A comprehensive Review on phishing website detection using machine learning and deep learning techniques.

Hajara, Musa[1], A. Y. Gital[2], Usman Ali A[3], A. M. Kwami[4]

[1] Department of Computer Science, Gombe State University, Gombe, Nigeria
[2,4]Department of Mathematical Sciences, Abubakar Tafawa Balewa University Bauchi, Nigeria
[3]Federal college of education Gombe, Nigeria

Abstract-*Nowadays there is extensive growth in the number of internet users, a lots of our daily life operations move to the cyber world such as communication, commerce, banking, registrations, applications, etc. and the criminal peoples (phishers) also move to cyber world and make their threats and crimes easily. To ensure the security and privacy of cyber data, technology must be used and organized carefully by using Cyber Security. This research focus on the reviewing the state of art techniques for phishing detection using different models in traditional and deep learning algorithms. It's also identified the solution for detecting phishing website, comparative analysis of detecting using different approach. This paper point out the approach proposed by different researchers. It then provides a discussion on the limitations of the techniques.*

Index term: machine leaning algorithms, deep learning algorithms, nature inspired algorithms, hybrid deep learning algorithms.

1.0 INTRODUCTION:

As most of human activities are being moved to cyberspace, phishers and other cybercriminals are making the cyberspace unsafe by causing serious risks to users and businesses as well as threatening global security and economy [30]. To ensure the security and privacy of cyber data, technology must be used and organized carefully by using Cyber Security.

Cyber security is the body of technologies about processes, networks, computers programs and data. Its aims to designed and protect these components of technologies from attack, damage and unauthorized access. According to [60] cyber security is the organization and collection of resources, processes and structures used to defend cyberspace and cyberspace enabled systems from events that misrelate by default ownership rights [60].

Phishing is a cyber-crime which involves the fraudulent act of illegally capturing private information like credit card details, usernames, password, account information by pretending to be authentic and esteemed in instant messaging, email and various other communication channels. The traditional approaches used by majority of the email filters for identifying these emails are static which make it weak to deal with latest developing patterns of phishing since the defrauders are dynamic in actions and keep on modifying their activities to dodge any kind of detection[38].

Phishing presents a diversified development trend, which poses new detection challenges. While phishers are pernicious and hide, security experts and researchers have dedicated many efforts in terms of phishing website detection. Phishing is a very popular method used in network attacks and leads to privacy leaks, identity theft and property damage[40]. The spread of phishing is no longer limited to traditional modalities such as e-mail, SMS, and pop-ups. Though the prosperity of the mobile Internet and social networks have brought convenience to users, they have also been employed to spread phishing, such as code phishing, spear phishing and spoof mobile applications [57], [58] and [59] etc.

Reducing the risk pose by phishers and other cybercriminals in the cyber space requires a robust and automatic means of detecting phishing websites, since the culprits are constantly coming up with new techniques of achieving their goals almost on daily basis. Phishers are constantly evolving the methods they used for luring user to revealing their sensitive information. The Main aim of the attacker is to steal banks account credentials. However, due to the dynamic nature of attackers and the challenging nature of the problem, it still lacks a complete solution [32]. In this paper, several different models are compared using machine learning, deep learning model and nature inspired algorithms.

1.1 Types of phishing attacks

I. *Pharming*

Pharming is the term given to hosts file modification or Domain Name System based phishing, hackers tamper with a company's host's files or domain name system so that requests for URLs or name service return a false address and thereby communications are directed to a forged site. The outcome: users are oblivious that the website where they are entering secret information is controlled by phisher and is probably not even in the same country as the justifiable website [35].

II. *Content-Injection Phishing*

It describes the situation where hackers replace part of the content of a legitimate site with false content designed to mislead/misdirect the user into giving up their confidential information to the hacker. For example, phisher may insert malicious code to log user's credentials or an overlay which can secretly collect information and deliver it to the phisher [64-65]

III. *Deceptive phishing attacks*
This attack refers to social engineering attacks where users receive messages or e-mails which redirect them to bogus websites (fake websites) with the aim to steal personal information such as bank account number, social insurance and security numbers, account user name and pass-words just to name some few [64]. Example hacker send messages alerting a problem to be solved rapidly while proposing to follow a link for solutions in cell phone.

IV. *Malware-based phishing attack*
Malware code is installed on users' PC when users' tries to open a malicious file attached to an email or download a file from a malicious website. This code could have the aim to compromised and join the corresponding PC on which it is installed to a botnet and the phisher as the botmaster will be able to conduct a general DoS, meaning allow attacker to access the device and its connection in order to steal data through victim machine) [66]

V. *Hosts File Poisoning*
When a user types a URL to visit a website it must first be translated into an IP address before it is transmitted over the Internet. The majority of SMB (small and medium business organizations) users' PCs running a operating system look up these "host names" in their "hosts" file before undertaking a Domain Name System (DNS) lookup. By "poisoning" the hosts file, hackers have a bogus address transmitted, taking the user unwillingly to a fake website where their information can be stolen phisher [64-65]

VI. *Man-in-the-Middle Phishing*
The attacker positions themselves between the user and the legitimate website or system. They record the information being entered but continue to pass it on so that users' transactions are not affected. Later they can sell or use the information or credentials collected when the user is not active on the system [64-65]

VII. *Web Trojans*
This attack pop-up invisibly when users are attempting to log in. They collect the user's credentials locally and transmit them to the phisher [64-65]

1.2 Detection Approaches:

1. URL-Based Approach uses only features extracted from a given URL to detect phishing. The URL protocol combine with other feature also helps in phishing website identification process as described by [61] and [62]. The URL-Based approach is used for detecting deceptive phishing websites as well deceptive emails.
2. Content-Based Approachfocuses on features extracted from a website HTML code or the content of an e-mail. Some URL features are still useful in the content-base approach when dealing with links extracted from the HTML code or email content. Web pages containing more external links than internal ones and password field input are classified as suspicious. This mean website content with more external links than internal links is an attempt to achieve some similarities and styles from external with the objective to steal user credential [63]. Another feature for content-Based approach is the website tag <form> that can help to confirm a web page is phishing. This tag is a means by which user's information could be leaked to phishers. Hence, in case an email contains a URL that leads to a website page containing the tag <form>, this page as well as the email is considered to be suspicious.
3. Combination of URL-Based and Content-Based Approach produce more efficient because it uses features selected from URLs and some URLs are well crafted and could not be quickly detected but their corresponding web pages contents which help to extract features that will help to classify them as phishing. Hence, though some research have succeeded to get good accuracy by using either each of these approaches, we strongly believe by combining the two approaches will lead to an efficient set of features that will help to get a high detection accuracy and efficiency [63].

2.0 proposed review on traditional and deep learning algorithms:
In order to review the approaches written by different researchers. A good number of recent research papers related to Phishing website detection from 2015 to 2022 are summarized

In 2015**,** [47] proposed an anti-phishing systemwhich is based on the development of the Add-on tool for the web browser. The performance of the proposed system is studied with four different data mining classification algorithms which are Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC). They have collected 7690 legitimate websites and 2280 phishing websites from the authorised sources like APWG database and PhishTank. The Bayesian classification is more accurate and showing fast response to the system. Also[37] Proposed a hybrid model to classify phishing emails using machine learning /algorithms with the aspiration of developing an ensemble model (Bayesian net ensemble with CART) for email classification with improved accuracy. The processed emails are provided as input to various machine learning classifiers. They have used the content of emails and extracted 47 features from it. It is observed and inferred that Bayesian net classification model when ensemble with CART gives highest accuracy of 99.32%. But the approach creates over-complex trees that do not generalize the data well is called overfitting.

