Technium Vol. 3, Issue 9 pp.19-35 (2021)
ISSN: 2668-778X
www.techniumscience.com

Phishing Detection Methods: A Review

Mafaz Alanezi

Dept. of Computer Science, College of Computer Science and Mathematics,
University of Mosul, Mosul, Iraq
Mafaz Alanezi, mafazmhalanezi@uomosul.edu.iq

Abstract. The web has turned into a principal part of our conventional social and financial activities. The web isn't significant for singular clients just yet additionally for associations, since associations that offer web-based exchanging can accomplish an upper hand by serving overall customers. Webworks arriving at clients all around the globe with no commercial center limitations and with successful utilization of internet business. Consequently, Internet customers may be defenceless against different kinds of web risks, that may cause financial damages, information forgery, brand reputation mischief, the sacrifice of private information, and loss of customers' confidence in online business and electronic banking. Thusly, the reasonableness of the Internet for business exchanges becomes dubious. Phishing is seen as a design of web peril which is classified as the forte of mimicking a website of a legitimate undertaking proposing to gain a client's private accreditations, for instance, usernames, passwords, and federal retirement aide numbers. In this paper, we present an overview of cutting-edge research on such attacks. Besides, we expect to perceive the cutting-edge advancements in phishing, furthermore, give a far-reaching overview and comparison of these kinds of literatures to understand the hole that is as yet prevailing around anti-phishing. This overview will be for the most types of online phishing detection techniques for the last 4 years.

Keywords. Phishing, Anti-phishing, Traditional and Non-Traditional Methods, Phishing Datasets.

1. Introduction

The way toward shielding the internet from cyber threats has come to be called as Cyber Security. Cyber Security is tied in with securing, forbidding, and recuperating all the assets that utilization the web from cyber threats [1]. According to [2] the 11 biggest cyber security threats in 2021, the Phishing threat tops the list, where Phishing meets COVID-19. Phishing, is a computerized message is shipped off fool individuals into clicking a connection within it. There are a few opportunities for pernicious entertainers to utilize such missions. Contingent upon the expectation of the entertainer, unsafe malware is introduced or delicate information is uncovered. Through the current Corona crisis, individuals stay at home more regularly. In addition, representatives are telecommuting like never before previously. This makes itself as incredible favourable place for digital hoodlums. Phishing threats are arrangement in a manner to transmit the casualties to websites with counterfeit data about the covid19. Customarily, these websites utilize the client's framework assets to procure digital currency like Bitcoin, all without the endorsement of the client [2].

The report of Phishlabs in 2018 on phishing patterns specifies that the objectives of phishing assaults moved from people to ventures. To exacerbate the situation, phishers currently approach free SSL certificates. Almost 50% of all phishing sites as of now use HTTPS, which was one of the significant

pointers of the authenticity of sites. The second report distributed by APWG at 2019 text on that the quantity of phishing assaults expanded by 30% from the past quarter and that the essential objectives were the SaaS (Software-as-a-Service) and email services [3].
The September 2020 "To Catch a Phish" [4] overview tracked down that 52% of email clients neglected to identify a genuine phishing email. This is a disturbing rate as phishing threats and other cybersecurity dangers are on the ascent. Through the primary quarter of 2021, 24.9% of phishing threats overall were coordinated towards monetary organizations. What's more, social media represented 23.6% of threats making these two the most elevated designated businesses when it came to phishing through this interval, figure 1, shows the online businesses most designated by phishing assaults as of the first 3 months of 2021[5].

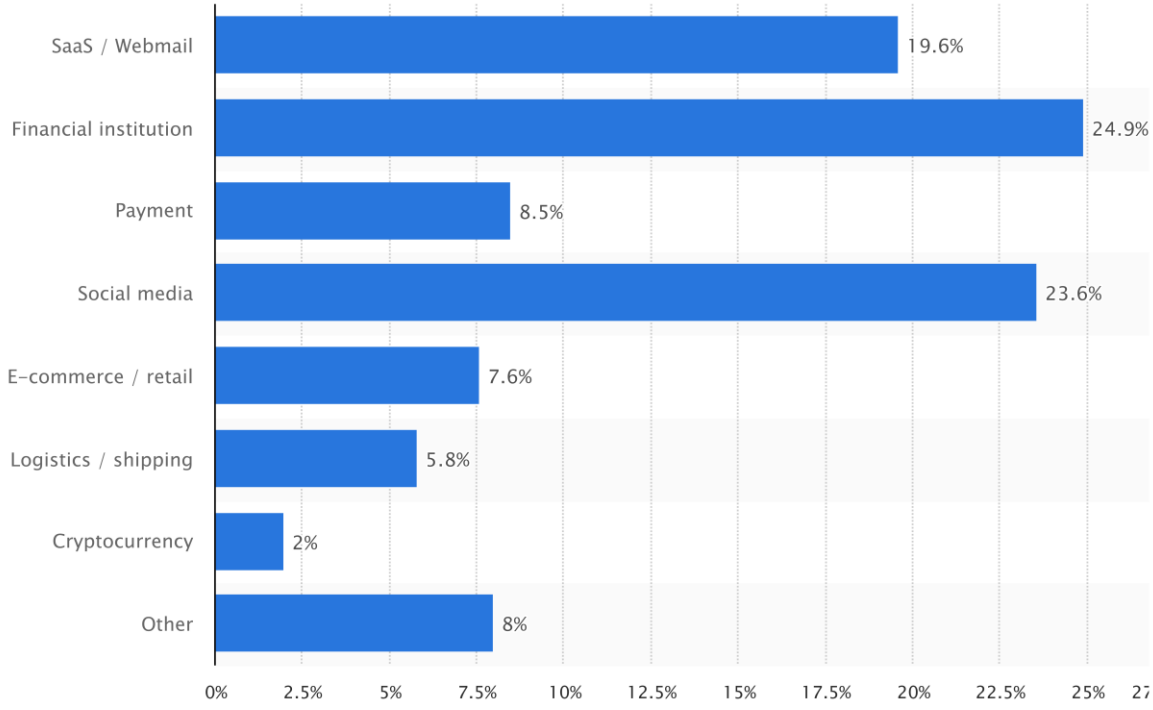

Figure 1: Percentage of phishing attacks [5].

