Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

A Review of the Current Status, Categorization, And Future Prospects of Web Phishing Detection Methods

Nana Yaw Ofosu Apea, Chunyong Yin
School of Computer and Software
Nanjing University of Information Science & Technology
No. 219 Ningliu Road, Jiangsu Province, China, 210044

Abstract: More than half of the world's population now lives online thanks to the internet. Unfortunately, as online transactions increase, so do cybercrimes, which are on the rise. Due to the anonymity of the internet, hackers try to trick end users by using phishing, malware, SQL injection, man-in-the-middle attacks, domain name system tunneling, ransomware, web trojans, and other techniques. Phishing is the most cunning assault among them because it preys on end-user weaknesses. Phishing frequently uses emails and rogue websites to trick users by seeming to be a reliable business. Many anti-phishing strategies have been proposed by security experts. There is currently no single solution that can eliminate all vulnerabilities. A taxonomy of automated web phishing detection is offered after a thorough analysis of current trends in web phishing detection approaches. This study's goal is to review the state of the art in automated web phishing detection research and assess its effectiveness. The research directions for further study are also discussed in this work.

Keywords–: Phishing; Web phishing; Taxonomy; Survey; Zero-hour

I. INTRODUCTION

The internet has revolutionized the world we live in today due to its flexibility and versatility. People can perform financial transactions such as banking and purchasing from any location and at any time, thanks to the advanced infrastructure provided by the internet. However, the internet also presents a unique set of security and privacy issues. Attackers take advantage of the internet's independent and anonymous structure to conduct cyber-attacks such as phishing, virus distribution, and privacy breaches, which pose serious threats to its users [1].

Phishing, the most prevalent and harmful cybercrime targeting individuals rather than computers, is an attack form that aims to deceive and steal sensitive information through social engineering tactics and technical gimmicks [2]. Phishing attackers employ fraudulent emails and malicious websites to trick users into disclosing their usernames, passwords, bank account details, credit card information, and more. The motivations behind phishing attacks vary, ranging from financial gain and identity theft to damaging business reputations or seeking notoriety [4]. Coined in the mid-1990s, the term "phishing" originated from its resemblance to luring unsuspecting individuals into a trap [5]. Initially emerging on America Online (AOL) in the early 1990s, phishing scams involved the creation of fake AOL accounts using fraudulent credit card details [6]. Over time, attackers turned phishing into a lucrative industry, impersonating various entities such as banks, credit card companies, online payment service providers (e.g., InstaReM), and social media platforms (e.g., Facebook). Notably, between 2013 and 2015, email phishing scams generated over $100 million from major internet players like Google and Facebook [7]. The Federal Bureau of Investigation is currently investigating a phishing attack that cost a Texas school district $2.3 million in 2020 [8]. According to Vade Secure, a leader in predictive email defense, over 550 million emails were exchanged in the first quarter of 2018 [9]. The APWG's phishing attack trend reports from the past three years [10-18] were analyzed, and Table 1 provides a summary. According to the APWG's Phishing Attack Trends Report [17] published in the second quarter of 2019, software as a service (SaaS) and webmail sites were still the top targets of phishing. Additionally, 55% of phishing attacks detected in the second quarter of 2019 used HTTPS encryption and secure sockets layer (SSL) certificates to deceive internet users.

According to the most recent APWG report [18], phishing attacks increased by 46% in the third quarter of 2019 compared to the second quarter. The table depicts current trends in phishing attacks from the first quarter of 2017 to the third quarter of 2019 in terms of the "number of detected unique phishing websites," the "number of detected unique phishing emails," the "top country hosting phishing sites," the "most targeted industry sectors," and the "most targeted top-level domain (TLD)." The table indicates that phishing attacks are becoming more dangerous every year. Their increasing numbers make it difficult to identify and block phishing websites daily. Despite the efforts of researchers, security professionals, and the industry to decrease phishing attacks, their rate has steadily increased for over 20 years.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

Consequently, phishing detection is still a popular topic for researchers, academics, business, and security professionals [1].

A. *Survey Methodology*

Each indexing database was subjected to a thorough, methodical search. The most recent information on web phishing detection was gathered. Based on the methodologies, the papers were categorized. Deep scanning was used to extract a taxonomy from the classified documents. The contributions described in this survey are comprehensive and include every recent advancement in this field.

B. *Research Contributions*

This paper's contribution can be summed up as follows:

- The presentation of a taxonomy of web phishing detecting methods.
- A thorough analysis of the most recent methods for detecting web phishing is provided.
- Based on evaluation metrics, a consistent comparison of the state-of-the-art systems is offered.
- Future research options are suggested by the identification and presentation of limitations in the most advanced detection methods.
- The analysis presented in this research will assist academia and business in determining the most effective anti-phishing method.

The remainder of the paper is organized as follows: The relevant research from this survey is examined in Section II. The taxonomy of phishing detection methods is presented in Section III. The categorization performance of various current anti-phishing solutions is assessed in Section IV.

II. RELATED WORK

Several surveys have been undertaken to assess the effectiveness of various phishing tactics and detection technologies. Workman [19] used user behaviour theories to conduct a theory-based examination of social engineering threats. According to the findings, successful social engineering attacks are more dependent on strong normative commitment, high continuing commitment, and high emotional commitment. This study, however, does not look at the links between individual factors and the effects of social engineering. To demonstrate the impact of the phishing attack, Wang et al. [20] sent actual spear-phishing emails to selected victims. According to the findings of this poll, phishing knowledge, sensitivity to visceral triggers, and awareness of phishing deception flags all have a substantial impact on phishing detection. Nonetheless, this report only looks at spear-phishing attacks and is based on a single round of surveys. Nevertheless, this study only examines spear-phishing attacks and relies on a single-round survey and a cognitive-effort measure. Alsharnouby et al. [21] conducted an empirical investigation against phishing attempts by providing users with access to browser security indicators to increase their awareness. After the study, it was discovered that participants had an average success rate of 53% in identifying phishing web-sites, which is equivalent to a random guess even in a controlled environment.

Khonji et al. [22] presented a survey on phishing mitigation strategies. When measuring the effectiveness of current detection approaches, this study does not take into account all of the relevant evaluation indicators, such as language independence, third-party independence, and zero-hour attack detection. Varshney et al. [23] con-ducted a survey on web phishing detection strategies, focusing primarily on detection methods and discussing assessment metrics such as language independence, third-party independence, and zero-hour attack detection. The study discovered that search engine-based detection methods are preferred, although they have a high false-positive rate (FPR) for zero-day life duration webpages. This survey did not ex-amine hybrid approaches of web phishing detection. Tewari et al. [24] provided an overview of various defenses against phishing efforts, although this publication just provides a synopsis of current web anti-phishing approaches and does not discuss all of the detecting techniques that can be employed to mitigate the threat.

While there are many surveys in the literature, to the best of our knowledge, no paper describes in detail all the web phishing detection methods currently in use. Most previous surveys only offer concept-oriented descriptions, but this study aims to cover all current methods for detecting web phishing and rates each method using a set of accepted assessment measures. This study also suggests a taxonomy for web phishing detection methods, which directs how the survey is presented. This thorough study will aid in understanding the benefits and downsides of the most recent developments in web phishing detection.

