Received: 22 March 2024 | *Revised:* 6 July 2024 | *Accepted:* 19 July 2024
DOI: 10.37917/ijeee.22.1.31 *Vol. 22 | Issue 1 | June 2026*

Iraqi Journal for Electrical and Electronic Engineering
Review Article

Open Access
IJEEE University of Basrah College of Engineering

Unraveling Phishing Attacks and Countermeasures: A Comprehensive Review

Ali Raheem Al-Hafiz*, Adnan J. Jabir
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Correspondance
*Ali Raheem Al-Hafiz
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq
Email: aliraheem2201m@sc.ubaghdad.edu.iq

Abstract
Recent advancements in communication and wireless technologies have greatly increased the number of internet users. These users often share personal information online, making it vulnerable to attackers. Phishing, a common type of online fraud, involves tricking people into giving their personal information through spam or other deceptive methods. Even though this threat has been around for a long time, it is still very active and successful. Attackers have improved their methods over the years to make their attacks more convincing and effective. Therefore, it is important to carefully study this type of attack to raise awareness among both users and cybersecurity researchers. This review paper explains the basics, types, and methods of phishing and presents a unified attack lifecycle framework to provide users and researchers with a clear understanding of phishing. Additionally, anti-phishing methods are thoroughly analyzed to determine their strengths and weaknesses. Researchers use different strategies to develop anti-phishing solutions, including blacklisting, whitelisting, heuristics, machine learning, and deep learning techniques. To help readers choose the best anti-phishing solution, research studies using these strategies are categorized, evaluated, and compared using specific criteria to show their strengths and weaknesses. Furthermore, the datasets used to develop anti-phishing models are discussed and reviewed. Finally, this paper provides a detailed overview of current phishing challenges and suggests future research directions in this area.

Keywords
Cybersecurity, Deep Learning, Machine Learning, Phishing Attacks.

I. Introduction

Over the last few years, there have been nearly 1.13 billion websites on the Internet, according to Forbes Advisor [1]. The number of Internet users has increased, since October 2023, with about 5.3 billion users accounting for 67.7% of the world's population. At the same time, the number of social media users increased reaching 4.5 billion users. Concurrently, the attacks on the websites has increased as evidenced by statistics [2] through the pandemic Covid-19, due to working and teaching becoming online, resulting in a significant increase in the work for email services, student platform websites, educations, banking services,..etc. Therefore, the targets being exposed to phishing attacks has been significantly increased [3, 4]. The Internet Crime Report Cybercrime (IC3) investigations show that phishing attacks have increased at the highest rate in the last five years; there have been 300,497 reports of phishing.

Consequently, it is predicted to have cost over $10 billion in losses overall, exceeding the $6.9 billion from 2021 [5]. According to Kaspersky's company, there were 500 million attacks in 2022, and that increased to double since before last year [6]. In addition, Anti-Phishing Working Group (APWG) reported that phishing attacks was 4.7 million and the number of phishing attacks increased to 150% at the start of 2019, furthermore, the report of the APGW in Q4 of 2022 [7] stated the highest priority is given by attackers to companies or organizations or brands to attack them to collect sensitive and useful information. According to the office of national

statistics. Recent research [8] found that Phishing has been the most popular technique employed by cybercriminals. It defined Phishing as a technique used to get sensitive and private information such as usernames, passwords, and payment card details, through fraudulent texts, emails, and websites. To trick users into doing things like clicking on a hyperlink that installs malware or steals personal data, the attackers use social engineering techniques [2].To launch phishing attacks, fraudsters mostly rely on fake emails where victims are tricked into providing the requested information. As a result, users may avoid online banking, shopping, and e-commerce. Businesses may also suffer from a decline in stock price, a loss of reputation, and a decline in customer confidence, Fig. 1 shows the targets that have received the most attacks

Most of the recent research introduced phishing attack as social engineering techniques to increase their probability of success [1]. Therefore, networks are seeing a sharp rise in social engineering attacks, which compromise cybersecurity. Cybercriminals attempt to influence people and organizations into disclosing sensitive and valuable information [9]. Whether a network has strong firewalls, encryption techniques, intrusion detection systems, or anti-virus software, social engineering poses a threat to network security. Compared to computers or other technologies, humans are more likely to trust other humans. As a result, they are the weakest link in the security chain. A person is psychologically persuaded to reveal private information or violate security protocols by malicious operations carried out through human interactions [10].

Social engineering assaults are the most potent since they pose a threat to all systems and networks because of these human interactions.

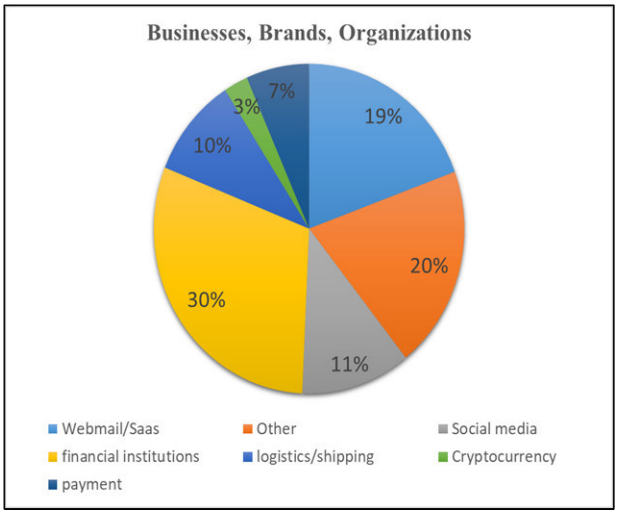

Fig. 1. Targets that have mostly been attacked

The U.S. Department of Justice lists social engineering attacks as one of the worst global concerns. According to a 2016 report by cyber security research firm Cyence, social engineering assaults targeted the United States the most, with the highest cost of attack, followed by Germany and Japan. These attacks are thought to have cost the US $121.22 billion. Particularly, hackers and cybercriminals from all over the world target and have an impact on American businesses. Another example in 2018, confidential customer data was stolen from the Equifax company after it was breached for several months. This business is a consumer credit reporting and monitoring agency that compiles information on private persons as well as commercial clients in order to keep an eye on their credit histories and stop fraud [11]. Therefore, phishing attacks have evolved in recent years due to high-tech-enabled economic growth worldwide. The rise in all types of fraud loss in 2019 has been attributed to the increase in deception scams and impersonation, as well as to sophisticated online attacks such as phishing.

There have been several well-known phishing attack types. This includes spoofing, malware-based phishing, email/spam, data theft, DNS-based phishing, and web-based phishing over the phone and package delivery [12].The global impact of phishing attacks will continue to intensify, and thus, a more efficient phishing detection method is required to protect online user activities. Several countermeasures methods for identifying phishing websites have been proposed in the literature like, Visual Similarity, Heuristic, Lists-Based, Machine Learning, and Deep Learning techniques. In this paper, the attacks types along with a phishing timeline is presented. In addition, the current research works dedicated to detect the phishing attacks are analyzed thoroughly to show their strengths and weaknesses. There have been several review articles [13] dedicated for the phishing and anti-phishing attacks; however, this review article has the following distinctive contributions.

- Presenting a unified flowchart capable of identifying numerous common phishing techniques, illustrating how attackers modify their approaches to target victims.
- The past research on phishing lacked critiques from fellow researchers. Therefore, this study extensively explains anti-phishing techniques and highlights the significant weaknesses of each method.
- Presenting an inclusive diagram detailing the steps to construct a phishing detection model for researchers, covering gap identification, dataset selection, classifier choice, data preprocessing and segmentation, classifier training, and testing.
- This study thoroughly explains datasets and their repositories from prior studies. It assists researchers in select-

ing suitable datasets for designing appropriate models.

The paper is organized as follows, Section II contains a detailed explanation of phishing, its history, the most important types, and its effects on individuals and institutions. Then we discuss the life cycle of phishing attacks, what steps the attacker takes to carry out the fraudulent link attack, and their motives. Section III presents the phishing types and the analysis of the current research works on the anti-phishing techniques. Section IV discusses the data set used in previous studies for the past five years and the methods of dividing and distributing it and adopting the repositories of the data set. Section V describes an analytical study of the distribution of different phishing scams. Section VI concludes with an analytical study and discussion of previous research and what this research paper has presented.

II. Attack Techniques

Cisco company referred to the Phishing as the act of sending fraudulent emails or other types of communications that seem to be from a reliable source. The intention is to either install malware on the victim's computer or steal sensitive data, such as credit card numbers and login credentials [14]. The phishing process involves the following elements:

1. Attacker: People or groups that execute phishing attempts with the intention of gaining a specific kind of advantage, such money or identity concealing (which, for example, describes the circumstance in which phishers do not utilize the identities they have stolen). According to AVG the Hacker (Attackers) can be classified into different types based on their motives, skills, and methods, such as: script kiddies, hacktivists, cybercriminals, cyberterrorists, and state-sponsored hackers [15].

2. Victims: Individuals or organizations who suffer the consequences of an attack, such as data loss, identity theft, financial damage, reputation harm, or physical injury. Victims can be targeted by attackers for various reasons, such as: personal vendetta, political agenda, financial gain, or random opportunity [16].

