifies the target domain name of the phishing webpage by analyzing the phishing relationship between the webpage and its related domain name, and identifies the target domain name of the phishing webpage by in-degree link association, and at the same time introduces a target verification algorithm to ensure that the identified target domain name attributes are correct, and the accuracy of the identification of this method is more than 99%.

What's more, literature [31] identifies phishing attacks existing in commercial and e-banking websites by analyzing the phase correlation between keywords, links, and visual designs in web pages, while the source web pages can be detected without training, and its accuracy is about 99.7%. While literature [32] builds a comprehensive dataset of URLs of seductive advertisement websites based on URL structure, text, and images, and identifies and detects them in a multi-class machine learning classification model, and detects them best with the XGBoost algorithm which has an accuracy and accuracy of 91% and 94%, respectively, where the decision tree algorithm has an accuracy of 90%. Intelligent algorithms such as machine learning have their own research on the identification of various types of phishing attacks, but the feature factors considered are monotonous and lack comprehensiveness. While decision trees have good adaptability to discrete and continuous features, can handle multi-classification and non-linear problems, and also have the advantages of anti-slip performance and scalability performance, but they are prone to overfitting [33]. It is believed that it can be improved to obtain good accuracy and precision performance in phishing attack recognition.

In this paper, a new feature selection algorithm is designed which selects the best feature set based on the newly defined f_Value index, decision tree and local search as the base classifier. In the first stage, features are encoded using evidence weights and a new feature importance evaluation metric, InVa, is introduced to evaluate the impact of different features on phishing detection more effectively, thus removing irrelevant features; in the second stage, a randomization strategy is used to optimize the search process for the optimal feature subset, and a simulated annealing-based improved bi-directional search algorithm, SA-BS, is proposed to further remove the associated redundant features. In order to evaluate the impact of different features on phishing detection, this paper proposes a new feature evaluation index, f_Value, and the new f_Value index is defined based on the Gini coefficient and decision tree.

2. Research on phishing detection based on feature selection

2.1. Bayesian-based detection models

Bayesian classifiers are built under probability to make classification decisions - a method. It is a simple and effective classifier, often used in applications such as phishing website detection, email filtering and so on. Bayesian classifier is a discriminative model: for a given sample to model the probability distribution, through the size of the posterior probability to decide the classification of the belonging. The Bayesian based phishing website prediction model is described as follows.

Given the feature vectors and $X=\{x_1,x_2,...,x_n\}$ categorical category $C=\{c_1,c_2,...,c_i\}$, compute the posterior probability corresponding to the category $P(C|X)$. The posterior probability $P(C|X)$ can be defined according to Bayes' theorem as:

$$P(C_i|X)=\frac{P(C_i)P(C_i|X)}{P(X)} \tag{1}$$

Since the conditional probability of a category is the joint probability of all features, it is difficult to obtain it directly from a finite training sample. To solve this problem, $P(C_i|X)$ of the plain Bayesian classifier can be defined again under the premise that the features are opposites of each other:

$$P(C_i|X)=\frac{P(C_i)}{P(X)}\prod_{i=1}^{d}P(x_i|C_i) \tag{2}$$

where d is the number of features and xi is the value of X on the ith feature. Since it is the same for all classification categories $P(x)$, the plain Bayesian classifier obtained from equation (2) can be defined as equation (3). where h_{nb} denotes the hypothesis trained by the Bayesian algorithm.

$$h_{nb}(X)=\arg\max_{C\in\gamma}P(C_i)\prod_{i=1}^{d}P(x_i|C_i) \tag{3}$$

For the detection of phishing websites, there are two classification categories C_1 (for phishing websites) and C_2 (for legitimate websites). When $P(C_1|X)\geq P(C_2|X)$, i.e., the probability of predicting a phishing website is greater than the probability of predicting a legitimate website, the website to be tested can be categorized as a phishing website and vice versa.

The Plain Bayesian detection model is characterized by using the value of the posterior probability to determine the classification of the website. The Plain Bayesian approach has the advantage of running fast and simple methods, and is more suitable for datasets with low correlation between features. However, the model has an obvious drawback. When the complexity of the dataset is high, the accuracy of the classification results is low.

2.2. URL and HTML Analysis

2.2.1. URL Analysis. The Uniform Resource Locator (URL) is used to represent the unique address of a resource on the Internet. It is composed of Protocol, Domain, File Path and Query. And, a URL represents a data object on the Internet. The information resources such as codes and pictures corresponding to URLs can be obtained through URLs, which provides the possibility of sensitive feature extraction of URLs. The structure of a URL can be shown in Figure 1.

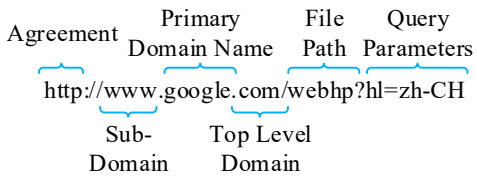

Fig. 1. URL segmentation.

2.2.2. HTML analysis. HTML (Hypertext Markup Language) was released as a working draft of the Internet Engineering Task Force (IETF). HTML tags each part of a web page with markup symbols that are to be displayed and parsed by the browser to display the content of the web page.The DOM tree structure is shown in Figure 2. The left side represents the source code tag content of the web page, and the right side represents the DOM tree structure generated with the HTML tag pairs. HTML DOM (Document Object Model) defines standard methods for accessing and manipulating HTML documents. HTML documents can be viewed as a collection of tagged nodes organized in a hierarchical structure, and these nodes mainly contain document nodes, element nodes, text nodes, and attribute nodes. In addition, some script codes (e.g., JavaScript) are also embedded in HTML pages. In feature extraction, this paper can obtain the corresponding HTML source code information through web crawling to extract the corresponding features of the web page.