Besides, specialized weaknesses (for example cache poisoning of Domain Name System (DNS)) may be utilized by assailants to develop undeniably more convincing socially-designed messages (for example utilization of authentic, however mock, domain names may be far more convincing than utilizing diverse domain names). This conveys phishing threats a layered issue, and a powerful detection should need resolving problems at the specialized also, human layers [6].
Powerful techniques for detecting phishing websites are critically expected to mitigate the dangers presented by phishing attacks. As active learning capacity from enormous datasets, the machine learning methods is broadly utilized to recognize phishing attacks. Nonetheless, in the phase of training datasets, numerous futile and little impact features will ambush the underlying classifier paradigm into the issue of over-fitting. The over-fitting issue ordinarily makes the trained paradigm that can't viably distinguish phishing websites. Through the selection algorithms, numerous futile and little impact features are truncated. Because no agitating impact from these overabundance provisions, the over-fitting issue of the basic classifier is reduced. In the meantime, these calculations are additionally ready to diminish the cost for time of the operations of phishing sites detection.

Because of the wide idea of the phishing issue, this phishing detection overview presenting a scientific categorization (i.e., taxonomy) of anti-phishing detection, and the kinds of literature review of anti-phishing detection methods based on artificial techniques.
The rest of paper is structured as follows: Section 2 shows the Taxonomy of Anti-Phishing Detection. Section 3 shows the comparison of literatures. Lastly, Section 4 concludes the paper and its benefits.

2. Taxonomy of Anti-Phishing Detection

Numerous strategies have been created to battle phishing as of late. In light of the method of battle phishing, two different methods of against phishing arrangements have been distinguished comprising *traditional and non- traditional*. *traditional* methods primarily incorporate Legal, Education and Awareness, Blacklist/Whitelist, Visual Similarity, and Search Engine. While *non-traditional* methods fundamentally include: Content-Based, heuristics-based, machine learning, Deep Learning, Fuzzy Rule-Based, Hybrid Learning, Data Mining, and Others Artificial Intelligence Methods, see figure 2. In this section we present an overview of the taxonomy and present the literature review for *non-traditional* methods.

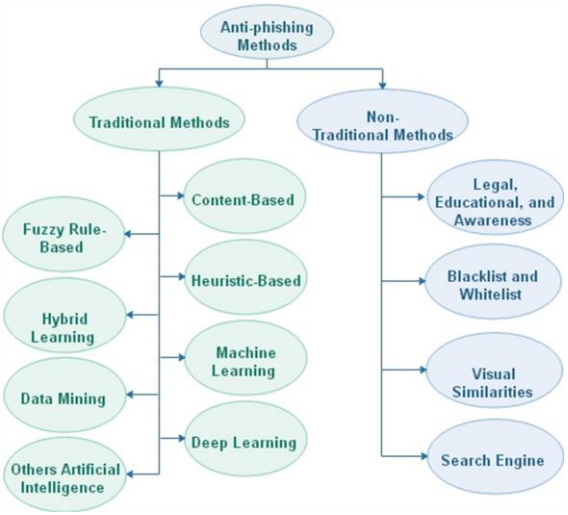

Figure 2: A Taxonomy of Anti-Phishing Detection.

2.1 Traditional Methods

2.1.1 Legal, Educational, and Awareness

Customarily, there are normal ways to deal with battle phishing, for example, legal, educational, and awareness programs. Administrators in nations like the USA, UK, Canada, and Australia have supported authoritative bills that implicate phishing, in which phishers can hope to confront genuine prison sentences. In any case, the lawful activities are not exceptionally viable in diminishing phishing, since a phishing site has a short life expectancy ± typically around two days ± which helps the phisher to vanish rapidly once the extortion has been submitted. Then again, while instructing clients may decidedly influence the worldwide endeavors of battling phishing, this methodology requests high financial expenses other than requiring clients to be outfitted with PC security information [7].

2.1.2 Blacklist and Whitelist Methods

These are lists containing phishing URLs, whereas the mentioned URL is contrasted with a predetermined phishing URL. The methodology's disadvantage is the blacklist as a rule can't cover all phishing sites since a recently made false site takes impressive time until it is being appended to the list. This hole on schedule between dispatching and appending the dubious site to the list might be sufficient for the phishers to accomplish their objectives. Subsequently, the detection operation ought to be

incredibly speedy, normally, when the phishing site is transferred and before the client begins presenting his qualifications [8].
Whitelist methods are something contrary to Blacklisting methods. For an obscure URL, it considers it is an authentic URL by coordinating it versus the "whitelist" Database. The "whitelist" Database mostly incorporates a list of well-known genuine URLs and their important data. As with Blacklist, it might require some investment to upload the new well-known URL. Through the upload time, the phisher can undoubtedly accomplish their objectives.

2.1.3 Visual Similarities Methods

Are methods recognizing phishing sites from authentic sites by the appearance of sites. To keep away from phishing detection strategies, attackers as a rule “embed pictures, Flash, ActiveX, and Java Applet instead of HTML text”. Visual similarities-based detection methods can quickly recognize as enrollment objects display in phishing site webpages. Visual similarity-based methods utilize a signature to distinguish phishing site pages. To make the signature by picking common components from the whole site as opposed to the loner website page. Thusly, one signature is enough to identify various designated webpages of the loner website or unique forms of a website. It tends to comprehensively characterize the visual similarities-based methodologies into HTML document object model (DOM) tree, Cascading Style Sheet (CSS) similarity, visual perception, visual features, pixel-based, and hybrid approaches [9].

2.1.4 Search Engine Methods

Are methods recognizing phishing sites through search engines. These methods depend on the personality data of sites that incorporate site logo, URL, text, etc. Ongoing techniques use search engine ways to deal with counter the phishing assaults as it accomplishes promising recognition precision. Yet, the limit of this methodology is that it bombs when a phishing page is hosted on the compromised server. Besides, it additionally brings about a low true negative rate when recently enlisted or non-well known domains are experienced [10].

2.2 Non-Traditional Methods

2.2.1 Content-Based Methods

Jain A. K. et al. [11] suggested a method to identify phishing cyberattacks which uncovered a client's qualifications to hostile substances prompting a penetrate of protection. The proposed approach includes utilizing a web search tool to coordinate with the space name of the website under a microscope with the destinations that surface because of our hunt inquiry. Here, their inquiry question is involved features which are domain name and catchphrases comprising of title, body text, and meta portrayal information. They at first perceive how ordinary TF-IDF “Term Frequency–Inverse Document Frequency” acts in arranging sites as phishing or real. Thereafter, at that point implement a weighted heuristic suggested exhaustively in their paper by allocating various weights to various label information and control the TF-IDF result likewise to work on the presentation of their phishing indicator.
Che H. et al. [12] depended on the point for distinguishing phishing messages is the things to which these messages' content alludes. These things are called the occasion, also a couple of occasions is called an occasion pair which addresses the connection between these two occasions, and the semantic web is used to change words in messages over to occasions. Another calculation for distinguishing phishing messages which depend on occasion matches is planned in their study. In this calculation, the initial segment is building the semantic web information base which gives the connections among words and occasion. The subsequent part is building the classification information base which is utilized to arrange phishing messages. The third part is the way to identify phishing utilizing the semantic web information base and class data set.
Sonowal G. and Kuppusamy K. S. [13] said that the single filter approach would be deficient to recognize various classifications of phishing endeavors. They gave a multi-facet model to identify phishing, named PhiDMA “Phishing Detection using Multi-filter Approach”. The PhiDMA model joins 5 layers: “Auto upgrade whitelist layer, URL features layer, Lexical signature layer, String matching

layer and Accessibility Score comparison layer". The people with visual weakness will get to a model execution with no boundary because the suggested PhiDMA model is worked with an available interface.

