Phishing Websites Detection Model based on Decision Tree Algorithm and Best Feature Selection Method

Dalia Shihab Ahmed [1], Assist. Prof. Dr. Karim Q. Hussein[2], Hanan Abed Alwally Abed Allah[3]
E-mail:[1]dalia_shihab@uomustansiriyah.edu.iq,[2]karim.q.h@mustansiriyah.edu.iq,
[3] hana.cs88cs@uomustansiriyah.edu.iq
Department of Computer Science, College of Science, Mustansiriya University, Baghdad, Iraq[1,2,3]

Abstract: The ongoing progress of network technology has a huge influence on their broad acceptance in many facets of our life in recent time. Phishing websites have suddenly emerged as a major cybersecurity concern. Phishing websites are counterfeit web pages created by hackers to replicate the web pages of legitimate websites in order to deceive users and steal personal information such as usernames and passwords. Despite the fact that several techniques for identifying phishing websites have been presented, phishers' strategies have evolved to circumvent detection. One of the most efficient approaches for identifying these harmful behaviours is machine learning. This is due to the fact that most phishing attacks exhibit features that machine learning algorithms can recognize. Accurate identification of phishing websites is a tough subject since it is based on various dynamic elements. This study proposes a Decision Tree (DT) classifier with optimal feature selection for phishing website detection, with the goal of improving the classification of phishing websites as phishing or legitimate websites. The experiments were conducted out using the publicly available phishing website dataset from the UCI Machine Learning Repositor, which comprises 4898 phishing websites and 6157 legitimate websites. We extract 30 features from this dataset. In addition, we selected 20 of the most significant features, such as wrapper and correlation-based feature selection. Ten-fold cross-validation was utilized for training, testing, and validation. The best experimental result was obtained by using 20 of the 30 features and submitting them to the classification algorithm. This study obtained 98.80% accuracy the wrapper-based features selection strategy, that is outperformed the DT classifier, with other feature selection method.

Keywords: Phishing websites, Wrapper-based features selection, Correlation-based feature selection, Decision Tree

1. Introduction

In recent years, phishing has been a big source of concern for security experts since it is quite simple to create a counterfeit website that seems to be similar to legitimate website. Although professionals can identify counterfeit websites, not all users can, and as a consequence, some people fall prey to phishing assaults. The attacker's primary purpose is to steal bank account information. Phishing attacks are growing increasingly effective as a result of a lack of user knowledge. It is difficult to combat phishing assaults since they take advantage of user vulnerabilities, yet it is vital to enhance phishing detection systems(Gupta, Singhal, & Kapoor, 2016)(Chawla, 2014).

The "blacklist" approach is a broad strategy for identifying phishing websites that involves updating the antivirus database with blacklisted URLs and Internet Protocol (IP). To circumvent blacklists, attackers use devious strategies such as obfuscation, as well as numerous more fundamental approaches such as fast-flux, in which proxies are automatically constructed to host the web page; algorithmic creation of new URLs, and so on. The method's primary shortcoming is that it cannot identify zero-hour phishing assaults.Heuristic-based detection, which contains features seen in legitimate-world phishing assaults and can identify zero-hour phishing attacks, but the traits are not guaranteed to be present in such attacks all of the time, and the false positive rate in detection is quite high(Khonji, Iraqi, Member, & Jones, 2013).

To overcome the inadequacies of blacklist and heuristic-based strategies, many security professionals are increasingly concentrating on machine learning methodologies. Machine learning is a set of algorithms that uses previous data to predict future data. This approach will be used by the algorithm to analyse a large number of prohibited and URLs and their properties in order to correctly detect phishing websites, including zero-hour phishing websites(Lokesh & Boregowda, 2020).