III. TAXONOMY OF WEB PHISHING DETECTION

User education and automated web phishing detection are the two methods for safeguarding end users against web phishing. End-user training teaches the user how to verify the authenticity of reliable websites. However, as the attackers swiftly devise new methods to seduce the users, end-user training is insufficient to safeguard the users. As a result, many academics suggested automated web phishing detection methods to shield end users from web phishing attacks. Phishing detection is automated using automated web phishing detection algorithms, with no human involvement. The taxonomy of automated web phishing detection methods shown in Fig. 1 is the result of a thorough analysis of the literature that has already been published. Based on the input parameters they employed, all web phishing detection methods in the literature were categorized. Based on the methods used, the approaches from the sub-class were further categorized.

Based on the input parameters, the automated web

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

phishing detection methods are divided into the following three groups.

- Web address-based assessment
- Evaluation based on similarity of content on websites
- Hybrid strategy

The URL is examined in web address-based evaluation techniques to identify web phishing. In the case of a web page, content/similarity-based assessment systems evaluate the web page contents, such as text features, HTML features, CSS features, and hyperlink characteristics, to determine the authenticity of the website. Web page content/similarity-based evaluation techniques and web address-based evaluation schemes are combined in hybrid approaches.

A. Web Address-Based Evaluation

The features of the URL are used in web address-based evaluation to assess the legitimacy of the websites. The protocol, domain name, domain extension, path, and file name make up a web address, also known as a URL (for example, https://www.google.com/). The majority of the time, hackers attempt to pass off phishing websites as legitimate ones by altering one or more characters in the web address (for example, https://www.paypal.com/in/ versus https://wwwpaypal.com/). The legitimate and fraudulent login user interfaces (LUIs) for the PayPal website are shown in Figs. 2a and 2b.

Due to the ease of obtaining a domain name these days, web address attacks are the most prevalent type of web phishing attack. Therefore, it is imperative to evaluate online addresses to lessen the impact of web phishing. This is done by comparing the URL and its components to the relevant authentic website. Based on the approaches used for URL classification, the three types of web address-based evaluation schemes are as follows.

- Approaches for list-based detection.
- Heuristic rule-based methods for detection.
- Strategies for detection based on learning.

List-Based Detection Techniques

A list-based detection safeguards the user from an online phishing scam by determining the validity of the URL using lists.

In this method, a list, or database, of URLs is kept up to date. Typically, it stores keywords, internet protocol (IP) addresses, and URLs. A whitelist, or collection of trustworthy URLs, is something that some researchers keep up with. The majority of researchers advise keeping a blacklist, which is a list of harmful URLs. This method compares the provided URL to the whitelist and blacklist and then takes the appropriate action. When using a whitelist technique, access is only given to the URLs that are on the whitelist, however when using a blacklist approach, access is given to any URL that isn't on the blacklist. Whenever necessary, steps are taken to update the list. The process flow for whitelist-based web phishing detection strategies is depicted in Fig. 3. The associated URL is compared against the system's local whitelist each time the user enters a website. Access to the website is permitted if the URL is included in the list; otherwise, it is processed further to verify its legitimacy or access is prohibited by a warning message.

To guard against phishing fraud, Google Safe Browsing [29] is a popular blacklist for online browsers that are installed in Google, Firefox, Safari, Vivaldi, and GNOME. Opera 9.1 makes use of PhishTank [30], an online database with millions of active malicious web addresses that are thought to be phishing. To avoid falling victim to phishing scams, many research studies have been carried out employing local whitelists and blacklists. Even though the list-based approach finds malicious URLs faster, the scheme's detection rate is lower than that of other schemes. This is a result of infrequent updates to the blacklists. In a specific phishing dataset, it was discovered that roughly 63% of the URLs persisted for less than two hours, while 43–83% of the URLs changed in the blacklist did so after twelve hours [31]. Therefore, using a blacklist-based strategy alone to defend against zero-day attacks is ineffective. To solve this problem, ML is used in conjunction with blacklists and whitelists to automatically identify harmful websites that are not included in the list. An automatic individual whitelist (AIWL)-based technique was suggested by Cao et al. [32] and keeps a local list of websites with the user's familiar login user interface (LUI) to warn the user whenever he tries to enter a new website with LUI. AIWL utilizes a naive Bayesian classifier to automatically add new websites to the list based on how frequently users access them. This strategy, however, is ineffective against viruses and trojan horses that affect only local computers.

To propose the auto-updated whitelist, Jain and Gupta [32] integrated the whitelist approach with heuristics and ML. In this study, a local whitelist is kept for initial filtering, and heuristics and ML algorithms are used to further process web pages that aren't on the whitelist. Later, to cut down on processing time wasted on pre-processing, feature extraction, and other activities, blacklists and whitelists are employed as filtering modules in many web phishing detection systems.

Accordingly, in recent research projects, this list-based detection method serves as a filtering mechanism to exclude dubious web pages before they can be detected.

Heuristic Rule-Based Methods for Detection

To ensure security, heuristic rule-based detection methods apply thumb rules to the suspected URL. Most of the time, a user cannot tell the difference between a valid and a phishing site address due to their minute variances. Attackers make every effort to create malicious websites that leave no room for the user to question their legitimacy. Web phishing has been carried out using a variety of techniques, including long URLs, links carrying the @symbol in the URL, URL alteration, false SSL or HTTPS, pop-up windows, redirect pages, website traffic, attaching IP addresses to URLs obscuring the links, and more [34]. Certain requirements found in a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

real web address can aid in spotting bogus websites. Heuristics developed based on these standards are applied by heuristic rule-based web address evaluation approaches to verify the validity of the suspected URL. Here are a few of them:

- To show the secure version of HTTP, the protocol should be HTTPS.
- Only two dots should be used to separate the domain name from the website address (for example, www.google.com).
- Most businesses (.com), organizations (.org), educational institutions (.edu), and nations (.in, uk, etc.) should use.com as their top-level domain name.
- Verification of SSL certificates, etc.

The process of the heuristic rule-based web address evaluation approach is shown in Fig. 4. Heuristic criteria are applied to the URL of a suspicious website based on the standards of the standard web address to assess the legitimacy of the website.

To identify URL-based web phishing attempts, Sahingoz et al. [35] used heuristics to extract natural language processing (NLP) properties from the URL. The heuristics are derived based on variables like the number of raw words, the number of short words, Alexa ranking, the number of comparable brand names, etc. Using various heuristics, Yukun Li et al. [36] checked the URL for anomalies including sensitive words, suspicious symbols (such @, _), https, URL length information, and the number of dots in the domain name. Rajsingh and Jeeva. [37] computed 14 heuristics, including the length of the host URL, the number of slashes in the URL, the number of terms in the hostname of the URL, special characters, IP addresses, Unicode in the URL, transport layer security, subdomains, specific keywords in the URL, top-level domains, the number of dots in the URL's path, the number of hyphens in the hostname of the URL, and the number of terms in the hostname of the URL. The associative rule mining algorithms are then provided with the extracted features. When a user accesses a website, a lightweight phish detector suggested by Varshney et al. [38] collects the domain name of the URL and the title of the webpage. To confirm the validity, a search engine is used to look up the retrieved URL domain name and the title page.

Heuristic rule-based algorithms can detect zero-day assaults, in contrast to list-based phishing detection schemes. Compared to list-based phishing detection techniques, it has a higher detection rate. On the other hand, the heuristics used completely determine how well the technique performs and is accurate.