To identify zero-day phishing attacks, Mishra A. and Gupta B. B. [14] proposed CSS and URI coordinating-based phishing detection framework (CUMP). Their framework depends on the idea of URI "uniform resource identifier" furthermore, the CSS "cascading style sheet" coordinating. This idea is utilized, like the phisher continuously attempts to mirror the URI design and visible plan with the expectation that even veteran clients cannot identify phishing sites by perception. To mirror the visual appearance, phishers by and large utilize the same CSS style. Without utilizing the same CSS, it is truly challenging to accomplish a similar plan. Their framework utilized the fundamental characteristics of any phishing attacks for URI and CSS coordinating, to safeguard in case of phishing sites threats particularly "zero-day" attacks.

2.2.2 Heuristic-Based Methods

These methodologies, whereas little features are picked up from the website to characterize it as for phishing or genuine. As opposed to the blacklist methods, a heuristic-based methods can recognize new phishing websites continuously. The viability of the heuristic-based techniques, some of the time called features-based strategies, relies upon picking a bunch of discriminative features that could help in distinctive the kind of website.

Rao R. S. and Ali S. T. [15] executed a PC app named PhishShield, that focuses on URL and Web Content of the phishing pages. The PhishShield accepts the URL as info and yields the situation with the URL as a phishing or real site. The heuristics used to identify phishing are footnote links with invalid worth, zero link within the body of HTML, copyright content, title content, and website identity. The app can identify zero-hour phishing attacks that blacklists are incapable to distinguish and it is quicker than visual-based appraisal procedures that are utilized in detecting phishing.

Rao R. S. et al. [16] proposed a heuristic strategy utilizing TWSVM (twin support vector machine) classifier to identify malignant enrolled phishing websites and furthermore websites that are facilitated on arrangement servers, to defeat the previously mentioned impediments. Their strategy distinguishes the phishing sites hosted on arrangement domains by looking at the sign-in webpage and a main webpage of the visited site. Identifying phishing websites that are malignant enlisted is done by utilized the hyperlink and URL - based features. For the arrangement of phishing websites, they have utilized a various variety of support vector machines (SVMs). They tracked down that twin support vector machine classifier (TWSVM) beats different adaptations.

Babagoli M. et al. [17] proposed a strategy for an anti-phishing website that uses a nonlinear relapse algorithm based on meta-heuristics as well as a feature selection technique. For approving the proposed strategy, they utilized a dataset that included 11055 genuine and phishing site pages, and select 20 features to be separated from the referenced sites. Their research uses two-element feature techniques: decision tree and covering to choose the optimum component sub-set, while the last brought about the recognition exactness rate to 96.32%. Then, two meta-heuristic algorithms are effectively carried out to anticipate and distinguish false sites: harmony search (HS) which was conveyed dependent on nonlinear relapse method, and support vector machine (SVM). The nonlinear relapse approach was utilized to order the sites, where the boundaries of the proposed relapse model were gotten utilizing the HS algorithm. The proposed HS algorithm employments a powerful pitch change rate and created new harmony. Subsequently, the examination tracks down that the nonlinear relapse-based HS brings about better execution contrasted with SVM.

Gupta B. B. and Jain A. K. [18] presented a web search tool based technique, which recognizes phishing site pages precisely paying little mind to the literary language utilized inside the site page. The proposed web search tool-based technique utilizes a lightweight, reliable and language autonomous pursuit inquiry to identify the legitimateness of the dubious URL. They have additionally coordinated five heuristics (Source Code Based Filtering, Input Tag Checking, Null and not Working URLs, Foreign Anchors, Fake Login Form) with the web search tool-based instrument to further develop the recognition

exactness, as some recently made legitimate sites may not show up in the web index. The proposed strategy can likewise effectively arrange the recently made legitimate sites that are not characterized by accessible internet searcher-based strategies.

2.2.3 Machine Learning Methods

Cuzzocrea A. et al. [19] proposed a machine learning-based strategy ready to recognize whether a website can perform phishing exercises. In this situation, a few web clients are communicating with phishing websites (at this point unclear, obviously), also, the objective of their structure is simply to recognize the phishing webpages and inform the clients. For this goal, feature extraction is responsible for extricating reasonable components to guide the machine-learning-based recognition stage. An ad-hoc Built-In dataset is populated along with this approach, to extricate the features. At last, execute the Decision Tree algorithms on the last dataset, and the web phishing occasion notice is at last answered for the web clients.

Chiew K. L. et al. [20] proposed a Hybrid Ensemble Feature Selection (HEFS), which is a feature selection structure for a machine learning-based phishing detection framework. In the 1st stage of HEFS, an original Cumulative Distribution Function gradient (CDF-g) calculation is taken advantage of to create essential feature sub-sets, that are the input into a data perturbation gathering to output optional feature sub-sets. The subsequent stage infers a bunch of pattern features from the optional feature subsets by utilizing a function disturbance set. General test outputs are recommended for HEFS works better when it is coordinated with Random Forest classifier than SVM, Naive Bayes, C4.5, JRip, and PART classifiers.

Yadollahi M. M. [21] created a dependable detection framework that can adaptively coordinate with the evolving climate and phishing sites. In dynamic environments, the Learning Classifier Systems (LCS) can be applied because of having a whiz capacity, its motivation from the overall rule of Darwinian evolution furthermore, cognitive learning. The XCS is the grinder of LCSs because of its adaptable design giving hopeful characteristics like internet learning capacity, commotion power, the over-simplification in the learning instrument, and nonstop transformation. They proposed a versatile detection framework that can identify phishing sites from URLs utilizing a bunch of the remarkable features of the web browser just and doesn't rely upon third-party services. Their framework has 2 fundamental segments called: 1) Feature Extractor: is liable for tracking down the legitimate list of features and 2) XCS: assumes a part as a phishing detector by perceiving the style of forthcoming phishing sites and advancing a standard set.

