Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777193

PHISHING WEBSITE DETECTION TECHNIQUES: A LITERATURE SURVEY

K.Subashini[1]*, Dr.V.Narmatha[2]
*[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India
[2]Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India
e-mail: * subaphdscholar@gmail.com; balaji.narmatha8@gmail.com

Abstract. Cybercrimes are becoming more widespread every day. Cybercrime techniques include phishing, where hackers try to steal personal data from users of websites by making websites that seem like authentic websites. Internet users are tricked into providing information about their identity, such as access credentials, credit or debit card information, and other details, through phishing attacks, which are unauthorised access. The anti-phishing field has developed a variety of strategies over the past few years. Even though, the problems still continue. An overview of several phishing attempts and information protection strategies is presented in this paper. The various strategies used by different authors over recent years will be thoroughly covered in this survey. This study looks for and highlights the top early techniques, such as supervised machine learning and deep learning, that may be utilised to build a hybrid model that can identify websites as benign or phishing with more precision and accuracy.

Keywords. Phishing; Legitimate; Websites detection; Machine-learning; Deep-learning.

1. Introduction

A few decades ago, phishing was a problem in the online world, and it is still a threat today. As phishers get more inventive in their planning and execution of attacks, the scope of their attacks has expanded and changed through time. Hence, it is mandatory to evaluate both historical and current phishing techniques [1]. In order to know about personal information, hackers pose as familiar websites so that deceive Internet users. Though numerous legitimate techniques, including blacklist or whitelist, visual similarity and heuristic -based methods have been presented to date, online users become prey to these phishing websites and they lose their personal information. In website phishing, the attacker creates a fake website that resembles the real thing and lures internet users there with advertisements like social networks i.e., Twitter, Facebook, etc. Some hackers can control phishing websites using security features like a green padlock, an HTTPS connection, etc. As a result, the validity of a website cannot be determined by an HTTPS connection [2].

Various organizations, including NSFOCUS and the “Anti-Phishing Working Group” (APWG), undertook surveys of these attacks. An international non-profit group called APWG examines attacks involving phishing reported by members, who include companies that make law enforcement organisations, security products, governmental organisations, service-oriented businesses, trade associations, communications companies and regional international treaties. Contributing members of APWG research cybercrime's dynamic characteristics and methods. The APWG has improved its reporting procedures with the help of this report. The phishing emails are collected and sent to APWG, they examine from URL it has been sent. In order to analysis the authenticity of phishing URLs [3].

Distinct phishing websites: A lot of people around the world report this kind of phishing. Based on the distinctive phishing emails and base phishing URL website, they are sent to the APWG repository. (Millions of modified URLs, all focusing on the same trick, are used to promote a single phishing site.) The APWG measures reported phishing sites more precisely by taking into account the methods used by phishers to create phishing URLs [3].

In March 2022, APWG reported 384,291 cyberattacks. The APWG marked 1,025,968 phishing hacks in the beginning of 2022. At the beginning of 2022, the quarterly total for phishing incidents exceeded one million during this quarter, which was the worst APWG had ever recorded. The previous high number of attacks was 888,585 in the last quarter of 2021[3].

In the beginning of 2022 APWG a new history was created as it exceeded the phishing rate of 1,025,968. In the second half of 2022 APWG recorded 1,097,811 phishing assaults which the higher witnessed phishing crime [4].

Most-Targeted Industry Sectors

During the first quarter of 2022, the phishing websites targeted the financial sector (i.e, banks). According to the APWG survey 23.6% targeted on these sectors. After the holiday shopping season, assaults on retail and e-commerce websites dropped from 17.3 to 14.6 percent, although attacks on email and cloud based services (SAAS) providers persisted. At the beginning of 2022, From 8.5 to 12.5 percent, phishing attempts on social networking websites surged. Phishing attacks against crypto currency targets including exchanges and wallet providers have been constant since late 2021, marginally rising from 6.5 to 6.6 percent in the most recent beginning. Opsec Security offers top-tier brand protection solutions, as shown in Figure 1 (a). [3].

One of the first APWG members is Opsec Security, claims that banks are the target of phishing websites, and that in Q2 2022, this source accounted for 27.6% of all malicious websites. E-mail and cloud based services (SAAS) suppliers continue to be frequently attacked, despite a drop in retail/ecommerce website attacks from 14.6 to 5.6 percent. Phishing attacks targeting social news outlets increased, reaching 8.5 percent of all assaults in the final quarter of 2021 and 15.5 percent in the second period of 2022. Attacks involving phishing continue and have increased than assaults against official websites, online gaming, and telecom services combined. These cyber-attacks targeted bit coin providers of wallets and exchanges. Opsec reported a 43 percent rise in phishing in Q1 2022 that are shown in Figure 1 (b) [4].

The paper is also discussed in this section. The background, various phishing attack types, phishing methodologies, and methods for phishing detection are all covered in Section 2. Section 3 presents the analysis of past literature studies for the identification of phishing attempts. The methodologies of different datasets, feature extractions, and feature selection are covered in Section 4. The approaches and analyses for phishing detection are stated in Section 5. The comparison between existing phishing website detection is provided in Section 6. Issues and difficulties are obtained in Section 7. Section 8 represents the conclusion of the survey.

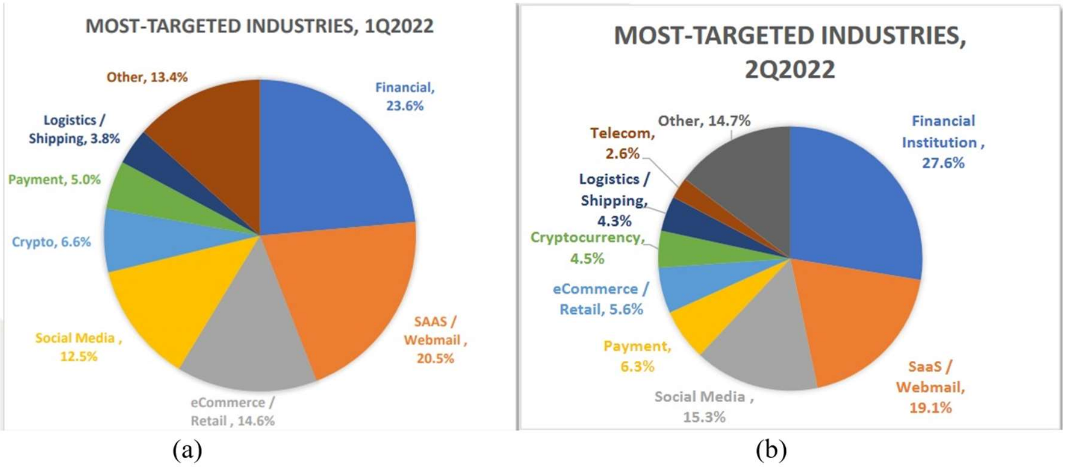

Figure 1 (a,b) Unique phishing sites detected during 2022 in an APWG survey report [3, 4]

2. Background Study

2.1 History

John Draper first used the term "phishing" in the 1970s. He developed the controversial Blue Box, which released detectable tones for telephone network hacking [5]. Online conartists began conducting social engineering attacks on ‘America Online’ (AOL) accounts in 1996 [6]. The word "phishing" comes from the comparison of a fishing operation. ‘ph’ derives from "phone phreaking," It were widely used in the 1970s as a kind of telephony assault [7]. The initial target of this attack was America Online Network (AOL) [7]. Furthermore, phishers aggressively impersonated a sizable number of payment access point, banking websites and social networking in addition to the AOL website.