In the proposed research, the wrapper-based features selection (WFS) and correlation-based feature selection (CFS) techniques were utilized to pick the most essential features to be used in efficiently predicting phishing websites. A decision tree (DT) classifier was employed with these important features selection methodologies to detect phishing websites. As a result, the number of features used in this model has been reduced to just20; this improves classification performance and efficiency, reduces noise caused by the inclusion of many features, and thus improves classification accuracy. The experimental results showed that using wrapper-based features selection improved the performance of the DT classifier and outperformed the DT classifier using other features selection methods.

The remainder of the paper is laid out as follows: Section 2 describes related work on phishing website detection; section 3 presents the Classification Requirements in our paper and describes our dataset, features, and algorithm; section 4 describes the methodology for detecting phishing websites; The experimental results are

described and analysed in section 5, the comparative analysis is presented in section 6, and the conclusion is presented in section7.

1. Related Work

In this part, we will look at some of the most current research projects on identifying phishing websites using machine learning techniques.

In**(Mahajan, 2018)** the authors proposed detection of phishing URLs examined several classifiers such as: Decision Tree, Random Forest and Support Vector Machine (SVM) algorithms are used to detect phishing websites. This model was implemented on the data set consists of total 36,711 URLs which include 17058 legitimate URLs and 19653 phishing URLs. URLs of legitimate websites were collected from www.alexa.com and The URLs of phishing websites were collected from www.phishtank.com. Dataset is divided into training set and testing set in different ratios. Experiments were conducted used extracted features. It was shown that the model provided a performance prediction accuracy of 96.71%, 97.14% and 96.51%, for Decision Tree, Random Forest, and SVM, respectively. The best classifications result was achieved using Random Forest algorithm with lowest false negative.In**(Kulkarni, Brown, Kulkarni, & Brown, 2019)** authors suggested a phishing website detection model that makes use of a number of classification techniques, including a decision tree, a Naive Bayes' classifier, SVM, and a Neural Network. This model was applied to a data set containing nine features from The University of California, Irvine Machine Learning Repository. Among such datasets it includes elements from 1353 URLs. 548 are legitimate, 702 are phishing attempts, and 103 are suspect. The data collection also includes nine features taken from each URL. Experiments were carried out for each classifier. The best classification result was obtained using DT, which had a classification accuracy of 90.39 %.In**(Shahrivari, 2020)** The authors suggested a model to categorize websites as phishing or legitimate by using a variety of classification methods, including Logistic Regression, Decision Tree, Support Vector Machine, Ada Boost, Random Forest, Neural Networks, KNN, Gradient Boosting, and XGBoost. This model was applied to a dataset of phishing websites obtained from the UCI Machine Learning Repository, which contains 6157 websites and 4898 phishing websites. Experiments with 30 features were carried out, and ten-fold cross-validation was employed for training, validation, and testing. For Decision Tree, Random Forest, and XGBoost, the model exhibited performance prediction accuracy of 96.59 %, 97.26 %, and 98.32 %, respectively. As a result, we obtained very good performance in ensemble classifiers such as Random Forest and XGBoost in terms of computation duration and accuracy.In**(Gadge, 2017)** The authors presented a technique for identifying phishing URL websites. This approach examines the sites and calculates heuristic values. Using the C4.5 decision tree approach, these features were utilized to determine if the site was phishing or not. Data from PhishTank and Google were used to create the dataset. This approach has two stages: pre-processing and detection. In the pre-processing step, features are collected using rules, and the features and their associated values are fed into the C4.5 algorithm, which obtained an accuracy of 89.40 %.

3.Classification Requirements

3.1 Data Collection

The first step in building the proposed phishing website detection model is choose an appropriate training dataset which consisting of both phishing and legitimate websites that are used to support and test the proposed system to evaluate its performance.

In this research, we tested the efficiency of the proposed phishing website detection method using a publicly accessible phishing website dataset from the UCI Machine Learning Repository (“Phishtank,”2016.). This dataset contains 4898 phishing websites and 6157 legitimate websites from which many website features were extracted. The phishing websites collection was mostly sourced from the Phishtank and MillerSmiles archives.Table (1) presents the key details of the phishing website dataset used in the tests and assessment.