Sahingoz O. K. et al. [22] presented a real-time phishing detection system, which utilizes seven unique classification algorithms and natural language processing (NLP) based features. Their framework has the accompanying distinctive properties from different studies: language freedom, utilization of a colossal mass of phishing and genuine information, real-time implementation, identification of recent sites, utilization of feature-rich classifiers, and freedom from 3rd-party services. To estimate the presence of the framework, the test outcomes are tried on another developed dataset. As indicated by the trial and relative outcomes from the carried-out classification algorithms, the best exhibition came from the Random Forest algorithm with just NLP-based features.

2.2.4 Deep Learning Methods

Smadi S. et al. [23] suggested a system to identify phishing assaults in the online mode for the first time, by joins a neural network together with reinforcement learning. Their suggested system can acclimate to deliver a new phishing email discovery structure that reflects changes in as of late examined rehearses, which is cultivated by embracing the possibility of reinforcement learning to improve the system powerfully after a period. Naturally, the suggested system appends more messages to the offline dataset in the online mode, to tackle the issue of the restricted dataset via. They suggested a new algorithm to examine any recent phishing practices in the novel dataset. Through thorough testing utilizing the notable informational collections, they presented that the suggested method can manage zero-day phishing assaults with predominant levels achieving high accuracy.

Wei W. et al. [24] proposed the technique for recognizing phishing sites dependent on just the URL address text by a deep neural network with convolutional layers. In spite of the past works, where URL or traffic measurements or web content are broken down, they examined just the URL text. Accordingly, the technique is quicker and recognizes zero-day attacks. The network they presented was properly upgraded so it tends to be utilized even on cell phones without altogether influencing its performance.
Pharming is an uncommon kind of phishing assault or DNS harming where the client is diverted to a phony site by alternating the IP address at the DNS server. Gajera K. et al. [25] proposed a two-stage system, the stage I use a feature engineering approach takes out significant factors. Then, at that point, the Artificial Neural Network is utilized to fabricate the system, at that point the system foretells a worth a value in the reach (0,1) to foresee the authenticity of the URL. Stage II identifies pharming where they inquiry a neighborhood and a worldwide DNS to acquire IP locations and examine if they equal. In the event that equivalent, this straightforwardly means no pharming, and they continue on for website page examination.
Yerima S. Y. and Alzaylaee M. K. [26] proposed a methodology that uses CNN (convolutional neural networks) for excellent exactness sorting to recognize phishing websites from certifiable destinations. They assess the frameworks utilizing a dataset acquired from 4,898 phishing sites and 6,157 certified. In view of the consequences of broad investigations, their CNN-based models end up being profoundly viable in distinguishing obscure phishing sites. Besides, the CNN-based methodology implemented superior compared to customary ML assessed on an equivalent dataset.
From the idea that deep learning methods are productive for image sorting and natural language. Adebowale M. A. et al. [27] assembled a hybrid sorting framework called the IPDS (intelligent phishing detection system) by utilizing the LSTM (short-term memory) algorithm and CNN (convolutional neural network). The blend of both LSTM and CNN was utilized to determine the issue of an enormous dataset and higher classifier expectation execution. Subsequently, joining the two strategies prompts a superior outcome as well as less preparing time for LSTM and CNN structure, whilst at the same time utilizing the picture, frame, and text features as a crossbreed for their framework recognition. To assemble the suggested framework, they utilized 1m widespread resource finders and more than 10,000 pictures to train the CNN and LSTM classifier. After that, the affectability of the suggested framework was controlled by using different variables like the kind of feature, number of wrong sorting, and split problems.
Zhu E. et al. [28] proposed OFS-NN, a viable phishing sites identification framework dependent on neural network and an ideal feature selection technique. The suggested OFS-NN was first acquainted with another index, feature validity value (FVV), and assessing the effect of delicate features on phishing sites identification. After that, considering the new FVV index, an estimation is planned to pick the best features from phishing websites. This algorithm can relieve the over-fitting issue of the fundamental neural network generally. The fundamental neural network was prepared by the chose ideal features, lastly, an ideal classifier was developed to recognize the phishing sites.
Adebowale M. A. et al. [29] designed and developed a deep learning-based phishing identification by using the Universal Resource Locator and site content, for example, pictures and frame components. To assemble a sorting framework, a CNN (Convolutional Neural Network) and the LSTM (Long Short-Term Memory) algorithm were utilized. The proposed IPDS utilized two DL layers to characterize phishing sites by utilizing LSTM on text and frame content and the CNN on pictures. Accordingly, the model can undoubtedly investigate the wealth of the words implanted in the site's Universal Resource Locator (URL) just as the pictures on the site.
Wang W. et al. [30] proposed a quick phishing site identification method considered PDRCNN which depends just on the site's URL. Not necessary that PDRCNN recover the substance of the objective site, and not utilizes any third-party services as past methods do. It encodes the data of a URL into a 2D tensor and inputs the tensor into a novella planned deep learning neural network to arrange the first URL. They initially utilize a bidirectional LSTM network to remove worldwide components of the belt tensor and give all string data to every character in the URL. Starting there ahead, they utilize a CNN to consequently decide what characters are assumed key parts in phishing identification, catch the vital

segments of the URL, and pack the removed components into a fixed-length vector space. Via PDRCNN consolidating the two sorts of networks, it accomplishes preferred execution.

2.2.5 Fuzzy Rule-Based Methods

Abuzuraiq A. et al. [31] suggested a framework with two phases. The 1st phase, diverse ML algorithms are executed to approve the picked dataset and using features choice techniques. The highest precision was accomplished by using Random Forest with just 20 features out of 48 provisions. Whilst in the 2nd phase, different FL (fuzzy logic) algorithms were applied to the equivalent dataset. Also, the test results from the utilization of FL algorithms were staggering. Whereas in executing the FURIA algorithm with just 5 features the precision rate was higher. At long last, they did an examination and conversation of the outcomes between executing ML algorithms and FL algorithms. Where the exhibition of utilizing FL algorithms surpasses the utilization of ML algorithms.

Adebowale M. A. et al. [32] suggested versatile Neuro-Fuzzy Inference System (NFIS) based strong plan utilizing the incorporated components of text, pictures, and frames for web phishing detection and insurance. They researched a concurrence to the authoritative features that ought to be utilized in phishing identification.

Zabihimayvan M. and Doran D. [33] applied FRS (Fuzzy Rough Set) hypothesis as a technique to choose the best features from 3 standard datasets. The chose features are input into 3 regularly utilized classifiers for phishing identification. The classifiers are trained by a different out-of-test dataset of 14,000 site tests, in order to assess the FRS feature selection in fostering a general phishing identification. Because there is no features from 3rd-part services in the general feature set, with no request from outside sources, they could acquire a quicker anti-phishing that is likewise vigorous toward zero-day assaults.