2.2 Types of Phishing Attacks

Phishing attacks come in a variety of types. These assaults' main objective is to steal critical information from end users. Various phishing attack types are shown in figure 2.

2.2a Email Phishing: Whenever a person opens on the email in this kind of phishing attack, all the information they have submitted will be forwarded to the attacker [8] with relation to any issue, update, or sensitive subject that needs to be changed right away. The most frequently phishing assault in the online world is this one. The malicious actor is attempting to deceive people into visiting a link or sending emails which appear to be from a trustworthy organisation and downloading a file. The hacker's target is to steal the user's personal information and install malicious code files on the end user's server. However, monitoring closely shortened links and regularly checking for spelling mistakes is the most ethical way to spot this attack. Attackers also create their mail IDs using unauthorised domains.

2.2b Spear Phishing: Although similar to email phishing, this assault also happens over web mail. In this kind of attack, phishers collect data on individuals from publicly accessible databases, such as social media or business websites, and then target them with emails that use actual names from the organisation. This gives the targeted individual the impression that he has received mail from a member of the organisation, leading to prompt responses to the correspondence. Looking at unexpected emails received from various organisational departments can help you spot this kind of phishing, though. Instead of targeting random programme users, attackers in this scenario target specific people or businesses. It's a more sophisticated form of phishing that necessitates in-depth familiarity with an organisation, including its hierarchy. Unlike phishing, emails are sent to specified recipients in this attack [8].

2.2c Whaling: It is denoted to as a whaling phishing. It is a form of spear phish in which hackers target prominent workers, like the CEO or CFO, in order to steal confidential data from a business. These personnel will have full access to sensitive information because they are in higher positions inside the company. Getting further details will be simple [9].

2.2d Smishing: Additionally, it is called SMS phishing. It is an instance of social engineering abuse used to acquire user passwords, contains credentials, financial information, and personal information. Smishing also seeks to use victim funds for money laundering. Scammers use SMS text messages to deliver phishing messages with a malicious link attached. When recipients click the fake link in the phishing email, they are taken to a phishing page where their personal data is collected [8][10].

2.2e Vishing: Another name for it is voice phishing. Victims' voice messages are used in this kind of phone scam to extort victims of their personal information or money. Automated speech recordings are used in phishing to entice victims. Vishing involves contacting a target and telling them their bank account has been hacked using an automated voice. The recipient is then instructed to contact a designated toll-free number in the voice message. When customers dial that toll-free number, the phone's keypad is used to collect the user's bank account number and other personal information [8][10][11].

2.2f Pharming: "Phishing without a lure" is another name for pharmacological warfare. When a user tries to browse to a website, their computer can either check a local hosts file—a defined mapping file—or a DNS server on the Internet to discover the IP address. A victim's hosts file on their computer is altered in two typical types of pharming: hosts file pharming and DNS poisoning [12].

2.2g Content-injection Phishing: In order to increase user trust and facilitate data entry, the content of the legitimate website is substituted with some random information and input fields that are similar to those on the authentic website [12].

2.2h Search Engine Phishing: It happens when phishers build websites with alluring offers and get them genuinely indexed by search engines. Users are tricked into providing their information on these websites while conducting routine searches for goods or services [12].

2.2i Angler Phishing: This attack, which usually occurs on social networking sites and combines smishing and vishing, involves sending the targeted person a voice message or a direct message to coerce him into taking the required action. As a result, it is always recommended to ensure that the social media user is a genuine individual and not just a fake.

2.2j Evil Twin: Hotspots are the target of this assault, in which the perpetrator sets up a false Wi-Fi hotspot, grants users access, and then steals their login information. In this situation, the attacker uses a Man-in-the-Middle assault and an evil twin. Users should never sign in to a hotspot without first entering their credentials.

2.2k Water Hole Phishing: The attacker keeps note of the websites a corporation uses and then changes the IP address of that site to the phoney site in this assault, which is likewise targeted at corporate individuals. Unaware of this action, the company's employee visits the website and, by downloading malicious files, falls victim to this attack. One approach to protecting yourself from this attack is to restrict browser restrictions. The system can also be protected from such server intrusions by using security firewalls.

2.2l Clone Phishing: The majority of the time, malicious individuals keep tabs on a person's typical clicking habits and use these services to their advantage in order to obtain them. Attackers conduct research about the services that a company invests in before focusing their mail on those companies. If you wish to be aware of this hazard, look for emails that ask for personal information in order to offer normal services.

2.2m Pop-Up Phishing. Pop-ups are generally regarded as those alert boxes that show up when a user visits a website. In this scenario, malicious actors insert codes inside the pop-up ad blocks to coerce people into clicking on them. Codes are immediately installed on the user's end after clicking the "Allow" button. The only way to stop these pop-up advertisements is to go to full screen mode.

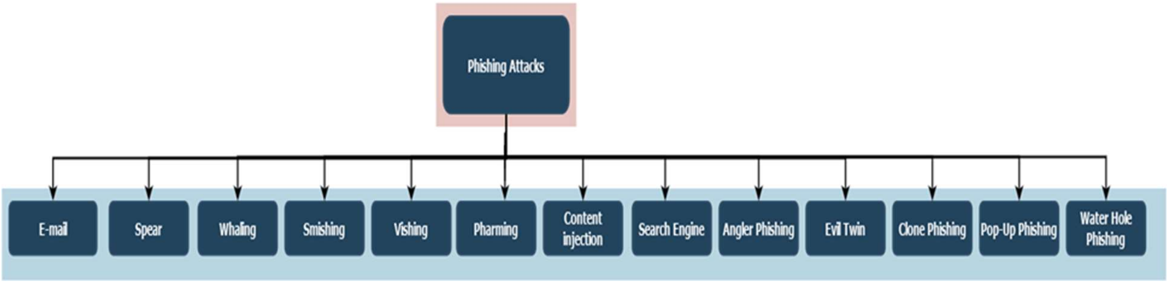

Figure 2. Phishing Attacks types

2.3 Phishing Techniques

Attackers employ a variety of methods to carry out different phishing schemes. Attackers can use these strategies to circumvent security measures and steal sensitive data from end users [13].

- Link Manipulation
- Website Forgery
- Pop ups

2.3a Link Manipulation: Phishing scams frequently employ the link manipulation method [13]. It is accomplished by tricking a person into visiting a link that leads to a false website. Attackers are currently using deceptive tactics to persuade people to click, including: misspelled URLs, utilisation of sub-domains, hidden URLs, and IDN (Internationalised Domain Name) homomorphic attacks.

2.3b Website Forgery: Another phishing method is website forgery [13], which involves creating a rogue website that pretends to be a legitimate one in order to deceive people into providing confidential information like account information, credit card numbers, passwords, etc. Cross-site scripting and website spoofing are the two basic methods for carrying out web forging.

2.3c Pop ups: One of the simplest ways to carry out successful phishing schemes is by using pop-up messages. By giving consumers pop-up messages and eventually directing them to bogus websites through these pop-ups, they enable hackers to obtain login data. When a user is doing online banking, a pop-up window impersonating a message from the bank called "in-session phishing" will appear [13].

2.4 Phishing Detection Techniques

In general, methods for detecting phishing include heuristic-based identification, machine learning-based identification, list-based identification, and deep learning-based identification. However, because there is no definite answer to completely eliminating all risks due to how complex the phishing problem is, different strategies are frequently used to stop particular attacks. As displayed in Figure 3, there are two basic categories of protection techniques: increasing user knowledge and using additional software. The phishing sites can be found using a variety of detection methods.

