WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

MLPhish: A Machine Learning Framework for Phishing URL Detection

S.SriRanjani[a*],K.DineshKumar[b] ,S.Yuvanraj[c] , K.Shamraj[d] M.Nithish[e]

[a]Student, ComputerScience & Engineering, Paavai College of Engineering, Namakkal
[bcdec]Student, Artificial Intelligence & Data Science, Paavai College of Engineering, Namakkal

ABSTRACT :

One of the most worrying things in a world that is changing all the time is phishing. Cybercrime, a new method of data theft, has emerged as a result of increased Internet usage. Cybercrime is the term for using computers to steal personal data and violate privacy. Phishing is the main method that is employed. One of the most popular forms of phishing involves the use of URLs (Uniform Resource Locators), and its main objective is to steal user data when the victim visits a malicious website. Finding a malicious URL is a difficult task. Through the use of machine learning algorithms that concentrate on the characteristics and behaviors of the recommended URL, this work seeks to give a method for identifying such websites. To identify harmful websites, the online security community has developed blacklisting services. These blacklists are produced using a range of techniques, including heuristics for site inspection and manual reporting. Many harmful websites unintentionally avoid blacklisting because of their recentness, lack of evaluation, or inaccurate evaluation. Algorithms like Support Vector Machine (SVM), Random Forests, Decision Trees, Light GBM, Logistic Regression, and Logistic Regression are used to build a machine learning model that determines whether a URL is malicious or not. The first stage is to extract features; the second is to apply the model.

Keywords: Phishing detection, Phishing dataset, Web technologies, Machine learning, Login

1. Introduction:

One of the most common forms of social engineering assaults is phishing, in which perpetrators fabricate a website in order to trick visitors and steal their passwords or other private information pertaining to a particular website or online service (Mohammad et al., 2015b). Phishing can happen via a variety of attack methods, such as instant chatting, emails, Short Message Service (SMS), and many more. Furthermore, one of the most significant vectors is the web page (Chiew, Yong et al., 2018, Gupta et al., 2018). In this scenario, attackers mimic a well-known company's website in order to gain user information, typically through a sign-up form or login. We are able to identify websites as the ultimate destination of phishing attempts because a Uniform Resource Locator (URL) that points to a website is present in many of these attack routes. In the most recent quarter, the Anti-Phishing Working Group (APWG) discovered up to 611,877 distinct phishing websites. According to the Anti-Phishing Working Group (2021) financial institutions accounted for the majority of those attacks (24.9%), with social media (23.6%), SAAS (Software As A Service) and webmail services (19.6%), and payment platforms (8.5%) following closely behind. Phishing efforts have a visible effect since the disclosed private data results in financial losses that harm both consumers and corporations, with millions of dollars at stake (Jain and Gupta, 2017, Shaikh et al., 2016). Corporations have an economic loss of over a million dollars (Bose & Leung, 2014).

In the most recent quarter, the Anti-Phishing Working Group (APWG) discovered up to 611,877 distinct phishing websites. According to the Anti-Phishing Working Group (2021) financial institutions accounted for the majority of those attacks (24.9%), with social media (23.6%), SAAS (Software As A Service) and webmail services (19.6%), and payment platforms (8.5%) following closely behind. Phishing efforts have a visible effect since the disclosed private data results in financial losses that harm both consumers and corporations, with millions of dollars at stake (Jain and Gupta, 2017, Shaikh et al., 2016). Corporations have an economic loss of over a million dollars (Bose & Leung, 2014).