Table (1): The description of the dataset of phishing websites utilized in the experiments

Attributes	**eulav**
Number of features (attributes)	30
Number of websites (instance)	11055
Number of phishing websites	4898
Percentage of phishing websites	44%
Number of legitimate websites	6157
Percentage of legitimate websites	56%

3.2 Features

Choosing the most suitable features in the experiment would provide a better outcome. Features have become an essential aspect of undertaking phishing website detection study. The following are some of the features of our dataset:

1. Having an IP Address: If an IP address is used in the URL instead of the domain name, such as http://217.102.24.235/sample.html.
2. Length of URL: Phishers may conceal the suspicious element of the URL in the address bar by using a lengthy URL.
3. URL Shortening Service: Provides access to a website with a lengthy URL. The URL http://sharif.hud.ac.uk/, for example, may be abbreviated to bit.ly/1sSEGTB.
4. Using the @ sambol:sign in the URL causes the browser to disregard anything before the @ symbol, and the true address often follows the @ symbol.
5. Double Slash Redirection: The presence of / in a URL indicates that the user will be redirected to another website.
6. Prefix Suffix: Phishers often add prefixes or suffixes separated by (-) to domain names in order to give visitors the impression that they are dealing with a reputable website. For instance, see http://www.Confirme-paypal.com.
7. Using a Subdomain: Using a subdomain in the URL.
8. SSL Status: Indicates whether or not a website employs SSL.
9. Domain Registration Length: Because a phishing website only exists for a brief time,
10. Favicon: A favicon is a visual image (icon) that is connected with a particular website. If the favicon is loaded from a domain different than the one displayed in the address bar, the site is most certainly a Phishing attempt.
11. Using Non-Standard Ports: It is much preferable to just open the ports that you need to regulate invasions. Several firewalls, proxy servers, and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only allow access to those that are explicitly allowed.
12. HTTPS token: Using a deceptive https token in the URL. For instance, http://www.mellat-phish.ir
13. Request URL: The request URL checks to see whether the external items on a site, like as photos, videos, and music, are loaded from another domain.
14 Anchor URL: An anchor is an element specified by the a > tag. This feature is processed in the same way as Request URL.
15. Links in Tags: websites often utilize Meta tags to provide information about the HTML content, Script tags to generate a client-side script, and Link tags to fetch external online resources.If the domain name in SFHs differs from the domain name of the site.
17. Submitting Information Via E-mail: A phisher may reroute the user's information to his email address.
18. Incorrect URL: It is derived from the WHOIS database. Identity is usually included in the URL of a legitimate website.
19. Website Redirect Count: If the redirection occurs more than four times, it is considered excessive.
20. Status Bar Customization: Utilize JavaScript to display a bogus URL to users in the status bar.
21. Disabling Right Click: This is the same as using on Mouse Over to conceal the Link.
22. Using Pop-up Window: Demonstrating the presence of pop-up windows on the website.
23. IFrame: An IFrame is an HTML element that displays an extra website inside the one that is now shown.
24. Domain Age: If the domain is less than a month old.
25. DNS Record: Possessing a DNS record
26. Web Traffic: The number of visits to a website is used to determine its popularity.
27. Page Rank: Page rank is a number that ranges from 0 to 1. PageRank attempts to determine the importance of a site on the Internet.
28. Google Index: This function determines whether or not a website is included in Google's index.
29. Number of Links Pointing To Page: The number of links that point to the web page.
30. Statistical Report: Determine whether or not the IP address belongs to the top phishing IP addresses.

3.3 Feature Selection

The selection of discriminating features in machine learning may be useful for lowering dimensionality, eliminating unnecessary data, enhancing learning accuracy, and improving result comprehensibility. As a result, the classifier's performance is improved.