In 2016**,** [35]they proposed a model for phishing website detection and preventing using modified SVM-PSO method. For feature extraction SVM classifier is used which is able to extract more feature (13 features) than the existing system (10 features) and for optimizing the feature set PSO (particular swarm optimization) is used. This develops acceptably classified phishing websites and legitimate website. The experimental results comparison among hierarchical clustering and SVM-PSO generates more value for

precision and recall parameter, the work outperforms than the existing system. Also[34] they proposed heuristic-based phishing detection technique that employs URL-based features. The system first extracts the features which clearly differentiate that whether website are benign or legitimate. Then they apply these features to machine learning techniques which identify website that are phish or legitimate. They used 10 URLs-based features with no specific number of dataset. The experiment shows that SVM has accuracy of 96% and very low false-positive rate. The proposed model can reduce damage caused by phishing attacks because it can detect new and temporary phishing sites. Additionally [36] they proposed the paper that compared different features assessment techniques in the website,thedatasets used 30 features which has been conducted using three known features selection methods. Experimental results on real phishing have been able to identify new clusters of features that when used together are able to detect phishing activities. Further, important correlations among common features have been derived.The problem of this approach can be hard to find a usable formal representation and it deals badly with quantitative measurements. More also [43] they presented the paper for a wide scope and fast phishing detection system. The models are constructed using both phishing and legitimate URLs including the features which have been extracted. Three classifiers are implemented through using WEKA (Waikato Environment for Knowledge Analysis). Classifiers should be trained using balanced datasets in order to get higher performance. They divide those URLs into three datasets. The proposed system can be integrated into such process in order to increase the detection performance in a real time. The datasets is collected from the Phishtank and OpenPhish, 46, 5461 URLs was collected from Phishtank, 4647 URLs have been collected from OpenPhish. To cover the diversity of benign websites, they are randomly collected 10,275 URLs from dmoz.org and 10,275 URLs from webcrawler.com. The best results are achieved by J48 classifier with accuracy of 93%. In addition[44] they proposed methodology uses the evolving spiking neural network which is very much adaptive if any changes happen in the input data then it easily learn. The network is very much flexible to adapt new changes in the environment. The method has two layers input layer and the output layer. The output layer is evolving in behaviour. The performance of the proposed eSNN (evolving spiking neural classifier) architecture are very much depends on the parameter tuning. The proposed method has around seven parameters need to tune to get the better results. These parameters are very much influence to the network performance. 200 phishing websites are collected from PhishTank (www.phishtank.com). In which 50% are phishing website's URLs and the rest are legitimate website's URLs. The eSNN perform better than the PNN with 92.5% and 89.5% respectively. The parameters selection and tuning are the major challenges of this network. Lastly in 2016 [50]they work to compared the fuzzy based anti-phishing system (Neuro Fuzzy based model) with other snit phishing system. The system has five neuron layers based feed forward network. The work was able to successfully design a simple and efficient fuzzy based anti-phishing website detection.They proposed an efficient non algorithmic anti- phishing system. They used UCI machine data to test inference system and found satisfactory results. Fuzzy logic has been used to successfully perform the task of phishing detection and categorization system with 96% accuracy.

In 2017 [55]theypresented a machine learning based novel anti-phishing approach that extracts the features from client side only. They have examined the various attributes of the phishing and legitimate websites in depth and identified nineteen outstanding features to distinguish phishing websites from legitimate ones. These nineteen features are extracted from the URL and source code of the website and do not depend on any third party, which makes the proposed approach that is random forest has fast, reliable, and intelligent. Compared to other methods (SVM, LR, NB and NN). They used 19 features set. They proposed approach has relatively high accuracy in detection of phishing websites as it achieved 99.39% true positive rate and 99.09% of overall detection accuracy. The approach has the problem of overfitting. And in [46] theyproposed the development of a Chrome Extension for identifying phishing websites. To counter this they used machine learning in trained the tool (RF, SVM and KNN) and categorize the new content it sees every time into the particular categories so that corresponding action can be taken. The dataset is obtained from UCI Machine Learning Repository which composed of 11055 entries of websites. It's classified as phishing and benign. These entries each have 30 features of the website used. The best result of 96.12% is obtained when the RF algorithm was used. But Random forests have been observed to overfit for some datasets with noisy classification tasks. Also Ali (2017**).** The author used a wrapper features selection method to detect phishing websites. There are 30 features that are recognized as key features and they are grouped in address bar-based features, abnormal based features, HTML and Javascript based features and domain-based features collected from UCI machine learning. As features selection method, author used wrapper features selection method which finds the best set of features for given machine learning classifier. Classifiers as Naive Bayes, Support Vector Machine, C4.5, k Nearest Neighbour and Random Forest are tested before and after features selection. Highest true value rates are achieved by Random Forest with 97.3% accuracy. More also [42] the authors combined NBTree, C4.5 and Random Forest to build an effective classifier for network intrusion detection. Random Tree outperformed other individual algorithms with accuracy 88.46%. NSL-KDD dataset with 41 features was used. Random Forest and NBTree achieved highest accuracy 89.24% when applied together. When a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data (overfitting). Additionally [45]the paper introduced a Case-Based Reasoning (CBR) Phishing Detection System (CBR-PDS). It mainly depends on CBR methodology as a core part. The proposed system is highly adaptive and dynamic as it can easily adapt to detect new phishing attacks with a relatively small data set in contrast to other classifiers that need to be heavily trained in advance. Experiments show that CBR-PDS system is proposed to predict phishing attacks with a high accuracy. The data set is collected from the PhishTank3 website. Experiments show that the CBR-PDS system accuracy exceeds to 95.62%.Similarly[48] they developed a framework, called "Fresh-Phish", for creating current machine learning data for phishing websites. Using 30 different website features that they query using python, they build a large labelled dataset and analysed SVM with a Gaussian kernel and an SVM with a linear kernel against the Fresh-Phish. The SVM against this dataset to determine which is the most accurate. They analysed not just the accuracy of the technique, but also how long it takes to train the model. They used there framework and gathered information of 6000 legitimate and 6000 phishing websites. And they created Fresh-Phish dataset subsequently and trained the classifier over this dataset. The results achieved an accuracy as high as 90% for the Fresh-Phish test data using a Gaussian kernel

SVM. In addition [51] developed automatic classification of a web-site into a phishing or non-phishing one based on aggregation of a set of predetermined features related to the content of the site. A classifier is developed based on Ant-Colony optimization, known as cAnt-MinerPB. The dataset contains a total of 11055 instances with extracted from UCI Repository and normalized features. The dataset is not divided into training and testing set. Thus, in the experiments, part of the dataset was used for training and another part was used for testing. cAnt-MinerPB has shown promising results compared to the well-known and well-established classification techniques. Furthermore, the results have shown that studied the aggregation levels is worth to be considered as many other pre-processing stages on different data mining applications and tasks.In contribution also [38] they developed the prediction of Ensemble Classifier of the five ML Algorithms Gaussian Naive Bayes, Bernoulli Naïve Bayes, Random Forest Classifier, K-Nearest Neighbours, and Support Vector Machines. The results show that for different Feature Groups based on the decisive values of the features, the algorithms that returned best accuracy is Random Forest with 96.07% accuracy. Random forests have been observed to overfit for some datasets with noisy classification tasks. The evaluation of model size is slow. Lastly [28] this paper presented a novel approach for detecting phishing websites based on probabilistic neural networks (PNNs). They also investigate the integration of PNN with K-medoids clustering to significantly reduce complexity without jeopardizing the detection accuracy. To assess the feasibility of the proposed approach, they conducted in depth study to evaluate various performance measures on a publicly available data set composed of 11 055 phishing and benign websites. The experimental results show that 96.79% accuracy is achieved with low false errors. This approach requires large memory spaces to store; the execution of network of this approach is slow.