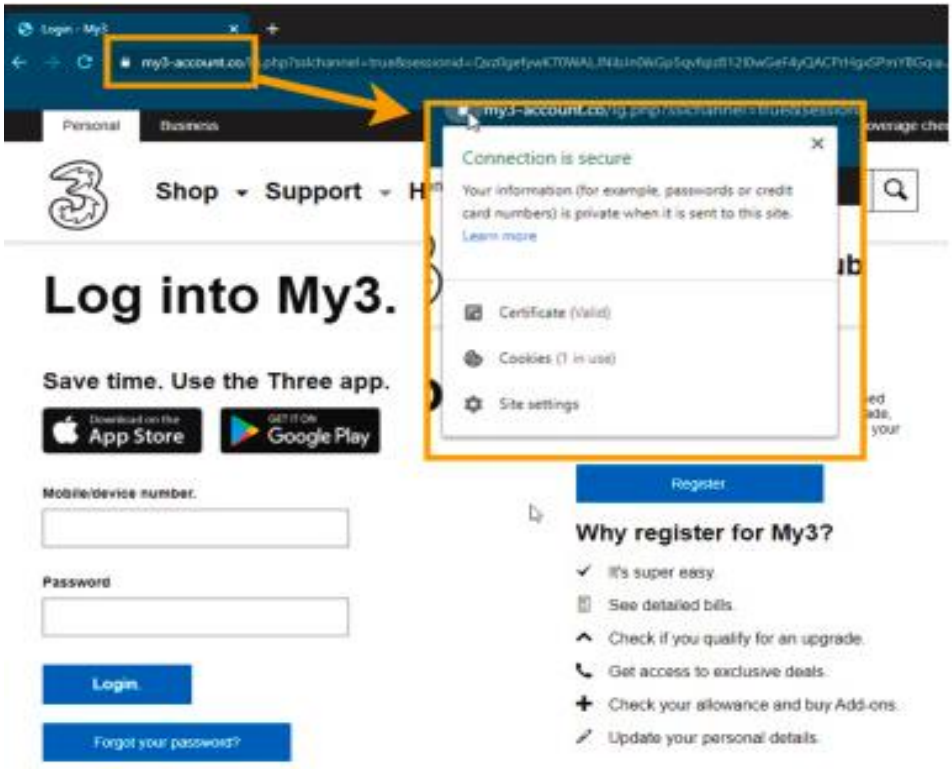

Figure1 : Phishing Website not yet Registreed

Previous studies (Adebowale et al., 2019, Sadique et al., 2020, Xiang et al., 2011) have also included features like the website ranking on Google Browsing Index or the domain registry age by utilizing third-party resources like Google or WHOIS. There are, however, at least four issues with these implementations. First, the real-time application is limited by their dependency on resources that may be unavailable or have a delayed reaction time. Second, some characteristics are flawed for a subset of websites. For example, the domain age function could inaccurately identify websites that are valid but have a limited lifespan. The lack of an updated, representative dataset for researchers to compare their findings with is one of the most pertinent concerns regarding the state-of-the-art methods for phishing detection. In order to compare phishing detection systems, it is necessary to use the same dataset and other approaches; otherwise, comparisons are subjective and may yield false results. A work that uses URL features typically presents a URL dataset; similarly, the majority of studies offered customized datasets to fit their methodology.

One of the most pertinent issues with the most recent methods for phishing detection is the lack of a representative, up-to-date standard dataset that would allow researchers to compare their findings. In order to compare phishing detection systems, it is necessary to use the same dataset and other approaches; otherwise, comparisons are subjective and may yield false results. A work that uses URL features typically presents a URL dataset; similarly, the majority of studies offered customized datasets to fit their methodology. The complexity of current works makes the first of these premises a laborious endeavor. The second one suggests employing the same strategy (URL, HTML, graphics, or particular data), as a dataset created just with URLs cannot be used in a work utilizing HTML code. We introduce and make publicly available4 PILWD-134K (Phishing Index Login Websites Dataset) in this work to address the shortcomings of published datasets and the dearth of offline datasets with raw data. 134,000 verified samples were gathered between August and September to make up this dataset. URLs, HTML code, screenshots, a copy of the website files, web technologies analysis, and additional metadata pertaining to the phishing reports are among the six categories of raw resources covered by PILWD-134K.5. This new dataset, which was not available prior to its compilation, can be used as a standard corpus for analyzing the results utilizing various phishing detection approaches across the same domains as well as a sizable sample of real-life websites.

The format of this document is as follows. A overview of relevant literature and works is presented in Section 2. The dataset, together with its features, structure, and quality filters, are introduced in Section 3. The technique, suggested features, and measurements are described in Section 4. The experiments and the outcomes are presented in Section 5. In Section 6, the conclusion, limitations, and future directions are discussed.