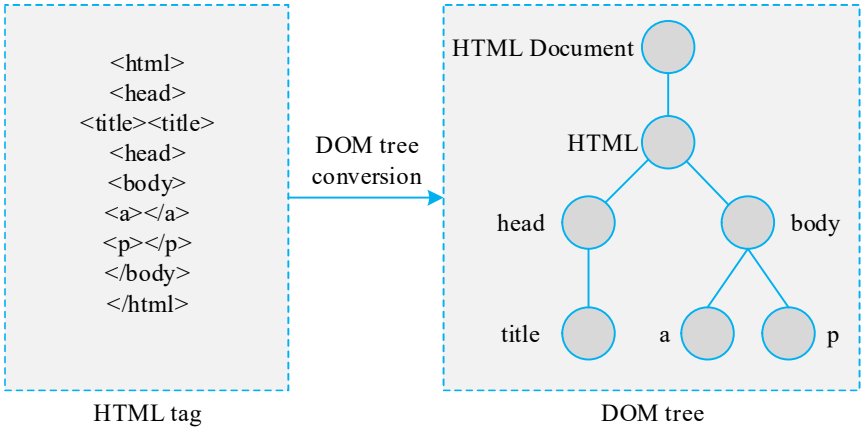

Fig. 2. Convert HTML tags to DOM trees.

2.3. Feature preprocessing

In the previous subsection a variety of features are extracted based on the URL and HTML of a website, and most of these features are discrete or continuous numerical features. In practice, different parts of such numerical features play different roles in improving phishing detection accuracy. In order to more accurately assess the magnitude of the impact of different features on the category labels as well as to train the phishing detection model more effectively, it is necessary to discretize the features. In this paper, we draw on the idea of the CART algorithm to divide the optimal feature segmentation points, and use the extended Gini index for phishing feature binning.

First, for phishing dataset D, its Gini index is calculated as shown in Equation (4), where K is the number of sample categories in phishing dataset D, and p_k is the proportion of category k samples to the total samples in D. Since phishing website detection is a binary classification problem, the samples in the dataset contain a total of two categories, i.e., phishing samples and legitimate samples. Therefore, the value of K in Equation (4) is fixed to 2.

$$Gini(D)=1-\sum_{k=1}^{K}p_k^2 \tag{4}$$

In this study, a certain feature f in the original dataset D is defined to have V possible values $a^1,a^2,\cdots,a^V$ and D is taken as the root node. According to the CART algorithm, when the root node D is divided into V sub-nodes using feature f, where the v th node contains all the samples of feature f that take the value of a^v, we denote the set of samples as D^v. Since D contains only two categories of phishing websites and legitimate websites, D is divided into D_1 and D_2 when a^v is chosen as the feature splitting point. The Gini index of the feature f in the dataset D when a^v is selected as the segmentation point is calculated as shown in Equation (5).

$$Gini\left(D,a^t\right)=\left(\left|D_1\right|Gini\left(\left|D_1\right|\right)+\left|D_2\right|Gini\left(\left|D_2\right|\right)\right)/\left|D\right| \tag{5}$$

where $\left|D_1\right|$ and $\left|D_2\right|$ are the number of samples in datasets D_1 & D_2, respectively, then the Gini index of feature f over all values taken can be calculated by Equation (5), where the feature value with the smallest Gini index is defined as the optimal segmentation point, as shown in Equation (6).

$$f\_a_*^v=\min_{a^v\in A}\mathrm{Gini}\left(D,a^v\right),v=1,2,\ldots,V \tag{6}$$

After dividing the samples in D using $f\_a_*^v$, for the set of samples on each branch node, these sets continue to be divided using the same method until the sample size in the set is less than the defined minimum sample size as shown in Equation (7). In this study, ρ is defined as the partition index of D and ρ is set to 0.1 based on experience.

$$\left|D^v\right|_{\min}=\left|D\right|\times\rho \tag{7}$$

Next, this paper designs an optimal binning algorithm for fishing features based on the Gini index in CART, assuming that each sample in the original fishing dataset D has m features, and now calculates the optimal binning $B_i=b_1,b_2,\ldots,b_d$ for any feature f_i, the specific steps of the algorithm are as follows:

1): first sort the V possible values of feature f_i from smallest to largest, assuming that the sorted result is $a^1,a^2,\cdots,a^V$, and then calculate the average of the two neighboring feature values as the segmentation point $b^1,b^2,\cdots,b^{V-1}$.

2): according to equation (5), calculate the Gini index when feature f_i uses b^v as the feature split point, where $v=1,2,\cdots,V-1$.

3): get the feature value $f\_b_*^v$ that minimizes the Gini index of feature f_i after division according to equation (6), and use this feature value as the optimal partition point to divide data set D into D_1 and D_2.

4): According to formula (7), calculate the current minimum sample size $\left|D^v\right|_{\min}$, if $\left|D_1\right|$ or $\left|D_2\right|$ is greater than $\left|D^v\right|_{\min}$, then repeat step (1) to continue to divide the subset, otherwise stop the division and output the optimal feature partition box B_i of the current feature.

2.4. Optimal feature selection

In this section, a two-stage feature selection algorithm is proposed to obtain the optimal set of artificial features, and Fig. 3 shows the flow of the algorithm. In the first stage, WoE is used to encode the features, and then an importance evaluation index $InVa$ for fishing features is designed, using

which irrelevant features can be effectively filtered out; in the second stage, an improved bi-directional search method SA-BS based on simulated annealing is proposed, which takes advantage of simulated annealing that can jump out of the locally optimal solution to further remove redundant features, and ultimately obtains a set of optimal subsets of features.