Pham C. et al. [34] utilized the URL features and web traffic features to recognize phishing sites in light of a planned neuro-fuzzy structure (named Fi-NFN). In light of the new methodology, fog computing as empowered by Cisco, planed an anti-phishing model to straightforwardly screen also, shield fog clients from phishing assaults. The trial aftereffects of their suggested approach, in light of a huge scope dataset gathered from genuine phishing cases, have shown that their framework can adequately forestall phishing assaults and work on the security of the network.

2.2.6 Hybrid Learning Methods

Ali W. and Ahmed A. A. [35] suggested a hybrid intelligent phishing site expectation utilizing deep neural networks (DNNs) with transformative algorithm-based feature selection and weighting strategies to upgrade the phishing site forecast. In these methods, the genetic algorithm (GA) is helpful in expanding the exactness of phishing site expectations by assigning heuristically to it the most powerful features and the ideal weights of site features, this step is for preparing DNNs to precisely anticipate phishing sites. The exploratory outcomes showed that the suggested hybrid intelligent phishing site expectation methods accomplished essentially high sorting precision, affectability, particularity, and mathematical mean in phishing site expectation.

Zhu E. et al. [36] suggested a DTOF-ANN ("Decision Tree and Optimal Features based Artificial Neural Network") handle over-fitting, which is a neural-network phishing recognition model dependent on decision tree and ideal feature selection. In the first place, was enhancing the customary K-medoids clustering algorithm with a steady choosing of introductory centers to eliminate the copy dots in the popular datasets. After that, an ideal feature choosing dependent on the novel feature include decision tree, assessment record, and neighborhood search strategy are intended to take off the futile and negative elements. At last, the ideal design of the neural network classifier is built by appropriately changing boundaries then trained by the chosen ideal features. Trial results have exhibited that DTOF-ANN displays good.

Suleman M. T. and Awan S. H. [37] utilized the Uniform Resource Locator (URL) based phishing detection method. Machine learning classifiers like Random Forest, Iterative Dichotomiser-3 (ID3), Decision Tree, K-Nearest Neighbor (KNN), and Naïve Bayes utilized for the sorting of authentic and

ill-conceived sites. This sorting could assist in the recognition of phishing sites. Nonetheless, further develop detection precision is done by utilizing Genetic Algorithms (GAs) for feature selection. The detection exactness of their test outcomes was further developed by the utilization of ID3 (Iterative Dichotomiser-3) alongside YAGGA (Yet Another Generating Genetic Algorithm).

Vrbančič G. et al. [38] presented a technique, named TDLHBA ("Tuning Deep Learning using Bat/Hybrid Bat Algorithm") that joins swarm intelligence methods for variables settings of deep learning neural networks. The fundamental objective is to consider whether a NN with variables set by using the suggested strategy will present higher sorting exactness than NN with normal settings. It is extremely clear that the fundamental benefit of the suggested strategy, is the use of different feed-forward NN topologies and distinctive datasets, with no need to physically look for the right learning variables.

Chin T. et al. [39] introduced PhishLimiter, a recognition, and relief method, where they initially suggested a new strategy of DPI (Deep Packet Inspection) then, influence it with SDN (Software-Defined Networking) to recognize phishing exercises by email as well as web-based communication. The suggested DPI method comprises 2 parts, phishing signature sorting, and real-time DPI. In light of the encoding of SDN, they foster SF (Store and Forward) mode and the FI (Forward and Inspect) mode to coordinate network traffic by utilizing an ANN (Artificial Neural Network) model to group phishing assault marks and plan the real-time DPI with the goal that PhishLimiter could deftly identify the elements of phishing assault in reality. Since PhishLimiter has a worldwide perspective on a network through SDN, it also gave better network traffic to the executives to contain phishing assaults. Moreover, they assess PhishLimiter utilizing a genuine world-tested climate and datasets comprising of true email as well as installed joins.

Chen W. et al. [40] proposed a strategy for joining PSO (Particle Swarm Optimization) and BP (Back Propagation) neural network to construct a phishing site recognition framework. In order to further develop the assembly execution of the neural network detection framework, they used PSO advances neural network boundaries. Exploratory outcomes indicate that this calculation could work on the forecast speed and the precision of identifying phishing sites by 3.7% contrasted with the traditional BP neural network algorithm.

Gupta S. and Singhal A. [41] suggested a strategy to characterize the URL into Phishing URL or Nonphishing URL. By utilizing PSO (particle swarm optimization) to train the ANN inorder to characterize URLs to work on the presentation of ANN. The suggested a strategy executed on various proportions of learning and distinctive activation function, a number of the hidden layers, and output layers. For assessing the ANN_PSO framework, they utilized precision and RMSE criteria. The ANN_PSO framework worked good training as far as precision as for Back Propagation Neural Network (BPNN).

2.2.7 Data Mining Methods

Smadi S. et al. [42] proposed a smart model for the detection of phishing messages which relies upon a preprocessing stage that takes out a bunch of features concerning diverse email parts. The separated features are ordered utilizing the J48 classification algorithm. They experimented with an aggregate of 23 features. Ten-fold cross-approval was applied for training, testing, and approval. The essential focal point of their paper is to improve the general measurements upsides of email characterization by concentrating on the preprocessing stage and decide the best algorithm that can be utilized in this field. The outcomes show the advantages of utilizing their preprocessing stage to remove features from the dataset. Their model accomplished higher precision for the random forest algorithm, which is the most noteworthy enlisted so far for an endorsed dataset.

Thabtah F. and Abdelhamid N. [43] analyzed various features evaluation procedures in the website phishing setting to decide the insignificant arrangement of features for identifying phishing exercises. Exploratory outcomes on real phishing datasets comprising of 30 features have been directed utilizing three known features selection techniques (Chi-Square, CFS, and IG). New features shorts have been distinguished after statistical examination using three data mining classifications algorithms (PART,

RIPPER, and C4.5). They have had the option to recognize new groups of features that when utilized together can distinguish phishing exercises. Further, significant relationships among normal features have been inferred.

Subasi A. [44] introduced a smart framework to recognize phishing attacks. They utilized various data mining strategies to choose classifications of websites: real or phishing. Various classifiers ("Artificial Neural Networks ANN, K-Nearest Neighbour k-NN, Support Vector Machine SVM, C4.5 Decision Tree, Random Forests RF") were utilized to develop an exact smart framework for phishing site recognition. They utilized to assess the exhibition of the data mining procedures: the classification accuracy, a region under the receiver operating characteristic (ROC) curves (AUC), and F- measure. The outcomes indicated that the Random Forest has outflanked optimum among the sorting strategies, also Random Forest runtimes are very quick, and it can handle various websites for phishing detection.