3. Tools for attack: Hardware or software tools that give attackers the ability to carry out various kinds of tasks, including after-exploitation, exploitation, scanning, and reconnaissance. Attacking tools can be either specially-made tools for specific attack, like malware, ransomware, or phishing kits, or they can be authentic tools used for malicious purposes, like Nmap, Metasploit, or Netcat while Vulnerability was refer a weakness in a system that can be exploited therefore the Exploit was taking advantage of the identified vulnerability [17].

According to their cybersecurity vulnerabilities, phishing attacks have been classified into several categories:

A. Email Phishing

Users are tricked into providing their credentials via a carefully crafted fake email, and a Phish-Me study [18] found that fake emails are the starting point for 91% of phishing attacks. Curiosity (13.7%), fear (13.4%), and urgency (13.2%) are the main reasons why people are vulnerable to these emails [19]. Phishing email is described as a message that usually appears to come from a well-known organization and requests your personal information [20]. The scammer targets two groups with phishing attacks: individuals and employees because they are trying to steal credit card number, social security number, account number, password, or private photo. Definitely the impact of this attack is huge as 3.4 billion spam emails are being sent daily throughout the world. Additionally, for every 100 internet users worldwide in 2021, there were 16.5 leaked emails.

B. Vishing

Is described as the act of using the telephone in an attempt to scam the user into surrendering private information that will be used for identity theft. When an attacker tries to get information from a victim over the phone, it is known as vishing, or voice phishing [21]. According to Gupta [22], phone numbers and personal information obtained from millions of users' social media accounts and caller ID can be utilized to execute these attacks. A caller posing as the victim's bank, for instance, can state that unexpected charges have been found on the victim's account before requesting that the victim confirm their credit card details. The target of that attack is individuals and the mostly goal is to get personal information about the victim by making fake calls or speaking on the phone with fraud sound. The impact of that attack was appeared in 2022; almost seven in ten respondents reported having encountered vishing attacks. This represents an increase from 54 percent in 2020.

C. Whaling

Refers to a specific, uncommon, and highly specialized form of phishing that targets only VIP clients. The term "whaling attacks" refers to phishing attempts that target high-ranking personnel, such as top executives in any organization and other important individuals in the business world. The higher profiles or any senior managerial level positions in a corporate firm are the primary targets of the whaling attack [23]. The primary aim of that attack is to steal money or sensitive information or gain access to their computer systems for criminal purposes. A whaling attack in 2022 was the most common type of attack against Asian organizations. One whaling attack costs the company a loss of $47 million [24–26].

D. SMISHING

The researchers defined the smishing in many form like is a type of phishing in which attackers send text messages that look to be from a reliable source, requesting that recipients click on a link or provide personal information via text messages rather than via emails. It also refers to a technique of cybercriminals using malicious links, phone numbers, or emails to target mobile Short Message Service (SMS) devices [27,28]. The goal of that attack is to gain personal information about a victim, the goal of that attack is to Gain personal information about a victim and using that sensitive information to get full access to bank account or get control access on computer. The impact for smishing was clearly appeared in April 2022, hackers sent an average of 2,649,564,381 smishing messages per week, moreover in 2023, less than 35% of people are familiar with smishing as stated in [29].

E. Spear Phishing

Is a different kind of attacks that exploit technical flaws in software, protocols, and machines. Spear-phishing attack's engineering might be described as social in nature as compared to technical. Spear-phishing is the practice of sending victim-specific emails that are specifically tailored to trick victims into doing something that will benefit the predator. Because of the nature of the attack, the attacker just needs a very basic understanding of technology [8]. The primary targets for that attack are government officials and high-profile individuals, business partners or suppliers. Therefore, the attack aims to obtain gain confidential information [30]. Undoubtedly, this attack was influential in 2022, 47% of spear-phishing worldwide were scams, making them the most common type of cyberattack in this category [31].

Fig. 2, shows most common phishing techniques and illustrates how attackers modify their approaches to target victims and we supposed that user try to login website of bank.

- Message Received: If you receive an Email, SMS, or social media message (ESS), ask yourself: Is it from someone I trust? If yes, great! Else the message is detecting as ESS phishing.

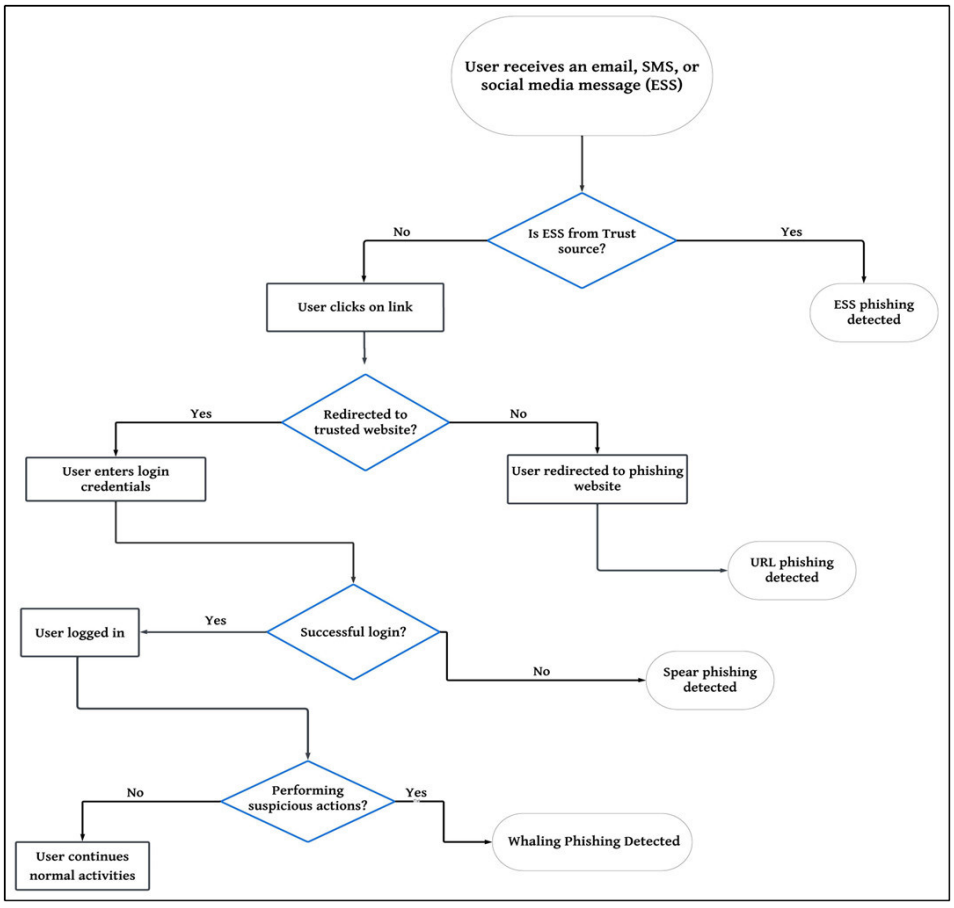

Fig. 2. Most common phishing technique

- Clicking Links: If you decide to click on a link in the message, watch out for two possible outcomes:
- Redirected to a Trusted Site: If the link takes you to a website you recognize and trust, like your bank's site, you're safe to enter your login details.
- Redirected to a Phishing Site: But if the link leads you to a suspicious website, such as a fake login page, be cautious. This is known as "URL phishing".
- Entering Login Details: If you do enter your login information, pay attention to what happens next:
- Suspicious Activity: If your account starts acting strangely, like sending odd emails or transferring money without your knowledge, it might be a "whaling".
- Login Failure: If your login fails unexpectedly, it could indicate a "spear phishing" attack.
- If nothing happens to them, the user can resume normal activities.

III. Phishing Timeline

In phishing, attackers use a phony website that appears to be a visual replica of an authentic website in order to obtain sensitive data from victims [32]. To clarify the phishing attack process the following scenario is assumed. The attacker creates a fraudulent website and sends convincing emails with a link to it to users of an online service in a generic/traditional phishing scenario, often known as mass-email phishing campaigns. The Data is collected by the server hosting the phony website, just when a user of the web service clicks on the link and inputs the personal information. As we shown in Fig. 3, the first stage is known as the reconnaissance or planning phase, during which the attackers select a phishing vector, communication channels, and possible victims [33, 34]. The attacker tries to collect public and private information using special tools, via social media, or using network tools, where sufficient information about the target must be known. The second stage, known as weaponization or setup, is when phishers get ready to spread phishing materials to their intended victims [35].

After completing the process of collecting information about the target (the victim) and knowing what the victim's desires are (the bait), the attacker then creates a fake web page using one of the tools found in "Kali Linux operating system". The subsequent phase is known as the "phishing phase," during which phishers begin to distribute to place baits on victims [36]. The attacker must choose the appropriate means (method) to send the fake URL or fake page to the victim using social media, email, or via text message.