2.4a List based: There are two variations of this technique: blacklisting and white-listing. These are also denoted to as conventional methods or database-oriented methods. They have very quick response times and excellent detection accuracy [14][15].

Black-Listing According to this method, URLs that are thought to be phishing sites are saved in a database. When a new URL is input, it is compared to the URLs in the database, and if they match, the browser blocks it and stores the matching URL in the database for later use. The drawback of this method is that zero-hour phishing assaults cannot be detected [15].

White-Listing According to this method, a database is used to hold valid URLs that are used to validate new URLs. In this method, whenever a new URL is input, the database is first examined to see whether it already exists. If not, the complete URL's information—including its domain names, SSL certificates, and linked-to website hypertext—is then verified before being recorded in the database. The disadvantage of the whitelisting strategy is whether the websites that have been registered as genuine are indeed authentic or just pretending to be legitimate. The drawback of listing approaches is their high space requirements [15].

2.4b Heuristic-Based Detection: It's a development of the listing approach. This method extracts website characteristics like URLs and content and uses them to compare other websites. These new websites are regarded as phishing sites if they match. These are superior to listing procedures and produce more accurate findings, although they respond slowly [16] [17][18][19] discusses an alternative method for detecting zero-hour phishing attacks.

URL-based detection algorithms are more common to increase the speed of the detection. If machine learning mixed with URL-based characteristics improves accuracy rates [20] [21]. This determines whether the websites are real or not by comparing their content to that of legitimate websites [22]. However, this identification approach fails when it comes to numerous websites that have little content. These days, pictures are used instead of webpage content [23].

2.4c Machine learning-based identification methods: Using a variety of machine learning classifiers, including the random forest, decision tree, kNN, and support vector machine, machine learning creates models from the numerous datasets collected from the different website properties [35]. Machine learning addresses the issue of zero hour phishing assaults by helping to forecast websites even before they are built. The classifier's performance varies according to the dataset's size and the kinds of characteristics used. Frameworks for phishing attack detection are also being developed [34][36][37].

2.4d Deep learning-based identification methods:

Artificial neural networks are used in deep learning, a type of machine learning, to carry out sophisticated computations on enormous amounts of data. It is a sort of supervised learning that is a function of the central nervous system and inspired by its structure. Machines are trained using deep learning algorithms by studying examples. Deep learning is frequently used in sectors like medical services, media, e-commerce, and marketing. Different algorithms are used in deep learning models. A few algorithms are more suited to certain problems than others; that is, recurrent neural networks, auto-encoders, deep neural networks, and long term short memory[46][47].

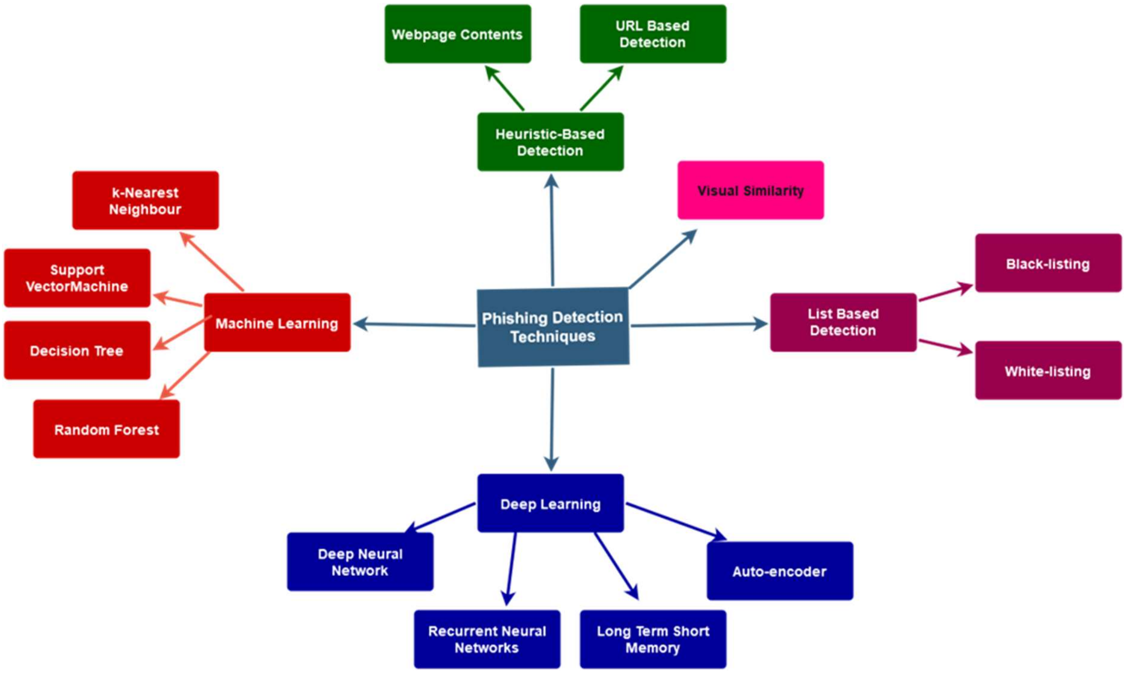

Figure 3. Phishing Detection Techniques

3. Literature review

Jain, Ankit Kumar, and B. B. Gupta [67] proposed that phishing is the course of identity theft, online users are duped into divulging their personal information, such as login passwords, bank account details, and other specifics. A number of anti-phishing strategies have been created by researchers recently, yet the issue is still present. This essay gives a systematic examination of phishing attack strategies and countermeasures.

Ojewumi, Theresa O., et al. [68] proposed the rule-based method to phishing detection used in this research, which was developed on a dataset of fourteen (14) characteristics using three machine learning approaches. The three algorithms used performed best when the random forest method was tested. During the development of PhishNet, web technologies including html, css, and Javascript were used.

Jalil, Sajjad, Muhammad Usman, and Alvis Fong [69] proposed that phishing is a kind of internet assault where the perpetrator poses as a credible website in order to trick the victim into disclosing personal information, including usernames, e-mails, passwords, and bank or credit card information. To prevent such assaults, several phishing detection methods have been offered. This research proposes a powerful machine learning system that anticipates phishing websites without visiting the website or utilising any third-party service. The suggested method is focused on URLs and makes use of the hostname, protocol scheme, entire URL, suspicious terms, brand name matching, and path and region of the URL.

Das Guptta, Sumitra, et al ., [70] an algorithm based on machine learning that uses hybrid characteristics focused on hyperlinks and URLs to identify phishing websites in actual time without the use of external systems. The results of our tests show that the proposed phishing method for detecting works is superior to conventional methods, with a detection accuracy of 99.17% using the XG Boost methodology.

Ramana, A. V., K. Lakshmana Rao [71] proposed that online users continue to become a target of phishing attempts despite the implementation of many anti-phishing strategies. In this study, we provide an intelligent model to identify fake websites using a collection of different feature selection techniques. Considering the outcomes of our experimental investigation, we are able to recognise data from UCI and Mendeley with an accuracy of 97.51%.

Ghaleb Al-Mekhlafi, Z., et al. [72] proposed a group classification algorithm's optimization for identifying phishing website. A genetic algorithm was used to optimise the recommended solution by modifying the parameters of many ensemble machine learning techniques., including random forest, lightgbm, adaboost, bagging, xgboost, and gradientboost. 4898 phishing websites and 6157 genuine websites made up the dataset used in the research in this paper.