In 2018[52]they discussed three approaches for detecting phishing websites. First is by analysing various features of URL , second is by checking legitimacy of website by knowing where the website is being hosted and who are managing it, the third approach uses visual appearance based analysis for checking genuineness of website. They make use of Machine Learning techniques and algorithms for evaluation of these different features of URL and websites. The algorithms that returned the best accuracy is Random Forest with 96.58%. In this case the model does better on training set than the test set, then we are likely overfitting. And [53]proposeda system which developed to detect URLs which are used in Phishing Attacks. In the proposed system some features have been taken out by using NLP techniques. The features are extracted and evaluated in two different groups. The first one is a person determined attribute that is thought to be distinctive to malicious URLs and legal URLs. The second group focuses on the usage of the words in the URL without performing any other operations by applying only the vectorization process. Experimental study is constructed over three different test scenarios, including tests for NLP-based features, tests for Word Vectors, and Hybrid approach tests for both of these features. During the tests, RF which is tree based algorithm, SMO which is a kernel based algorithm, and NB is a statistical based algorithm is used. According to the results obtained, the tests made for the hybrid approach were more successful than the other tests and experimental results showed that Random Forest algorithm has a very good performance. In the collected data set, there are 73,575 URLs including 37,175 malicious URLs and 36,400 legal URLs. 40 features are used for Testing NLP Features, 238 features are used for Testing Word Vectors and 278 features are used for Hybrid Tests. The Random Forest Algorithm was observed to be more successful than the other algorithms tested with 97.2% success rate. Also [41]they proposed a learning-based aggregation analysis mechanism to decide page layout similarity, which is used to detect phishing pages. Firstly they checked and filtered those invalid pages manually. After that they excluded those pages whose layout elements are too small and whose layout appearance is totally different from their target. They selected 13 target pages and 102 suspicious pages to test their approach. The experiment results shows that the approach is accurate and effective in detecting phishing pages. They integrate and analyse a few of potential learning algorithms. Support Vector Machine (SVM) is a widely used classification algorithm due to its good performance. The basic idea of SVM is to maximize the margin between two classes' closest points and find an optimal separating hyperplane between them. Decision Tree (DT) learning is one of the predictive modelling algorithms. It takes a decision tree as the predictive model and determines an item's target value (represented as a leaf) according to the observations about the item (represented in branches). They collected phishing websites from phishtank.com and alexa.com. According the experiment results, two classifiers both have more than 93% accuracy and more than 95% precision, which demonstrates that our approach can make an effective detection in phishing websites, but it does not perform well when the data set has more noise (overlapping) and also when large data set used because the required training time is higher. More also [39] they proposed the system which identify whether an URL is either phishing or legitimate. In order to compare the efficiency of the different algorithms, they used both Artificial Neural Network (ANN) and Deep Neural Network (DNN) approaches for training and testing the system with the help of Tensorflow framework. And experimental results showed that the proposed approaches produce very good accuracy rates for detecting phishing URLs. For training the system and catch abnormal request by analysing the URL ofweb pages.In order to train the system they have used a dataset which contains about 74,000 items (the dataset contains 37,175 phishing and 36,400 legitimate web pages to train the system in both these type. The proposed approaches DNN gives better accuracy rate with 96% than ANN. But DNN is extremely expensive to train due to the complex data model. In addition [54]theyproposed a novel approach using deep neural network (DNN) algorithms to more accurately detect malware and phishing web certificates. Using these algorithms, they improved the feature engineering process by allowing the model to automatically uncover the hidden patterns in the malicious web certificates. This new proposed algorithm is able to leverage detection by more effectively analysing text data, in addition to the other features they included. Using a deep learning model, they were able to outperform other results. In case of malware. The high success rate of the classification in both cases demonstrates the strength of the proposed model.They used a database of 1,000,000 legitimate, 5,000 phishing and 3,000 malware certificates obtained by crawling the internet. The results show that system is capable of identifying malware certificates with an accuracy of 94.87% and phishing certificates with an accuracy of 88.64%. More also [33]The aimed of the research is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm. The algorithms used are Random forest algorithms, Decision tree and SVM. URLs of benign websites were collected from www.alexa.com which include 17058 and The URLs of phishing

websites were collected from www.phishtank.com which include 19653. The data set consists of 36,711 benign and phishing. Result shows that Random forest algorithm gives better detection accuracy of 97.14% with lowest false negative rate than decision tree and support vector machine algorithms. More over [56]they employed two different approaches to measure each feature. The initial approach, used a binary definition for features to create the dataset. They assigned a value based on if it was believed to be legitimate (+1), or phishing (-1). For non-binary features like (Age of Domain or Links Pointing to Page) they used a threshold to convert it to a binary value. Secondly they created another dataset with original values of 59 feature without using thresholds to convert non-binary values to binary ones. For example, they use an integer length for the URL in the model. They used 2500 clean websites and 2500 phishing websites and created our Fresh-Phish dataset subsequently they trained our classifier over this dataset. They achieved an accuracy of 93% on binary dataset and 96% on non-binary dataset which is acceptable rate. Also, it shows that the idea of using non-binary values helps to improve the accuracy. Butit has several key parameters that need to be set correctly to achieve the best classification results for any given problems. Additionally [32] they proposed a new phishing detection model based on Extreme Gradient Boosted Tree (XGBOOST) algorithm. Experimental results demonstrated that XGBOOST-based phishing detection model is promising by returning an accuracy of 97.27% which outperformed both probabilistic Neural Network (PNN) and Random forest (RF) that returned accuracies of 96.79% and 95.66% respectively. The datasets is collected from UCI Repository which contains 30 features with 11055 datasets. It is clearly showed that the predictive performance of phishing website detection model using XGBOOST algorithm is optimized to 97.29%. also [19] The research covers the development of phishing website model based on different algorithms with different set of six feature categories (address bar based feature, abnormal based feature, domain based feature, feature selection, HTML and javascript based features, full dataset) in order to investigate the most significant features in the dataset and test the effect of the dataset size, feature selection is important because dataset may contain irrelevant noisy and redundancy feature in which if they are included (incorporated), it will surely affect the model negatively. The results demonstrated that using full dataset is better because it generate and returned high accuracy performance which indicate that the combination of all features is important. The datasets is collected from UCI Repository which contains 30 features with 11055 datasets. The result obtained after applying feature selection method utilizing six (6) categories of subset features. The results shows that the collection of address bar based XGBOOST attained >91% accuracy while that of PNN is >87% accuracy. Using feature selection with nine (9) subsets features, the performance of XGBOOST achieved >94% accuracy while that of PNN returned 92% accuracy. But incase of HTML and javascript based feature both XGBOOST and PNN has very poor performance results with 57.46% and 56.99% respectively. But using full dataset is better because it generate and returned high accuracy performance which indicate that the combination of all features is important.