Wrapper-based Feature Selection (WSF)

The importance of the features subset is evaluated using an inductive classifier in wrapper-based techniques. To reduce duplicate and unnecessary features, the inductive classifier is trained individually with numerous subsets. The classification error rate of the classifier model is then used to calculate the score for each subgroup. A search method is utilized in the wrapper-based evaluation to search over the space of potential features and assess each subset by running a model on the subset. Because each subset trains a new classifier, wrapper-based assessment approaches are often computationally costly for big datasets. Wrapper-based algorithms, on the other hand, often give the most influential features set and obtain the highest performance for that specific kind of classifier(Chandrashekar & Sahin, 2014)(Kohavi & John, 2015).

Correlation-based Feature Selection (CFS)

CFS uses a correlation-based heuristic evaluation function to rank feature subsets. The evaluation function is biased in favour of subsets with features that are substantially associated with the class but uncorrelated with one another. Irrelevant features should be discarded since their correlation with the class will be poor. Because they will be substantially associated with one or more of the remaining traits, redundant features should be filtered out. The degree to which a feature predicts classes in portions of the instance space not previously predicted by other features will determine its acceptability(Hall, 1999).

3.4 Machine Learning Classification

To distinguish phishing websites from legitimate websites, a variety of classification approaches are utilized. To learn classifiers, termed training datasets, a collection of features is employed, and then the output is predicted. In this case, the techniques learn the features of phishing and legitimate websites to categorize them as phishing or legitimate. The suggested approach in this study employs the DT classification algorithm, which is described as follows:

Decision Tree (DT)

A decision tree is a decision-making aid that employs a tree-like graph or model of options and their potential outcomes, such as chance event outcomes, resource costs, and utility. It's one approach to show an algorithm made up entirely of conditional control statements.Decision trees are frequently used prediction models in domains such as machine learning, statistics, and data mining. A decision tree is a supervised learning approach that is non-parametric. A tree structure, also known as a predictive model, is built by inferring rules from data points made up of feature vectors. Leaf nodes reflect projected classes, whereas inside nodes represent "decisions." The majority of algorithms construct a decision tree from the top down. An algorithm begins at the root node and makes the "best" choice recursively, after which the remaining instances are divided into child nodes until there are no more features to select from or the node is pure.(Survey, Rokach, & Maimon, 2005).Table (2) provides the parameters that the decision tree employed for the experimental investigation.

Table 2: Experiment Parameters for Decision Tree

Parameter	**value**
Confidence threshold for pruning	0.25
Minimum number of instances per leaf	2
Number of Instances	1105
Number of Leaves	169
Size of the tree	297
Number of cross-validation folds	10-fold

4. Methodology Detection of the Phishing Website

The purpose of the study is to distinguish phishing websites from legitimate websites so that the receiver is not harmed by the phishing website.

The suggested research starts by looking at the accuracy of detecting phishing websites using a set of features. Then, by selecting the best features, you may save time and space while extracting and using these features. The performance of the DT classification algorithm is then evaluated using the Wrapper Features Selection (WFS) and Correlation-based Feature Selection (CFS) techniques on the collection of extracted features and chosen features. Finally, the performance of the classifier is compared using two feature selection strategies to decide which is the best.Figure (1) shows the phishing websites detection architecture based on machine learning approach.

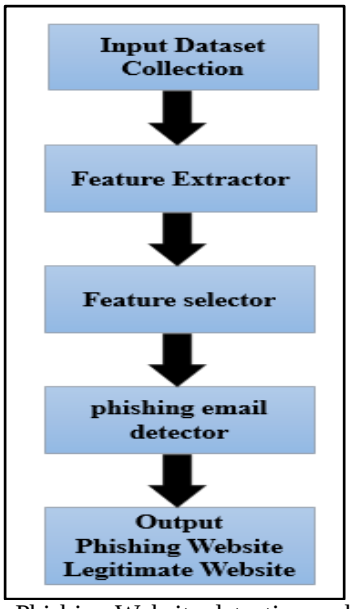

Figure 1: Phishing Website detection architecture.