According to Singh and Charu [73], the review advises readers to practice phishing protection by increasing their knowledge of phishing attacks, assisting them in identifying phishing attempts, and more. In phishing, to target a specific individual, scammers use SMS or email as a weapon, sending a group of people a URL link that will trick them. Many evaluations suggest using machine learning to identify phishing attacks.

Harinahalli Lokesh, Gururaj [74] proposed a phishing classification system that pulls characteristics designed to thwart popular phishing detection techniques. Machine learning is the scientific study of algorithms and data analysis that has demonstrated success in combating phishing pages when differentiated. The use of ML approaches to recognise phishing assaults is analysed in this study, together with the pros and cons.

Sánchez-Paniagua, Manuel, et al [75] suggested that due to the vast number of organisations participating in online transactions and services, phishing assaults are among the most difficult social engineering intrusions. In these attacks, hackers utilise a login form that looks just like the genuine website to steal user credentials or sensitive information before sending it to a hostile server. According to test findings, a LightGBM classifier can identify phishing websites with a 97.95% accuracy rate.

Khan, Sohail Ahmed, and Wasiq Khan [76] proposed that phishing attacks, which have been hurting people and companies all over the world, are the most prevalent kind of cyber assaults used to gain personal information. In recent years, the deployment of machine intelligence has led to the proposal of a number of strategies for specifically identifying phishing assaults. According to the statistical findings, artificial neural networks and random forests perform better than other categorization systems.

Sameen, Maria, and Kyunghyun Han [77] Deep neural networks were suggested as a mechanism for creating phishing URLs that might be used to launch attacks. We created PhishHaven, a collective detection method using machine learning, to stop this sort of assault. With 100% accuracy, it can tell the difference between phishing URLs created by humans and those created by AI. We also present a solution to the open problem of handling small URLs: the URL Hit method.

Faris, Humam, and Setiadi Yazid [78] proposed that the goal of phishing is to deceive consumers into submitting critical information to the scammer by using phony websites that seem just like legitimate web pages. Between 2015 and 2020, phishing websites will likely become more prevalent. In this study, the authors assessed a number of approaches and put forth rule-based software programmes that are more effective in phishing detection.

Korkmaz, Mehmet, and Ozgur Koray Sahingoz [79] proposed that the majority of security administrators employ learning algorithms that are trained on pre-collected datasets by making use of attributes found in web page URLs and content. By conducting a comparative examination of the literature, the goal of this work was to analyse the elements that had previously been employed in the classification of web pages. Through this study, it is anticipated to produce a universal survey resource for academics working on network security or categorising websites.

Tang, Lizhen, and Qusay H.Mahmoud [80] proposed that phishing, a significant type of fraud that includes tricking fake websites with risky links, is a growing problem. Getting user personal details, and using that information to pretend to log into associated accounts in order to steal money. This report provides a cutting-edge analysis of techniques for the identification of fraudulent websites. It starts out by discussing about the phishing life span, then goes on to discuss about identifying phishing approaches, then mostly concentrates on recognizing scams, and has a solid understanding of artificial learning-based solutions.

Somesha, M., et al., [81] propose that phishers is a dishonest practice and a sort of information security in which private information is obtained by mimicking trustworthy websites. Various anti-phishing solutions are already available, as well as those focused on heuristic characteristics, blacklists or whitelists, and visual similarities. The authors of this study provide brand-new phishing URL detection models that utilise just 10 characteristics and one feature from a third-party service.

4. Methodology

4.1 Dataset.

Phishing detection is an important topic of study, yet the absence of large sample datasets has limited the discovery of actual issues and the development of efficient solutions. Data is the source of any method and has a significant impact on results. Importing public datasets and directly getting URLs from the website are the two methods for obtaining data. Table 1 displays a variety of important data sources. Every row's data object in these three released datasets has numerous characteristics taken from a website and a label of classes. Implementing accessible APIs or data mining programmes on websites might help obtain the original URL strings. Datasets are crucial in anticipating fraudulent websites. Data will be collected from multiple sources in practices such as machine learning, neural networks, and deep learning. A model will be created using the collected data, and the platform's ability to correctly identify urls will be tested. The names of legal sites that are likely to be used for phishing are included in Alexa and Common Crawl [26][27]. End-users submit suspicious URLs to Phish-tank and Open-Fish to assess if these websites are phishing scams [28][29]. The dataset is acquired and saved in the UCI-Machine learning repository from various sources. This dataset is mostly used for research [30]. The dataset in Majestic comprises domains with referring subnets [31]. Kaggle is a web-based data repository that holds a large number of data points obtained through different sources [32]. Such data sources are valuable for the training set.

4.2 Feature Extraction

Using software solutions for feature extraction reduces time by replacing the manual procedure and thereby improves phishing detection quality. Because training cannot be conducted on strings, Make an extracted features which translates labelled input Links into encapsulated features. The effects of a website are used to classify all of the features. There are several ways for detecting features. They are URL features and Content features.

4.2a URL features:

Machine learning detects phishing URL characteristics. Host-based features and lexical features are examples of URL-derived characteristics [62][63]. URLs are simply separated into subparts that include a host name, a route, a protocol, or a scheme. The correctness of a site's authenticity can be accessed based on any combination of these components [62]. There are 30 different phishing website elements.

IP address. In phishing, instead of the web address of the website, the IP address is utilised.

URL length. In order to hide suspicious areas, phishers use long URLs in the web address.

Tiny URL. On the "internet," a method, URL shortening, in which the URL is regarded as being less in length, is used. Phishers deceive consumers by using "Tiny URLs," in which a lengthy URL is used to link the tiny URL [64]. URLs with the "@" sign are used by phishers to deceive users.

Redirecting using "//". By including "//" in the URL route, the user is sent to a different website. The number of pages in a valid document is fewer than two; in a suspicious document, it is between two and four; otherwise, it is termed "phishing" [64].

Adding (-) sign to the Domain. A dash sign is used in normal URLs to make harmful URLs that trick users [65].

Dots in URL. A legal or phishing website's URL's dot count. If the number of dots is larger than the legitimate number, it is dubbed "Phishing" [66].

HTTP with SSL. Legitimate websites utilise HTTPS to transfer sensitive data. It requires a certificate for use and a minimum age of two years.

Domain registration length. While legitimate domains are bought years in advance, phishing websites are only up for a brief period of time.

Favicon. A graphic picture known as a favicon is formed on a web page. Phishing is committed when the URL and the domain's favicon differ. [65] [64].

Using non-standard port. A certain server's service is up and down. User data is at risk if all ports are open. It is recommended that only required ports be opened to control infiltration.

"HTTPS" Token in the Domain Part. To deceive individuals, the HTTPS sign is appended to the URL.

Table 1. Phishing and legitimate website datasets

S.No.	Dataset Name and Sources	Size of Dataset	Type	Remarks
1	Alexa (https://www.alexa.com/)	1 Million URLs	Website	Legitimate URLs
2	PhishTank (https://phishtank.org/)	68,40,198 URLs	Website	Valid Phishing URLs
3	OpenPhish (https://openphish.com/)	11,100,075 URLs	Website	Valid Phishing URLs
4	University of california irvine (http://www.uci.edu/)	2,14,748 URLs	Published Dataset	Valid Phishing URLs
5	Mendeley (https://data.mendeley.com/)	2 Millions URLs	Published Dataset	Phishing + Legitimate URLs
6	CommonCrawl (https://commoncrawl.org/)	940 Millions URLs	Website	Legitimate URLs
7	Kaggle (www.kaggle.com/)	11,430 URLs	Published Dataset	Phishing + Legitimate URLs
8	Majestic Million (https://majestic.com/)	Million URLs	Website	Legitimate URLs
9	Phishstorm (https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset)	96,018 URLs	Published Dataset	Phishing + Legitimate URLs
10	DMOZ (https://dmoz-odp.org/)	3,861,202 URLs	Website	Legitimate URLs

4.2b Content based features:

These properties, such as pictures, anchor links, and favicons, are retrieved from the loaded webpage content.