In 2019[29]they developed the methods of defence utilizing various approaches to categorize websites. Specifically, they have developed a system that uses machine learning techniques to classify websites based on their URL. They used four classifiers: the decision tree, Naïve Bayesian classifier, support vector machine (SVM), and neural network. In detecting phishing URLs, there are two steps. The dataset is collected from the University of California, Irvine Machine Learning Repository. It contains 9 features from 1353 URLs. The results of the experiments show that the classifiers were successful in distinguishing real websites from fake ones with the highest accuracy of 91.5%. And[19]they focuses on design and development of a deep learning based phishing detection solution that leverages the Universal Resource Locator and website content such as images and frame elements. A Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) algorithm were used to build a classification model. This type of algorithm has a high probability of detecting newly generated phishing URLs and, moreover, does not need manual feature engineering. The dataset consist of one million URLs taken from PhishTank and a legitimate site from Crawl as well as over 10,000 images from legitimate and phishing websites. The experimental results showed that the proposed model achieved an accuracy rate of 93.28%.The challenges while training the model using CNN and LSTM are Overfitting, exploding and class imbalance. Also [40]they proposed a multidimensional feature phishing detection approach based on a fast detection method by using deep learning (MFPD).The approach reduced the detection time for setting a threshold. Testing on a dataset containing millions of phishing URLs and legitimate URLs. In the first step, character sequence features of the given URL are extracted and used for quick classification by deep learning, and this step does not required third-party assistance or any prior knowledge about phishing. In the second step, they combine URL statistical features, webpage code features, webpage text features and the quick classification result of deep learning into multidimensional features. A dynamic category decision algorithm (DCDA) is proposed by revising the output judgment conditions of the softmax classifier in the deep learning process and setting a threshold, the detection time can be reduced. They build a real dataset by crawling a total of 1 021 758 phishing URLs as positive samples from phishtank.com, and a total of 989 021 legitimate URLs as negative samples from dmoztools.net.After conducted a series of experiments on a dataset containing millions of phishing and legitimate URLs. From the MFPD approach achieves the highest precision and F1 of 99.41 and 990 respectively.

In 2020[27]in this research, deep learning algorithm (DNN) used to check the phishing attack website based on the behaviour of the website. From the model, it has been clearly observed that the algorithm can predefined segments the URL based on the training and testing samples and various process like forward and backward approach of the algorithm different features are extracted. They used 30 features for the experiment. Experimental results shows DNN has the highest accuracy of 94.3%. DNN outperformed other model CNN, RNN and NB. And [22] in this work, they address the problem of phishing websites classification. Three classifiers were used with the feature selection methods from Weka. They obtained the results by feature selection and machine learning methods. Several feature selection methods were applied, and their results were compared to find the attributes with highest impact to the result. The classification algorithms, such as KNN, Decision Tree and Random Forest were applied to initial and reduced dataset in which RF the achieved accuracy of 100%. From the total number of samples there are 1 185 non-fraudulent, while 10 030 of them are categorized as phishing. RF the achieved accuracy of 100%. Also [26] they designed an ensemble machine learning-

based detection system called PhishHaven to identify AI-generated as well as human-crafted phishing URLs. To the best of their knowledge is the first study to consider detecting phishing attacks by both AI and human attackers. PhishHaven employs lexical analysis for feature extraction. To speed up the ensemble-based machine learning models, PhishHaven employs a multi-threading approach to execute the classification in parallel, leading to real-time detection. During the experiments, they analyze their solution with a benchmark dataset of 100,000 phishing and normal URLs. Theoretically analysed that the solution can detect tiny URLs as well as future AI-generated Phishing URLs based on their selected lexical features with 100% accuracy. More also [24] They use the model with different Machine Learning Algorithms, namely Logistic Regression, Decision Trees, K-Nearest Neighbours and Random Forests, and compare the results to find the most efficient machine learning framework. K-Nearest Neighbours gave an accuracy of 93.7% in detecting phishing web pages. Based on our results, we would recommend using the KNN algorithm to identify phishing websites. The accuracy is depends on the quality of data, when the data large the prediction stage might be slow and also is require high memory because it need to store all the training data. In addition [25] They proposed a fast deep learning-based solution model, which uses character-level convolutional neural network (CNN) for phishing detection based on the URL of the website, is proposed. The proposed model does not require the retrieval of target website content or the use of any third-party services. It captures information and sequential patterns of URL strings without requiring a prior knowledge about phishing, and then uses the sequential pattern features for fast classification of the actual URL. For evaluations, comparisons are provided between different traditional machine learning models and deep learning models using various feature sets such as hand-crafted, character embedding, character level TF-IDF, and character level count vectors features. They collected URLs from different sources (Alexa, openphish, spamhaus.org, techhelplist.com, isc.sans.edu and phishtank) of 318,642 datasets. The proposed model achieved an accuracy of 95.02% on our dataset and an accuracy of 98.58%, 95.46%, and 95.22% on benchmark datasets when compared others research work. More over [12]this study proposed 3 meta-learner models based on Forest Penalizing Attributes (ForestPA) algorithm. ForestPA uses a weight assignment and weight increment strategy to build highly effcient decision trees by exploiting the prowess of all attributes (non-class inclusive) in a given dataset This indicates that the ForestPA algorithm is effectively detects website types with very high accuracy with a bias to the majority class and with very little false alarm rate. Further, with the superiority of the proposed models over other existing methods. The datasets is collected from UCI Repository which contains 30 features with 11055 datasets. From the experimental results, the proposed meta-learners (ForestPA-PWDM, Bagged-ForestPA-PWDM, and Adab-ForestPA-PWDM) are highly effcient with the accuracy of 96.26%, 96.58% and 97.40% respectively. Furthermore [17] the proposed study is an endeavour toward the detection of phishing by using random forest and binary long short-term memory (BLSTM) classifiers. The proposed study are promising in phishing detection, and the study reflects the applicability of the proposed algorithms in the information security. This high recognition rate for the BLSTM-based model reflects the applicability of the proposed model for phishing detection. This dataset has 30 different keywords and 2456 varying instances. The experimental results show that the BLSTM-based phishing detection model is prominent in ensuring the network security by generating a recognition rate of 95.47% compared to the conventional RF-based model that generates a recognition rate of 87.53%.Also[18] the research presented a data-driven framework for detecting phishing webpages using deep learning approach. More precisely, a multilayer perceptron, which is also referred as a feed-forward neural network is used to predict the phishing webpages. They used MLPClassifier function from sklearn.neural_network. The number of default hidden layer was 100. The default hidden layer's number was applied throughout the whole experiment. The default iteration number was 200. They increased the number of iterations by 1000, which increased the training and test accuracy. The default alpha parameter was 0.0001, which was applied for 200 and 1000 iterations. The dataset was collected from Kaggle and it contains 10 features with 1700 datasets. The proposed model has achieved 95% training accuracy and 93% test accuracy. [23] Also proposed two models and compared. The first model is neural network without using PCA (Principal Component Analysis), and the second model is neural network using PCA. The first and the second model will be compared based on accuracy and computational time. This study uses back-propagation algorithm based on neural network method and PCA based on feature selection to reduce large attributes into small attributes. This paper compares neural network model without using PCA and neural network using PCA. This paper uses dataset from UCI Machine Learning Repository. The dataset has 11055 training set and 31 attributes. The result shows that neural network using PCA has better accuracy in 55.67% and neural network without using PCA only reaches 54.43% accuracy. However neural network without using PCA has faster computing time than neural network using PCA. This study can be used as a phishing protection technique. Lastly [31]the authors focus on studying various features employed in different phishing attacks. So many studies have been conducted on single feature to have high accuracy for attack detection while others advanced on the use of many features to detect different attack behaviours with high accuracy. Researchers have advanced the study to the adoption and standardization of thirty (30) features to be examined in phishing attack in order to achieve high accuracy of detection. They examined all the features used so far and used XGBOOST classification model to categories the features into different kinds to detect important features. The analysis revealed that some features hampers on the accuracy and are unfruitful which also contributes in slowing the whole detection process. The model helps us to select useful features and weeds out the useless features. This yields higher accuracy and less time in detection process. The dataset is extracted from UCI Machine Learning Repository. The dataset has 11055 training set and 31 attributes. Total phishing website is 4898 and total legitimate website is 6157. This dataset has been proven effective in predicting phishing website. The experiment show the higher accuracy from using only the few selected features with accuracy of 97.41% while a full features gives the accuracy of 97.29%