The phishing website detection architecture is made up of five components that serve as a task assembly. The following are the functions of each module:

Input Dataset Collection: This module is in charge of acquiring the phishing and legitimate website datasets obtained from the UCI Machine Learning Repository ("UCI Machine Learning Repository: Phishing Websites Data Set.," 2016). This dataset contains 4898 phishing websites and 6157 legitimate websites from which many website features were extracted.

Feature Extractor: To identify phishing websites from legitimate ones, many features may be gleaned from a website. The quality of the retrieved features is critical to the performance of phishing website detection techniques.
The UCI Machine Learning Repository ("UCI Machine Learning Repository: Phishing Websites Data Set.," 2016) has a dataset of phishing websites that contains 30 essential features of websites that have been shown in (Mccluskey & Mccluskey, 2012) to be effective and influential in predicting phishing and legitimate websites. The following table covers the important features that may aid in the successful prediction of phishing websites.Table (3) presents the main features that can contribute in the effective prediction of phishing sites.

Table 3: The main features that can contribute to the effective prediction of phishing sites

Feature Group	Features Names
Address features	Using the IP Address, Long URL to Hide the Suspicious Part, Using URL Shortening Services "TinyURL", URL's having "@" Symbol, Redirecting using "//", Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domains, HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer), Domain Registration Length, Favicon, Using Non-Standard Port, and The Existence of "HTTPS" Token in the Domain Part of the URL
Abnormal features	URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Server Form Handler (SFH), Submitting Information to Email and Abnormal URL
HTML and JavaScript features	Website Forwarding, Status Bar Customization, Disabling Right Click, Using Pop-up Window, and IFrame Redirection,
Domain features	Age of Domain, DNS Record, Website Traffic, Page Rank, Google Index, Number of Links Pointing to

Feature Selector: It is the process of determining which features are meaningful from the retrieved features. Certain features are more significant than others, since some of the other features have little or no influence. As a result, attribute selection is crucial in our machine learning architecture.
The feature set is selected using two feature selection methods: Wrapper Features Selection (WFS) and Correlation-based Feature Selection (CFS).

As previously stated, wrapper-based features selection typically results in the highest performing features set for that specific kind of classifier. As a result, the wrapper-based features selection method is employed in this study to choose the most significant features that may be used to identify phishing from legitimate websites. In the wrapper-based features selection, the machine learning classifier is considered the main part used to evaluate the goodness of all the selected features subsets. The wrapper technique does a search in space for all feasible features subsets and uses a machine learning classifier to evaluate the features subsets. The optimal features subset is chosen based on the highest score to be utilized in the machine learning classifier's training. As a result, the highest-scoring 20 elements for detecting phishing websites were selected.

The most important elements are chosen using the Correlation-based Feature Selection (CFS) method to distinguish between phishing and legitimate websites. CFS calculates the value of a subset of features by weighing the predictive potential of each item individually as well as the degree of redundancy between them. The core component of the CFS algorithm, as defined in Equation (4), is a heuristic for assessing the utility or quality of a subset of features. Individual features, as well as their degree of intercorrelation, are useful in predicting the class label in this heuristic.

$$Merit\ s = \frac{\mathrm{k}\overline{rcf}}{\sqrt{k+k(k-1)\overline{rff}}} \qquad (1)$$

where, Merits is the heuristic merit of an attribute subset S with k features, rcf denotes the average attribute class correlation, and rff denotes the average attribute-attribute correlation.The goal of the heuristic is to get rid of any unneeded or duplicated features that are inefficient class predictors.

Phishing Email Detector: The suggested classification method, DT, is applied to the collection of features in this module. DT method will use 30 features derived from the data set to determine if a website's synchronization is a phish or not.