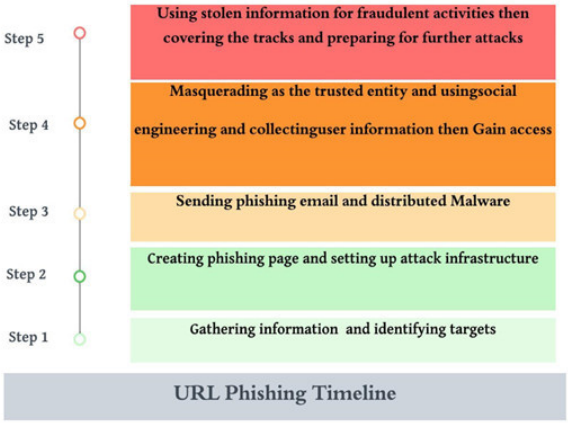

Fig. 3. Phishing attack steps

Sometimes the victim will doubt the fake link and will not click on it. Therefore, the attacker must use social engineering to convince the victim to click on the fake link and even more to convince the victim that it is a trustworthy site. The next phase is known as the "exploitation" or "penetration phase", during which phishers take advantage of victims' vulnerabilities to trick them into disclosing sensitive information [37]. When the victim clicks on the fake URL, the fake website will appear which is similar to the original website. Here, several options appear for an attacker to either use the malicious programs like Trojan while loading the website, or click on the poisoning link, which enables him to penetrate the victim's computer and obtain all his sensitive information, or steal private information of the victim like (name, work, birth, bank number, and password). Before to the last phase, the phishing operation has been effective, and the hackers have managed to get the data they had planned to steal. Phishers may choose to take additional steps in order to profit financially or may choose to use the information they have gathered for other objectives [33]. In the final stage, the attacker must erase the traces that indicate an address on the Internet or use a hidden browsing method or a network (VPN), as we show in Fig. 3.

IV. Phishing Detection Techniques

A. List-Based

These techniques are used by browsers such as Google Chrome, Microsoft Edge, and Firefox to identify phishing websites. It comes in two varieties: whitelisting and blacklisting [38]. A white list consists of a list of authorized, valid domain names or URLs. The details of websites that are trustworthy and that the user wants to visit are listed on a white list. A black list, on the other hand, is a list of websites that the user does not want to visit and that are fraudulent. In addition, compared to the black-list, the white-list data is more precise and smaller [39].

Furthermore, Whitelist-based phishing which was a website detection technique had worked on some features such as logs and the IP address and login user interface (LUI) data of every URL a user visits to detect phishing websites [40]. Additionally, Google company used API called it "block list" which is a database that includes websites and IP addresses which Google, other search engines and antivirus software providers have marked as unsafe for use. That allows users to check if a specific URL address is included in the blacklist of phishing websites on Google. However, drawbacks have appeared in that technique like their susceptibility to minor URL modifications that can bypass the system and provide a high rate of false positives. Moreover, that technique fails in preventing zero-day phishing sites [41]. Therefore, researchers suggest that these lists must be regularly updated to detect the phishing website.

B. Visual Similarity

Theses techniques compare the visual features of the target website with the suspected website to identify phishing attempts. The features can be Global features and local features, and Visual attributes like including CSS, text layout, source code, logos, and screenshots. Because users are now smarter than ever, skilled phishers always design websites that are "visually similar" or even identical to the target websites. However, aside from HTML text, there are other ways to create a "visually similar" or similar web page, including photos, flashes, movies, and more [42–44]. Additionally, [45] suggested a technique for identifying phishing websites using some features like CSS files in web pages to detect if phishing websites or legitimate. In addition, [46] proposed the Link Guard algorithm to determine whether a URL is suspicious and used an image-based page-matching approach to obtain similarity between the target pages and pages in phishing websites. Moreover, [47] proposed a detection method based on the URL and web page similarity to detect phishing websites or not. Nonetheless, these techniques compare the suspect web page to previously visited or saved web pages, it's ineffective against zero-hour phishing attacks and it is complex, also slower by nature furthermore it is expensive and needs computing resources [48]. In contrast [49] proposed solutions for drawbacks of visual similarity technique by combining visual similarity with machine learning or URL base, or by using machine learning or deep learning techniques in phishing detection.

C. Heuristic

This approach makes use of several features that are gathered from the website and utilized to determine if it is legitimate or illegitimate. There are two methods in the heuristic-based approach, referred to as content-based and non-content-based. While features based on host information and URLs are employed in non-content-based approaches to identify phishing sites, content-based approaches use the website's content to determine the legitimacy of the website [50–54]. The heuristic techniques examine the text and URL structure of phishing websites, and extract the characteristics of phishing, then create a model to identify phishing sites based on the extracted attributes. Additionally, most researchers present methods for phishing detection such as [55]. presented the 'Phish Net' technique, which proposed lists of the basic URL-based phishing websites based on five heuristic guidelines. In addition, [56]. presents an FSM technique that monitors web page forms and corresponding reactions to assess the behavior of a web page to detect phishing websites. Consequently, when comparing heuristic to list-based techniques, this method is faster and has fewer false positives or false negatives, but it is less accurate. Once an attacker learns the heuristic approach, he can get past the heuristic filter and accomplish his purpose of stealing credentials [57]. To mitigate the drawbacks in the heuristic technique [58] proposed to increase accuracy using a Hybrid-rule base for phishing detection.

D. Machine Learning (ML)

Is the scientific study of statistical models and algorithms used by computer systems that perform certain tasks without the need for explicit programming. Machine learning has been used in several everyday applications. A learning algorithm that has figured out how to rank web pages is one of the reasons that every time a web search engine like Google is used to search the internet, it functions so well. These algorithms are used for various purposes like data mining, image processing, predictive analytics, security detection, etc. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically [59]. ML can be classified into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, the data has already been classified because the system was trained on labeled data, but in unsupervised learning, the algorithm is trained on unlabeled data, which means that the data is not classified. Semi-supervised learning combines supervised and unsupervised learning. Also, the Reinforcement learning algorithm learns by interacting with its environment. The most ML classifiers are Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes, Hidden Markov Models (HMM), Decision Tree, K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), Artificial Neural Network (ANN) [60]. Authors in [61] explained how ML is used in phishing detection, by looking at things like web addresses, how websites were built, and the code used on the page. This information helped to make datasets that could recognize fake sites. The work

in [62] presents how ML dealt with a dataset, teaching machine learning computers to spot these fake sites using these datasets. The work in [63] used ANN classifier to achieve an accuracy of 83.38%. The work in [64] proved that ML worked well with big and complicated data, and it was proven to be super accurate, with success rates of more than 99%. However, some drawbacks appear in ML it cannot detect all Zero-day attacks, and another challenge is the skill of fraudsters, which makes it difficult to detect fraud [65]. Therefore [46, 66] suggested using ensemble or Hybrid rather than single classifier to get more accurate in detecting phishing attacks. Other research [67, 68] was proposed using feature extraction and feature selection to get a high accuracy rate. Table I, states the recent works that utilized the ML for phishing detection.

E. Deep Learning

A subset of machine learning algorithms called deep learning algorithms attempt to identify numerous levels of distributed representation. It is a new strategy that has been widely used in standard artificial intelligence fields such as semantic parsing, transfer learning, natural language processing, computer vision, and many more [99]. The idea of Deep learning is to train a neural network to find specific information [19]; Therefore, neural networks was used to analyze large amounts of data, detecting subtle and complex phishing patterns in emails, websites, or other content. Several deep learning techniques [78] have been presented recently to deal with various artificial intelligence challenges like Deep Neural Networks, Recurrent Neural Networks, and others. Authors in [100] suggested identifying phishing URLs, using an LSTM model. This approach encodes the URL string using the one-hot encoding method first and then feeds each encoded character vector into the LSTM neurons for testing and training. Also [101] used LSTM-based method for phishing page identification. The work in [102] explained how to

TABLE I.
PRIMARY STUDIES ON MACHINE LEARNING TECHNIQUES

Ref	Classifiers	Accuracy	Outline of research	Dataset	Dataset ratios	Publisher	Year
[63]	ANN	83.38%	The ANN was used to classify the phishing/phishing websites using 18-URL features.	UCI	TR=80% TS= 20%	IOP science	2018
[69]	SVM, NV-Bayes	90%	Two machine learning classifiers were utilized and compared using 14 different features to distinguish phishing websites from legitimate websites.	PhishTank	NA	Springer	2018
[70]	RF, NLP	97.98%	A real time ant phishing model was proposed based on NLP features and machine learning classifiers.	PhishTank, Yandex	NA	Elsevier	2019
[71]	Stacking 1 is RF, KNN, Bagging	97.4%	Several feature extraction methods were used with several stacking ensemble models were used to identify the phishing website.	Kaggle	NA	Emerald	2019
	Stacking 2 is KNN, RF, Bagging	97.2%					
[72]	ANN,RF, SMO	95%	Fuzzy Rough Set (FRS) method was used to identify nine common features throughout the three data sets to detect the phishing websites.	UCI1,UCI2, Mendeley	TR=50% TS= 50%	IEEE	2019
[73]	ADBOOST	99%	A feature selection based on the correlation to the other features and to the class label was used with the ADBOOST to increase the ability to identify phishing.	PhishTank, MillerSmile, Google	Dataset is classified from 50%-90%	JJCIT	2020
[74]	RF, SVM, GLM, GAM, RP, RT	98.34%	Several types of classifiers in machine learning was used with 30 selected distinct features to identify phishing with high accuracy.	UCI	NA	EUDL	2020
[75]	RF, NV- Bayes, KNN, LR, DT	98.03%	Several types of machine learning classifiers were utilized. Their performance were compared using the common metrics.	GitHub	TR=70% TS= 30%	IEEE	2020
[76]	RF, PCA, DT	97% 91% 94%	Several machine learning classifiers was utilized and evaluated to identify the phishing websites. The PCA method was used to select the important features.	Kaggle	NA	ICSSIT	2020
[77]	RF, Chi-Squared Pearson	96.2% 97.8%	Several machine learning classifiers was utilized and evaluated to identify the phishing websites. The Chi-Squared method was used to select the important features.	UCI, Mendeley	TR=50% TS= 50%	CONECCT	2020