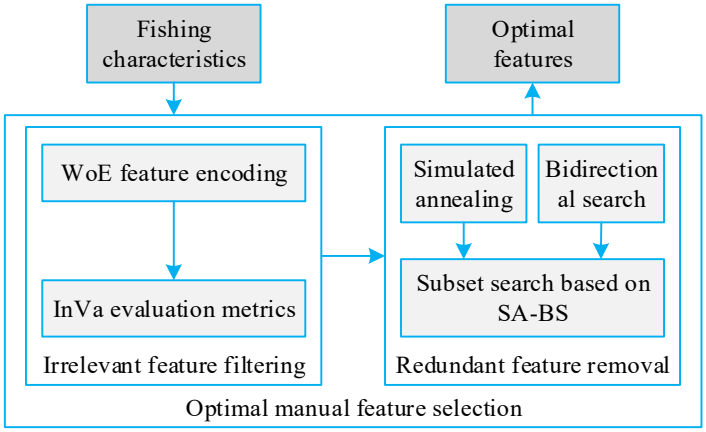

Fig. 3. The flow diagram of our proposed feature selection algorithm

2.5. Optimal feature set selection

When feature extraction is performed based on URL and HTML, it is found that some features are present in both phishing and legitimate websites, which makes these features phishing website detection model useless in training, defining too many extracted features, there may also be some features that have a negative effect. Therefore, in order to train an optimal phishing website detection model, this paper requires feature selection before training classification.

Firstly, Gini index and feature evaluation index (f_Value) based on decision tree CART algorithm are defined to measure the importance of all features in the dataset in phishing detection. Then, based on the feature evaluation metrics (f_Value), the algorithm for constructing the best set of features is designed by the local search method to obtain the best feature vector. This vector will be used as an input to the underlying neural network classifier.

2.5.1. Feature selection based on information gain. Define a new f_Value-index based on the Gini coefficient. As an impurity decomposition method, the Gini coefficient is widely used in decision tree algorithms. In general, the Gini coefficient can be defined as follows.

Suppose the sample set of phishing sites is $D=\{(X_1,y_1),(X_2,y_2),\cdots,(X_N,y_N)\}$, where $X_i\in R^n$, $i=1,2,\cdots,N$, for phishing site detection $y_i\in\{1,-1\}$, $i=1,2,\cdots,N$, X_i are the i th feature vectors, also known as instances, and y_i is the class labeling for X_i. (X_i,y_i) is the sample point. The data set feature attribute $A=\{a_1,a_2,\cdots,a_d\}$, assuming that attribute a has V possible values $\{a^1,a^2,\cdots,a^v\}$ and if the sample data set D is partitioned using attribute a, V branching nodes are created, where the v th branching node contains all the samples in D that take the value of a^v on attribute a and is denoted as D^v. For the sample set D, which has a Gini index that can be expressed by Equation (8), the attribute a Gini index of the partitioned sample set D can be expressed by equation (9).

$$Gini(D)=\sum_{m-1}^{|y|}\sum_{m\neq m}p_m p_m=1-\sum_{m-1}^{|y|}p_m^2 \tag{8}$$

$$Gini_index(D,a)=\sum_{v\text{-}1}^{v}\frac{|D^v|}{|D|}Gini\left(D^v\right) \tag{9}$$

Based on the formula, the Gini index can be calculated for all feature $A=\{a_1,a_2,\cdots,a_d\}$ in the sample set. Then, the feature with the smallest Gini index is found as the dividing attribute. It is shown as follows

$$a_{sclect}=\min_{a\in A}\left\{Gini_index_{(D,a)}\right\} \tag{10}$$

In CART algorithm, the bisection recursion technique is implemented as a result of dividing the current sample space into two disjoint subsets. With this algorithm, a binary tree is generated in which each non-leaf node has two branches. A portion of the decision tree is on the phishing dataset using the CART algorithm. The branch node with the maximum extended Gini coefficient is considered as the root of the whole binary tree. Based on the Gini coefficient, a new f_Value index is defined in the definition.

2.5.2. Definition of characterization indicators. To better evaluate the importance of different features in phishing detection, the new index f_Value of feature a is defined as shown in Equation (11). Where a is the root of the decision tree; $\mathrm{N_a}$ is the number of sample points in the root node a; $Gini\_index_l$ and $Gini\_\_index_r$ are the $Gini$ coefficients of the left and right child levels of the root node a; N_l and N_r are the number of sample points in the corresponding nodes; and $|D|$ specifies the total number of sample points in the sample D space.

$$\begin{aligned} &f\_Value(a)= \\ &\left(N_a*Gini\_index(D,a)-N_l*Ginin_{index_l}-N_r*Gini_{index_r}\right)/|D| \end{aligned} \tag{11}$$

In this paper, f_Value indexing is used as a criterion for selecting the best feature. The larger the f_Value index value of the feature parameter, the more important this feature is in phishing detection. If the f_Value index of a feature is 0, the feature is considered as a useless feature. As defined in the formula, the value of f_Value index is not less than 0, and the value of negative features is also greater than 0. All features have index values basically greater than 0. Therefore, negative features cannot be determined by the index value alone.

2.5.3. Algorithms for Constructing the Optimal Feature Set. In order to effectively differentiate between negative and positive features, this paper sets an inter-value ρ with the help of f_Value index. If the threshold value of a feature is greater than the defined value, the feature is negative and anyway it is a positive feature. Then traverse all the features to get the best set of features. The threshold value is calculated as shown in equation (12):

$$\rho=\frac{\left|Times_{Tem}-Times_{opt}\right|}{acc_{Tem}-acc_{opt}} \tag{12}$$

In this equation, $Times_{opt}$ and acc_{opt} denote the time and accuracy corresponding to the training of the initial feature set, respectively, and $Times_{Tem}$ and acc_{Tem} are the time and accuracy obtained from the training after a feature is added to the initial feature set. The data increase after adding a feature to the initial feature set, then the value of $Times_{Tem}-Times_{opt}$ is greater than 0. Therefore, once the threshold value is less than 0, it means that $acc_{Tem}-acc_{opt}$ is less than 0. It means that the initial feature set makes the accuracy decrease when a feature is added, which means that this feature is a negative feature. On the contrary, it is a positive feature.

In the new algorithm, based on the f_Value index defined in Eq. (11), the CART decision tree algorithm first calculates the index values of the entire collection of features. Then, the set of features with the highest f_Value index value is selected as the initial best feature set. Then, a local search method based on KNN (K nearest neighbor) is used to obtain the accuracy of the initial best feature set. The accuracy of the initial best feature set is considered as the initial solution of the local search algorithm.