Strategies for Detection Based On Learning

Artificial intelligence is now being used in practically all study fields. Most recent efforts to identify web phishing have used learning algorithms.

Based on the information collected from the URL, learning methods such as ML and deep learning are utilized to detect the attacks. In learning-based web phishing detection, statistical and NLP information from URLs is retrieved and fed into machine learning (ML) methods like support vector machines (SVM), decision trees, naive Bayes algorithms, random forests, etc. for additional classification. Fig. 5 illustrates how learning-based web phishing detection works. Based on the inference drawn from the training data, the classifier builds a model. The classifier's model is used to evaluate the suspicious URL.

Sahingoz et al. [35] tested the performance of seven different machine learning (ML) algorithms on the features extracted from the URL, including naive Bayes, random forest, k-nearest neighbour (k-NN), Adaboost, K-star, sequential minimum optimization, and decision tree. A deep learning strategy was put forth by Yang et al. [39] to organically extract information from URLs and identify web phishing attacks. In this method, the long short-term memory (LSTM) network is utilized to learn the sequential dependency from the character sequences, and the convolutional neural network (CNN) is used to extract the correlation features. A neural network-based method for detecting web phishing was proposed by Zhu et al. [40].

ML algorithms offer a quicker detection time and can detect zero-day attacks. The performance of this strategy, however, is feature-sensitive and depends on the features of the ML algorithm used.

B. Evaluation Based on Similarity of Content on Websites

Phishing can also be carried out by webpage spoofing in addition to URLs. Webpage spoofing is the practice of building fake websites that imitate the look and feel of legitimate websites to deceive users. A faked website will typically mimic the typefaces, layout, photos, and logos to make the site appear as genuine as feasible. By extracting several features from the webpage, itself, webpage content or similarity evaluation is carried out to find this. Content-based detection techniques and layout-based detection schemes are two categories under which web page content/similarity-based evaluation can be categorized. While the layout of the website is taken into account in layout-based detection systems, the content of the web page is the primary criterion used to categorize phishing websites in content-based detection schemes. The earlier content-based technique measured a webpage's accessibility by extracting keywords from a suspicious webpage and feeding them into the search engine [41].

The current method of evaluating page similarity involves applying heuristics or machine learning (ML) techniques to the web page's retrieved HTML and CSS data. The following two categories are used to present the web page content/similarity-based phishing detection algorithms.

- Calculation of website similarity using heuristic rules.
- Comparison of websites using machine learning.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

Evaluation of Webpage Similarity Using Heuristic Rules

To enable a protected environment against phishing schemes, keywords and attributes are retrieved from the suspect webpage and compared to the intended webpage utilizing search techniques in heuristic-based webpage similarity calculations. The HTML features that were extracted include things like the quantity of internal and external links, empty links, login forms, alarm windows, redirections, hidden/restricted information, consistency between title and URL brands, consistency between most frequent link brands and URL brands, internal and external resources, the number of times the URL brand name appears in HTML, and more [36]. includes CSS characteristics like the colour property, padding concerning the paragraph's element, font size, border, font family, margin, etc. [42].

Tan et al. [43] proposed a phishing webpage detection method comprising four modules: identity keywords extraction, search engine lookup, target domain name finder, and three-tier identity matching. The extraction of identity keywords is accomplished using a weighted URL token system based on the N-gram model. These keywords are then utilized in a search engine to identify the target domain name. The compromise programming technique is employed to extract the target domain name from the search results by considering identity-related factors such as keyword density, domain name frequency in search results, and domain name frequency on a query web page. The three-tier identity matching system, consisting of full string matching, country code TLD matching, and IP alias matching modules, is subsequently utilized to analyze the query webpage's status based on the input of the target domain name and the actual domain name.

By separating the CSS features from the web page's underlying design, Mao et al. [42] proposed a phishing alarm. To categorize the web pages, page similarity computations are used in conjunction with the extracted features. A client-side phishing detection program created by Marchal et al. [44] provides improved privacy, real-time protection, useful alerts, and resistance to dynamic phishing. This method detects phishing websites using target identification methods and phish detectors.

Comparison of Websites Using Machine Learning

Later, various researchers suggested ML-based methods for detecting web page similarity. This method extracts HTML, XML, JS, and CSS features from the webpage's source code and feeds them to machine learning (ML) algorithms for additional classification. The workflow of ML-based webpage similarity evaluation is shown in Fig. 7.

A content-based approach to detecting web phishing was proposed by Xiang et al. [45] by extracting features from URLs (embedded domain, IP address, number of dots in the URL, suspicious URL, number of sensitive words in the URL, and out-of-position TLD), HTML (bad forms, bad action fields, non-matching URLs, and out-of-position brand name), and web (age of the domain, page in top search results, page rank and to decrease false positives (FP) and shorten run times, two filters are used. a phishing detector that utilizes hashing to detect phish that is nearly identical. The second is a filter that only allows access to websites containing login forms. Two methods—randomized evaluation and time-based evaluation utilizing a Bayesian network—are used to assess the proposed strategy. Mao et al.'s [46] proposal used ML techniques to determine layout similarities based on learning. To categorize the similarity of the webpages, SVM and decision trees are employed. A brand-new feature selection approach for ML-based phishing detection systems was put forth by Chiew et al. [47]. To produce a small set of primary features, a novel cumulative distribution function gradient algorithm is created as an automatic feature cut-off rank identifier. These primary features are then subjected to data perturbation and function perturbation techniques to produce the hybrid ensemble features. To evaluate the effectiveness of the feature set, classifiers such as SVM, random forest, naive Bayes, C4.5, JRip, and PART are utilized. By removing hyperlinks from websites, Jain and Gupta [48] proposed a novel web phishing detection approach. The suggested method has extracted 12 specific hyperlink features, including the total feature, no feature, internal and external hyperlinks, null hyperlinks, internal and external CSS, internal and external redirection, internal and external error, login form link, internal and external favicons, and internal and external error. The collected characteristics are then used as input for machine learning (ML) techniques such as logistic regression, naive Bayes, random forest, SVM, Adaboost, and neural networks. All ML algorithms' performance was measured and reported.