External Favicon. The website's resources, such as favicons, graphics, and logos, are fetched from a private host or a linked hostname, respectively. The favicons on certain spoofing sites, however, are loaded from an outside host; as a result, in these circumstances, the value is assigned to 1; otherwise, it is assigned to 0.

Request urls. To load the whole webpage, the genuine page requests resources (images, html, logos, and favicons) from the same domain as the visited page. Furthermore, the genuine page may contain a small amount of URLs from other sites making requests. As a result, set as 1 if the webpage has 22% external request URLs. However, if the external request URLs are between 22 and 61%, the value is set to 0, suggesting suspicious activity. If the proportion of external request URLs is greater than 61, the value is set to 1.

Link Anchor. This function is comparable to requests URLs in that it recognises outside anchoring links, i.e., links anchor referring to external domains, rather than external request URLs. If the proportion of external anchor links is less than 31, the function is disabled. If the proportion falls between 31 and 67, the value is set as 0. (Suspicious) If the proportion of external anchor connections is greater than 67, the value is set to 1.

Anchor Links in tags. In contrast to the hostname in the Web address, the links anchor from meta, script, and links are extracted. When the anchor links in the tags match the domain URL, they are considered local domains. If the percentage of local domains is less than 17%, the website is deemed authentic, and the value is set to 1. If the percentage of local domains is between 17 and 81, the webpage is regarded as suspicious and the value is set to 0, otherwise set as 1.

Server form handler. Actions like a local domain webpage or an outside subdomain webpage, a blank page, or an empty string are included in the server form. Normally, the information in the form is delivered to the local domain; so, if the server form contains the local domain, set as 1. If the server form includes empty or about: blank, the page is regarded as suspicious and the value is set to 0; otherwise, the value is set to 1 since the server form has a domain.

Emails submit. For the existence of phishing activity, this feature examines if the server form uses the mail() or mailto() methods. The value is assigned to 0 if the server form uses the following methods to submit user data, otherwise it is assigned to 1.

Abnormal url. Reputable websites include their product in the URL; scammers, however, may spell it incorrectly, omit the hostnames or use a different name in the URL. If the brand name does not appear in the URL, the value is set to 1, otherwise set as 1. The number of connections that forward from websites. Redirect links are used more frequently on phishing websites to trick users. If there are less than two divert links, the value is set to one. If the divert links are between 2 and 4, the value is set to 0, otherwise set as 1. (i.e., redirect links greater than 4).

Status bar with mouse over. If the URL in the web address and taskbar match, a link will show when your cursor is over it; otherwise, the value will be assigned to 0.

Pop-up window. The website is considered fraudulent and the value is assigned to 1 if a pop-up window is utilized to collect sensitive information; otherwise, it is assigned to 0.

Right click status. Right-clicking is restricted by the intruder since it could reveal the website's program code.

As a consequence, the value is assigned to 1 if right click is disabled on a website, otherwise to 0.

iframe redirection. iframes are utilized by attackers to display requests from other websites. As a consequence, the value is assigned to 1 if the website utilise iframe and to 0 otherwise.

Links pointing to page. The presence of internet connections leading to the hostname is determined by this attribute. Authentic websites have certain backlinks that point to the hostname, while scammers don't have any. As a result, the value is assigned to 1 when there are no connections leading to the website. The domain is marked as suspicious and set to 0 if it has either one or two links pointing to it; otherwise, it is marked as 1.

4.3 Feature Selection

Using just vital information and excluding unnecessary information, In order to reduce the model's input variable, feature selection is an approach. It is the technique of immediately selecting suitable characteristics for the model of machine learning depends on the kind of issue that are trying to resolve. By doing this, critical elements may be included or eliminated without affecting them. It improves reduce the size of inputted data and the volume of unnecessary data. The following are the feature selection algorithms used in the proposed model.

Information Gain. When changing the dataset, information gain affects the amount of entropy that is reduced.

By computing the data gain of each parameter with regard to the target variable, it may be utilised as a feature selection strategy.

Relief-F. The importance of each characteristic is updated depending on the selected instance and nearest neighbour instance pairings of the same and opposite class. If a difference in feature value is found between a feature value and a neighbouring instance of the same class (hit), the feature score will be lower. Similarly, if a difference in feature value is found between a neighbouring instance of a different class (miss), the feature score increases.

Fisher Score. One of the most used strategies for selecting supervised features is the Fisher's score. In decreasing order, it returns the variable's rank according to Fisher's criterion. Following that, one may choose the variables with a high fisher's score.

Recursive Feature Elimination (RFE). This method is a wrapper-type feature selection methodology, in which machine learning is utilised at the core and wrapped by RFE to choose the appropriate features. During the training phase, RFE evaluates all features and effectively eliminates the least performed or least significant features until getting the desired number of features. The RFE works with the specified ML (say, Random Forest), which then ranks features by relevance and updates the feature set by deleting the least important features. The updated features are refitted into the model, and the procedure is repeated until the required number of features is obtained.

5. Phishing Detection Techniques Analysis

About 20% of phishing assaults launched on the system server can be detected using any of the usual methods now in use. Software developers frequently use this technique for identifying threats using ML techniques, despite the fact that it takes time and typically only works on a limited dataset. In addition to ML, other deep learning-based approaches are also demonstrated. Since DNN designs make it simple to execute detection and classification strategies that utilise several layers of the underlying architecture, deep learning techniques frequently implement them. To identify the same, ML and DL algorithms are applied in contrast to heuristic-based processes. An overview of three detection methods is provided in this section.

5.1 Machine Learning based Detection Technique

In order to develop an optimised model, machine learning (ML) in malware detection must first effectively identify and categorise labels. ML techniques are utilised in phishing to target websites that are in charge of injecting such assaults. The collected dataset has to be educated so that it can recognise all of the characteristics of the phishing website. This is a pictorial representation illustrating the typical process used by machine learning to identify and categorise phishing.

The authors of [82] used a mix of binary classifiers and random forest as a classifier to carry out phishing detection procedures. The scientists used a trained random forest approach to run the detection algorithm after obtaining the dataset from the Mendeley website. The algorithms used for feature selection selected effective features, which were then categorised using the same techniques.

"A., Supramaniam, N., & Jhanjhi. (2019)" in [83] used the concepts of bagging, boosting, and stacking to implement the detection approach. Using this technique, an ensemble learning model was created that could pull out 30 features from the dataset. To get the requisite accuracy, further implementation was done once the dataset was taken from the UCI source.

The study in [84] also used ML approaches to distinguish between authentic files and those used in phishing attacks. The authors used SVM, ANN, and random forest classifiers to carry out their research. It was found that RF classifiers worked more effectively and provided better results.