In fact, local search algorithms are usually used to find the best solution to an optimization problem. The basic idea of the algorithm is to iterate over adjacent solutions so that the objective function can be progressively optimized until it can no longer be optimized. By sequentially analyzing the remaining collected features, the local search algorithm constructs the final optimal feature vector for the underlying neural network classifier.

In this algorithm, step (1) computes the f_Value index values of all the features in the D sample using the CART algorithm. Step (2) obtains the initial best feature and stores it in the X_{opt} set. By step (3), the remaining features are placed in the X_{rem} set. Step (4) Train the initial best feature using KNN and store the training accuracy to be employed. Step (5) Update the optimal feature set X_{opt} by selecting positively influenced features sequentially from set X_{rem}. In fact, if the newly added feature improves the accuracy of the original X_{opt}, this feature is positive and is put into set X_{opt}. Otherwise, this feature is negative and will not be put into the set. Step (6) Construct the best feature vector based on the selected best feature set. This vector will be used as an input to the base neural network classifier.

3. Fishing detection training experiment results and analysis

3.1. Experimental Validation of CART Phishing Detection System

The following experiments are deployed on Windows operating platform (Windows 7 operating system) with a cpu of 3.3GHz, 16GB of RAM, and Python version 3.8.1.

According to the improved method proposed in this paper, a new neural network phishing detection model based on decision tree and optimal feature selection is constructed - CART phishing detection system. In order to verify the effectiveness of the improvement method of CART classifier for decision tree algorithm, experiments were designed to analyze the evaluation indexes of CART classifier. The experiments still use datasets from UCI and phishtank to test the adaptability of the CART classifier. Since the CART classifier needs to split the training set before performing the classifier training for segmentation enhancement training. Therefore the training set is divided equally into 10 equal parts. The first dataset has 1275 data items and after making equal parts there are 5 sets of 127 sample sets with 5 sets of 128 sample sets. The second dataset has 9880 data items and after aliquoting there are 10 groups of 988 sample sets. The experiment is started after the training set is partitioned.

3.1.1. First set of experimental data. The CART classifier needs to calculate the information gain value of each feature attribute of the dataset, by calculating the information gain distribution of the two datasets. In this case, the information gain distribution of the feature attributes of the first data set is shown in Fig. 4.

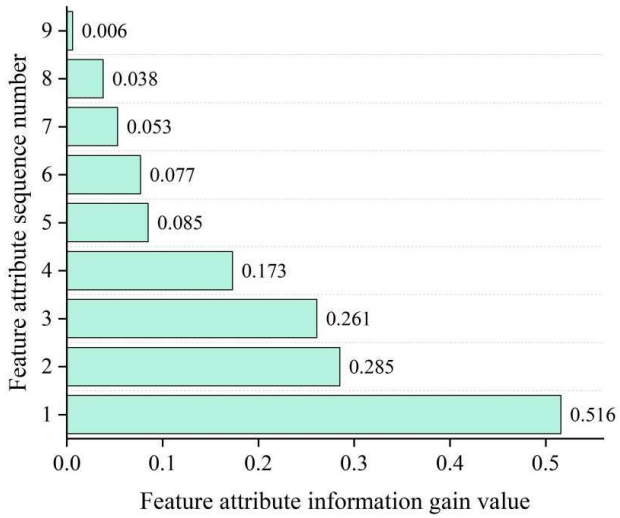

Fig. 4. Information gain of first data sets

According to Fig. 4, for the first data set, the first feature attribute has the greatest impact on the classification results, and the feature attributes are ranked according to their information gain value from the largest to the smallest. The feature selection is performed on the first data set to determine the optimal number of features from the first data set through experiments according to the number of feature selections n from small to large. Since no single attribute can completely divide the dataset, the experiments are performed sequentially with n increasing from 2. The results of the training time consuming experiments for feature selection for the first set of datasets are shown in Fig. 5 and the results of the evaluation of each performance of the classifier are shown in Fig. 6.