In 2021[14]they proposed an effective phishing website detection approach, which can call HinPhish. HinPhish extracts various link relationships from webpages and uses domains and resource objects to construct a heterogeneous information network. HinPhish applies a modified algorithm (random forest) to leverage the characteristics of different link types in order to calculate the phish-score of the target domain on the webpage. Moreover, HinPhish not only improves the accuracy of detection, but also can increase the phishing cost for attackers. Extensive experimental results demonstrate that HinPhish can achieve an accuracy of 98.56%.and[20] the research surveys the features used for detection and detection techniques using machine learning. Phishing is

popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents. They have tested two machine learning algorithms on the Phishing Websites Dataset and reviewed their results. They selected the best algorithm based on it is performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. The dataset consists of different features that are to be taken into consideration while determining a website URL as legitimate or phishing. They have collected unstructured data of URLs from Phishtank website, Kaggle website and Alexa website, etc they have detected phishing websites using Random Forest algorithm with an accuracy of 97.31%. More also [15] they proposed system that could provide significant benefits to current traditional techniques against phishing attacks. The proposed system has two subsystems. Firstly: collection data corresponds to normal and phishing websites. Secondly: distributed framework to apply different neural networks with different machines cores (logistic regression and linear regression). The result of the proposed system achieved high accuracy and detection rate. This application can be extended with real time detection. Data has been collected as unstructured data of URLs from Phishtank website, and Alexa website for the normal Data. PhishTank is an online platform dedicated to fighting phishing. The results shows that LR (logistic regression) has 99.97% accuracy. Furthermore [16] theydeveloped an efficient phishing website detection plugin service using machine learning technique (random forest) based on the prevalent phishing threat while using existing web browsers in critical online transactions. The proposed model was also modelled using use-case and sequence diagrams to test its internal functionalities. The dataset consisting of 11,000 data points with 30 features downloaded from phishtank. The result revealed that the proposed model had an accuracy of 96%. In addition [8]This paper discussed the machine learning and deep learning algorithms and apply all these algorithms (DT, LR, KNN, XGBoost, Ada Boost, And RF) on their dataset and the best algorithm having the best precision and accuracy is selected for the phishing website detection. This work can provide more effective defences for phishing attacks of the future. The dataset comprises 95911 rows and 12 columns of phishing and legitimate website data. Decision Tree model provides the best and highest accuracy with 95.50%. Moreover [9]they proposed a novel Convolutional Neural Network (CNN) with self-attention named self-attention CNN for phishing Uniform Resource Locators (URLs) identification. Specifically, self-attention CNN first leverages Generative Adversarial Network (GAN) to generate phishing URLs so as to balance the datasets of legitimate and phishing URLs. Then it utilizes CNN and multi-head self-attention to construct new classifier which is comprised of four blocks, namely the input block, the attention block, the feature block and the output block. Finally, the trained classifier can give a high-accuracy result for an unknown website URL. They collected 68030 legitimate URLs from 5000 Best Websites and 12003 phishing URLs from PhishTank. The experiments indicate that self-attention CNN achieves 95.6% accuracy, which outperforms CNN-LSTM, single CNN and single LSTM by 1.4%, 4.6% and 2.1% respectively.Similarly[11]in this research, an improved spotted hyena optimization algorithm (ISHO algorithm) is proposed to select proper features for classifying phishing websites through support vector machine. The proposed ISHO algorithm outperformed the standard spotted hyena optimization algorithm with better accuracy. The proposed ISHO algorithm is implemented in MATLAB environment and its accuracy is compared with SHO algorithm. Then the ISHO and SHO algorithms are used for feature selection and their results are compared. The proposed algorithm is also compared with a number of classification algorithms (SVM-SHO, RSVM, LSVM, KNN and Novel NN) proposed before on the same dataset. They used 30 features with 11055 datasets. The result of the proposed model achieved 98.64% accuracy for detection.Additionally [3] In this proposed work, an improved version of Binary Bat namely Swarm Intelligence Binary Bat Algorithm is used for designing the neural network which categorized the network URL websites similar to classification approach. It is utilized for the initial moment in this domain of relevance to the preeminent of understanding. The number of samples collected for phishing websites detection is 11055. The phishing websites consist of 6157 samples and the legitimate websites is 4898 samples. the experimental results shows that deep learning based Adam optimizer reaches high classification accuracy as 94.8% in phishing websites attack detection based on SI-BBA.In contribution again [21]They proposed a model which used three detection models that are combined with each other, namely (decision tree, random forest and support vector machine), to investigate the problem of phishing on sites in addition to using the forms separately for the purpose of comparison with the proposed model. The proposed method enhanced the site security as anti-phishing technology. The phishing detection used three classification algorithms, which are the decision tree; the supporting vector machine and the random forest were combined into one system that was proposed in this paper for the purpose of obtaining the highest accuracy in detecting phishing sites. They used 30 features. The results of the proposed algorithm showed 98.52% higher accuracy than others. In addition [49] in this paper, they proposed a feature-based phishing detection technique that uses uniform resource locator (URL) features. This paper focuses on the extracting the features and then classified based on their effect within a website. The feature groups include address- bar related features, abnormal- based features, HTML – JavaScript based features and domain based features. They used machine learning and implemented some classification algorithms using random forest, naïve Bayes, decision tree, and support vector machine. They compared the performance of these algorithms on their own dataset. The dataset consists of 11,055 entries with 6157 phishing instances and 4898 legitimate instances. Each instance consists of 30 features comprising of various attributes typically associated with phishing or suspicious webpages. It showed a high accuracy of 95.89%.Also[12]this paper proposed an optimized stacking ensemble method for phishing website detection. The optimisation was carried out using a genetic algorithm (GA) to tune the parameters of several ensemble machine learning methods, including random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM. The optimized classifiers were then ranked, and the best three models were chosen as base classifiers of a stacking ensemble method. The experiments were conducted on three phishing website datasets that consisted of both phishing websites and legitimate websites. To conduct the experiment, the Dataset 1 consist of 30 features, 11055 datasets extracted from UCI. Dataset 2 includes 48 features extracted from 5000 phishing websites and 5000 legitimate websites, while Dataset 3 includes 111 features extracted from 30,647 phishing websites and 58,000 legitimate websites. They obtained accuracy reached 97.16%.More also [7]The author combines the key points of phishing website detection based on decision tree and optimal feature selection to study such as URL feature and HTML feature analysis, website

application feature analysis, K-Medoids cluster analysis, and feature set screening. The author uses simulation experiments to complete the website performance check. The purpose of this article is to optimize the performance of phishing website detection and improve the security of the website's operating environment. The sample data of the normal website comes from the basic data collected in the Common Crawl system. The phishing sample data comes from the commonly used sample data included in the Phish Tank system. The accuracy of the new detection system can reach 97.3%.Lastly[10]this paper introduced a possible solution to avoid such attacks by checking whether the provided URLs are phishing URLs or legitimate URLs. They provided 2000 phishing and 2000 legitimate URL dataset. They consider the Random Forest Algorithm due to its performance and accuracy. They consider 9 features, 2000 phishing and 2000 legitimate URL dataset. This approach follows the Random Forest algorithm where the accuracy is 86% for the proposed system.