In this paper, we use 10-fold cross-validation to train and assess our classifier. For the 10-fold cross-validation process, divide the data set into 10 parts; 9 of the 10 parts are used to train the classifier, and the information gained from the training phase is used to validate (or test) the 10th part; this is repeated 10 times, with each part serving as both training and test data at the end of the training and testing phase. The accuracy of each run is calculated. As a result, the ultimate accuracy of learning from this dataset is the average of the n accuracies for all runs. The use of cross-validation ensures that the training and test data are both varied. In machine learning, the cross-validation method is well-known for providing a very accurate estimate of a classifier's generalization error.

Output: Based on the selection of features and classification techniques employed, this module generates results. The outcome is generated using phishing website detection accuracy, which is used to identify the unclassified website as either legitimate = 1 or phishing = -1.

5. Experiment

The results of the performance assessment of this effort to tackle the phishing issue are presented in this part.

5.1. Evaluation Metrics

To show the performance of our suggested technique, we employed four generally used evaluation measures, namely Precision, Recall, F-measure and Accuracy, which are defined as follows[16]:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad (2)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (3)$$

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

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{TP+FP+TN+FN} \qquad (5)$$

Table (1) shows the simplified definitions for True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).

Table 1. Descriptions for Parameters FP, FN, TN, and TP

Parameter	Description
True Positive (TP)	The number of phishing websites that were correctly identified.
False Negative (FN)	The number of phishing websites that were mistaken for legitimate websites.
False Positive (FP)	The number of non-phishing websites that were mistakenly identified as phishing websites.
True Negative (TN)	The number of benign websites discovered as benign websites.

5.2 Experimental Results

We evaluated the performance of the suggested technique in this part by comparing the information generated from the experiments. (30) features were used to train and evaluate the suggested method.

The performance of machine learning classifiers with wrapper-based features selection in phishing website detection was evaluated using ten-fold cross validation in this research. Their results were also compared to those of another prominent feature selection approach, Correlation-based Feature Selection (CFS).

The performance metrics outlined in the preceding section were used to assess the findings. Initially, all features (30 features) are subjected to Experiment (1). Experiment (2) is carried out on features chosen using the wrapper-based features selection approach. Finally, the features chosen using the Correlation-based Feature Selection (CFS) approach were used in Experiment (3).

It is clear from the results of Experiment (2) that the DT classifier with the wrapper-based features selection which is outperformed the DT classifier with Correlation-based Feature Selection (CFS) method in Experiment (3) and results of all features in Experiment (3). Table (5) shows the classification results of DT algorithm with three experiments. The classification results of the DT method with three experiments are shown in Table (5).

Table 5: Results of DT with two feature selections

	Precision	**Recall**	**F-Measure**	**Accuracy**
Experiment 1	95.46	96.09	95.77	95.76%
Experiment 2	98.04	97.83	97.93	**98.80%**
Experiment 3	96.98	97.61	97.29	97.28%

The DT classifier with wrapper-based features selection, in particular, had the greatest accuracy. This is because the wrapper-based features selection uses a machine learning classifier as an evaluation function to assess the quality of all chosen feature subsets.

When the wrapper-based features selection was utilized, the suggested method achieved the maximum accuracy score of 98.80 %, proving the capacity of the proposed technique to identify legitimate from phishing websites. Furthermore, the suggested technique received a 98.80 % accuracy score, which shows the %age of correctly classified websites to all websites.

As we shall explain in the following part, the findings of our model attain excellent accuracy rates for categorizing phishing websites and surpass comparable suggested classification techniques.

6. Comparative Analysis

In this section, we compare our proposed model to a number of previously suggested phishing detection methods. Table (6) lists four past studies on the topic, as well as the classification methods used and the accuracy of the classification findings.