TABLE I.
PRIMARY STUDIES ON MACHINE LEARNING TECHNIQUES (*Continued*)

Ref	Classifiers	Accuracy	Outline of research	Dataset	Dataset ratios	Publisher	Year
[78]	RF, SVM, C4.5, DT, PCA, KNN	99.2%	Several machine learning classifiers was utilized and evaluated to identify the phishing websites. The PCA method was used to select the important features. In addition, various dataset repository was used to train the systems in order to increase the detection accuracy.	Ebbu2017, UCI	NA	Springer	2020
[79]	KNN, RF ANN, RF RFC, C4.5	97.33% 97.16% 96.36%	Ensemble learning with several machine learning classifiers were used to classify the phishing websites.	UCI	NA	IEEE	2020
[80]	RF	99.36%	Introduce a newly method, named EPDB which focused on extracting the important features from URL and used the random forest classifier to detect the phishing websites.	PhishTank	TR=56% TS=44%	Springer	2020
[81]	RF, DT, K-NN, SVM	96.87%	Several machine learning classifiers with wrapper-based feature selection techniques were used to detect the phishing websites.	Millersmile, PhishTank	NA	Taylor Francis	2021
[82]	SVM, ISHO	98.64%	an improved spotted hyena optimization algorithm was used to select the important features with SVM classifier.	UCI	NA	Springer	2021
[83]	LR, Ad Boost, GB	85.6%	Different machine learning algorithms were compared with ensemble learning techniques. The stacking classifier has shown the best accuracy.	PhishTank, OpenDNS	TR= 80% TS= 20%	IEEE-GCAT	2021
[84]	ANN, DT	97.40%	ANN and DT were used to detect the data patterns in the URL.	Alexa, PhishTank	TR1=50% TS1=50% TR2=70% TR2=30%	ICDABI	2021
[85]	RF DT,XGB	97.51% 98.45%	Different machine learning classifiers were used with two dataset repositories.	UCI, Mendeley		Springer	2021
[86]	SVM, DT, RF	97%	Several types of classifiers in machine learning were used with two feature selection methods, then the final fusion was decided.	UCI	TR=55% TS=45%	ICAIS	2021
[87]	SVM, NVB, LR, RF, DT	73.95% 79.80% 94.48% 96.61% 94.09%	A general scheme for building reproducible and extensible datasets for website phishing detection was proposed. Several methods were used to select the relevant features and tested with different types of machine learning classifiers.	Alexa, Yandex, Phishtank, Openphish	NA	Elsevier	2021
[88]	KNN	85.08%	The KNN classifier as utilized with a correlation techniques to find the best feature set.	Kaggle	NA	JAIMLNN	2021
[89]	LR	98.42 %	The logistic regression classifier was used to classify website as phishing or not.	Kaggle, Github, PhishTank	NA	Conference COINS	2021
[90]	Locally-SVM, SVM, BDT, AP, LR, NN, DT	99.9%	Seven machine learning classifiers with three different datasets with different size was used to detect the phishing emails. The effect of the number of features used for the classification process was also analyzed.	NA	TR=70% TS=30%	Springer	2022
[91]	XGB	96.44%	An XGBoost ensemble model combining Random Forest and K-Nearest Neighbors was used to detect the phishing websites.	Kaggle	TR=0.67% TS=0.33%	IEEE	2022
[92]	XGB, RF	97%	An adaptive approach was used to detect the phishing websites.	Mendeley	TR= 80% TS= 20%	IEEE-IDSTA	2022
[93]	K-means, RF, DT, Cat Boost, Light GB, Ad Boost	98.61%	The SHAP values was used for features selection, then several well-known machine learning classifiers were used with the SMOTE method for data normalization. In addition, various dataset repository were used to train model for phishing website detection.	Kaggle, Alexa, UCI, PhishTank	NA	IEEE- CCICT	2022

TABLE I.
PRIMARY STUDIES ON MACHINE LEARNING TECHNIQUES (*Continued*)

Ref	Classifiers	Accuracy	Outline of research	Dataset	Dataset ratios	Publisher	Year
[94]	Ad Boost, GBM	98.63% 97.05%	Two novel feature selection strategies was utilized to find and examine the key properties required for identifying spoof websites.	Mendeley, Rami et al	NA	Springer	2022
[95]	RF, DT, XG Boost, SVM, LR	99.17%	A hybrid feature set was built based on URL and hyperlinks to be used with the machine learning classifiers in order to increase the system accuracy.	Alexa, PhishTank	TR=80% TS= 20%	Springer	2022
[96]	LR, RF, NV-Bayes, DT, XGB, Extra Tree	95.99%	Several machine learning classifiers were used with three ensemble methods to detect the phishing websites.	Kaggle (D1,D2)	NA	IC2IE	2022
[97]	MLP, KNN, RF, LR, XGB at layer 1, RF,XGB, MLP at layer 2 and XGB as meta learner.	97.76%, 98.9%, 96.79%, 98.43%	A multi-layer stacked ensemble model for the detection of phishing websites, named MLSELM, was proposed, such that the output of each learning layer become input to the next layer to increase the detection accuracy.	UCI(D1), Mendeley 2018(D2), Mendeley 2020 (D3,D4)	NA	IEEE	2022
[98]	RF, XGB	RF= 97% XGB= 97%	Three factors that affecting the classifiers were studied and analyzed, like dataset balancing, hyper-parameters optimization, and feature selection.	UCI, Mendeley	NA	MDPI	2023
	GB, XGB	GB= 98% XGB=98%					

extract the representation of the URL of the phishing website using the auto encoder. The work in [103] used two classifiers DNN and SVM and employed a two-label dataset with 28 features to achieve an accuracy of 96%. Authors in [104] used a multi-classifier with 14 features, achieving an accuracy of 90%. The work in [105] used a new method called (PDR-CNN) which was extremely fast and encoded information from URLs. It used two classifiers CNN and LSTM with an accuracy of 97%. A Hybrid between machine learning and deep learning classifier RF and NLP is used in [70] with feature selection methods (CFS subset eval) to achieve an accuracy of 97.98%. authors in [71] divided the dataset into two groups, weak and strong features, to increase accuracy; therefore, it had two results of accuracy because it used two different classifiers in the same model. The first one used (RF with NN) with an accuracy result of 97.4%, and the second one used (KNN, RF, and NN) with an accuracy of 97.2%. Authors in [79] used various deep learning methods with a single machine learning classifier (KNN and RF) (ANN and RF) (C4.5 and RFC) to achieve accurate results of 97.33%, 97.16%, and 96.36%. [74] Used various machine learning models (RF, SVM, Generalized Linear Model, Generalized Additive Model, Recursive Partitioning, Regression Trees), and only one dataset repository to achieve an accuracy of 98.34%. For instance, [75,81,85–87,92,95,96,98] used Multi-classifier to improve the timely detection of phishing attacks and provide a robust model for website security. Nonetheless, researchers turned to using feature selection methods such as (Chi-Squared and Pearson or PCA, SHO, SOMTE, TF-IDF) with various machine learning models [72, 76, 78, 82, 93, 106] to increase the accuracy of phishing detection. The evolution of anti-phishing detection by using machine learning or deep learning was continuous; therefore, the researchers introduced novel methods to improve the accuracy of phishing detection such as (EPDB, MLSELM, and PhiKitA) [80, 97, 107]. While incredibly powerful, Deep learning does come with certain drawbacks or challenges such as that requiring a vast collection of data for high-efficiency output. In comparison to the other traditional machine learning approaches, the other one has high computational costs including powerful GPUs and large amounts of memory, furthermore a dependence on data quality; meaning the model's performance will suffer if the data is biased, noisy, or incomplete Table II provides a detailed comparison among the research works that utilized the deep learning techniques.

V. DATASET

It is a collection of information comprising various characteristics of phishing websites, URLs, labels, HTML code, screenshots, and phishing kits—tools used by attackers to create and deploy phishing websites. Such datasets are valuable for machine learning purposes, aiding in the training and testing of algorithms designed to detect phishing websites. Essentially, a dataset is a compilation of webpage elements gathered from numerous websites. This section will elaborate on the datasets sources and sizes. Dataset size refers to the total count of legitimate and phishing webpages utilized within the dataset, while the source column indicates where authors acquired these webpages. Most phishing datasets are downloaded from public repositories or private resources, as depicted in the Fig. 4.