In 2022[6]they presented a method for evaluating phishing detection models in adversarial situations by adversarial sampling attacks. This adversarial nature makes standard evaluations less useful in predicting model performance in such adversarial situations. They found some limitations such as the exclusion of domain modifications and non-applicability for models that utilize the URL directly. All the studied models did not perform well in the evaluation. This may be because the attacker was unrestricted in the proposed threat model, as the attacker had unlimited access to the prediction function. To address these limitations, they proposed a more restricted adversarial scenario where the attacker has limited access to the prediction function. To evaluate this adversarial scenario, they presented a parameterized text-based mutation strategy used for generating adversarial samples. These parameters tune the attacker's restrictions. They proposed phishing detection solution based on Convolutional Neural Network (CNN) model, they referred as PUCNN model. They focused on text based mutation due to their focus on URL-exclusive models. The PUCNN model generally showed robustness and performed well, and also when the parameters were low is indicated a more restricted attacker. PUCNN had the best results of 95.99% when the mutations were exclusive to the domain. It can be seen from the results that PUCNN performed well when the mutations number and generation number are small. And [5]this paper proposes an effectual Hybrid Deep Learning (HDL)-centric Phishing Detection System (PDS) using the Modified crow search-based deep learning neural network (MCS-DNN) classifier. The datasets used comprise 30 URL features, which are taken as of various legitimate as well as phishing URLs gathered from the publicly available University of Huddersfield website. The experiment outcomes evinced that the proposed methods rendered a better accuracy level on considering the existing techniques, and it attained 96.54% accuracy. Also [4] they proposed the multidimensional phishing susceptibility prediction model (MPSPM) to implement the prediction of user phishing susceptibility. They constructed two types of emails: legitimate emails and phishing emails. They gathered 1105 volunteers to join their experiment by recruiting volunteers. They send these emails to volunteers and collected their demographic, personality, knowledge experience, security behaviour, and cognitive processes by means of a questionnaire. They then applied 7 supervised learning methods (GBDT, SVM, LR, DT, RF, XGBoost and AdaBoost) to classify these volunteers into two categories using multidimensional features: susceptible and nonsusceptible. The experimental results indicated that some machine learning methods have high accuracy in predicting user phishing susceptibility, with a maximum accuracy rate of 89.04%. In terms of accuracy, GBDT correctly predicted 89.04%.More also[2] they proposed a new approach to solve the anti-phishing problem. The new features of their approach is represented by URL character sequence without phishing prior knowledge, various hyperlink information, and textual content of the webpage, which are combined and fed to train the XGBoostclassifer. One of the major contributions of their work is the selection of different new features, which are capable enough to detect 0-h attacks, and these features do not depend on any third party services. Particularly they extracted character level Term Frequency-Inverse Document Frequency (TF-IDF) features from noisy parts of HTML and plaintext of the given webpage. Theycreated their own data set with 60,252 webpages to validate the proposed solution. This data contains 32,972 benign webpages and 27,280 phishing webpages. For evaluations, the performance of each category of the proposed feature set is evaluated, and various classification algorithms are employed. From the empirical results, it was observed that the proposed individual features are valuable for phishing detection. However, the integration of all the features improved the detection of phishing sites with significant accuracy. The proposed approach achieved an accuracy of 96.76% with only 1.39% false positive rate on our dataset, and an accuracy of 98.48% with 2.09% false-positive rate on benchmark dataset, which outperformed the existing baseline approaches. Lastly in [1] this paper proposed a malicious URL detection method based on a bidirectional gated recurrent unit (BiGRU) and attention mechanism. The method is based on the BiGRU model. A regularization operation called a dropout mechanism is added to the input layer to prevent the model from overfitting, and an attention mechanism is added to the middle layer to strengthen the feature learning of URLs. Finally, the deep learning network DA-BiGRU model is formed. In this experiment, 65,536 benign URLs and 65,536 malicious URLs were randomly selected, for a total of 131,072 URLs. The experimental results demonstrate that the proposed method can achieve better classification results with 97.92% in phishing URL detection, which has high significance for practical application.

Conclusion:

Based on the literature reviewed, it can be seen as it is evident that the existing solutions have not achieved the expected decrease of phishing attacks due to the fact that the human security factors that phishers exploit often have not received an easy to use and identify phishing email. Users fall for this attack as ordinary web browsing users are not aware of how phishing attacks start or how to visually recognise phishing websites to differentiate them from non-phishing ones [67]. The existing solutions are either residing in the servers or installed in the users' system and what the systems do are not known to the user, only the decision of the system would determine whether the user will continue or not, such as blacklist and whitelist which checks the requested URL by comparing it to what is listed in. However, with the identified downside of Blacklist, it cannot detect correctly if the URL is not listed and in

such cases, the users still believe this system because the decision of the system is not visible to them [68]. The expert was introduced to help novice users to be aware of the circumstances of phishing attacks, that they may be able to minimise or avoid this risk, perhaps stop it as early as possible, also has a limitation in that users' knowledge retention on what is taught about phishing attack and how to protect themselves from such attack. Therefore, phishing detection research should be geared towards users ease of use and identify phishing attack by developing a system that can display originality and malicious nature of both email and website [68]. This paper reveals website phishing solution in phishing attack detection and provides a literature analysis of different existing phishing detection approaches.