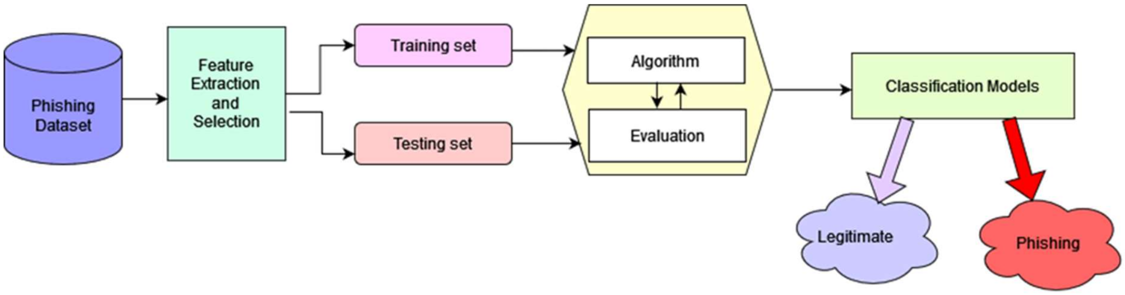

Figure 3. Machine Learning-based Detection Technique workflow diagram

5.2 Deep Learning based Detection Technique

Intrusion detection systems also make use of deep learning principles. Numerous studies also found that a neural network's workflow outperformed machine learning methods and aided in achieving higher levels of accuracy. The operating principles of RNN and LSTM are typically included in deep learning algorithms. Additionally, it has been demonstrated that employing DL to execute detection and classification procedures outperforms using ML. A traditional RNN has several layers, including hidden layers in the center that are in charge of the primary classification and detection steps. The typical process of a DL model for malware detection and classification is shown below.

Pages of a website were worked on by authors in [85]. In order to identify phishing assaults, they concentrated on building a website layout for comparison considerations. Page phishing classifiers like SVM and decision trees were employed to accomplish this goal.

The authors of a different study [86] provided their findings on a related method of conducting phishing page classifiers. However, in this instance, they used more than two datasets to perform their trials, which comprised information on the payment gateway and connections to 1530 phishing websites.

Finally, authors published a thorough and in-depth analysis of attack tactics and detection approaches in another survey [87]. They used feature extraction techniques as well to get more accuracy.

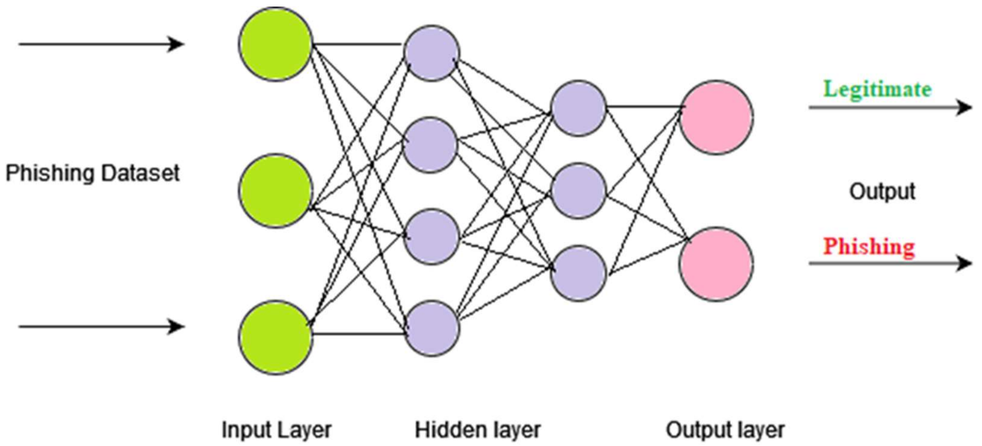

Figure 4. Deep Learning-based Detection Technique workflow diagram

5.3 Performance Evaluation for Phishing Website Detection

In our experiments, using the following metrics, we assess how effectively the phishing detection methods perform: True positive rate (TPR), False positive rate (FPR), precision, f-score, recall, and accuracy (ACC). According to the equations below, the metrics were calculated.

False negative rate (FNR). The number of phishing websites that are improperly categorised is shown below in a formula (1).

$$FNR = \frac{FNR}{FNR+TPR} \qquad (1)$$

False positive rate (FPR). In this formula below, FPR stands for the fraction of legal websites that are mistakenly labelled as phishing sites. FP called as phishing website's. Where TN as legitimate websites identified accurately as shown in formula (2).

$$FPR = \frac{FPR}{FPR+T} \qquad (2)$$

Recall. The ratio of correctly predicted rumour tweets (True Positives) to all other tweets (True Positives + False Negatives) as shown in formula (3).

$$Recall. = \frac{TPR}{TPR+FNR} \qquad (3)$$

Precision. This measures the percentage for correctly predicted rumor tweets (True Positives) to all previously identified rumour tweets (True Positives + False Positives) as shown in formula (4).

$$Precision = \frac{TPR}{TPR+FPR} \qquad (4)$$

F-score. This has precision and memory and is symmetrical. It achieved a compromise between evaluations of recall and precision as shown in formula (5).

$$f - score = 2 \times \frac{precision \times recall}{precision+rec} \qquad (5)$$

Accuracy (ACC). ACC refers to the proportion of websites with the proper classification, including those that are legitimate websites and those that are accurately identified as phishing websites as shown in formula (6).

$$Acc. = \frac{TPR+TNR}{TPR+TNR+FPR+FN} \qquad (6)$$

An overview of existing technologies for detecting phishing websites

Author	Year	Techniques used	Dataset used	Pros	Cons
Ojewumi, Theresa O., et al	2022	Random Forest	Phishtank	PhishNet is a simple way to lessen the danger of phishing.	Less data gathered. Use blacklist from PhishTank
Jalil, Sajjad, Muhammad Usman, and Alvis Fong	2022	Random Forest	Kaggle	Higher accuracy using TF-IDF technique.	• Phishing sites misclassified as valid. • The URL string may not identify the target webpage.
Das Guptta, Sumitra, et al.	2022	XG Boost	Phishtank	Anti-phishing technique based on hybrid features that includes extracts from client-side URL and hyperlink data.	Third-party reliant features will get more difficult.
Ramana, A. V., K. Lakshmana Rao, and Routhu Srinivasa Rao.	2021	Random Forest, Decision tree, and XGBoost	UCI ML Repository and Mendeley	Outperformed individual feature selection ensemble.	Further improve the performance of the model
Ghaleb Al-Mekhlafi, Z., et al	2022	Genetic Algorithm, random forest, adaboost, gradientboost, and lightgbm.	UCI ML Repository	Optimizing an ensemble prediction model for greater accuracy.	Limited dataset (4898 phishing websites and 6157 legitimate websites)
Sánchez-Paniagua, Manuel, et al.	2022	LightGBM classifier	Alexa and Phishtank	Phishing detection contains real scenarios and crafted with Phishing Index Login Websites Dataset.	Because there are so many samples in the dataset, individual false positives or negatives are feasible.
Tang, Lizhen, and Qusay H. Mahmoud	2022	RNN-GRU Algorithm	Phishtank and Kaggle	Minimize the number of false alarms and computation times by using whitelist filtering and blacklist interception.	Tiny URL are not supported. URLs of more than 200 characters will lose some of their features.

7. Issues and Challenges

Awareness about phishing. Every time academics offer a technique for identifying and restoring phishing websites, the fraudsters undermine the solutions. As a result, the researcher and the phisher are in a rigorous race. Attacks involving phishing are more successful since people are unaware of them. Therefore, encouraging people to defend themselves against phishing is one of the primary issues in terms of security.

Reduce false positives. A confusion matrix is provided by machine learning in categorization problems. Certain classifiers have a high false positive rate, which indicates that even if the websites are legitimate, the model labels them as fake. As a outcome, consumers are unable to access the official portal. If they can be reduced, consumers shouldn't have difficulty accessing trusted sites.