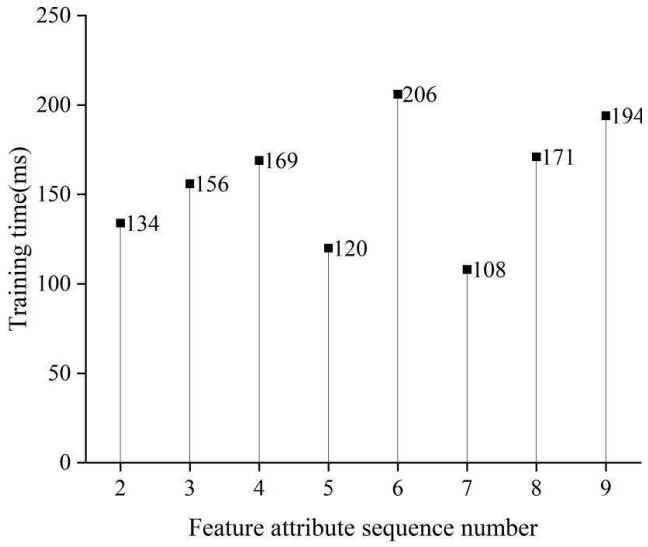

Fig. 5. The first data sets features selection number training time results

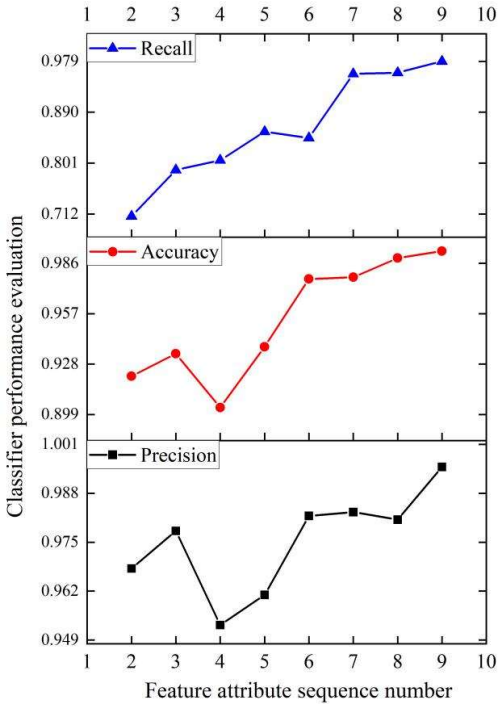

Fig. 6. Performance evaluation results of the classifier on first data sets

From Fig. 5, it is clear that training takes less time when the number of feature selections is 2, 5 and 7. From Fig. 6, it is clear that the CART classifier has higher correctness, precision and recall when the number of feature selections is 7, 8 and 9 than the classifiers using other number of feature selections. However, when the number of feature selections is 8 and 9, the training time is relatively long, while the classifier with the number of feature selections of 7 has only a slightly lower recall compared to the classifier using all feature attributes, while the training time is shorter by 82 ms. Therefore, we believe that this loss in the evaluation of recall is somewhat lower is acceptable, and therefore 7 is chosen as the optimal number of feature selections for the first dataset.

3.1.2. Second set of experimental data. The distribution of feature attribute information gain for the second dataset is shown in Figure 7 below.

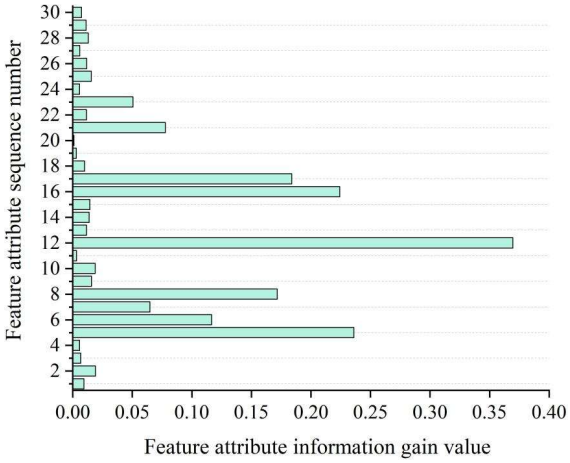

Fig. 7. Information gain of second data sets

As can be seen from Figure 7 for the second dataset, most of the specialty attributes have less information gain due to the high number of features the dataset possesses. However, there are still

individual specialty attributes that show high information gain. For example, the 5th, 8th, 12th, 16th and 18th feature attributes. Therefore, experiments were conducted in increasing order starting from 4 in selecting the optimal number n of feature attributes. The results of the training time-consuming experiments for feature selection and the results of the evaluation of each performance of the classifier for the second dataset are shown in Fig. 8 and Fig. 9, respectively.

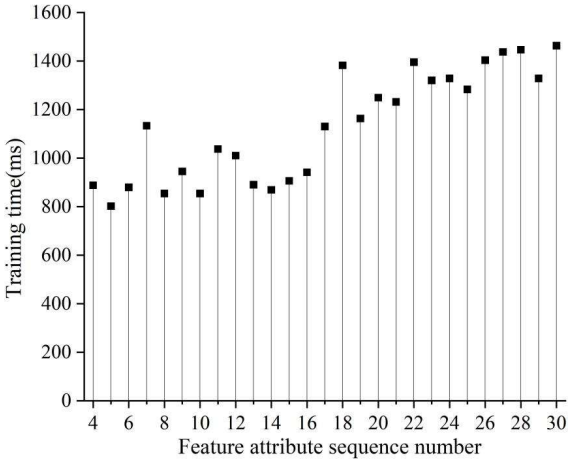

Fig. 8. The second data sets features selection number training time results

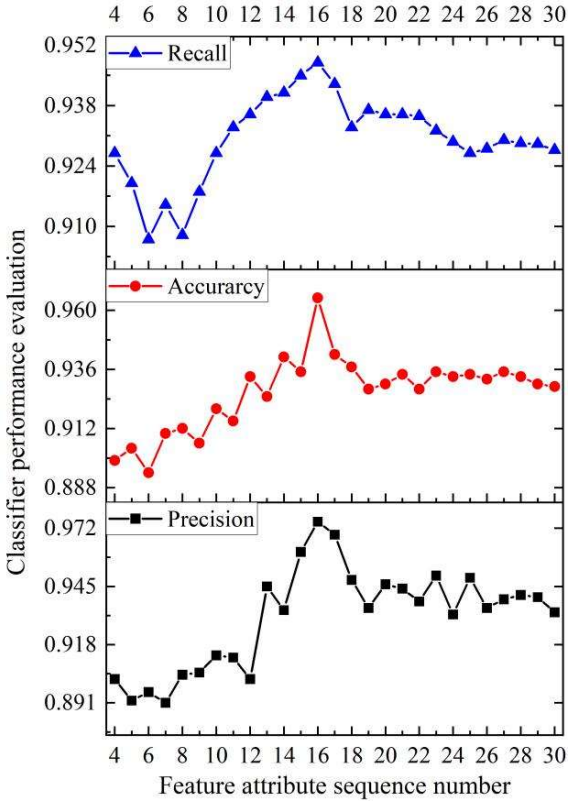

Fig. 9. Performance evaluation results of the classifier on second data sets

As can be seen from Fig. 9, the CART classifier is in an upward trend in the number of feature selections less than 16, and when the number of feature selections is greater than 17, although the correctness rate rises in some of the values, the overall trend is still in a downward trend. Meanwhile, the precision and recall fluctuate greatly when the number of feature selections is less than 16. This indicates that the classification model does not simulate the dataset well when the number of feature

selections is small, resulting in unstable performance of the classifier. As the number of feature selections increases, the dataset contains more information that makes the classification model simulate the dataset more and more well, which is reflected in the increase of all indicators of the classifier. When the number of feature selections is greater than 19, the performance of the classifier becomes more and more stable. At this point, it can be assumed that the classification model has been trained quite well, and there is no need to select more features to be added to the training. The reason is that as the number of selected features increases, the larger the dataset becomes, the amount of computation required for the training of the classification model will also increase, resulting in a longer training time. We can find that the correctness, precision and recall of the classifier are highest when the number of feature selections is 16. In contrast, although the training time of the classifier is shorter when the number of feature selections is 14, which is 72ms shorter, the performance of the classifier decreases more, especially the precision decreases by 4.16%. Therefore, 16 was chosen as the optimal number of feature selections for the second dataset.