2. Related Works

Zhang et al. (2007) developed CANTINA, a phishing detection system that pulls five signature terms from a website and feeds them into the Google Search engine using the TF-IDF (Term Frequency - Inverse Document Frequency) technique. A website was deemed legitimate if its domain name appeared in the top 30 results after analysis. Later on, Xiang et al. (2011) introduced CANTINA, an improved version that included two filters and features (PageRank, copyright, and WHOIS) pertaining to URLs, HTML, and the web. Using a Bayesian network on an 8000 sample dataset, the suggested approach obtained 92%. Two sets of features were implemented by Moghimi and Varjani (2016). The first set comprised nine heritage features from earlier phishing detection works, and they had to do with the URL and detecting a set of keywords within the route, the URL query, and the domain name. Typosquatting features between the URL and the source of the elements loaded in the website (CSS, JavaScript, pictures, and links), as well as the

protocol used to load those resources, were added in the second set of eight proposed improvements. They combined the two feature sets and using a dataset of 1707 e-banking phishing attempts, they obtained a 98.65% accuracy rate.

In 2017, Rao and Pais introduced FeedPhish, a program designed to identify phishing attempts based on the way a phony credential submission is handled. The authors developed three modules using Selenium WebDriver to mimic user interaction on the login form. The HeuristicsCheck module examines the content of the response after the FeedFakeCredentials module introduces a random account and credential and the LoginCheck module first confirms whether the website has a login form. In the end, TargetDomainCheck uses the anchor links to determine the domain's identification. The suggested heuristics achieved 96.38% accuracy on a dataset of 2342 samples, according to the results. Rao and Pais (Rao & Pais, 2019a) presented 16 heritage and innovative characteristics in later investigations, which were categorized into three sets: hyperlink-based, third-party-based, and URL obfuscation.

Li et al. (2019) combined the GBDT, XGBoost, and lightGBM models with 20 features taken from the URL and HTML code to propose a stacking model. Furthermore, they suggested new capabilities including the ability to detect brands, maintain consistency in the URL domain and title, and extract a string embedding from HTML using Word2Vec. Using a dataset of 50,000 samples, the initial implementation, which had 20 features, achieved 97.11% accuracy. Additionally, they used a tiny CNN (Convolutional Neural Network) to investigate visual aspects using website screenshots, increasing the accuracy of their model to 98.60%.

Three sets of features were used by Adebowale et al. (2019) to propose an ANFIS (Adaptative Neuro-Fuzzy Inference System): (i) properties from website images; (ii) frame features, which identify website behavior such as pop-up windows, redirections, and disabled right-click among other things; (iii) and a text subset primarily related to the URL and third-party services like Google Page Rank and WHOIS. Together with a self-collected image dataset, they achieved a 98.30% accuracy on the frame and text sets of the Mohammad et al. (2015a) dataset. Three different kinds of features were presented with an Extreme Learning Machine (ELM) model by Yang et al. (2021): (i) Surface features make use of the URL's data. They specifically made use of handmade URLs, Domain Name System (DNS) features pertaining to the target domain's registration date and DNS records, and (ii) topological features pertaining to the website's structure. Eventually, the text, image, and general similarity were used to extract deep features (iii). On a dataset of 60,000 samples, they achieved 97.5% accuracy by combining these features with the ELM classifier.

A methodology for real-time phishing detection employing four sets of URL features was provided by Sadique et al. (2020): (i) Lexical features, which count the characters, dots, and symbols that appear in various sections of the URL; (ii) Host-based features, which pertain to the server and IP address where the website is housed; (iii) WHOIS features, which count the days that pass between the registration date and the expiration date; and (iv) GeoIP-based features, such as the Autonomous System Number (ASN), the nation, or the city in which the website is housed. Using 98,000 samples from Phishtank—where valid samples are also selected from false positives gathered at PhishTank—a total of 142 distinct attributes were assessed. Using the suggested method, they achieved a 90.51% accuracy on an RF classifier.