Sentürk S. et al. [45] created and analyzed a prevention system against email phishing attacks profoundly by utilizing machine learning and data mining strategies. By utilizing machine learning techniques a few features have been separated in the email and grouping has been refined dependent on characterized models. They utilized a decision tree (J48 algorithm) that is fabricated hierarchical from a root node and containing parceling the data into subsets which include examples as well as comparative fineness (homogenous). To compute the homogeneity of a sample, the suggested strategy used entropy. If the entropy is zero its means that the sample is fully homogeneous, and if entropy is one its means that the sample is similarly partitioned. In view of the information gain and entropy ideas, data mining procedures have been applied with the semi-automatic finding of patterns, relationships, changes, abnormalities, rules, and measurably huge constructions in data. WEKA instrument has been productively utilized for data mining processes of an email. By WEKA, the seek for relationships and rules permitting one to make expectations about the future from a lot of information has been finished.

2.2.8 Others Artificial Intelligence Methods

Ali W. and Malebary A. S. [46] suggested an intelligent phishing website detection utilizing particle swarm optimization-based feature weighting to upgrade the phishing of phishing websites. The suggested method recommends using PSO to weigh different sites' features viably to accomplish better exactness when recognizing phishing sites. In view of how significant the features contribute towards perceiving phishing from genuine sites, the suggested PSO-based site feature weighting is utilized to separate among the different features in sites. The trial outcomes showed that the suggested PSO-based accomplished exceptional upgrades as far as classification exactness of the ML frameworks utilizing just minimal sites features used in the anti-phishing sites.

Feng F. et al. [47] suggested a phishing discovery framework dependent on a new neural network grouping technique. The identification framework could accomplish better precision and had great speculation capacity by configuration hazard minimizing rule. The training process of the identification framework was done by the Monte Carlo algorithm, so it is basic and stable. In light of checking a bunch of amiable and phishing sites, they noticed that this novel phishing identification model accomplishes the best exactness equivalent to different models.

3. Comparison of Literatures

Basically, for a total and exhaustive investigation of the features, the researchers used different and broad datasets, in any case, the feature significance might be one-sided or off base. Also, as email is a famous assault transporter for conveying phishing URLs or malware, we likely overviewed URLs and site pages as well as emails.

Table 1 presents the outline of non-traditional methods for phishing websites detection, the techniques used for detection, the Used Datasets and their outcomes accuracy.

Table 1: The literature review of non-traditional methods for phishing websites detection.

Anti-Phishing Method	Authors	Techniques	Dataset	Accuracy

Content-Based	Jain A. K. et al. [11]	Modified TF-IDF	Alexa dataset[48], OpenPhish [49], Phish Tank [50]	89%
	Che H. et al. [12]	semantic web and fuzzy control	A case study on emails	-
	Sonowal G. and Kuppusamy K. S. [13]	PhiDMA framework incorporates five layers	Phishload, 2016. Legitimate URL dataset [51]	92.72%
	Mishra A. and Gupta B. B. [14]	CSS and URI matching-based	Target list from Phish Tank [50]	-
Heuristics-based	Rao R. S. and Ali S. T. [15]		Real phishing web sites	96.57%
	Rao R. S. et al. [16]	TWSVM	PhishTank [50], Alexa dataset[48]	98.05%
	Babagoli M. et al.[17]	meta-heuristics (HS, SVM)	UCI phishing Datasets[52], [53]	92.80%
	Gupta B. B. and Jain A. K. [18]	search engine-based	PhishTank [50], OpenPhish[49]	99.05%
Machine Learning	Cuzzocrea A. et al. [19]	Decision Tree	PhishTank [50]	-
	Chiew K. L. et al. [20]	Cumulative Distribution Function gradient (CDF-g), Random Forest, SVM, Naive Bayes, C4.5, JRip, and PART	UCI phishing Datasets [52], [53]	94.6%
	Yadollahi M. M. [21]	XCS	Real URLs	98.39%
	Sahingoz O. K. et al. [22]	Random Forest with NLP	PhishTank [50], Yandex[54]	97.98%
Deep Learning	Smadi S. et al. [23]	Reinforcement Learning, Neural Network	PhishingCorpus[55], SpamAssassin[56], PhishTank[50]	97%
	Wei W. et al. [24]	convolutional neural networks	PhishTank [50], Common Crawl Foundation [57]	99.98%
	Gajera K. et al. [25]	Artificial Neural Network(ANN)	PhishTank [50], fcsit.unimas[58] [59]	98.77%
	Yerima S. Y. and Alzaylaee M. K. [26]	convolutional neural networks (CNN)	UCI Machine Learning Repository [53]	98.2%
	Adebowale M. A. et al. [27]	CNN and LSTM	PhishTank [50], Common Crawl[57]	93.28%
	Zhu E. et al. [28]	feature validity value (FVV) and neural network	UCI phishing [53], PhishTank[50]	98.49%

	Adebowale M. A. et al. [29]	CNN and LSTM	PhishTank [50], Common Crawl[57]	93.28%
	Wang W. et al. [30]	LSTM and CNN	Alexa [48], PhishTank [50]	97%
Fuzzy Rule-Based	Abuzuraiq A. et al. [31]	Random Forest and FURIA	Phish Tank [50], OpenPhish [49], Alexa[48], Common Crawl[57]	99.98%
	Adebowale M. A. et al. [32]	Adaptive Neuro-Fuzzy Inference System (ANFIS)	UCI phishing Datasets [52], [53]	98.3%
	Zabihimayvan M. and Doran D. [33]	Fuzzy Rough Set (FRS)	UCI1 [53], UCI2 [60], Mendeley	95% F-measure
	Pham C. et al. [34]	neuro-fuzzy, Fog computing, Cloud computing	PhishTank [50], DMOZ [61]	98.36%
Hybrid Learning	Ali W. and Ahmed A. A. [35]	deep neural networks (DNNs) and genetic algorithm (GA)	UCI phishing websites [60]	91.13
	Zhu E. et al. [36]	Decision Tree and Optimal Features based Artificial Neural Network, K-medoids clustering algorithm	UCI[52][53], PhishTank[50], Alexa[48]	95.76%
	Suleman M. T. and Awan S. H. [37]	Iterative Dichotomiser-3 (ID3) and Yet Another Generating Genetic Algorithm (YAGGA)	UCI machine learning website[52], [53]	95%
	Vrbančič G. et al. [38]	bat algorithm (BA) and hybrid bat algorithm (HBA)	UCI [53]	96.5%
	Chin T. et al. [39]	Deep Packet Inspection (DPI), Software-Defined Networking (SDN) and ANN	UCI[60]	98.39%
	Chen W. et al. [40]	Particle Swarm Optimization (PSO) and BP neural network	Yahoo Directory[62], Phishtank [50]	98.95%