3.2. Comparison experiment of different fishing detection models

3.2.1. Performance metrics analysis on the phishtank dataset. This experiment will be proposed in this paper based on the improved decision tree algorithm model and the existing based on the random forest (RF), deep forest (DF), logistic regression (LR), XGBoost, AdaBoost five kinds of models for experimental testing, the different models performance experimental results are shown in Table 1.

Table 1. Experimental results of different model performance on phishtank

Model	Precision	Accuracy	Recall	F1	AUC
CART	0.927	0.965	0.883	0.871	0.910
Random Forest	0.850	0.913	0.835	0.820	0.857
Deep Forest	0.834	0.932	0.849	0.831	0.884
Logistics Regression	0.878	0.879	0.787	0.845	0.847
XGBoost	0.902	0.859	0.819	0.806	0.835
AdaBoost	0.874	0.839	0.736	0.842	0.841

According to Table 1, it can be seen that the model based on the improved decision tree algorithm performs the best in five evaluation indexes: precision, correctness, recall, F1 and AUG, with a precision of 92.7%, a correctness of 96.5%, a recall of 88.3%, with an F1 of 87.1% and an AUG of 91.0%. Analyzed from the perspective of accuracy, the XGBoost model also performs well with an accuracy of over 90%. Analyzed from the point of view of correctness and recall, the models based on random forest and deep forest performed well, with correctness above 90%, while the logistic regression model and the Adaboost model performed poorly, with a recall of only 78.7% and 73.6%.

The ten-fold cross-validation method is used for the evaluation, and the boxplots comparing the working performance of different models are shown in Fig. 10. The ROC curves for each model are shown in Fig. 11.

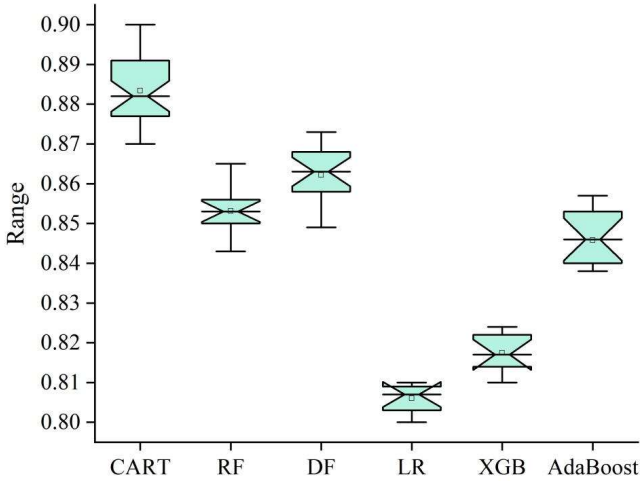

Fig. 10. Comparison box diagram of the performance of each model

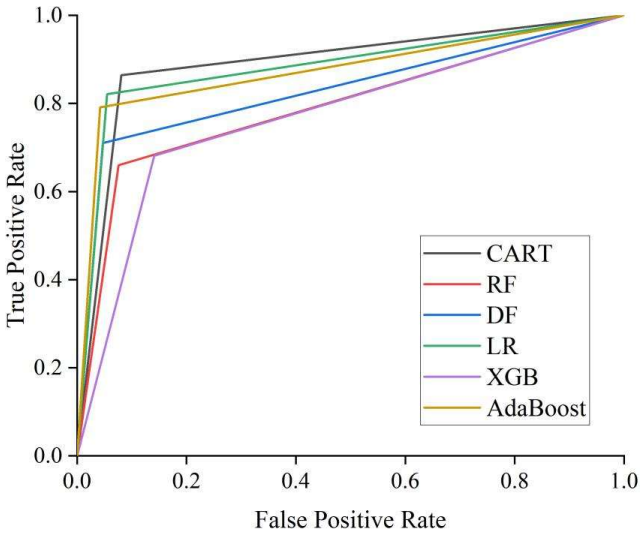

Fig. 11. ROC curves of each model

In the ROC curve graph, the ROC curve of the Logistics Regression model is closer to that of the CART model, but there is still a certain gap.The classification advantage of the CART model is more obvious, and the classification performance is optimal.The Logistics Regression model and the Adaboost model also perform well in classification.

3.2.2. Performance metrics analysis on the Vrbanˇciˇc dataset. The Vrban ˇ ci ˇ c dataset contains two datasets: Vrban ˇ ci ˇ c-small is a balanced dataset, and Vrban ˇ ci ˇ c-full is an unbalanced dataset, and both have larger sample sizes. In order to verify whether the model based on the decision tree algorithm is effective when facing the unbalanced dataset and the network data with larger sample size, the decision tree algorithm is used to process the data and conduct the comparison test on the classifiers.The experimental results of the performance of the different models in the Vrban ˇ ci ˇ c dataset are shown in Table 2.

Table 2. Experimental results of different model performance on Vrban ˇ ci ˇ c

Model	Data set	Precision	Accuracy	Recall	F1	AUC
CART	Vrban ˇ ci ˇ c-small	0.983	0.991	0.995	0.989	0.987
	Vrban ˇ ci ˇ c-full	0.981	0.996	0.983	0.982	0.975
Random Forest	Vrban ˇ ci ˇ c-small	0.947	0.946	0.920	0.952	0.961
	Vrban ˇ ci ˇ c-full	0.926	0.931	0.918	0.917	0.933
Deep Forest	Vrban ˇ ci ˇ c-small	0.921	0.928	0.915	0.966	0.933
	Vrban ˇ ci ˇ c-full	0.952	0.940	0.965	0.946	0.960
Logistics Regression	Vrban ˇ ci ˇ c-small	0.852	0.869	0.847	0.869	0.858
	Vrban ˇ ci ˇ c-full	0.834	0.841	0.826	0.829	0.862
XGBoost	Vrban ˇ ci ˇ c-small	0.923	0.966	0.967	0.943	0.969
	Vrban ˇ ci ˇ c-full	0.925	0.928	0.969	0.938	0.964
AdaBoost	Vrban ˇ ci ˇ c-small	0.923	0.942	0.962	0.964	0.917
	Vrban ˇ ci ˇ c-full	0.969	0.967	0.961	0.933	0.924