Dataset Description

This section explains how the Phishing Index Login Websites Dataset (PILWD) was created using more than 134,000 confirmed samples that were gathered between August and September. It contains unprocessed data from both trustworthy and fraudulent websites, enabling researchers to identify patterns and create phishing detection tools. It includes many different types of raw data, such as URLs, source code, screenshots, analyses of various technologies, and an offline version of the website. Every sample also has additional metadata, such as filters, the time of remembrance, and, in the event of a phishing class, PhishTank information. A web crawler was created using Python 3 and Selenium WebDriver to visit both the reported phishing URLs and the genuine domains in order to gather the samples.

Modern papers typically gather reputable webpages from a variety of sources. They select the most popular domains from numerous sources, including Alexa Topsites, DMOZ, and more. The Majestic Million9 and Quantcast Top Sites8 served as our crawler's sources. The domains with the greatest referring subnets are supplied by these providers. Since there aren't enough websites pointing to phishing websites, we may thus assume that it is difficult for them to get listed in these lists. We created a list of 150,000 domains sorted by visitor count from the first source, and we appended Majestic's top million domains to it. Once the two lists were combined, we eliminated the duplicate domains.Present-day phishing detection systems (Aljofey et al., 2020, Li et al., 2019, Rao et al., 2020) input their algorithms with phishing samples (Fig. 2(c)) and authentic homepages (Fig. 2(a)) without further probing the domain. However, most homepages lack login forms, and phishing assaults concentrate on stealing credentials. Therefore, we chose to gather the real-case scenario (Sanchez-Paniagua et al., 2022) where the user is unsure if a login page (Fig. 2(b)) is authentic or phishing (Fig. 2(c)), rather than only the official homepage. The similarities between the authentic login page and the phishing one are illustrated in Fig. 2, particularly with regard to the URL and the output HTML structure. Regarding the legal.

By leveraging user votes to validate reported URLs, Phishtank increases the likelihood that the samples are authentic phishing attempts. We get hourly JSON reports while the samples were being gathered. In addition, we asked the API to confirm the samples that were gathered and to extract other data from the JSON report, such as the IP address, the impacted brand, the server location, and the network that made the announcement. We did not interact with the website in order to locate the login form or remove the cookies banner, in contrast to the lawful dataset gathering. We simply gathered the reported land page as a result. Since the great majority of samples on Phishtank were offline websites or forums for downloading malware, they were excluded and not collected. Additionally, samples that were flagged as spam or prohibited from voting were not included in the collection.