	Gupta S. and Singhal A. [41]	ANN_PSO, BPNN	UCI [60], PhishTank[50]	99.8%, 96.7%
Data Mining	Smadi S. et al. [42]	J48 algorithm and C4.5 algorithm	PhishingCorpus[55], SpamAssassin[56]	98.87%
	Thabtah F. and Abdelhamid N. [43]	C4.5, PART, RIPPER	Yahoo directory[62], Phishtank[50]	91.26% PART
	Subasi A. [44]	Random Forest	UCI [53], WEKA[63]	97.36%
	Sentürk S. et al. [45]	Decision tree with J48 algorithm	WEKA[63]	89%
Others Artificial Intelligence	Ali W. and Malebary A. S. [46]	PSO-based website feature weighting with BPNN, SVM, kNN, C4.5, RF, and NB	UCI [53]	96.83% for RF
	Feng F. et al. [47]	Monte Carlo algorithm	UCI [53]	97.71%

4. Conclusion

Quite possibly the latest online dangers are phishing, which has made huge misfortunes for online customers, electronic organizations, and monetary establishments. A typical method of phishing is mimicking the websites to bamboozle online clients and take their monetary data. In this review, a precise survey of the latest things in anti-phishing is done and a scientific taxonomy of anti-phishing is suggested depending on the used methods. The review covered the most cited paper in anti-phishing for the last 4 years. Moreover, this paper presented the dataset used by researchers in this field. The investigation is given here will assist the scholarly community and businesses to recognize the recent status of phishing recognition and guide them to discover a novel plan to foster the best web phishing detection strategies.

Acknowledgment

We are grateful for the support from the Department of Computer Science, College of Computer Science and Mathematics, University of Mosul.

References

[1] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, Jan. 2021, doi: 10.1007/s11235-020-00733-2.

[2] "11 Biggest cyber security threats in 2021 | G DATA." https://www.gdatasoftware.com/blog/biggest-security-threats-2021 (accessed Aug. 05, 2021).

[3] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020, doi: 10.1109/ACCESS.2020.2969780.

[4] "How Well Can You Catch a Phish?," *GreatHorn*, Aug. 20, 2020. https://www.greathorn.com/blog/how-well-can-you-catch-a-phish/ (accessed Aug. 06, 2021).

[5] "Phishing: most targeted industries 2021," *Statista*. https://www.statista.com/statistics/266161/websites-most-affected-by-phishing/ (accessed Aug. 07, 2021).

[6] M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 4, pp. 2091–2121, 2013, doi: 10.1109/SURV.2013.032213.00009.

[7] N. Abdelhamid, F. Thabtah, and H. Abdel-jaber, "Phishing detection: A recent intelligent machine learning comparison based on models content and features," in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, Jul. 2017, pp. 72–77. doi: 10.1109/ISI.2017.8004877.
[8] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, Aug. 2014, doi: 10.1007/s00521-013-1490-z.
[9] A. K. Jain and B. B. Gupta, "Phishing Detection: Analysis of Visual Similarity Based Approaches," *Secur. Commun. Netw.*, vol. 2017, pp. 1–20, 2017, doi: 10.1155/2017/5421046.
[10] R. S. Rao and A. R. Pais, "Jail-Phish: An improved search engine based phishing detection system," *Comput. Secur.*, vol. 83, pp. 246–267, Jun. 2019, doi: 10.1016/j.cose.2019.02.011.
[11] A. K. Jain, S. Parashar, P. Katare, and I. Sharma, "PhishSKaPe: A Content based Approach to Escape Phishing Attacks," *Procedia Comput. Sci.*, vol. 171, pp. 1102–1109, 2020, doi: 10.1016/j.procs.2020.04.118.
[12] H. Che, Q. Liu, L. Zou, H. Yang, D. Zhou, and F. Yu, "A Content-Based Phishing Email Detection Method," in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, Prague, Czech Republic, Jul. 2017, pp. 415–422. doi: 10.1109/QRS-C.2017.75.
[13] G. Sonowal and K. S. Kuppusamy, "PhiDMA – A phishing detection model with multi-filter approach," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 1, pp. 99–112, Jan. 2020, doi: 10.1016/j.jksuci.2017.07.005.
[14] A. Mishra and B. B. Gupta, "Intelligent phishing detection system using similarity matching algorithms," *Int. J. Inf. Commun. Technol.*, vol. 12, no. 1/2, p. 51, 2018, doi: 10.1504/IJICT.2018.089022.
[15] R. S. Rao and S. T. Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach," *Procedia Comput. Sci.*, vol. 54, pp. 147–156, 2015, doi: 10.1016/j.procs.2015.06.017.
[16] R. S. Rao, A. R. Pais, and P. Anand, "A heuristic technique to detect phishing websites using TWSVM classifier," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5733–5752, Jun. 2021, doi: 10.1007/s00521-020-05354-z.
[17] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," *Soft Comput.*, vol. 23, no. 12, pp. 4315–4327, Jun. 2019, doi: 10.1007/s00500-018-3084-2.
[18] B. B. Gupta and A. K. Jain, "Phishing Attack Detection using a Search Engine and Heuristics-based Technique:," *J. Inf. Technol. Res.*, vol. 13, no. 2, pp. 94–109, Apr. 2020, doi: 10.4018/JITR.2020040106.
[19] A. Cuzzocrea, F. Martinelli, and F. Mercaldo, "A machine-learning framework for supporting intelligent web-phishing detection and analysis," in *Proceedings of the 23rd International Database Applications & Engineering Symposium on - IDEAS '19*, Athens, Greece, 2019, pp. 1–3. doi: 10.1145/3331076.3331087.
[20] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019, doi: 10.1016/j.ins.2019.01.064.
[21] M. M. Yadollahi, F. Shoeleh, E. Serkani, A. Madani, and H. Gharaee, "An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features," in *2019 5th International Conference on Web Research (ICWR)*, Tehran, Iran, Apr. 2019, pp. 281–286. doi: 10.1109/ICWR.2019.8765265.
[22] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019, doi: 10.1016/j.eswa.2018.09.029.