Selecting and utilising features. A website's numerous characteristics, to identify the website, its Urls, pages, contents characteristics, domains characteristics, source code, and others are utilised. It is challenging to choose the characteristics that may be used to develop a model capable of improving detection performance. The prediction results may not be reliable when only one characteristic is used for detection. Using a webpage with unique properties offers more details about the website that can assist in identification.

Time to response for real-time systems. Third-party resources based on an URL and rule parsing are used in rule-based frameworks. Because they only take each client request's unique URL as an input, they require a real-time prediction system to have a rather slow response time. Phishing attacks have moved across a variety of communication channels and target gadgets, including laptops and other smart devices. Supporting all devices with a single solution is a significant barrier for developers. Language and runtime platform isolation should be considered to lessen the difficulty of host construction and the expense of late maintenance.

Dark net monitoring for phishing websites. Identification of phishing websites on the dark net. Using the dark net, criminals obscure the details of fake sites. DNS records, HTTPS certificates, and WHOIS records are all hidden on dark net sites.

Detection of a tiny url. Rule-based algorithms for selecting features do not provide the correct web address, source location, or search criteria; they may be ineffective for short urls. It is difficult to convert small URLs created by many providers to their websites that are unique. Small urls also have fewer characters, which makes it challenging to retrieve feature properties from them using natural language processing. Small urls are prone to creating false or missed alerts if they are not carefully treated during data purification and pre-processing. Customer experience is crucial when it comes to internet goods, and consumers are hypersensitive to erroneous warnings produced by encryption software.

Best quality datasets. In order to discover novel principles and create machine learning algorithms, reliable phishing identification programmes need to continuously mix current information. There is always a conflict between phishing and anti-phishing. Attackers will modify the creation of phishing links in accordance with the techniques and guidelines for fighting phishing that have been made public. Similarly, anti-phishing models and algorithms need to be improved based on fresh phishing data. Furthermore, the volume and accuracy of the training dataset, as well as its quality, have a significant impact on how well machine learning-based solutions work. Small datasets that don't meet the requirements of deep learning algorithms make up the published datasets. The power rule states that deep learning efficiency grows with learning data volume. Consequently, obtaining both genuine and phishing URLs from websites.

8. Conclusion

This study covered an in-depth examination of several popular techniques for detecting website phishing. Even though there have been prior survey publications, they generally focus on general phishing detection techniques whereas we focused on identifying specific website features. The APWG survey results and the advent of phishing were the first topics we discussed. The discussion then turned to phishing attacks, phishing techniques, and phishing detection systems. Third, we go over the analysis of past evaluations of the literature for phishing attack detection. The diverse datasets, feature extraction, and feature selection are covered in the fourth paragraph. Number five is the analysis of phishing detection techniques and performance evaluation metrics. In the sixth paragraph, a comparison of current phishing website detection is given. The article concludes with ML and DL detection methods that may help prevent future, even more effective phishing websites from operating on legitimate websites by helping to identify them earlier. By resolving one or more of the issues raised in this study, we want to create a unique deep learning-based phishing website detection model in future work.

References

Chiew Kang Leng Kelvin Sheng Chek Yong and Choon Lin Tan 2018 A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications. 106: 1-20

Rao Routhu Srinivasa and Alwyn Roshan Pais 2019 Detection of phishing websites using an efficient feature based machine learning framework. Neural Computing and Applications. 31(8): 3851-3873

https://docs.apwg.org/reports/apwg_trends_report_q1_2022.pdf

https://docs.apwg.org/reports/apwg_trends_report_q2_2022.pdf

Kay R 2004 Sidebar: The Origins of Phishing. [Online]. Available: http://www.computerworld.com/s/article/89097/Sidebar_The_Origins_of_Phishing

Mohmoud khonji Youssef Iraqi and Andrew Jones 2013.Phishing detection: A Literature Survey. IEEE communications Systems and Tutorials.99:1-31

Jakobsson M and S Myers 2006 Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft. New Jersey. United States: john wiley and sons

https://www.itgovernance.eu/blog/en/the-5- most-common-types-of-phishing-attack

https://searchsecurity.techtarget.com/definition /whaling

https://cyware.com/news/smishing-and-vishingwhats-the-difference-between-them-4f55d408/

https://www.imperva.com/learn/applicationsecurity/phishing-attack-scam/

https://www.pcworld.com/article/135293/article.html

http://www2.deloitte.com/content/dam/Deloitte/sg/Documents/risk/searisk-cyber-101-part10.pdf

G J W Kathrine P M Praise A A Rose and E C Kalaivani 2019 Variants of phishing attacks and their detection techniques 3rd International Conference on Trends in Electronics and Informatics (ICOEI). 255-259

Rao R S Pais A R 2019 Detection of phishing websites using an efficient feature-based machine learning framework. Neural Comput & Applic. 31: 3851–3873

Rao R S Pais A R and Anand P 2020 A heuristic technique to detect phishing websites using TWSVM classifier. Neural Comput & Applic

Rao R S Pais A R 2019 Detection of phishing websites using an efficient feature based machine learning framework. Neural Comput & Applic. 31: 3851–3873

S Roopak A P Vijayaraghavan and T Thomas 2019 On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection. 1st International Conference on Advanced Technologies in Intelligent Control. Environment, Computing & Communication Engineering (ICATIECE):172-175

A Nakamura and F Dobashit 2019 Proactive Phishing Sites Detection. IEEE/WIC/ACM International Conference on Web Intelligence (WI). 443-448

F Tajaddodianfar J W Stokes and A Gururajan 2020 Texception: A Character/Word-Level Deep Learning Model for Phishing URL Detection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2857-2861

K Althobaiti G Rummani and K Vaniea 2019 A Review of Human and Computer Facing URL Phishing Features. IEEE European Symposium on Security and Privacy Workshops. 182-191

Carlo Marcelo Revoredo da Silva Eduardo Luzeiro Feitosa Vinicius Cardoso Garcia 2020 Heuristic based strategy for Phishing prediction: A survey of URL-based approach. Computers & Security, 101613

Athulya A A and K Praveen 2020 Towards the detection of phishing attacks. 4th international conference on trends in electronics and informatics (ICOEI)(48184). IEEE

Sahar Abdelnabi Katharina Krombholz and Mario Fritz 2020 VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity. Association for Computing Machinery. 1681–1698

S Haruta H Asahina and I Sasase 2017 Visual Similarity-Based Phishing Detection Scheme Using Image and CSS with Target Website Finder. IEEE Global Communications Conference. pp. 1-6

https://www.alexa.com/topsites

http://index.commoncrawl.org/

https://www.phishtank.com/developer_info.php

https://openphish.com/

https://archive.ics.uci.edu/ml/datasets/phishing+websites

https://majestic.com/reports/majestic-million

https://www.kaggle.com/datasets

Yang P Zhao G Zeng P 2019 Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. IEEE Access 7: 15196–15209

C Singh and Meenu 2020 Phishing Website Detection Based on Machine Learning: A Survey. Advanced Computing and Communication Systems. pp. 398-404

N N Gana and S M Abdulhamid 2019 Machine Learning Classification Algorithms for Phishing Detection: A Comparative Appraisal and Analysis. IEEE Nigeria Computer Chapter. pp. 1-8

M M Yadollahi F Shoeleh E Serkani A Madani and H Gharaee 2019 An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features. Web Research. pp. 281-286