Table 6: Comparison of our approach with previous work

Paper Reference	**Classification Algorithms**	**Accuracy**
[5]	Decision Tree, random forest, SVM	97.14%
[6]	decision tree, Naïve Bayes' classifier, Support Vector Machine (SVM), and the Neural Network	90.39%
[7]	Logistic Regression, Decision Tree, Support Vector Machine, Ada Boost, Random Forest, Neural Networks, KNN, Gradient Boosting, and XGBoost	96.59%
[8]	c4.5 decision tree	89.40%
Our Approach	DT	**98.80%**

7. Conclusion

The continuing advancement of network technology has aided in the widespread adoption of e-commerce, electronic banking, social media, e-health, and e-learning in many facets of our life. With financial institutions continuing to incur huge financial losses and phishing websites becoming more difficult to recognize, it is vital to create more effective strategies for detecting them.

The accuracy of the phishing website detection model was tested using the DT classifier algorithm on all features and chosen features in this study. Finally, a comparison of the two scenarios of chosen features was done (wrapper and Correlation-based Feature Selection methods). We utilized a 10-fold cross-validation strategy to train and evaluate this model to prevent overfitting. According to the findings, the choice of efficient features has an impact on the accuracy of the work of phishing website detection. As a result, when we employed the DT classifier based on the wrapper features selection approach, we got the maximum accuracy of 98.80%.

Despite the fact that the wrapper-based features selection technique takes longer and requires more computing overhead with the classifier, it is often employed just once to offer the most significant features. In order to increase the effectiveness and flexibility of the phishing website detection systems, the classifier should be retrained with these chosen features on a regular basis throughout the update process.

ACKNOWLEDGMENT

Thanks to the Department of Computer Science at Mustansiriyah University for their assistance with this project, which authors gratefully acknowledge.

secnerfeR

Authors, F. (2016). A hybrid firefly and support vector machine classifier for phishing email detection, (0368-492X). https://doi.org/10.1108/K-07-2014-0129

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, M. (2014). A Survey of Phishing Attack Techniques. *International Journal of Computer Applications*, *93*(0975 – 8887). https://doi.org/10.5120/16197-5460

Gadge, L. M. J. (2017). Phishing sites detection based on C4.5. *In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India*, 1–5.

Gupta, S., Singhal, A., & Kapoor, A. (2016). A Literature Survey on Social Engineering Attacks : Phishing Attack, 537–540.

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning, (April).

Khonji, M., Iraqi, Y., Member, S., & Jones, A. (2013). Phishing Detection : A Literature Survey, *15*(4), 2091–2121.

Kohavi, R., & John, G. H. (2015). Wrappers for Feature Subset Selection, *97*(0004-3702), 273–324.

Kulkarni, A. D., Brown, L. L., Kulkarni, A., & Brown, L. L. (2019). Phishing Websites Detection using Machine Learning.

Lokesh, G. H., & Boregowda, G. (2020). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, *5*(2374-2917), 1–14. https://doi.org/10.1080/23742917.2020.1813396

Mahajan, R. (2018). Phishing Website Detection using Machine Learning Algorithms, (October), 14–17. https://doi.org/10.5120/ijca2018918026

Mccluskey, T. L., & Mccluskey, L. (2012). An assessment of features related to phishing websites using an automated technique. *Conference Proceedings*, (January), 492–497.

Shahrivari, V. (2020). Phishing Detection Using Machine Learning Techniques.

Sonowal, G. (2020). Phishing Email Detection Based on Binary Search Feature Selection. *SN Computer Science*, *1*(2662-995X), 1–14. https://doi.org/10.1007/s42979-020-00194-z

Survey, C. A., Rokach, L., & Maimon, O. (2005). Top-Down Induction of Decision Trees, *35*(4), 476–487.

UCI Machine Learning Repository: Phishing Websites Data Set. RetrievedMay 9, 2016, from https://archive.ics.uci.edu/ml/datasets/Phishing+Websites