C. Hybrid Methods

Recent studies have shown that a hybrid approach performs better for web phishing detection, which involves merging current strategies. Gowtham and Krishnamurthi [49] proposed a method that uses a pre-approved site identifier, login form finder, and SVM ML algorithms. Li et al. [36] utilized URL features, HTML source code features, and HTML string embedding, as well as a stacking model of gradient-boosted decision trees to boost system speed. Yang et al. [39] presented a hybrid strategy that merged the approaches of URL evaluation, web page similarity method, and content-based approach. Rao and Pais [50] proposed a two-level filtering technique, and Li and Wang [51] proposed the PhishBox approach, which involves an ensemble model and active learning. All these methods use various algorithms, including XGBOOST, random forest, extra tree, and CNN-LSTM, to detect and identify web phishing attacks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

D. Figures, Tables and Schemes

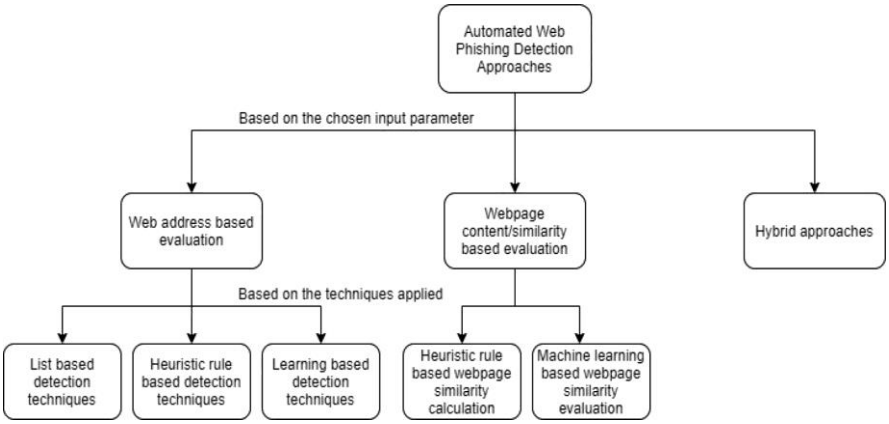

Figure 1: Web phishing detection taxonomy [40].

Table 1: Displays trends in phishing attacks.

Parameters	1H2017	3Q2017	4Q2017	1Q2018	2Q2018	3Q2018	4Q2018	1Q2019	2Q2019	3Q2019
no. of detected unique phishing websites	291,096	190,942	180,757	263,538	233,040	151,014	138,328	180,768	182,465	266,378
no. of detected unique phishing emails	592,335	296,208	233,613	262,704	264,483	270,557	239,910	112,393	112,163	118,260
top country hosting phishing websites	US	US	US	US	US	US	US	NA	NA	NA
most targeted industry sectors	Payment 45%	Payment 41.99%	Payment 42%	Payment 39.4%	Payment 36%	Payment 38.2%	Payment 33%	SaaS/web mail 36%	SaaS/web mail 36%	SaaS/web mail 33%
most targeted TLD	Legacy gTLDs	.COM 53%	NA	.COM 48.6%	NA	.COM	.COM	NA	.COM	.COM

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

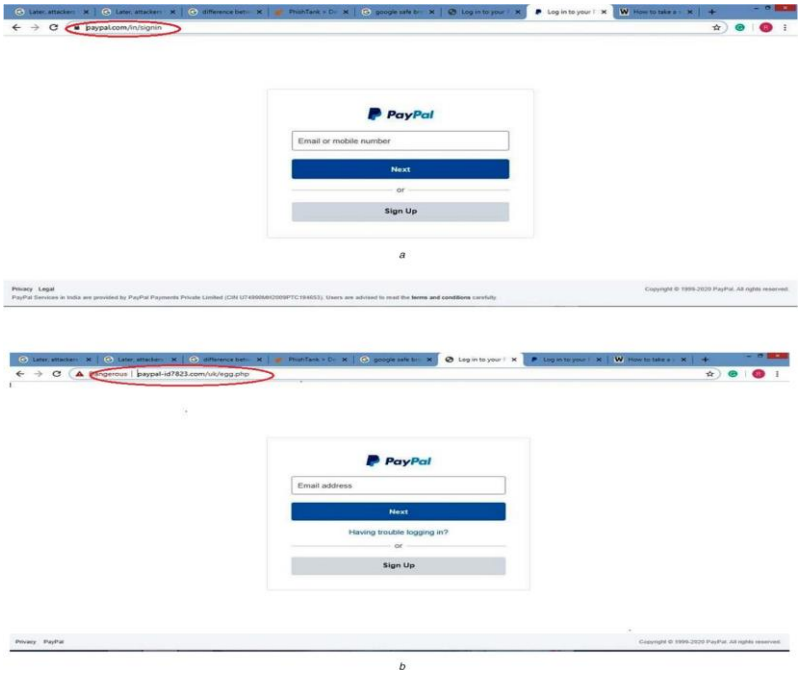

Figure 2: An example of phishing. (a) LUI for the PayPal official website, (b) LUI for the PayPal fake website [55]

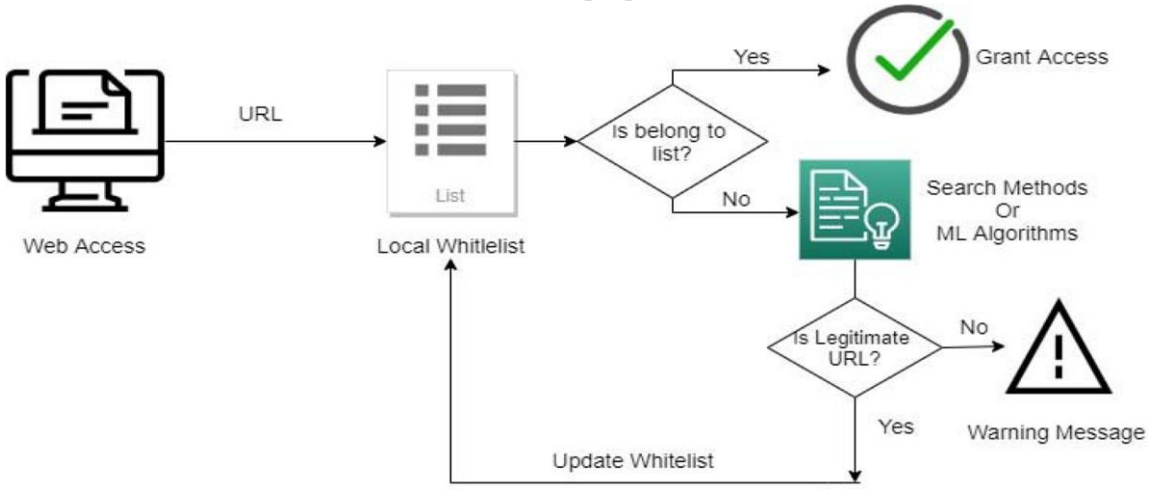

Figure 3: Web phishing detection with whitelists [1]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

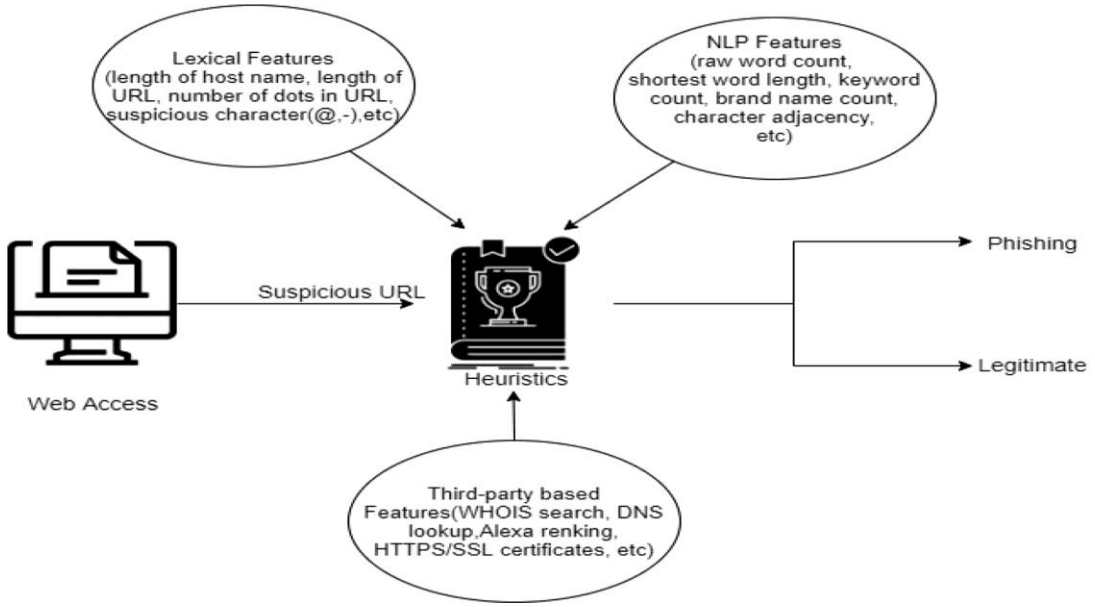

Figure 4: Web phishing detection using heuristics [55]