Jain AK Gupta B B 2019 A machine learning based approach for phishing detection using hyperlinks information. J Ambient Intell Human Comput 10. 2015–2028

J Kumar A Santhanavijayan B Janet B Rajendran and B S Bindhumadhava 2020 Phishing Website Classification and Detection Using Machine Learning. Computer Communication and Informatics. pp. 1-6

M Korkmaz O K Sahingoz and B Diri 2020 Detection of Phishing Websites by Using Machine Learning Based URL Analysis. Computing, Communication and Networking Technologies. pp. 1-7

Doyen Sahoo Chenghao Liu and Steven C H Hoi 2017 Malicious URL detection using machine learning: A survey. arXiv preprint arXiv:1701.07179

M N Alam D Sarma F F Lima I Saha R E Ulfath and S Hossain 2020 Phishing Attacks Detection using Machine Learning Approach. Smart Systems and Inventive Technology. pp. 1173-1179

H Shirazi K Haefner and I Ray 2017 FreshPhish: A Framework for Auto-Detection of Phishing Websites. IEEE Information Reuse and Integration. pp. 137-143

Abdulhamit Subasi Emir Kremic 2020 Comparison of Adaboost with MultiBoosting for Phishing Website Detection. Procedia Computer Science 168. pp 272-278

Adebowale M A Lwin K T and Hossain M A 2020 Intelligent phishing detection scheme using deep learning algorithm. Journal of Enterprise Information Management. 9

Sahingoz O K Baykal SI Bulut D Phishing detection from urls by using neural networks

Somesha M Pais A R Rao R S et al 2020 Efficient deep learning techniques for the detection of phishing websites. Sadhana 45. 165

https://www.kdnuggets.com/2020/02/deepneural-networks.html

I Saha D Sarma R J Chakma M N Alam A Sultana and S Hossain 2020 Phishing Attacks Detection using Deep Learning Approach. Smart Systems and Inventive Technology. pp. 1180-1185

https://en.wikipedia.org/wiki/Feedforward_neural_network

Aljofey A Jiang Q Qu Q Huang M Niyigena JP 2020 An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. Electronics 9. 1514

S Y Yerima and M K Alzaylaee 2020 High Accuracy Phishing Detection Based on Convolutional Neural Networks Computer Applications & Information Security. pp. 1-6

Y Huang Q Yang J Qin and W Wen 2019 Phishing URL Detection via CNN and Attention Based Hierarchical RNN. IEEE Access. pp. 112-119

Le H Pham Q Sahoo D and Hoi SC 2018 URLNet: Learning a URL representation with deep learning for malicious URL detection. arXiv preprint arXiv:1802.03162.

M Arivukarasi and A Antonidoss 2020 Performance Analysis of Malicious URL Detection by using RNN and LSTM. Computing Methodologies and Communication. pp. 454-458

Y Su 2020 Research on Website Phishing Detection Based on LSTM RNN. IEEE Access. pp. 284-288

https://en.wikipedia.org/wiki/Recurrent_neural_network

https://en.wikipedia.org/wiki/Long_short_term_memory

https://en.wikipedia.org/wiki/Capsule_neural_network

Y Huang J Qin and W Wen 2019 Phishing URL Detection Via Capsule-Based Neural Network. Anti-counterfeiting, Security, and Identification. pp. 22-26

Feng J et al 2019 A Phishing Webpage Detection Method Based on Stacked Autoencoder and Correlation Coefficients. J. Comput. Inf. Technol. 27: 41-54

R Priya 2016 An Ideal Approach for Detection of Phishing Attacks using Naive Bayes Classifier. Computer Trends and Technology

Arun Kulkarni Leonard L 2019 Phishing Websites Detection using Machine Learning. Advanced Computer Science and Applications

Aron Blam Brad Wardman Thamar Solorio and Gary Warner 2010 Lexical feature based phishing URL detection using online learning. Security and Artificial Intelligence

Rami M Mohammad Fadi Thabtah and Lee McCluskey 2014 Phishing Websites Features

Tyagi I Shad J Sharma S Gaur S and Kaur G 2018 A Novel Machine Learning Approach to Detect Phishing Websites. Signal Processing and Integrated Networks

Singh P Maravi Y P S and Sharma S 2015 Phishing websites detection through supervised learning networks. Computing and Communications Technologies

Jain Ankit Kumar and B B Gupta 2022 A survey of phishing attack techniques defence mechanisms and open research challenges. Enterprise Information Systems. 16(4): 527-565

Ojewumi Theresa O et al 2022 Performance evaluation of machine learning tools for detection of phishing attacks on web pages. Scientific African. 16: e01165

Jalil Sajjad Muhammad Usman and Alvis Fong 2022 Highly accurate phishing URL detection based on machine learning. Journal of Ambient Intelligence and Humanized Computing: 1-19

Das Guptta Sumitra et al 2022 Modeling Hybrid Feature Based Phishing Websites Detection Using Machine Learning Techniques. Annals of Data Science: 1-26

Ramana A V K Lakshmana Rao and Routhu Srinivasa Rao 2021 Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Social Network Analysis and Mining. 11(1): 1-9

Ghaleb Al-Mekhlafi Z et al 2022 Phishing websites detection by using optimized stacking ensemble model. Computer Systems Science and Engineering. 41(1): 109-125

Singh Charu 2020 Phishing website detection based on machine learning: A survey. IEEE Advanced Computing and Communication Systems

Harinahalli Lokesh Gururaj and Goutham BoreGowda 2021 Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology. 5(1): 1-14

Sánchez Paniagua Manuel et al 2022 Phishing websites detection using a novel multipurpose dataset and web technologies features. Expert Systems with Applications. 207: 118010

Khan Sohail Ahmed Wasiq Khan and Abir Hussain 2020 Phishing attacks and websites classification using machine learning and multiple datasets (a comparative analysis). Springer, Cham

Sameen Maria Kyunghyun Han and Seong Oun Hwang 2020 PhishHaven—an efficient real-time ai phishing URLs detection system. IEEE Access. 8: 83425-83443

Faris Humam and Setiadi Yazid 2021 Phishing Web Page Detection Methods: URL and HTML Features Detection. IEEE Access

Korkmaz Mehmet Ozgur Koray Sahingoz and Banu Diri 2020 Feature selections for the classification of webpages to detect phishing attacks: a survey. IEEE Access

Tang Lizhen and Qusay H Mahmoud 2021 A survey of machine learning-based solutions for phishing website detection. Machine Learning and Knowledge Extraction. 3(3): 672-694

Somesha M et al 2020 Efficient deep learning techniques for the detection of phishing websites. Sādhanā. 45(1): 1-18

Joshi A Pattanshetti P and Tanuja R 2019 Phishing attack detection using feature selection techniques. Communication and information processing

Ubing A A Jasmi S K B Abdullah A Jhanjhi N and Supramaniam M 2019 Phishing website detection: An improved accuracy through feature selection and ensemble learning. Advanced Computer Science and Applications. 10(1), 252–257

Subasi A Molah E Almkallawi F and Chaudhery T J 2017 Intelligent phishing website detection using random forest classifier. IEEE electrical and computing technologies and applications. pp. 1–5

Mao J Bian J Tian W Zhu S Wei T Li A et al 2018 Detecting phishing websites via aggregation analysis of page layouts. Procedia Computer Science. 129: 224–230

Jain A K and Gupta B B 2018 Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems. 68(4): 687–700

Shie E W S 2020 Critical analysis of current research aimed at improving detection of phishing attacks. Selected computing research papers. pp. 45