Dataset sizes range from 100 to 1000 phishing websites for small datasets, to hundreds of thousands for larger ones

TABLE II.
PRIMARY STUDIES ON DEEP LEARNING TECHNIQUES

Ref	Classifiers	Accuracy	Outline of research	Dataset	Dataset ratios	Publisher	Year
[104]	SVM, DBN	90%	Two types of features were used which re original features and interaction features. DBN with SVM were used for website classification.	Real flow data from ISP	NA	Hindwai	2018
[103]	DNN, SVM	93% 96%	A fresh-phish framework was proposed based on DNN and SVM using 28 different website features.	PhishTank, Alexa	TR=50% TS= 50%	IGI Global	2018
[104]	DNN	86.63%	A lightweight phishing URL model detection was proposed based on DNN. The model was implemented using raspberry Pi with low energy consumption.	Real time flow data	NA	MDPI	2019
[105]	LSTM, CNN	97%	A fast phishing detection, named PDRCNN was proposed. It used LSTM to extract global features from URL and CNN for phishing website classification.	Alexa, PhishTank	Classified dataset into 8:1:1	Hindawi	2019
[108]	RF, GBTF, DNN	96.4%	Several machine learning and deep learning methods were used with some URL features, like lexical, host-based, and content-based.	PhishTank, Majestic repository	NA	IEEE	2020
[109]	DNN, LSTM, CNN	99.52% 99.57% 99.43%	A deep learning model was proposed with 10 features only to increases the speed of phishing detection.	PhishTank, Alexa	TR=75% TS=25%	Springer	2020
[110]	DNN	96.25%	A deep neural network with Adam optimizer was utilized to detect the phishing URL using 30-feature vector.	UCI	TR=66.% TS=34%	Springer	2021
[111]	RNN, LSTM	94.3% 93.6%	Employed the RNN with LSTM to detect the phishing websites.	PhishTank, Alexa rank	NA	PLOS ONE	2021
[112]	CNN-LSTM	93.28%	A hybrid classification model of CNN and LSTM was proposed for website phishing detection using URL and website contents as classification features.	PhishTank, WHOIS	TR=70% TS=30%	Emerald	2023
[113]	RF, CNN, FCNN, LSTM	96.94%, 91.38%, 90.13%, 89.73%	More than one datasets from different sources were used and an ensemble model with more than one phishing method and deep learning models was proposed.	UCI_2015, Mendeley_2018 Mendeley_2020	NA	IEEE	2022
[114]	RF, DT, LR, KNN, ANN, Max Vote	97.73%	The focus was on data preprocessing to extract the most features required to increase the model accuracy.	UCI	Classified dataset into 9:1	IJISAE	2022

focusing on phishing attacks. These datasets include information extracted from websites, such as IP addresses, HTTP, HTTPS, DNS records, and more. A primary feature (target) distinguishes legitimate (0) from illegitimate (1) websites, is crucial in identifying phishing websites. The differences in feature size between legitimate and phishing datasets in some instances are substantial. This observation may lead us to infer that the anti-phishing community still lacks consensus regarding the optimal dataset size. Consequently, there are varied sources for datasets. The largest phishing dataset available is Phish Tank [115] which is the most widely used resource for phishing datasets. Over 24% of the research works primarily utilized PhishTank as a source for phishing content. Example for Phish tank dataset is which has introduced a newly method to preserve the existing user experience while improving the security. The dataset contains 30 features and 11055 samples which was divided into training data=56% and test data=44%.

The work in [69] used 14 different features and 33000 samples to distinguish phishing websites from legitimate websites as we shown in Table III.

Authors in [73] used Phish tank dataset with different dataset repository (PhishTank,, MillerSmiles, Google searching) and divided into 50% -90% train and test data. The work in [83] used (Open-dns) dataset with phish tank and divided into 80% train and 20% test. Moreover, authors in [84] used various dataset repositories (Alexa and Phishtank) and divided dataset into Train1=50% and Test1=50% and Train2=70% and Test2=30%. Several researches only use highly ranked datasets or widely visited websites for their experiments such as Kaggle. One of dataset in Kaggle repository can be found in [116], which includes 247950 instances, of which 128541 are from phishing URLs and 119409 are from legitimate URLs. It includes 41 features and 1 target variable (0=legitimate, 1=phishing).

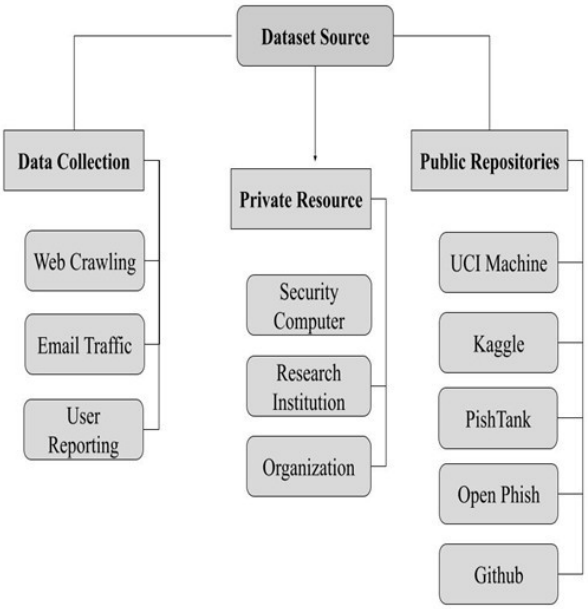

Fig. 4. Dataset sources.

TABLE III.
PHISHTANK DATASET SUMMARY

Ref	Dataset -Source	Samples	features
[81]	PhishTank	11055	30
[70]	PhishTank	33000	14

Furthermore [76, 88, 96] used Kaggle dataset without classified dataset while [91] used Kaggle dataset and classified it into Train=0.67% and Test=0.33%, as shown in Table IV.

Other research works deal with different dataset like UCI [117], which contains several dataset for phishing website that include URL structure, content, and external services. Several works used only UCI dataset, like [86] used UCI dataset and divided into Train=55% and Test=45%. While the work in [78] used various dataset repositories (Ebbu2017, Scenario, Role-based UCI). In contrast, the ambiguous size and classification of a dataset for each research was not mentioned. Other research works [2, 22] used Mendeley [118] dataset for identifying phishing website with 87,111 number of features and different size of simple we shown in Table IV. The research work [13, 84, 87, 93] that utilized the above datasets are summarized in Table V.

On the other hand, some research work proposed new datasets. For example, the work in [119] created two dataset variations that contain 58,645 and 88,647 websites labeled as legitimate or phishing. Another example is the work in [107] proposed PhiKitA, it was a novel dataset, which consists of phishing kits, and phishing websites generated using these kits. These datasets were created to help researchers and users in detecting the phishing websites. Fig. 5, shows the percentage of use of each of the dataset repositories.

TABLE IV.
KAGGLE DATASET SUMMARY

Ref	Dataset Source	Samples		Features	
[88]	Kaggle	1353		10	
[96]	Kaggle	651191	11056	11	32

TABLE V.
MULTI-SOURCE PHISHING DATASET

Ref	Dataset -Source	Samples		Features	
[84]	Alexa, Phishtank	58645	88647	111	111
[7, 18]	UCI , Mendeley	10000	11055	31	49
[97]	Mendeley	58645	88647	111	111
[2]	UCI, Mendeley, D1, D2	10000	11055	31	49
[119]	Mendeley D1, D2	58645	88647	111	111
[93]	Kaggle, Alexa, UCI, Phishtank	11055		32	
[70]	PhishTank, Alexa Rank	7900 , 5800		NA	
[92]	Mendeley	11430		87	
[77]	Alexa, Yandex, PhishTank, OpenPhish	11430		87	
[87]	PhishTank, Yandex	73,575		102	

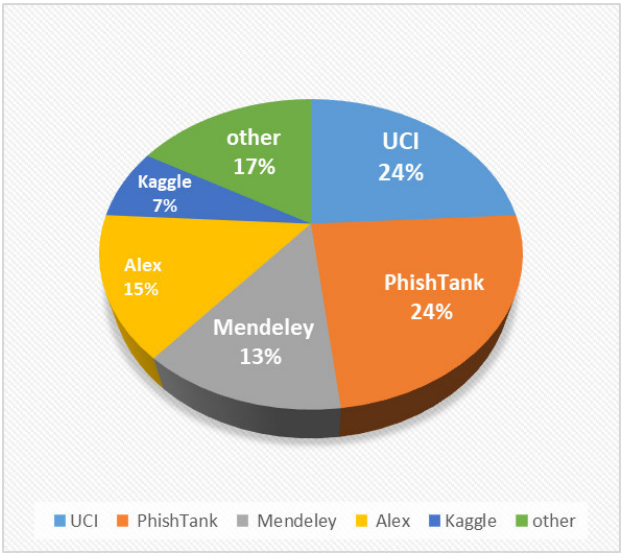

Fig. 5. Dataset repositories.