Table 2 shows that the classification performance of the decision tree-based model on both Vrban ˇ ci ˇ c-small and Vrban ˇ ci ˇ c-full datasets is excellent. In the Vrban ˇ ci ˇ c-small balanced dataset, the classification effect of CART is optimal on five evaluation indexes, and all indexes are more than 98%; in the Vrban ˇ ci ˇ c-full unbalanced dataset, the classification effect of CART still exceeds other models, and the Accuracy value reaches 99.6%. The random forest model performs well on the Vrban ˇ ci ˇ c-small, but its classification effect decreases by 2.0% on the Vrban ˇ ci ˇ c-full unbalanced dataset; the logistic regression model performs poorly, with less than 90% on both datasets, which may be due to the fact that the Vrban ˇ ci ˇ c dataset has a larger sample size and more complex feature information, and logistic regression cannot adapt to this type of dataset. The reason may be that the Vrban ˇ ci ˇ c dataset has a larger sample size and more complex feature information, and logistic regression cannot adapt to such datasets. The CART model has an obvious advantage in dealing with unbalanced data, because the multi-granularity scanning structure in the improved decision tree algorithm can sample the unbalanced data in the same proportion to ensure the balance of each training set, so as to achieve a good classification effect.

The comparative box line plots of the working performance of different models on two datasets, Vrban ˇ ci ˇ c-small and Vrban ˇ ci ˇ c-full, are shown in Fig. 12 and Fig. 13, respectively.

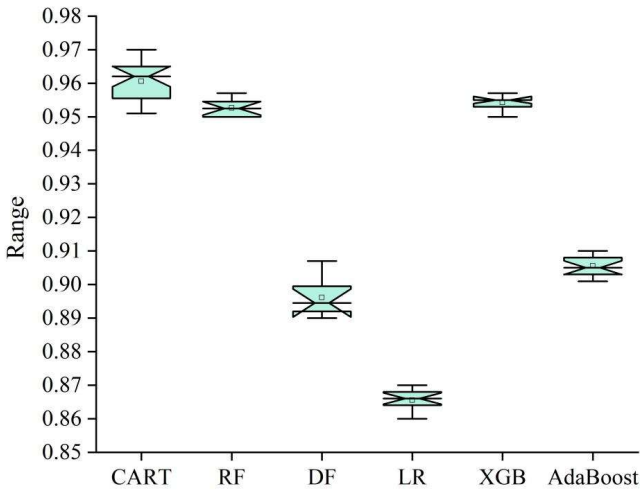

Fig. 12. Box diagram of the performance on Vrban ˇ ci ˇ c-small

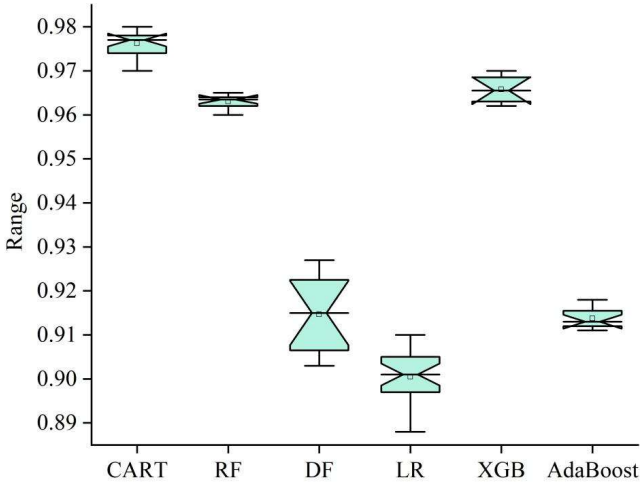

Fig. 13. Box diagram of the performance on Vrban ˇ ci ˇ c-full

It can be seen that the CART model performs the best among the models, with an average working performance of 0.962 and 0.978 on the Vrban ˇ ci ˇ c-small and Vrban ˇ ci ˇ c-full datasets, respectively, which are better than the other model performances.

4. Conclusion

In this paper, a phishing detection model based on decision tree and optimal feature selection is proposed. The adaptability of the CART classifier is tested by testing it on the phishtank dataset. It is also tested on the Vrban ˇ ci ˇ c dataset in a comparative experiment with five existing models based on random forest, deep forest, logistic regression, XGBoost, and AdaBoost.

On the first data set of phishtank with fewer data items, the optimal feature selection number is 7, its training time is 108ms, and the precision, accuracy, and recall are 0.973, 0.968, and 0.943, respectively, with the optimal overall performance in all aspects; in the second data set with 9,980 data items, most of the special attributes have less information gain, and 16 is selected as the optimal feature selection number, at this time the training time consumed is 889ms, the precision, accuracy and recall are the highest among the feature selection numbers, which can simulate the dataset well.

In phishtank, Vrban ˇ ci ˇ c-small and Vrban ˇ ci ˇ c-full three kinds of datasets on CART's phishing detection model performance are optimal, and its average performance is 0.883, 0.962 and 0.978, respectively, which reflects that the phishing detection model based on the decision tree and the best feature selection is more obvious compared to the other methods of modeling and classification. The advantages are more obvious.

Funding

This work was supported by Beijing Troy Cloud Data Technology Co., Ltd.

References

[1] Bhavsar, V., Kadlak, A., & Sharma, S. (2018). Study on phishing attacks. International Journal of Computer Applications, 182(33), 27-29.

[2] Ghazi-Tehrani, A. K., & Pontell, H. N. (2022). Phishing evolves: Analyzing the enduring cybercrime. In The New Technology of Financial Crime (pp. 35-61). Routledge.

[3] Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications, 106, 1-20.