[23] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," *Decis. Support Syst.*, vol. 107, pp. 88–102, Mar. 2018, doi: 10.1016/j.dss.2018.01.001.
[24] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer, and M. Woźniak, "Accurate and fast URL phishing detector: A convolutional neural network approach," *Comput. Netw.*, vol. 178, p. 107275, Sep. 2020, doi: 10.1016/j.comnet.2020.107275.
[25] K. Gajera, M. Jangid, P. Mehta, and J. Mittal, "A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, Jun. 2019, pp. 196–200. doi: 10.1109/ICECA.2019.8822053.
[26] S. Y. Yerima and M. K. Alzaylaee, "High Accuracy Phishing Detection Based on Convolutional Neural Networks," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, Riyadh, Saudi Arabia, Mar. 2020, pp. 1–6. doi: 10.1109/ICCAIS48893.2020.9096869.
[27] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms," *J. Enterp. Inf. Manag.*, vol. ahead-of-print, no. ahead-of-print, Jun. 2020, doi: 10.1108/JEIM-01-2020-0036.
[28] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, "OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network," *IEEE Access*, vol. 7, pp. 73271–73284, 2019, doi: 10.1109/ACCESS.2019.2920655.
[29] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection," in *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Island of Ulkulhas, Maldives, Aug. 2019, pp. 1–8. doi: 10.1109/SKIMA47702.2019.8982427.
[30] W. Wang, F. Zhang, X. Luo, and S. Zhang, "PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks," *Secur. Commun. Netw.*, vol. 2019, pp. 1–15, Oct. 2019, doi: 10.1155/2019/2595794.
[31] A. Abuzuraiq, M. Alkasassbeh, and M. Almseidin, "Intelligent Methods for Accurately Detecting Phishing Websites," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, Apr. 2020, pp. 085–090. doi: 10.1109/ICICS49469.2020.239509.
[32] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text," *Expert Syst. Appl.*, vol. 115, pp. 300–313, Jan. 2019, doi: 10.1016/j.eswa.2018.07.067.
[33] M. Zabihimayvan and D. Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, New Orleans, LA, USA, Jun. 2019, pp. 1–6. doi: 10.1109/FUZZ-IEEE.2019.8858884.
[34] C. Pham, L. A. T. Nguyen, N. H. Tran, E.-N. Huh, and C. S. Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 15, no. 3, pp. 1076–1089, Sep. 2018, doi: 10.1109/TNSM.2018.2831197.
[35] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting," *IET Inf. Secur.*, vol. 13, no. 6, pp. 659–669, Nov. 2019, doi: 10.1049/iet-ifs.2019.0006.
[36] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, "DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features," *Appl. Soft Comput.*, vol. 95, p. 106505, Oct. 2020, doi: 10.1016/j.asoc.2020.106505.
[37] Muhammad Taseer Suleman and Shahid Mahmood Awan, "Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms," *Autom. Control Comput. Sci.*, vol. 53, no. 4, pp. 333–341, Jul. 2019, doi: 10.3103/S0146411619040102.
[38] G. Vrbančič, I. Fister, and V. Podgorelec, "Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites Classification," in

Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, Novi Sad Serbia, Jun. 2018, pp. 1–8. doi: 10.1145/3227609.3227655.

[39] T. Chin, K. Xiong, and C. Hu, "Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking," *IEEE Access*, vol. 6, pp. 42516–42531, 2018, doi: 10.1109/ACCESS.2018.2837889.

[40] W. Chen, X. A. Wang, W. Zhang, and C. Xu, "Phishing Detection Research Based on PSO-BP Neural Network," in *Advances in Internet, Data & Web Technologies*, vol. 17, L. Barolli, F. Xhafa, N. Javaid, E. Spaho, and V. Kolici, Eds. Cham: Springer International Publishing, 2018, pp. 990–998. doi: 10.1007/978-3-319-75928-9_91.

[41] S. Gupta and A. Singhal, "Phishing URL detection by using artificial neural network with PSO," in *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, Noida, India, Aug. 2017, pp. 1–6. doi: 10.1109/TEL-NET.2017.8343553.

[42] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Kathmandu, Nepal, Dec. 2015, pp. 1–8. doi: 10.1109/SKIMA.2015.7399985.

[43] F. Thabtah and N. Abdelhamid, "Deriving Correlated Sets of Website Features for Phishing Detection: A Computational Intelligence Approach," *J. Inf. Knowl. Manag.*, vol. 15, no. 04, p. 1650042, Dec. 2016, doi: 10.1142/S0219649216500428.

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

[45] S. Senturk, E. Yerli, and I. Sogukpinar, "Email phishing detection and prevention by using data mining techniques," in *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, Oct. 2017, pp. 707–712. doi: 10.1109/UBMK.2017.8093510.

[46] W. Ali and S. Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection," *IEEE Access*, vol. 8, pp. 116766–116780, 2020, doi: 10.1109/ACCESS.2020.3003569.

[47] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Wang, "The application of a novel neural network in the detection of phishing websites," *J. Ambient Intell. Humaniz. Comput.*, Apr. 2018, doi: 10.1007/s12652-018-0786-3.

[48] "Alexa - Top sites." https://www.alexa.com/topsites (accessed Sep. 02, 2021).

[49] "OpenPhish - Phishing Intelligence." https://openphish.com/ (accessed Sep. 02, 2021).

[50] "PhishTank | Join the fight against phishing." https://www.phishtank.com/index.php (accessed Aug. 27, 2021).

[51] "Phishload - Download." https://www.medien.ifi.lmu.de/team/max.maurer/files/phishload/download.html (accessed Aug. 27, 2021).

[52] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing Websites Features," p. 7.

[53] "UCI Machine Learning Repository: Phishing Websites Data Set." https://archive.ics.uci.edu/ml/datasets/phishing+websites (accessed Sep. 02, 2021).

[54] "Yandex.XML — Yandex Technologies." https://yandex.com/dev/xml/ (accessed Sep. 02, 2021).

[55] "Index of /~jose/phishing." https://monkey.org/~jose/phishing/ (accessed Sep. 02, 2021).

[56] "Index of /old/publiccorpus." https://spamassassin.apache.org/old/publiccorpus/ (accessed Sep. 02, 2021).

[57] "Common Crawl." https://commoncrawl.org/ (accessed Sep. 02, 2021).

[58] K. L. Chiew, E. H. Chang, C. L. Tan, J. Abdullah, and K. S. C. Yong, "Building Standard Offline Anti-phishing Dataset for Benchmarking," *Int. J. Eng. Technol.*, vol. 7, no. 4.31, Art. no. 4.31, 2018.

[59] “Phishing Dataset.” https://www.fcsit.unimas.my/phishing-dataset (accessed Sep. 02, 2021).
[60] “UCI Machine Learning Repository: Website Phishing Data Set.” https://archive.ics.uci.edu/ml/datasets/Website+Phishing (accessed Sep. 02, 2021).
[61] “DMOZ - RDF Data.” https://dmoz-odp.org/docs/en/rdf.html (accessed Sep. 02, 2021).
[62] “Yahoo Directory.” http://dir.yahoo.com (accessed Sep. 02, 2021).
[63] “WEKA.” http://www.cs.waikato.ac.nz/ml/weka/ (accessed Sep. 02, 2021).