VI. Phishing Detection Framework

According to the above analysis, the phishing detection process can be put in a general framework. The proposed framework encompasses the five fundamental steps involved in crafting a phishing detection model, as shown in Fig. 6.

Initially, the process begins with defining the research gap and identifying its sources, whether hypothetical, derived from prior global research publications, or real-world issues obtained from cybersecurity-specialized companies, or institutions. Moving to the second phase, the emphasis lies on determining dataset sources and sizes and evaluating the number of features and the number of samples available, Following this, the third phase entails formulating an action strategy, selecting a detection model or combining multiple models, and choosing suitable classifiers aligned with dataset size and information for enhancing accuracy. Proceeding to the penultimate phase, dataset preparation involves preprocessing to eliminate distorted or irrelevant data, ensuring uniform formatting suitable for training and testing the classifier. Each model designed for phishing detection involves reducing the number of features (feature selection) or transforming original features into new sets (feature extraction). Feature selection aims to enhance a machine learning model's performance in terms of speed, accuracy, and efficiency by choosing relevant features from a dataset. Numerous methods, such as filters, wrappers, and embedded methods, are utilized for feature selection in machine learning models. Filter methods rely on statistical measures like correlation, information gain, or chi-square tests to select top-ranked features. In contrast, wrapper methods employ machine learning algorithms to select the optimal subset for maximizing model performance. Embedded methods integrate the feature selection process within machine learning algorithms, such as NN or DT [77, 120]. Finally, in the last phase, the dataset undergoes division into two segments: the training dataset utilized for classifier training, and the test dataset employed to assess classifier performance and determine the accuracy of results.

VII. Outstanding Concerns and Challenges

In many studies, people have suggested different ways to stop phishing attacks. But none of these ways completely stops phishing. Over time, phishing attacks are getting more common and are becoming a popular way to commit online crimes. Whenever researchers come up with a solution to stop phishing, the attackers change their tactics to get around it. So, it's like a close competition between the attackers and the researchers. Phishing scams happen in two main ways: through tricking people with fake emails or websites, or by using harmful software. Solutions to stop phishing are based on these ways of attacking. The most global challenges are:

1. Blacklisting and whitelisting methods have a low success rate of approximately 20% in detecting zero-hour phishing attacks and incur network communication overhead.
2. Machine learning and Deep learning approaches are more effective but are time-consuming, especially with small datasets, and require frequent updates, adding to operational costs.
3. User education plays a critical role in mitigating phishing risks.
4. Improving user interface design with clear warnings and automated malicious message detection can complement user education efforts against phishing attacks.
5. The research used in this article revealed several challenges encountered by the researchers, which can be succinctly summarized like:
    - Runtime analysis problem.
    - Dataset Size.
    - Dataset Splitting.
    - Model (classifier) Selection.
    - Feature selection Techniques.

VIII. Conclusion

This review aims to provide a comprehensive overview of phishing and anti-phishing techniques, bringing together scattered information into one article. Through a systematic review of the literature, we analyzed how well different phishing site detection methods work, examining 122 studies that detail the datasets and algorithms researchers have used over the past five years. Our research included articles and reports from prestigious international cybersecurity organizations specializing in artificial intelligence. The results of this investigation have led to a strategic roadmap to help researchers detect phishing. This roadmap outlines important steps for detecting phishing, focusing on key data sources and model considerations. This paper has presented the most important and common methods used by attackers and the target groups for each type, ensuring the success of the attack depending on the target group. Additionally, we created a timeline that shows the steps attackers take to carry out phishing attacks. From this detailed study, we conclude that although significant progress has been made in understanding and combating phishing, continued efforts are necessary to keep pace with evolving threats. The strategic roadmap and attacker timeline presented in this paper serve as valuable tools for researchers

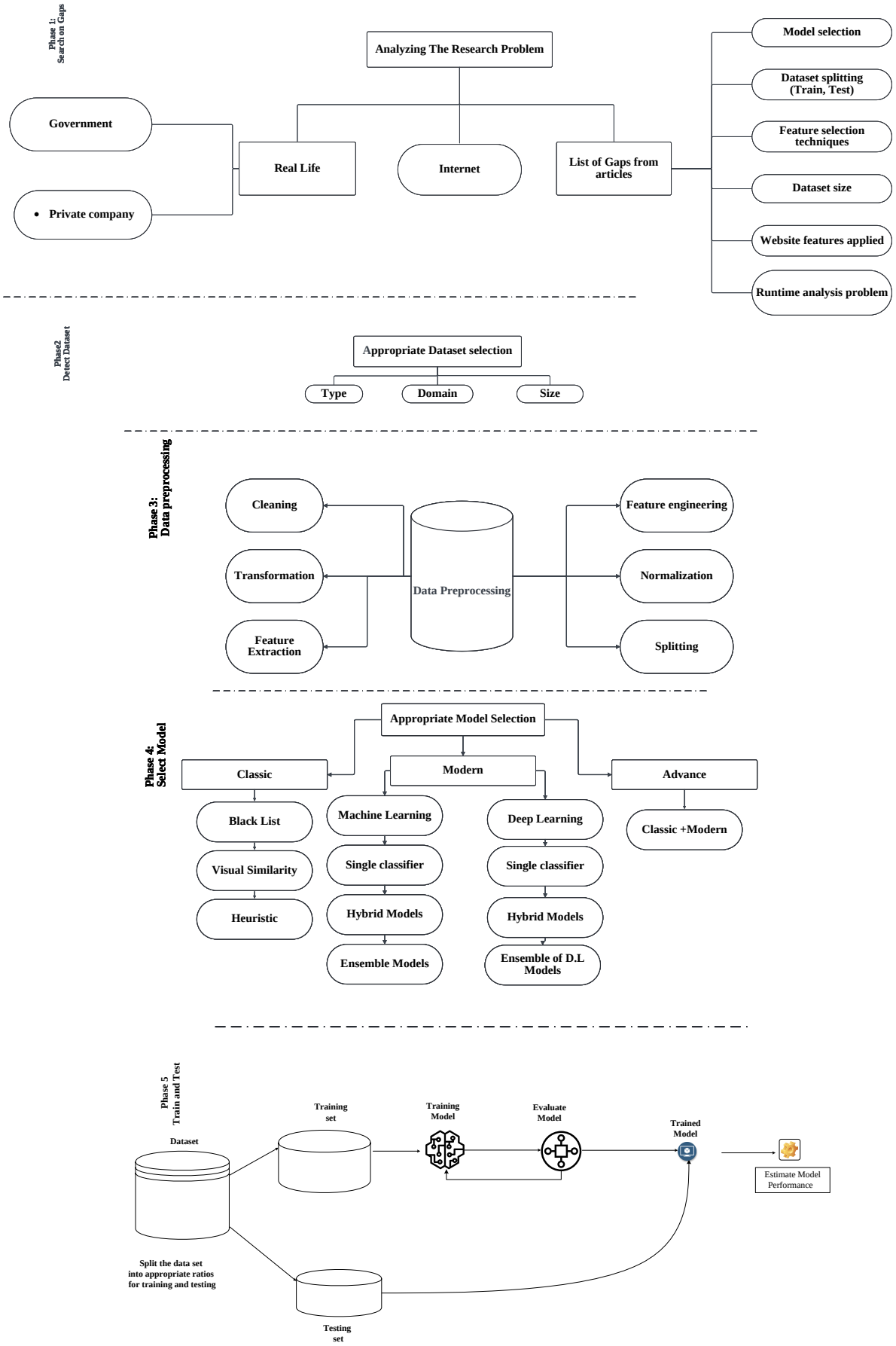

Fig. 6. Phishing detection framework.

and cybersecurity professionals, guiding them in developing more effective anti-phishing solutions. Future research should focus on improving detection methods, exploring new data sources, and using advanced machine learning and deep learning algorithms to enhance the effectiveness of anti-phishing measures.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] K. Haan, "Top website statistics for 2024," 2024. Available online: https://www.forbes.com/advisor/business/software/website-statistics/.

[2] Statista, "Number of internet and social media users worldwide as of july 2024," 2023. Available online: https://www.statista.com/statistics/617136/digital-population-worldwide.

[3] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, "Perceptual representation of spam and phishing emails," *Applied Cognitive Psychology*, vol. 33, no. 6, pp. 1296–1304, 2019.

[4] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," *International Journal of Security and Its Applications*, vol. 10, no. 1, pp. 247–256, 2016.

[5] C. I. Internet Crime Complaint, "2023 internet crime report," 2023. Available online: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf [Accessed on 27 June 2024].

[6] Kaspersky, "The number of phishing attacks doubled to reach over 500 million in 2022," 2023. Available online: https://www.kaspersky.com/about/press-releases/ the-number-of-phishing-attacks-doubled -to-reach-over-500-million-in-2022 [Accessed: 12 Feb. 2023].

[7] G. Anti-Phishing Working, "Phishing activity trends reports," 2023. Available online: https://apwg.org/trendsreports [Accessed on 13 December 2023].

[8] Y. Al-Hamar, H. Kolivand, M. Tajdini, T. Saba, and V. Ramachandran, "Enterprise credential spear-phishing attack detection," *Computers & Electrical Engineering*, vol. 94, p. 107363, 2021.