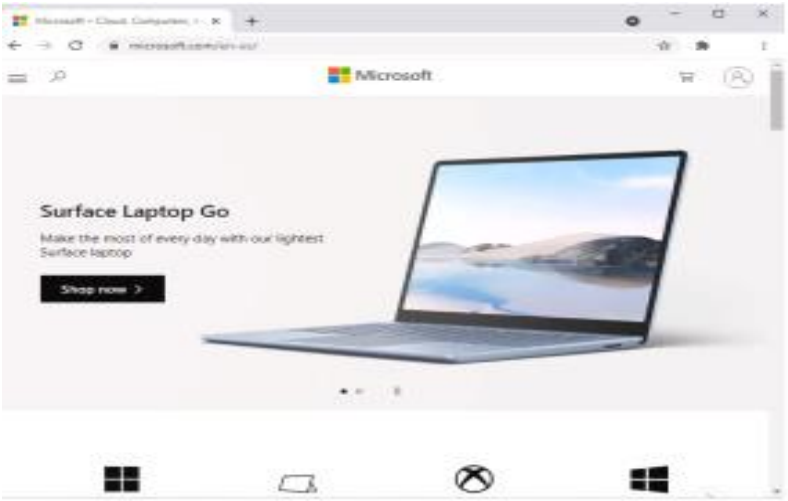

Figure 2: Legitimate Homepage

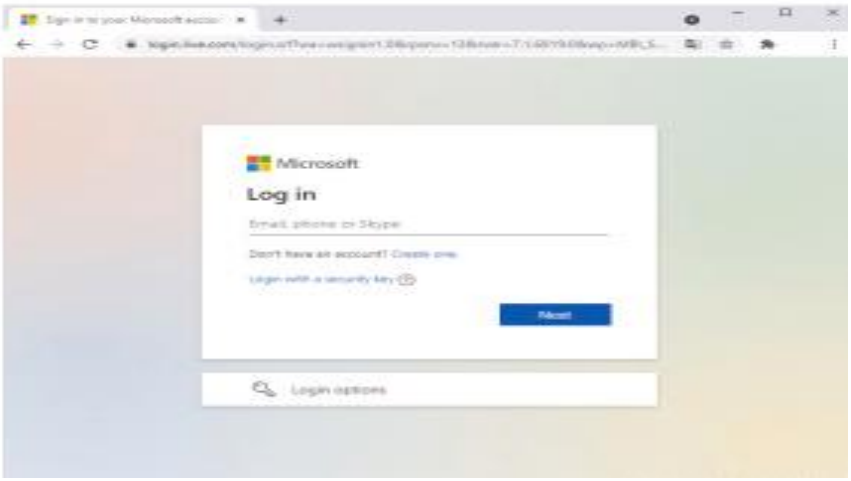

Figure 3: Legitimate Login Page

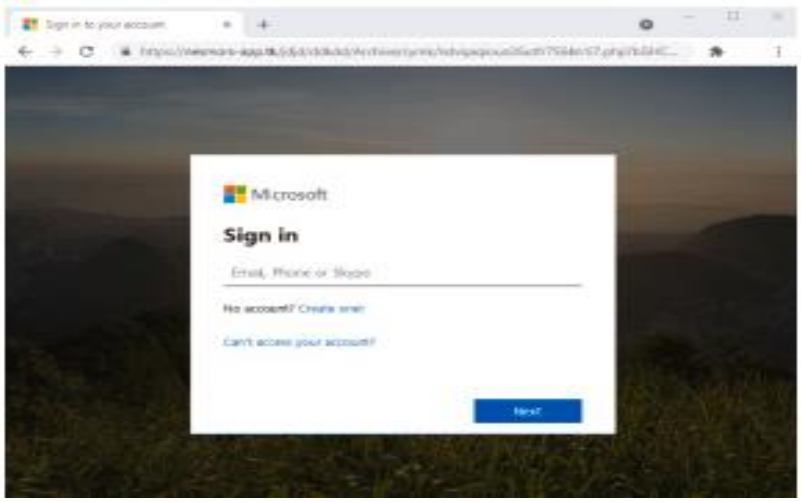

Figure 4 :Phishing Login Page

Table 1 :Phishing Webpages

Domain name	Samples	Banned
Total samples	186,000	–
*.000webhostapp.com	5702	No
*.weebly.com	4044	No
docs.google.com	2849	No
*.godaddysites.com	1055	No
*.appspot.com	903	No
*.umbler.net	799	No
storage.googleapis.com	720	No
firebasestorage.googleapis.com	642	No
2m.ma	1680	Yes
www.google.com	1312	Yes
www.imdb.com	698	Yes
www.paypal.com	626	Yes
drive.google.com	386	Yes
www.amazon.co.jp	705	Yes
login.microsoftonline.com	356	Yes
www.icloud.com	325	Yes

3. Conclusion

In addressing the issue of phishing detection, this study makes the following contributions to the field: First, we describe a new phishing detection paradigm using PILWD, a high-quality offline dataset that primarily consists of authentic samples from login sites. Second, a new set of features built on website technologies increased the detection system's robustness and improved the accuracy of detection. Third, in addition to the other cutting-edge features, we suggested adding more unique ones that were centered on the URL and HTML. Ultimately, the right mix of the assessed variables for characterizing webpages and the chosen LightGBM classifier creates a novel approach that enables phishing detection with a competitive 97.94% accuracy in extremely realistic circumstances.