[4] Carroll, F., Adejobi, J. A., & Montasari, R. (2022). How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society. SN Computer science, 3(2), 170.

[5] Abusaimeh, H., & Alshareef, Y. (2021). Detecting the phishing website with the highest accuracy. Tem Journal, 10(2), 947.

[6] Das, K., Gethner, E., Dincelli, E., & Jafarian, J. H. (2021). D3cyt: Deceptive camouflaging for cyber threat detection and deterrence. In Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 1 (pp. 756-771). Springer International Publishing.

[7] Das, A., Baki, S., El Aassal, A., Verma, R., & Dunbar, A. (2019). SoK: a comprehensive reexamination of phishing research from the security perspective. IEEE Communications Surveys & Tutorials, 22(1), 671-708.

[8] Drury, V., Lux, L., & Meyer, U. (2022, August). Dating phish: An analysis of the life cycles of phishing attacks and campaigns. In Proceedings of the 17th International Conference on Availability, Reliability and Security (pp. 1-11).

[9] Cui, Q., Jourdan, G. V., Bochmann, G. V., Couturier, R., & Onut, I. V. (2017, April). Tracking phishing attacks over time. In Proceedings of the 26th International Conference on World Wide Web (pp. 667-676).

[10] Liu, M., Zhang, Y., Liu, B., Li, Z., Duan, H., & Sun, D. (2021, December). Detecting and characterizing SMS spearphishing attacks. In Proceedings of the 37th Annual Computer Security Applications Conference (pp. 930-943).

[11] Khan, A. I., & Unhelkar, B. (2024). An Enhanced Anti-Phishing Technique for Social Media Users: A Multilayer Q-Learning Approach. International Journal of Advanced Computer Science & Applications, 15(1).

[12] Choi, K., Lee, J. L., & Chun, Y. T. (2017). Voice phishing fraud and its modus operandi. Security Journal, 30, 454-466.

[13] Qbeitah, M. A., & Aldwairi, M. (2018, April). Dynamic malware analysis of phishing emails. In 2018 9th International Conference on Information and Communication Systems (ICICS) (pp. 18-24). IEEE.

[14] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.

[15] Bitaab, M., Cho, H., Oest, A., Zhang, P., Sun, Z., Pourmohamad, R., ... & Ahn, G. J. (2020, November). Scam pandemic: How attackers exploit public fear through phishing. In 2020 APWG Symposium on Electronic Crime Research (eCrime) (pp. 1-10). IEEE.

[16] Moody, G. D., Galletta, D. F., & Dunn, B. K. (2017). Which phish get caught? An exploratory study of individuals′ susceptibility to phishing. European Journal of Information Systems, 26(6), 564-584.

[17] Das, S., Nippert-Eng, C., & Camp, L. J. (2022). Evaluating user susceptibility to phishing attacks. Information & Computer Security, 30(1), 1-18.

[18] Lee, J., Lee, Y., Lee, D., Kwon, H., & Shin, D. (2021). Classification of attack types and analysis of attack methods for profiling phishing mail attack groups. IEEE Access, 9, 80866-80872.

[19] Manoharan, S., Katuk, N., Hassan, S., & Ahmad, R. (2022). To click or not to click the link: the factors influencing internet banking users' intention in responding to phishing emails. Information & Computer Security, 30(1), 37-62.

[20] Sharma, P., Dash, B., & Ansari, M. F. (2022). Anti-phishing techniques–a review of Cyber Defense Mechanisms. International Journal of Advanced Research in Computer and Communication Engineering ISO, 3297, 2007.

[21] Ahmed, N. S. S., Acharjya, D. P., & Sanyal, S. (2017). A framework for phishing attack identification using rough set and formal concept analysis. International Journal of Communication Networks and Distributed Systems, 18(2), 186-212.

[22] Bhardwaj, A., Al-Turjman, F., Sapra, V., Kumar, M., & Stephan, T. (2021). Privacy-aware detection framework to mitigate new-age phishing attacks. Computers & Electrical Engineering, 96, 107546.

[23] Al-Hamar, Y., Kolivand, H., Tajdini, M., Saba, T., & Ramachandran, V. (2021). Enterprise Credential Spear-phishing attack detection. Computers & Electrical Engineering, 94, 107363.

[24] Hossain, S., Sarma, D., & Chakma, R. J. (2020). Machine learning-based phishing attack detection. International Journal of Advanced Computer Science and Applications, 11(9).

[25] Pandey, A., Gill, N., Sai Prasad Nadendla, K., & Thaseen, I. S. (2020). Identification of phishing attack in websites using random forest-svm hybrid model. In Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2 (pp. 120-128). Springer International Publishing.

[26] Butt, U. A., Amin, R., Aldabbas, H., Mohan, S., Alouffi, B., & Ahmadian, A. (2023). Cloud-based email phishing attack using machine and deep learning algorithm. Complex & Intelligent Systems, 9(3), 3043-3070.

[27] Assegie, T. A. (2021). K-nearest neighbor based URL identification model for phishing attack detection. Indian Journal of Artificial Intelligence and Neural Networking, 1(2), 18-21.

[28] Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. Electronics, 9(9), 1514.

[29] Azeez, N. A., Salaudeen, B. B., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2020). Identifying phishing attacks in communication networks using URL consistency features. International Journal of Electronic Security and Digital Forensics, 12(2), 200-213.

[30] Ramesh, G., Gupta, J., & Gamya, P. G. (2017). Identification of phishing webpages and its target domains by analyzing the feign relationship. Journal of Information Security and Applications, 35, 75-84.

[31] Jain, A. K., & Gupta, B. B. (2018). Detection of phishing attacks in financial and e-banking websites using link and visual similarity relation. International Journal of Information and Computer Security, 10(4), 398-417.

[32] Shaukat, M. W., Amin, R., Muslam, M. M. A., Alshehri, A. H., & **e, J. (2023). A hybrid approach for alluring ads phishing attack detection using machine learning. Sensors, 23(19), 8070.

[33] Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. International Journal of Computer Applications, 163(8), 15-19.