Reference:

1. Wu, T., Wang, M., Xi, Y., Zhao, Z. (2022): *Malicious URL Detection Model Based on Bidirectional Gated Recurrent Unit and Attention Mechanism*. Appl. Sci. 2022, 12, 12367. https://doi.org/10.3390/ app122312367
2. Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, 1–19. https://doi.org/10.1038/s41598-022-10841-5
3. Pavan Kumar, P., Jaya, T., &Rajendran, V. (2021). SI-BBA – A novel phishing website detection based on Swarm intelligence with deep learning. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.178
4. Yang, R., Zheng, K., Wu, B., Li, D., Wang, Z., & Wang, X. (2022). Predicting User Susceptibility to Phishing Based on Multidimensional Features. *Computational Intelligence and Neuroscience*Vol 2022, Article ID 7058972, 11 pp 1-11 https://doi.org/10.1155/2022/7058972 *2022*.
5. Anitha, J. (2022). A new hybrid deep learning-based phishing detection system using MCS-DNN classifier. *Neural Computing and Applications*, *0123456789*. https://doi.org/10.1007/s00521-021-06717-w
6. Al-ahmadi, S. (2022). Robust Phishing Detection Against Adversaries. *WSEAS TRANSACTIONS on COMPUTER RESEARCH*. vol10 pp 1–8. https://doi.org/10.37394/232018.2022.10.1
7. Zhang, Q. (2021). Practical Thinking on Neural Network Phishing Website Detection Research Based on Decision Tree and Optimal Feature Selection. *Journal of Physics: Conference Series*, *2031*(1). https://doi.org/10.1088/1742-6596/2031/1/012062
8. Selvakumari, M., Sowjanya, M., Das, S., &Padmavathi, S. (2021). Phishing website detection using machine learning and deep learning techniques. *Journal of Physics: Conference Series*, *1916*(1). https://doi.org/10.1088/1742-6596/1916/1/012169
9. Xiao, X., Xiao, W., Zhang, D., Zhang, B., Hu, G., Li, Q., & Xia, S. (2021). Phishing websites detection via CNN and multi-head self-attention on imbalanced datasets. *Computers and Security*, *108*. https://doi.org/10.1016/j.cose.2021.102372
10. Ravindra, S. S., Sanjay, S. J., Gulzar, S. N. A., &Pallavi, K. (2021). Phishing Website Detection Based on URL. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. https://doi.org/10.32628/cseit2173124
11. Sabahno, M., &Safara, F. (2021). ISHO: improved spotted hyena optimization algorithm for phishing website detection. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-10678-6
12. Al-Sarem, M., Saeed, F., Al-Mekhlafi, Z. G., Mohammed, B. A., Al-Hadhrami, T., Alshammari, M. T., Alreshidi, A., &Alshammari, T. S. (2021). An optimized stacking ensemble model for phishing websites detection. *Electronics (Switzerland)*, *10*(11). https://doi.org/10.3390/electronics10111285
13. Alsariera, Y. A., Elijah, A. V., &Balogun, A. O. (2020). Phishing Website Detection: Forest by Penalizing Attributes Algorithm and Its Enhanced Variations. *Arabian Journal for Science and Engineering*, *45*(12), 10459–10470. https://doi.org/10.1007/s13369-020-04802-1
14. Guo, B., Zhang, Y., Xu, C., Shi, F., Li, Y., & Zhang, M. (2021). HinPhish: An Effective Phishing Detection Approach Based on Heterogeneous Information Networks. *Applied Sciences*, *11*(20). https://doi.org/10.3390/app11209733.
15. Alzahrani, S. M. (2021). Phishing Attack Detection Using Deep Learning. *International Journal of Computer Science and Network Security*, December 2021. *21*(12), pp. 213-218 https://doi.org/10.22937/IJCSNS.2021.21.12.31
16. John-otumu, A. M., Rahman, M., &Oko, C. U. (2021). An Efficient Phishing Website Detection Plugin Service for Existing Web Browsers Using Random Forest Classifier. *American Journal of Artificial Intelligence*5(2), 66–75. https://doi.org/10.11648/j.ajai.20210502.13
17. Wang, S., Khan, S., Xu, C., Nazir, S., &Hafeez, A. (2020). Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers. *Complexity*, *2020*. https://doi.org/10.1155/2020/8694796
18. Saha, I., Sarma, D., Chakma, R. J., Alam, M. N., Sultana, A., & Hossain, S. (2020). Phishing attacks detection using deep learning approach. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Icssit*, pp. 1180–1185. https://doi.org/10.1109/ICSSIT48917.2020.9214132
19. Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2019). Deep learning with convolutional neural network and long short-term memory for phishing detection. *2019 13th International Conference on Software, Knowledge, Information Management and Applications, SKIMA 2019, August*. https://doi.org/10.1109/SKIMA47702.2019.8982427
20. Deshpande, A. Pedamkar, O., Chaudhary, N. &Borde, S. D. (2021). Detection of Phishing Websites using Machine Learning. *International Journal of Engineering Research & Technology (IJERT) 10(05)*, pp. 430–434.
21. Abusaimeh, H., &Alshareef, Y. (2021). Detecting the Phishing Website with the Highest Accuracy. *TEM Journal*, *10*(2) pp.947-953, https://doi.org/10.18421/TEM102-58

22. Mehanović, D., &Kevrić, J. (2020). Phishing website detection using machine learning classifiers optimized by feature selection. *Traitement Du Signal*, *37*(4). https://doi.org/10.18280/TS.370403
23. Deyanara T, &Antoni W. (2020). Phishing Website Detection using Neural Network and PCA based on Feature Selection. *International Journal of Recent Technology and Engineering*, *8*(6). https://doi.org/10.35940/ijrte.d4532.038620
24. Ramalingam, V. V, Yadav, P., & Srivastava, P. (2020). Detection of Phishing Websites using an Efficient Feature-Based Machine Learning Framework. *International Journal of Engineering and Advanced Technology (IJEAT)* ISSN: 2249 – 8958, Vol.9 No.3, February, 2020 *pp.* 2857–2862. https://doi.org/10.35940/ijeat.C5909.029320
25. 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 (Switzerland)*, *9*(9). https://doi.org/10.3390/electronics9091514
26. Sameen, M., Han, K., Hwang, S. O. U. N., & Member, S. (2020). PhishHaven — An Efficient Real-Time AI Phishing URLs Detection System. *National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant 2020R1A2B5B01002145.8*. pp. 83425- 83443. https://doi.org/10.1109/ACCESS.2020.2991403
27. Hema, R., Ramya, V., Sahithya, K., &Sekharan, R (2020). Detecting of Phishing Websites using Deep Learning. *Journal of Critical Reviews*, *7*(11), 3606–3613.
28. El-Alfy, E. S. M. (2017). Detection of Phishing Websites Based on Probabilistic Neural
29. Networks and K-Medoids Clustering. *The Computer Journal*, 1-15.
30. Kulkarni, A., & Brown, L. L. (2019). Phishing websites detection using machine learning. *International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019* pp 8-13
31. Musa, H., Aminu, A. A., & Muhammad, A. N. (2019). Investigation the Effect of Dataset Size on the Performance of Different algorithm in Phishing Website Detection, *International Journal of Artificial Intelligence and Applications (IJAIA), Vol.10, No.3, May 2019pp*, 53–59. https://doi.org/10.9790/7388-0901025359
32. Musa, H., Gital, A. Y., Bitrus, M. G., Juma, F., &Balde, M. A. (2020). *Boosting the Accuracy of Phishing Detection with Less Features Using XGBOOST*. *8*(2), 81–90.
33. Musa, H., Gital, A. Y., Zambuk, F. U., Umar, A., Umar, A. Y., &Waziri, J. U. (2019). A comparative analysis of phishing website detection using XGBOOST algorithm. *Journal of Theoretical and Applied Information Technology*, *97*(5).
34. Mahajan, R., &Siddavatam, I. (2018). Phishing Website Detection using Machine Learning Algorithms. *International Journal of Computer Applications*, *181*(23), 45–47. https://doi.org/10.5120/ijca2018918026
35. Solanki, J., and Vaishnav, R. G. (2016). Website Phishing Detection using Heuristic Based Approach, 2044–2048.
36. Suryavanshi, N., and Jain, A. (2016). Phishing Detection In Selected Feature Using ModifiedSVM-PSO, 5, 208–214.
37. Thabtah, F., and Abdelhamid, N. (2016). Deriving Correlated Sets of Website Features for Phishing Detection: A Computational Intelligence Approach, 15(4), 1–17. doi:10.1142/S0219649216500428.
38. Vaishnaw N., Tandan, S. R., Scholar, M. T., and Bilaspur, C. G. (2015). Development of Anti-Phishing Model for Classification of Phishing E-mail, 4(6), 39–45. doi:10.17148/IJARCCE.2015.4610.
39. Yadav, D. P., Paliwal, P., KumaD, D., and Tripathi, R. (2017). A Novel Ensemble Based Identification of Phishing E-Mails, 2–6.
40. Sahingoz, O. K., Işılay Baykal, S., &Bulut, D. (2018). *PHISHING DETECTION FROM URLS BY USING NEURAL NETWORKS*. https://doi.org/10.5121/csit.2018.81705.
41. Yang, P., Zhao, G., & Zeng, P. (2019). Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access*, *7*. https://doi.org/10.1109/ACCESS.2019.2892066
42. Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2018). Detecting Phishing Websites via Aggregation Analysis of Page Layouts. *Procedia Computer Science*, *129*. https://doi.org/10.1016/j.procs.2018.03.053
43. Kevric, J., Jukic, S., Subasi, A. (2017). An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Computing and Applications, 28(S1): 1051-1058. https://doi.org/10.1007/s00521-016-2418-1
44. Daeef, A. Y., Ahmad, R. B., Yacob, Y., &Phing, N. Y. (2017). Wide scope and fast websites phishing detection using URLs lexical features. *2016 3rd International Conference on Electronic Design, ICED 2016*. https://doi.org/10.1109/ICED.2016.7804679
45. Arya, A. S., Ravi, V., Tejasviram, V., Sengupta, N., &Kasabov, N. (2016). Cyber fraud detection using evolving spiking neural network. *11th International Conference on Industrial and Information Systems, ICIIS 2016 - Conference Proceedings*, *2018-January*. https://doi.org/10.1109/ICIINFS.2016.8262948
46. Abutair, H. Y. A., &Belghith, A. (2017). Using Case-Based Reasoning for Phishing Detection. *Procedia Computer Science*, *109*. https://doi.org/10.1016/j.procs.2017.05.352
47. Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017). Malicious web content detection using machine leaning. *RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, *2018-January*. https://doi.org/10.1109/RTEICT.2017.8256834
48. Shukla, P. K., Gandhi, R., &Vishwavidyalaya, P. (2015). System Design, Investigation and Countermeasure of Phishing Attacks using Data Mining Classification Methods and its Analysis. *June*. https://doi.org/10.14257/ijast.2015.78.03
49. Shirazi, H., Haefner, K., & Ray, I. (2017). Fresh-Phish: A framework for auto-detection of phishing websites. *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, *2017-January*. https://doi.org/10.1109/IRI.2017.40
50. Kadam, S. (2021). Feature based Phishing Website Detection using Random Forest Classifier. *International Journal for Research in Applied Science and Engineering Technology*, *9*(VI). https://doi.org/10.22214/ijraset.2021.35400