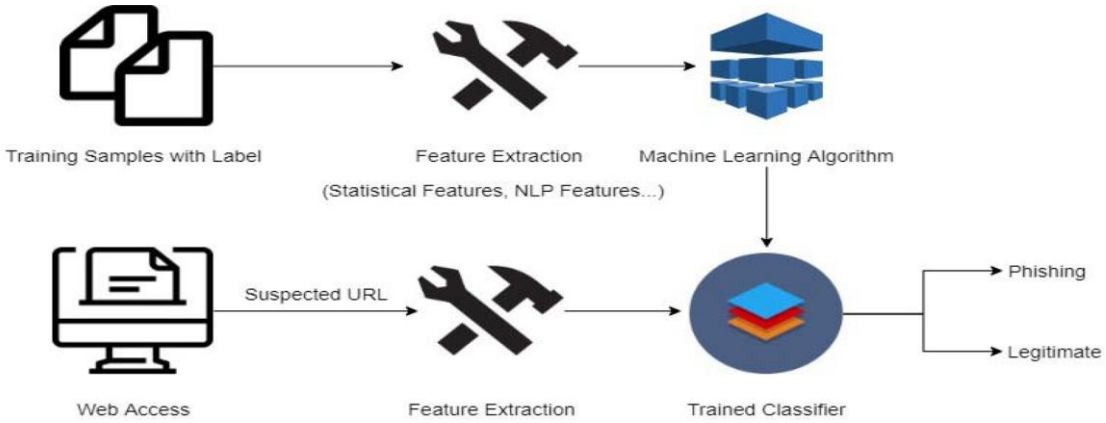

Figure 5: Detection of web phishing using learning [40]

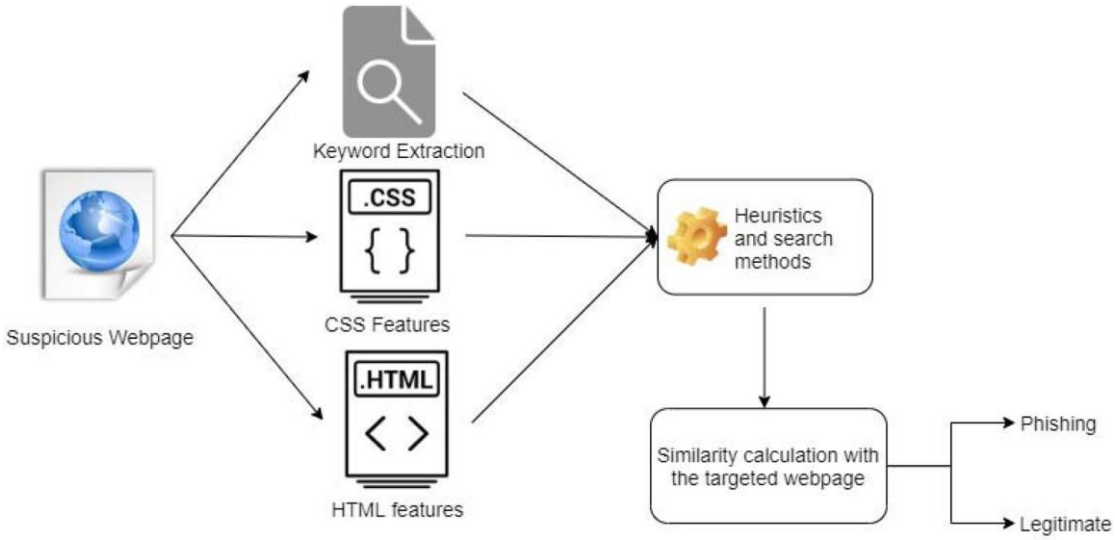

Figure 6: Calculation of website similarity using heuristic rules [55]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

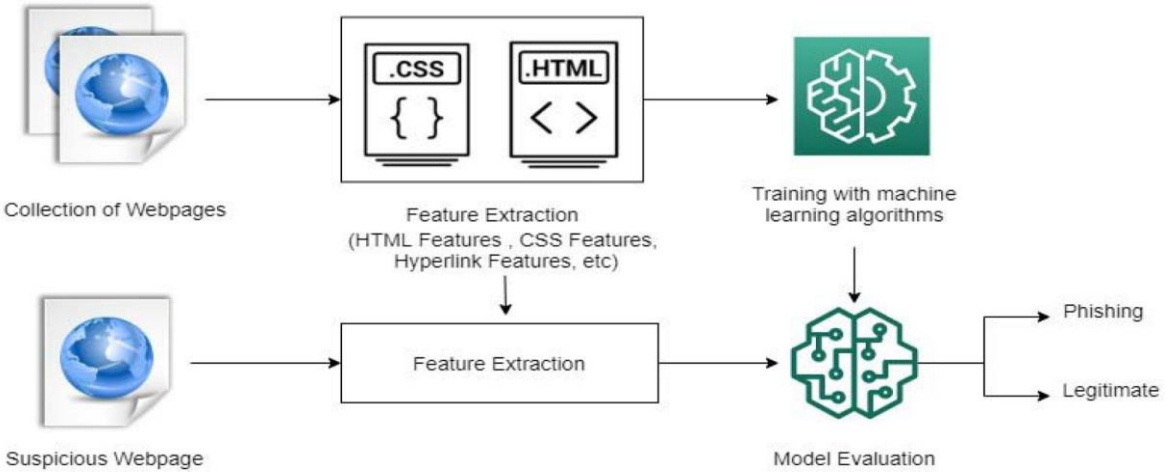

Figure 7: Comparison of websites using machine learning [40]

IV. PERFORMANCE EVALUATION

Modern methods for detecting web phishing were gathered and carefully examined. These detection methods' effectiveness is examined and provided. The detection methods were divided into three categories based on the taxonomy that was derived: web address-based evaluation, webpage content/similarity-based evaluation, and hybrid approach. The study was carried out using a framework that was being developed and includes common phishing detection metrics. All the most recent research on automated phishing has been organized, rated, and presented by category.

The following evaluation measures were used to assess the methods falling into each of these categories:

- Methods employed.
- Detection of an attack at zero hour.
- Independence from language.
- Independent services from third parties.
- Utilized dataset.
- The classification results: False-negative rate (FNR), accuracy, True-positive rate (TPR).
- Limitations.