[9] R. Kalniňš, J. Puriňš, and G. Alksnis, "Security evaluation of wireless network access points," *Applied Computer Systems*, vol. 21, no. 1, pp. 38–45, 2017.

[10] N. N. Pokrovskaia and S. O. Snisarenko, "Social engineering and digital technologies for the security of the social capital's development," in *2017 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, pp. 16–18, IEEE, 2017.

[11] M. A. Chargo, "You've been hacked: How to better incentivize corporations to protect consumers' data," *Transactions: Tennessee Journal of Business Law*, vol. 20, p. 115, 2018.

[12] G. J. W. Kathrine, P. M. Praise, A. A. Rose, and E. C. Kalaivani, "Variants of phishing attacks and their detection techniques," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 255–259, IEEE, 2019.

[13] A. A. Zuraiq and M. Alkasassbeh, "Phishing detection approaches," in *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)*, pp. 1–6, IEEE, 2019.

[14] Cisco, "What are the most common cyberattacks," 2023. Available online: www.cisco.com/c/en/us/products/security/commoncyberattacks.html#$\sim$types-of-cyber-attacks.

[15] B. Gorman, "Different types of hackers: White hat, black hat, gray hat, and more," 2023. Available online: https://www.avg.com/en/signal/types-of-hackers.

[16] Kaspersky, "I'm a phishing victim! what do i do now?," 2023. Available online: https://usa.kaspersky.com/resource-center/threats/handling-phishing-attacks.

[17] NCSC, "Understanding vulnerabilities," 2023. Available online: https://www.ncsc.gov.uk/information/understanding-vulnerabilities.

[18] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B. Alouffi, and A. Ahmadian, "Cloud-based email phishing attack using machine and deep learning algorithm," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 3043–3070, 2023.

[19] A. K. Jain and B. Gupta, "A survey of phishing attack techniques, defence mechanisms and open research

challenges," *Enterprise Information Systems*, vol. 16, no. 4, pp. 527–565, 2022.

[20] J. Rastenis, S. Ramanauskaitē, J. Janulevičius, A. Čenys, A. Slotkienē, and K. Pakrijauskas, "E-mail-based phishing attack taxonomy," *Applied Sciences*, vol. 10, no. 7, p. 2363, 2020.

[21] F. Maggi, "Are the con artists back? a preliminary analysis of modern phone frauds," in *2010 10th IEEE International Conference on Computer and Information Technology*, pp. 824–831, IEEE, 2010.

[22] S. Gupta, P. Gupta, M. Ahamad, and P. Kumaraguru, "Abusing phone numbers and cross-application features for crafting targeted attacks," *arXiv preprint arXiv:1512.07330*, 2015.

[23] S. Yadav and B. Bohra, "A review on recent phishing attacks in internet," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 1312–1315, IEEE, 2015.

[24] R. McCurdy, "The biggest phishing breaches of 2022 and how to avoid them for 2023," 2022. Available online: https://securityboulevard.com/2022/11/the-biggest-phishing-breaches-of-2022-and-how-to-avoid-them-for-2023.

[25] Fortinet, "19 types of phishing attacks," 2023. Available online: https://www.fortinet.com/resources/cyberglossary/types-of-phishing-attacks.

[26] F. times, "Statistic," 2023. Available online: https://firewalltimes.com/category/statistic.

[27] A. R. Mahmood and S. M. Hameed, "Review of smishing detection via machine learning," *Iraqi Journal of Science*, vol. 64, no. 8, pp. 4244–4259, 2023.

[28] Z. H. Ali, H. M. Salman, and A. H. Harif, "Sms spam detection using multiple linear regression and extreme learning machines," *Iraqi Journal of Science*, vol. 64, no. 10, pp. 6342–6351, 2023.

[29] Helpnetsecurity, "Phishing attacks hit all-time high in december 2021," 2022. Available online: https://www.helpnetsecurity.com/2022/03/03/phishing-attacks-december-2021.

[30] B. Parmar, "Protecting against spear-phishing," *Computer Fraud & Security*, vol. 2012, no. 1, pp. 8–11, 2012.

[31] X. Liu, H. Lu, and A. Nayak, "A spam transformer model for sms spam detection," *IEEE Access*, vol. 9, pp. 80253–80263, 2021.

[32] G. Sonowal and K. Kuppusamy, "Phidma–a phishing detection model with multi-filter approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 1, pp. 99–112, 2020.

[33] D. Goel and A. K. Jain, "Mobile phishing attacks and defence mechanisms: State of art and open research challenges," *Computers & Security*, vol. 73, pp. 519–544, 2018.

[34] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018.

[35] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of knowledge (sok): A systematic review of software-based web phishing detection," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2797–2819, 2017.

[36] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, "A keyword-based combination approach for detecting phishing webpages," *Computers & Security*, vol. 84, pp. 256–275, 2019.

[37] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," *Neural Computing and Applications*, vol. 28, pp. 3629–3654, 2017.

[38] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, 2023.

[39] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, vol. 2016, pp. 1–11, 2016.

[40] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *Proceedings of the 4th ACM Workshop on Digital Identity Management (DIM '08)*, (New York, NY, USA), pp. 51–60, Association for Computing Machinery, 2008.

[41] L. Yang, J. Zhang, X. Wang, Z. Li, Z. Li, and Y. He, "An improved elm-based and data preprocessing integrated approach for phishing detection considering comprehensive features," *Expert Systems with Applications*, vol. 165, p. 113863, 2021.

[42] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 4, pp. 301–311, 2006.

[43] K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phishing with discriminative keypoint features," *IEEE Internet Computing*, vol. 13, no. 3, pp. 56–63, 2009.

[44] Y. Zhou, Y. Zhang, J. Xiao, Y. Wang, and W. Lin, "Visual similarity based anti-phishing with the combination of local and global features," in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 189–196, IEEE, 2014.

[45] A. K. Jain and B. B. Gupta, "Two-level authentication approach to protect from phishing attacks in real time," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, pp. 1783–1796, 2018.

[46] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, 2019.

[47] N. M. Shekokar, C. Shah, M. Mahajan, and S. Rachh, "An ideal approach for detection and prevention of phishing attacks," *Procedia Computer Science*, vol. 49, pp. 82–91, 2015.

[48] Z. Liu, B. Yang, J. An, and C. Huang, "Similarity evaluation of graphic design based on deep visual saliency features," *The Journal of Supercomputing*, vol. 79, no. 18, pp. 21346–21367, 2023.

[49] A. K. Jain and B. B. Gupta, "Phishing detection: analysis of visual similarity based approaches," *Security and Communication Networks*, vol. 2017, no. 1, p. 5421046, 2017.

[50] A. Abunadi, O. Akanbi, and A. Zainal, "Feature extraction process: A phishing detection approach," in *2013 13th International Conference on Intelligent Systems Design and Applications*, (Selangor, Malaysia), pp. 331–335, 2013.

[51] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in *2012 International Conference for Internet Technology and Secured Transactions*, (London, UK), pp. 492–497, 2012.

[52] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, pp. 443–458, 2014.

[53] L. A. T. Nguyen and H. K. Nguyen, "Developing an efficient fuzzy model for phishing identification," in *2015 10th Asian Control Conference (ASCC)*, (Kota Kinabalu, Malaysia), pp. 1–6, IEEE, 2015.

[54] J. Solanki and R. G. Vaishnav, "Website phishing detection using heuristic based approach," in *Proceedings of the Third International Conference on Advances in Computing, Electronics and Electrical Technology*, pp. 87–92, 2015.

[55] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in *2010 Proceedings IEEE INFOCOM*, (IEEE), pp. 1–5, 2010.

[56] H. Shahriar and M. Zulkernine, "Trustworthiness testing of phishing websites: A behavior model-based approach," *Future Generation Computer Systems*, vol. 28, no. 8, pp. 1258–1271, 2012.

[57] R. S. Rao and S. T. Ali, "Phishshield: a desktop application to detect phishing webpages through heuristic approach," *Procedia Computer Science*, vol. 54, pp. 147–156, 2015.

[58] Y. Mourtaji, M. Bouhorma, D. Alghazzawi, G. Aldabbagh, and A. Alghamdi, "Hybrid rule-based solution for phishing url detection using convolutional neural network," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–24, 2021.

[59] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.

[60] T. H. Yousiaf and M. S. Al-Tamimi, "The role of artificial intelligence in diagnosing heart disease in humans: A review," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 1, pp. 321–338, 2023.

[61] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, vol. 53, pp. 231–242, 2016.

[62] H. Shirazi, *Unbiased phishing detection using domain name based features*. PhD thesis, Colorado State University, Fort Collins, CO, USA, 2018.

[63] M. Pratiwi, T. Lorosae, and F. Wibowo, "Phishing site detection analysis using artificial neural network," *Journal of Physics: Conference Series*, vol. 1140, no. 1, p. 012048, 2018.

[64] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," *Soft Computing*, vol. 23, no. 12, pp. 4315–4327, 2019.