51. Paliwal, S., Anand, D., & Khan, S. (2016). Application of Rule based Fuzzy Inference System in Prediction of Internet Phishing. *International Journal of Computer Applications (0975 – 8887) Volume 148 – No.14, August 2016.* https://doi.org/10.5120/ijca2016911334
52. Sarhan, A. Al, Jabri, R., &Sharieh, A. (2017). Website Phishing Detection Using Dom-Tree Structure and Cant-MinerPB Website Phishing Detection Using. *American Journal of Computer Science and Information Engineering*, American Journal of Computer Science and Information Engineering 2017; 4(4): 38-42 http://www.aascit.org/journal/ajcsie ISSN: 2381-1110 (Print); ISSN: 2381-1129.
53. Patil, V., Thakkar, P., Shah, C., Bhat, T., &Godse, S. P. (2018). Detection and Prevention of Phishing Websites Using Machine Learning Approach. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*. https://doi.org/10.1109/ICCUBEA.2018.8697412.
54. Buber, E., Diri, B., &Sahingoz, O. K. (2018). NLP Based Phishing Attack Detection from URLs. *Advances in Intelligent Systems and Computing*, 736, pp. 608–618, 2018. https://doi.org/10.1007/978-3-319-76348-4_59
55. Torroledo, m, I., Camacho, L. D., &Bahnsen, A. C. (2018). Hunting malicious tls certificates with deep neural networks. *Proceedings of the ACM Conference on Computer and Communications Security*. https://doi.org/10.1145/3270101.3270105.
56. Jain, A. K., & Gupta, B. B. (2017). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, *68*(4). https://doi.org/10.1007/s11235-017-0414-0
57. Shirazi, H., Bezawada, B., & Ray, I. (2018). Kn0w thy doma1n name: Unbiased phishing detection using domain name based features. *Proceedings of ACM Symposium on Access Control Models and Technologies, SACMAT*. https://doi.org/10.1145/3205977.3205992.
58. Ahmad A. Y, Selvakumar M., Mohammed A., Mohammed A. and Samer A. S. (2016). TrustQR: A New Technique for the Detection of Phishing Attacks on QR Code, *Adv. Sci. Lett.*, vol. 22, no. 10, pp. 2905-2909, Oct. 2016.
59. Inez C. C. and Baruch F. (2018). Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk, *Risk Anal.*, vol. 38, no. 4, pp. 826-838, Apr. 2018.
60. Diksha G. and Kumar J. A. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges, *Comput. Secur.*, vol. 73, pp. 519-544, Mar. 2018.
61. Craigen, D., Nadia Thibault, P., and Diakun, R.(2014) "Defining Cybersecurity," October 2014. [Online]. Available: http://www.timreview.ca/sites/default/files/article_PDF/ Craigen_et_al_TIMReview_October2021.pdf. [Accessed 2021].
62. Ghotaish, M., A., and Abdullah, O., B.,(2011) Phishing websites detection based onphishing characteristics in the webpage source code, *International Journal of Information and Communication Technology Research* 1. (6) 5-7
63. Huang, H., Qian, L., and Wang, Y., (2012). A svm-based technique to detect phishingurls, Information Technology Journal 11, no. 7, 921.
64. Basnet, R., B., Sung, A., H., & Liu, Q., (2011). *Rule-based phishing attack detection*,International Conference on Security and Management , Las Vegas.
65. Huang, H., Tan, J., and Liu, L., (2009). Countermeasure techniques for deceptive phishing attack, New Trends in Information and Service Science. NISS'09. International Conference on, IEEE, 2009, pp. 636-641.
66. Bargadiya, M., Mohd, V., Khan, I. &Verma B. (2010). The web identityprevention:factorsto consider in the anti-phishing design, *international journal of engineering science and technology*, 2. (7) 6960–6964.
67. Yin, C., Zou, M., Iko D., and Wang, J., Botnet detection based on correlation of malicious behaviors, International Journal of Hybrid Information Technology 6 (2013), no. 6, 291-300.
68. Qabajeh I., Thabtah F., &Chiclana F. (2018) A recent review of conventional vs. automated cybersecurityn anti-phishing techniques.
69. Nmachi, W. P., & Win, T. (2021). *Mitigating Phishing Attack in Organisations: A Literature Review*. 75–83. https://doi.org/10.5121/csit.2021.110105.