The most recent study was thoroughly examined and provided in Tables 2-4. The effectiveness of measures for detecting URL-based web phishing is shown in Table 2. The effectiveness of webpage content-based web phishing detection methods is shown in Table 3. Table 4 summarizes the effectiveness of hybrid techniques.

A. Methods Employed

Web phishing detection systems can be classified into three categories: list-based schemes, heuristics-based schemes, and learning-based schemes. List-based techniques rely on predefined lists for phishing detection. Heuristics-based systems utilize heuristics and search techniques to identify phishing attempts. Learning-based methods employ machine learning algorithms to classify phishing websites. By using this metric, one can explore the techniques employed in the literature for phishing detection.

B. Detection of an Attack at Zero-Hour

Zero-hour threats, which exploit vulnerabilities in a system even before their creators become aware of them, pose a significant risk. Detecting such threats requires a faster approach, as fraudulent websites associated with zero-hour attacks often exist for only 2-4 hours. Relying solely on blacklist/whitelist approaches is inadequate for effectively detecting these attacks due to their low detection rates. To combat zero-hour attacks, current research efforts combine blacklists and whitelists with other detection methods, such as heuristics and machine learning.

C. Language Dependency

Among the most popular websites on the internet, besides English, languages such as Russian, German, Spanish, French, Japanese, Portuguese, Italian, and Persian are also available in this multilingual cyberspace [54]. The prevalence of websites offering multiple languages is rapidly

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

increasing [55]. To assess the effectiveness of web phishing detection methods, various metrics can be employed, including language independence. Linguistic independence ensures the ability to identify web phishing regardless of the language used. Two types of web page similarity detection methods exist: content-based and layout-based. However, the majority of content-based techniques do not support language independence. On the other hand, layout-based approaches provide support for language independence. Among all the contemporary web anti-phishing techniques, Varshney et al. [38] and Yang et al. [39] are identified as language-dependent, as indicated in the table.

D. *Independent Services From Third Parties*

A third party is a person who acts on behalf of another system, entity, or organization. Web phishing detection uses some third-party services, such as DNS lookup, WHOIS search, search engine evaluation, SSL certificate verification, etc. It's crucial to create an anti-web phishing system that is independent of outside services if you want to attain real-time performance. To increase the detection rate, a few strategies used in [33, 39, 40, 43, 45, 49] rely on other services.

E. *Dataset Used*

There are many datasets available for detecting web phishing. We may obtain the benchmark datasets from various sources, including Kaggle and the ML repository at the University of California, Irvine (UCI). However, the majority of recent studies have made use of real-world data that was gathered from online phishing databases like PhishTank, Phishload, OpenPhish, and others. The most popular resource for online communities is PhishTank, which provides millions of verified and active phishing URLs for research.

F. *Classification Outcome*

Binary classification and multiclass classification were the two categories used to classify classification procedures. The most popular classification method in machine learning (ML) is binary classification, which predicts the outcome of an unknown dataset as either true or false. The confusion matrix, the output matrix for any binary classifier, has four outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- TP - number of samples that were correctly identified as phishing websites
- TN - the percentage of accurately identified samples from reliable websites
- FP - the number of phishing websites that were wrongly identified as samples
- FN - number of false positives for samples misidentified as valid websites

Accuracy, precision, TPR, and FNR are used to assess the classification performance of modern web phishing detection methods.

Accuracy

Accuracy refers to the number of samples that were successfully predicted (TP + TN) divided by the total number of samples that were forecasted (TP + FP + TN + FN) to determine the model's accuracy. Equation (1) shows the mathematical calculation of Accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

True Positive Rate (TPR)

As shown in Equation (2), TPR is characterized as the ratio of the number of correctly predicted phishing samples (TP) to the total number of actual phishing samples (TP + FN).

$$\text{TPR} = \frac{TP}{TP+FN} \quad (2)$$

False Negative Rate (FNR)

Equation (3) also shows the FNR, which is defined as the percentage of legitimate samples (FN) that were mistakenly predicted out of all the actual phishing samples (TP + FN).

$$\text{FNR} = \frac{FN}{FN+TP} \quad (3)$$

G. *Limitations*

To evaluate the resilience of cutting-edge web phishing detection systems against phishing attacks, an investigation is conducted on their detection time, detection rate, and storage complexity. However, the majority of current web phishing detection methods face a drawback in feature selection mechanisms as they heavily rely on custom features for attack identification.

With the aforementioned parameters in mind, extensive research is conducted on automated phishing, focusing on the most up-to-date discoveries in URL-based web phishing detection. Various researchers' approaches are examined, assessed, and summarized in Table 2.

Table 3 presents a metrics-based analysis of the recent advancements in phishing detection, specifically focusing on webpage similarity

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

evaluation.

Furthermore, Table 4 provides an analysis and display of the effectiveness of hybrid approaches in combating phishing attacks.

4.8 Tables

Table 2(a): Evaluation of the effectiveness of URL-based web phishing detection

Method	Dataset	Techniques Used	Attacks Detection	Independency
[33]	Alexa, StuffGate, PhishTank	List, Search and Heuristics	Zero-Hour	Language
[52]	PhishTank	List, Search and Heuristics	Zero-Hour	Third-party and language
[38]	PhishTank and Alexa	Search and Heuristics	Zero-Hour	Third-party
[37]	StuffGate, PhishTank, Alexa	Heuristics and ML	Zero-Hour	Third-party and language
[44]	PhishTank	List, Search and Heuristics	Zero-Hour	Third-party and language
[53]	PhishTank and Alexa	Heuristics and ML	Zero-Hour	Third-party and language
[35]	PhishTank	Heuristics and ML	Zero-Hour	Third-party and language
[40]	PhishTank and Alexa	Heuristics and ML	Zero-Hour	Language

Table 2(b): Evaluation of classification outcomes and limitations of URL-based web phishing detection

Method	TPR	FNR	ACC	Limitations
[33]	86.02	1.48	89.38	The threshold value is explicitly chosen
[52]	97.2	0.043	96.57	Phish shields ignore phishing websites.
[38]	99.5	0.02	95.95	The system is influenced by the T search keyword.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

[37]	95	0.07	93	The number of used instances affects how difficult the computation is.
[44]	95.1	0.45	99.9	The computation of HTML source code features might not be precise.
[53]	98.3	2.6	93.98	Computational Cost
[35]	98.10	0.89	97.98	Most URLs with just a single domain name cannot be recognized by the NLP-based characteristics.
[40]	98.7	1.86	99.3	Feature-Sensitive collection is required to execute OFS on more features.

Table 3(a): Evaluation of the similarity of webpage performance

Method	Dataset	Techniques Used	Attacks Detection	Independence
[45]	PhishTank and Alexia	Heuristic and ML	Zero-Hour	None
[43]	PhishTank OpenPhish and Alexia	Search and Heuristics	Zero-Hour	Language
[48]	Alexa, StuffGate and PhishTank	Heuristics and ML	Zero-Hour	Language and Third-Party
[42]	PhishTank	List, Search and Heuristics	Zero-Hour	Language and Third-Party

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

[54]	PhishTank	Heuristics and ML	Zero-Hour	Language and Third-Party
[47]	UCI	Heuristics and ML	Zero-Hour	Language and Third-Party

Table 3(b): Evaluation of classification of outcomes and limitations of webpage performance

Method	TPR	FNR	ACC	Limitations
[45]	97.1	0.47	95.9	Sensitive to cross-site scripting (XSS), and impractical when phishing URLs contain images.
[43]	99.68	0.32	96.10	Ineffective against cloning, pharming and DNS poisoning techniques
[48]	98.42	1.61	98.42	Misclassifies pages containing embedded objects and depends on web page source code
[42]	97.92	0.54	98.02	The effectiveness of the method is impacted by the similarity rates of websites.
[54]	95.7	0.67	93	Requires a large dataset for high-accuracy results.
[47]	95.2	0.76	94.6	The suggested feature selection framework requires more

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

				computing than current methods.