[65] G. K. Kulatilleke, "Challenges and complexities in machine learning based credit card fraud detection," *arXiv preprint*, vol. arXiv:2208.10943, 2022.

[66] M. Al-Sarem, F. Saeed, Z. Al-Mekhlafi, B. Mohammed, T. Al-Hadhrami, M. Alshammari, A. Alreshidi, and T. Alshammari, "An optimized stacking ensemble model for phishing websites detection," *Electronics*, vol. 10, no. 11, p. 1285, 2021.

[67] H. Zuhair, A. Selamat, and M. Salleh, "Feature selection for phishing detection: a review of research," *International Journal of Intelligent Systems Technologies and Applications*, vol. 15, no. 2, pp. 147–162, 2016.

[68] A. A. Ubing, S. K. B. Jasmi, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019.

[69] A. Jain and B. Gupta, "Phish-safe: Url features-based phishing detection system using machine learning," in *Cyber Security* (M. Bokhari, N. Agrawal, and D. Saini, eds.), vol. 729 of *Advances in Intelligent Systems and Computing*, Singapore: Springer, 2018.

[70] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[71] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum, and M. Hamdani, "Phishing web site detection using diverse machine learning algorithms," *The Electronic Library*, vol. 38, no. 1, pp. 65–80, 2020.

[72] 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), pp. 1–6, 2019.

[73] A. Odeh, I. Keshta, and E. Abdelfattah, "Phiboost-a novel phishing detection model using adaptive boosting approach," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 7, no. 01, 2021.

[74] A. Suryan, C. Kumar, M. Mehta, R. Juneja, and A. Sinha, "Learning model for phishing website detection," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 27, pp. e6–e6, 2020.

[75] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. Bindhumadhava, "Phishing website classification and detection using machine learning," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, IEEE, 2020.

[76] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, and S. Hossain, "Phishing attacks detection using machine learning approach," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1173–1179, IEEE, 2020.

[77] S. R. Sharma, R. Parthasarathy, and P. B. Honnavalli, "A feature selection comparative study for web phishing datasets," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6, IEEE, 2020.

[78] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of ai-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.

[79] A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, "A novel ensemble machine learning method to detect phishing attack," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pp. 1–5, IEEE, 2020.

[80] M. G. HR and A. MV, "Development of anti-phishing browser based on random forest and rule of extraction framework," *Cybersecurity*, vol. 3, no. 1, pp. 1–14, 2020.

[81] G. Harinahalli Lokesh and G. BoreGowda, "Phishing website detection based on effective machine learning approach," *Journal of Cyber Security Technology*, vol. 5, no. 1, pp. 1–14, 2021.

[82] M. Sabahno and F. Safara, "Isho: improved spotted hyena optimization algorithm for phishing website detection," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 34677–34696, 2022.

[83] V. K. Nadar, B. Patel, V. Devmane, and U. Bhave, "Detection of phishing websites using machine learning approach," in *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pp. 1–8, IEEE, 2021.

[84] Q. A. Al-Haija and A. Al Badawi, "Url-based phishing websites detection via machine learning," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 644–649, IEEE, 2021.

[85] A. Ramana, K. L. Rao, and R. S. Rao, "Stop-phish: an intelligent phishing detection method using feature selection ensemble," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 110, 2021.

[86] A. Lakshmanarao, P. S. P. Rao, and M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 1164–1169, IEEE, 2021.

[87] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104347, 2021.

[88] T. A. Assegie, "K-nearest neighbor based url identification model for phishing attack detection," *Indian Journal of Artificial Intelligence and Neural Networking*, vol. 1, pp. 18–21, 2021.

[89] R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy, and T. R. Gadekallu, "Malicious url detection using logistic regression," in *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pp. 1–6, IEEE, 2021.

[90] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, "An intelligent cyber security phishing detection system using deep learning techniques," *Cluster Computing*, vol. 25, no. 6, pp. 3819–3828, 2022.

[91] J. Gu and H. Xu, "An ensemble method for phishing websites detection based on xgboost," in *2022 14th International Conference on Computer Research and Development (ICCRD)*, pp. 214–219, IEEE, 2022.

[92] M. Atari and A. Al-Mousa, "A machine-learning based approach for detecting phishing urls," in *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 82–88, IEEE, 2022.

[93] N. Puri, P. Saggar, A. Kaur, and P. Garg, "Application of ensemble machine learning models for phishing detection on web networks," in *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pp. 296–303, IEEE, 2022.

[94] B. Alotaibi and M. Alotaibi, "Consensus and majority vote feature selection methods and a detection technique for web phishing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 717–727, 2021.

[95] S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Annals of Data Science*, pp. 1–26, 2022.

[96] F. Hossain, L. Islam, and M. N. Uddin, "Phishrescue: A stacked ensemble model to identify phishing website using lexical features," in *2022 5th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 342–347, IEEE, 2022.

[97] L. R. Kalabarige, R. S. Rao, A. Abraham, and L. A. Gabralla, "Multilayer stacked ensemble learning model to detect phishing websites," *IEEE Access*, vol. 10, pp. 79543–79552, 2022.

[98] S. R. Abdul Samad and et al., "Analysis of the performance impact of fine-tuned machine learning model for phishing url detection," *Electronics*, vol. 12, no. 7, p. 1642, 2023.

[99] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[100] W. Chen, W. Zhang, and Y. Su, "Phishing detection research based on lstm recurrent neural network," in *Data Science* (Q. Zhou, Y. Gan, W. Jing, X. Song, Y. Wang, and Z. Lu, eds.), vol. 901 of *Communications in Computer and Information Science*, Springer, 2018.

[101] M. Nivaashini and R. Soundariya, "Deep stacked autoencoder based feature representation for phishing urls detection," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 9, no. 6, pp. 904–916, 2017.

[102] H. Le, Q. Pham, D. Sahoo, and S. C. Hoi, "Urlnet: Learning a url representation with deep learning for malicious url detection," *arXiv:1802.03162*, 2018.

[103] H. Shirazi, K. Haefner, and I. Ray, "Improving auto-detection of phishing websites using fresh-phish framework," *Int. J. Multimed. Data Eng. Manag.*, vol. 9, no. 1, pp. 1–14, 2018.

[104] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," *Wirel. Commun. Mobile Comput.*, vol. 2018, p. 4678746, 2018.

[105] W. Wang, F. Zhang, X. Luo, and S. Zhang, "Pdrcnn: Precise phishing detection with recurrent convolutional neural networks," *Sec. Commun. Networks*, vol. 2019, pp. 1–15, 2019.

[106] A. R. Mahmood and S. M. Hameed, "A smishing detection method based on sms contents analysis and url inspection using google engine and virustotal," *Iraqi J. Sci.*, pp. 6276–6291, 2023.

[107] F. Castano, E. F. Fernández, R. Alaiz-Rodríguez, and E. Alegre, "Phikita: Phishing kit attacks dataset for phishing websites identification," *IEEE Access*, 2023.

[108] S. Singhal, U. Chawla, and R. Shorey, "Machine learning & concept drift based approach for malicious website detection," in *2020 Int. Conf. COMmunication Syst. NETworkS (COMSNETS)*, pp. 582–585, IEEE, 2020.

[109] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, "Efficient deep learning techniques for the detection of phishing websites," *Sādhanā*, vol. 45, pp. 1–18, 2020.

[110] L. Lakshmi, M. P. Reddy, C. Santhaiah, and U. J. Reddy, "Smart phishing detection in web pages using supervised deep learning classification and optimization technique adam," *Wirel. Pers. Commun.*, vol. 118, no. 4, pp. 3549–3564, 2021.

[111] A. K. Dutta, "Detecting phishing websites using machine learning technique," *PloS One*, vol. 16, no. 10, p. e0258361, 2021.

[112] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms," *Journal of Enterprise Information Management*, vol. 36, no. 3, pp. 747–766, 2023.

[113] Y. Wei and Y. Sekiya, "Sufficiency of ensemble machine learning methods for phishing websites detection," *IEEE Access*, vol. 10, pp. 124103–124113, 2022.

[114] A. Chawla, "Phishing website analysis and detection using machine learning," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 1, pp. 10–16, 2022.

[115] R. Alabdan, "Phishing attacks survey: Types, vectors, and technical approaches," *Future internet*, vol. 12, no. 10, p. 168, 2020.

[116] S. Tiwari, "Phishing dataset for machine learning," 2021. Available online: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning.

[117] I. UC, "Browse datasets," 2024. Available online: https://archive.ics.uci.edu/datasets.

[118] Mendeley, "Mendeley data," 2023. Available online: https://data.mendeley.com/research-data/?type=DATASET&search=phishing.

[119] G. Vrbančič, I. Fister Jr, and V. Podgorelec, "Datasets for phishing websites detection," *Data in Brief*, vol. 33, p. 106438, 2020.

[120] Y. Wei and Y. Sekiya, "Feature selection approach for phishing detection based on machine learning," in *Proceedings of the International Conference on Applied CyberSecurity (ACS) 2021* (H. Ragab Hassen and H. Batatia, eds.), vol. 378 of *Lecture Notes in Networks and Systems*, pp. 61–70, Cham: Springer, 2022.