Table 4(a): Performance monitoring of a hybrid strategy

Method	Dataset	Techniques Used	Attacks Detection	Independence
[48]	PhishTank, OpenPhish and Alexia	List, Heuristics and ML	Zero-Hour	Language and Third-Party
[55]	PhishTank	Heuristics and ML	Zero-Hour	Language and Third-Party
[51]	Alexa and PhishTank	Heuristics and ML	Zero-Hour	Language and Third-Party
[36]	URL and HTML Codes	Heuristics and ML	Zero-Hour	Language and Third-Party
[39]	PhishTank	Heuristics and ML	Zero-Hour	Language
[50]	Google, PhishTank	List, Heuristics and ML	Zero-Hour	Language and Third Party

Table 4(b): Evaluation of classification outcomes and limitations of a hybrid strategy

Method	TPR	FNR	ACC	Limitations
[48]	99.6	0.34	99	The reliability of the login form finder module is purely reliant on the extracted keywords.
[55]	99.14	0.86	99.65	Websites that employ a combination of photos for phishing cannot be identified.
[51]	92.9	0.4	95	Low rate of detection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

[36]	97.7	1.54	97.3	This method is unable to learn how to embed any new HTML string that has never been used in the training corpus.
[39]	98.6	0.69	98.99	High detection time is caused by multi-dimensional feature selection, and the threshold value is established through trial and error.
[50]	98.5	1.48	98.72	When there is little similarity between the web pages, the FPR rises.

V. DISCUSSION

Various methods can be employed to detect web phishing attacks based on the sources utilized to access websites or web pages. However, these methods often suffer from low detection accuracy and high false alarm rates, particularly when attackers employ innovative phishing strategies [2]. Techniques relying on blacklist and whitelist mechanisms, which require regular updates, fail to reliably identify emerging phishing attacks. Meanwhile, visual similarity approaches encounter challenges of computational complexity and spatial constraints. To overcome these issues and achieve real-time performance, the development of lightweight similarity-based phishing detection algorithms becomes imperative [2].

Recent advancements in web phishing detection heavily rely on Machine Learning (ML) techniques. However, the heuristics employed as features in these algorithms solely depend on site addresses or webpages. The primary objective of all anti-phishing measures is to minimize the impact of phishing attacks. To achieve this goal, recent research has integrated diverse methodologies to create lightweight web phishing detection techniques that offer high detection rates, independence from external entities, zero-hour detection, and language independence. Therefore, the utilization of hybrid approaches emerges as the preferred alternative for effectively combating modern phishing scams [2].

Furthermore, there exists an opportunity for further research in web phishing detection, specifically exploring deep learning-based approaches that focus on webpage content or similarity-based phishing detection. Consequently, lightweight and efficient techniques incorporating hybrid approaches and integrating deep learning exhibit promise for the future of web phishing detection.

VI. CONCLUSION

In this study, current trends in web phishing detection are systematically reviewed, and based on the input parameters selected, a taxonomy of web phishing detection is provided. Modern web phishing detection methods are examined and their

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

effectiveness is provided in detail. To guide future studies, this report also examined the shortcomings of the web phishing detection algorithms already in use. The study presented in this paper will assist academia and business in understanding the state of web phishing detection and inspire them to generate fresh concepts for the finest web anti-phishing method(s).

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant No. 61662039, and in part by the Jiangxi Key Natural Science Foundation under no. 20192ACBL20031, in part by the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (NUIST) under Grant no. 2019r070, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

VII. REFERENCES

[1] Vijayalakshmi, M., et al. "Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions." *Iet Networks* 9.5 (2020): 235-246.

[2] Akanbi O.A. Amiri I.S. Fazeldehkordi E.: ' *A machine-learning approach to phishing detection and defence* ' (Syngress, 2014, 1st Edition), pp. 1– 8

[3] Phishing Activity Trends Reports, 2020. Available at https://apwg.org/trendsreports, accessed on 13 April 2023

[4] Goel D. Jain A.K.: 'Mobile phishing attacks and defence mechanisms: state of art and open research challenges ', *Comput. Sec.*, 2018, **73**, pp. 519– 544

[5] Phishing, 2020. Available at https://en.wikipedia.org/wiki/Phishing#History, accessed on 13 April 2023

[6] Stephen Moramarco, Phishing Definition and History. Available at https://resources.infosecinstitute.com/category/enterprise/phishing/phishing-definition-and-history/, accessed on 13 April 2023

[7] Gibbs S.: ' Facebook and Google were conned out of $100M in phishing scheme '. Available at https://www.theguardian.com/technology/2017/apr/28/facebook-google-conned-100m-phishing-scheme, accessed on 13 April 2023

[8] Trend Micro: ' Texas School District loses $2.3 million to phishing scam ', BEC, 2020. Available at https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/texas-school-district-loses-2-3-million-to-phishing-scam-bec, accessed 13 April 2023

[9] Hadley E.: ' Vade secure discovers new phishing attack targeting 550 million email users globally '. Available at https://www.vadesecure.com/en/phishing-attack-targets-550-million/, accessed on 13 April 2023

[10] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_h1_2017.pdf, accessed on 13 April 2023

[11] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_q3_2017.pdf, accessed on 13 April 2023

[12] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_q1_2018.pdf, accessed on 13 April 2023

[13] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_q2_2018.pdf, accessed on 13 April 2023

[14] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_q3_2018.pdf, accessed 13 April 2023

[15] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org//reports/apwg_trends_report_q4_2018.pdf, accessed 13 April 2023

[16] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org/reports/apwg_trends_report_q1_2019.pdf, accessed 13 April 2023

[17] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf, accessed on 13 April 2023

[18] Phishing Activity Trends Reports, 2020. Available at https://docs.apwg.org/reports/apwg_trends_report_q3_2019.pdf accessed on 13 April 2023

[19] Workman M.: 'Wisecrackers: a theory-grounded investigation of phishing and pretext social engineering threats to information security ', *J. Am. Soc. Inf. Sci. Technol.*, 2008, **59**, (4), pp. 662– 674

[20] Wang J. Herath T. Chen R. *et al.*: 'Research article phishing susceptibility: an investigation into the processing of a targeted spear phishing email ', *IEEE Trans. Prof. Commun.*, 2012, **55**, (4), pp. 345– 362

[21] Alsharnouby M. Alaca F. Chiasson S.: 'Why phishing still works: user strategies for combating phishing attacks ', *Int. J. Hum.-Comput. Stud.*, 2015, **82**, pp. 69– 82

[22] Khonji M. Iraqi Y. Jones A.: 'Phishing detection: a literature survey ', *IEEE Commun. Surv. Tutor.*, 2013, **15**, (4), pp. 2091– 2121

[23] Varshney G. Misra M. Atrey P.K.: 'A survey and classification of web phishing detection schemes ', *Sec. Commun. Netw.*, 2016, **9**, (18), pp. 6266– 6284

[24] Tewari A. Jain A.K. Gupta B.B.: ' Recent survey of various defence mechanisms against phishing attacks', *J. Inf. Priv. Sec.*, 2016, **12**, (1), pp. 3– 13

[25] Jain A.K. Gupta B.B.: 'Phishing detection: analysis of visual similarity based approaches', *Sec. Commun. Netw.*, 2017, pp. 1– 20

[26] Dou Z. Khalil I. Khreishah A. *et al.*: 'Systematization of knowledge (SOK): a systematic review of software-based web phishing detection', *IEEE Commun. Surv. Tutor.*, 2017, **19**, (4), pp. 2797– 2819

[27] Chiew K.L. Yong K.S.C. Tan C.L.: 'A survey of phishing attacks: their types, vectors and technical approaches', *Expert Syst. Appl.*, 2018, **106**, pp. 1– 20

[28] Qabajeh I. Thabtah F. Chiclana F.: 'A recent review of conventional vs. automated cybersecurity anti-phishing techniques', *Comput. Sci. Rev.*, 2018, **29**, pp. 44– 55

[29] Google Safe Browsing, 2020. Available at https://safebrowsing.google.com/, accessed on 13 April 2023

[30] PhishTank, 2020. Available at https://www.phishtank.com/, accessed on 13 April 2023

[31] Sheng S. Wardman B. Warner G. *et al.*: 'An empirical analysis of phishing blacklists'. Proc. 6th Conf. on Email Anti-Spam, Mountain View, California, USA., July 2009, pp. 59– 78

[32] Cao Y. Han W. Le Y.: 'Anti-phishing based on automated individual white-list. Proc. 4th ACM Workshop on Digital Identity Management, Alexandria, Virginia, USA., 31 October 2008, pp. 51– 60

[33] Jain A.K. Gupta B.B.: 'A novel approach to protect against phishing attacks at client side using auto-updated white-

list, *EURASIP J. Inf. Sec.*, 2016, (1), p. 9

[34] Gautam S. Rani K. Joshi B.: 'Detecting phishing websites using rule-based classification algorithm: a comparison'. Information and Communication Technology for Sustainable Development, 8 November 2017, pp. 21– 33 Available at: https://www.worldcat.org/title/information-and-communication-technology-for-sustainable-development-proceedings-of-ict4sd-2016-volume-2/oclc/1012347167 accessed on 13 April 2023

[35] Sahingoz O.K. Buber E. Demir O. *et al.*: 'Machine learning based phishing detection from URLs', *Expert Syst. Appl.*, 2019, **117**, pp. 345– 357

[36] Li Y. Yang Z. Chen X. *et al.*: 'A stacking model using URL and HTML features for phishing webpage detection', *Future Gener. Comput. Syst.*, 2019, **94**, pp. 27– 39

[37] Jeeva S.C. Rajsingh E.B.: 'Intelligent phishing URL detection using association rule mining', *Hum.-Centric Comput. Inf. Sci.*, 2016, **6**, (1), pp. 1– 19

[38] Varshney G. Misra M. Atrey P.K.: 'A phish detector using lightweight search features', *Comput. Sec.*, 2016, **62**, pp. 213– 228

[39] Yang P. Zhao G. Zeng P: 'Phishing website detection based on multidimensional features driven by deep learning, *IEEE Access*, 2019, **7**, pp. 15196– 15209

[40] Zhu E. Chen Y. Ye C. *et al.*: 'OFS-NN: an effective phishing websites detection model based on optimal feature selection and neural network', *IEEE Access*, 2019, **7**, pp. 73271– 73284

[41] Zhang Y. Hong J.I. Cranor L.F.: 'Cantina: a content-based approach to detecting phishing websites'. Proc. 16th Int. Conf. on World Wide Web, Banff, Alberta, Canada, 8 May 2007, pp. 639– 648

[42] Mao J. Tian W. Li P. *et al.*: 'Phishing-alarm: robust and efficient phishing detection via page component similarity', *IEEE Access*, 2017, **5**, pp. 17020– 17030

[43] Tan C.L. Chiew K.L. Wong K.: 'PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder', *Decis. Support Syst.*, 2016, **88**, pp. 18– 27

[44] Marchal S. Armano G. Gröndahl T. *et al.*: 'Off-the-hook: an efficient and usable client-side phishing prevention application', *IEEE Trans. Comput.*, 2017, **66**, (10), pp. 1717– 1733

[45] Xiang G. Hong J. Rose C.P. *et al.*: 'Cantina+ a feature-rich machine learning framework for detecting phishing web sites', *ACM Trans. Inf. Syst. Sec.*, 2011, **14**, (2), pp. 1– 28

[46] Mao J. Bian J. Tian W. *et al.*: 'Detecting phishing websites via aggregation analysis of page layouts', *Procedia Comput. Sci.*, 2018, **129**, pp. 224– 230

[47] Chiew K.L. Tan C.L. Wong K. *et al.*: 'A new hybrid ensemble feature selection framework for machine learning-based phishing detection system', *Inf. Sci.*, 2019, **484**, pp. 153– 166

[48] Jain A.K. Gupta B.B.: 'A machine learning based approach for phishing detection using hyperlinks information', *J. Ambient Intell. Humaniz. Comput.*, 2019, **10**, (5), pp. 2015– 2028

[49] Gowtham R. Krishnamurthi I.: 'A comprehensive and efficacious architecture for detecting Phishing webpages', *Comput. Sec.*, 2014, **40**, pp. 23– 37

[50] Rao R.S. Pais A.R.: 'Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach', *J. Ambient Intell. Humaniz. Comput.*, 2019, **11**, pp. 3853– 3872

[51] Li J.H. Wang S.D.: 'Phishbox: an approach for phishing validation and detection'. 2017 IEEE 15th Int. Conf. on Dependable, Autonomic and Secure Computing, 15th Int. Conf. on Pervasive Intelligence and Computing, 3rd Int. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, USA., 6 November 2017, pp. 557– 564

[52] Rao R.S. Ali S.T.: 'Phishshield: a desktop application to detect Phishing webpages through heuristic approach', *Procedia Comput. Sci.*, 2015, **54**, pp. 147– 156

[53] Yuan H. Chen X. Li Y. *et al.*: 'Detecting phishing websites and targets based on URLs and webpage links'. 24th Int. Conf. on Pattern Recognition (ICPR), Beijing, China, 20 August 2018, pp. 3669– 3674

[54] Mao J. Bian J. Tian W. *et al.*: 'Phishing page detection via learning classifiers from page layout feature', *EURASIP J. Wirel. Commun. Netw.*, 2019, **2019**, (1), p.43

[55] Moghimi M. Varjani A.Y.: 'New rule-based phishing detection method', *Expert Syst. Appl.*, 2016, **53**, pp. 231– 242

[56] Languages used on the Internet, 2020. Available at https://en.wikipedia.org/wiki/Languages_used_on_the_Internet, accessed on 13 April 2023

[57] Joe Lobo, 7 reasons why you should get a multilingual business website, 2018. Available at https://wpml.org/community/2018/07/multilingual-business-website/, accessed on 16 April 2023