In order to help researchers assess their recommendations in this sector, PILWD, a vast and updated dataset of phishing websites, is offered. It contains up to 324,000 total samples and 134,000 validated samples. It is a more thorough benchmark for evaluating different strategies, giving researchers a simple way to assess their work and conclusions against the same dataset. The gathered samples' URL, HTML, screenshots, technology analysis, files, and other metadata are all included in the information. The availability of this data will also spare academics from having to spend time gathering data.

Our discoveries led us to develop a new paradigm for phishing identification since we believe that a large number of authentic login pages should be present in the datasets used to test any proposed solutions. Since the main goal of phishing websites is to obtain user information through sign-up or login forms, we came to the conclusion that they must be included in the legitimate class of any phishing dataset. As a result, 40.80 login forms belong to the phishing class and 61.93% belong to the genuine class in our dataset. This method is easily adaptable to real-world applications, such as alerting users before requiring their login credentials on a website.

There are drawbacks to our method. Isolated false positives or negatives may occur because of the dataset's large sample size. We therefore put into practice the filters discussed in this paper, which are able to detect the majority of compromised cases. Furthermore, we are unable to ensure that every phishing sample has been correctly verified because the verification procedure is totally reliant on PhishTank users and services. Another drawback would be that we suggested a number of novel hybrid characteristics for website identification that make use of copyright, title, and domain name. Even in cases where no brand list was used, this strategy can nevertheless have an impact on websites that are unsupported by firms or trademarks.

REFERENCES :

1. Johny JHB, Nordin WAFB, Lahapi NMB, Leau YB. SQL Injection prevention in web application: a review. In: Communications in computer and information science, vol. 1487 CCIS, no. January. 2021. p. 568–585. https://doi.org/10.1007/978-981-16-8059-5_35.
2. Alghawazi M, Alghazzawi D, Alarifi S. Detection of sql injection attack using machine learning techniques: a systematic literature review. J Cybersecur Privacy. 2022;2(4):764–77.
3. Han S, Xie M, Chen HH, Ling Y. Intrusion detection in cyber-physical systems: techniques and challenges. IEEE Syst J. 2014;8(4):1052–62.

4. Dasmohapatra S, Priyadarshini SBB. A comprehensive study on SQL injection attacks, their mode, detection and prevention. 2021. p. 617–632. https://doi.org/10.1007/978-981-16-3346-1_50.
5. Hu J, Zhao W, Cui Y. A survey on SQL injection attacks, detection, and prevention. In: ACM international conference on proceeding series, no June. 2020. p. 483–488. https://doi.org/10.1145/3383972.3384028.
6. Blog. What is SQL injection attack? Definition & FAQs|Avi networks.
7. Imperva. SQL (structured query language) injection. Imperva. 2021.
8. Deepa G, Thilagam PS, Khan FA, Praseed A, Pais AR, Palsetia N. Black-box detection of XQuery injection and parameter tampering vulnerabilities in web applications. Int J Inf Secur. 2018;17(1):105–20. https://doi.org/10.1007/s10207-016-0359-4.
9. Dizdar A. SQL injection attack: real life attacks and code examples. 2022.
10. Pan Y, et al. Detecting web attacks with end-to-end deep learning. J Internet Serv Appl. 2019. https://doi.org/10.1186/s13174-019-0115-x.
11. Anti-phishing Working Group (APWG) Phishing Activity Trends Report 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf
12. FBI Internet Crime Report 2020, https://www.ic3.gov/Media/PDF/AnnualReport/2020 IC3Report.pdf
13. Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/reports/2020-databreachinvestigations- report.pdf
14. World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security
15. Ye Cao, Weili Han, and Yueran Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management-DIM 08, pp. 51-60, 2008
16. M. Sharifi, and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008
17. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no.13, pp. 5948-5959, 2014
18. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide WebWWW 05, pp. 1060-1061, 2005
19. C. L. Tan, K. L. Chiew et al., "Phishing website detection using url assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), IEEE, pp. 054-059, 2014
20. K. L. Chiew, E. H. Chang, W. K. Tiong et al., "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16-26, 2015
21. K. M. kumar, K. Alekhya, "Detecting phishing websites using fuzzy logic," International Journal of Advanced Research in Computer Engineering Technology(IJARCET), vol. 5, no. 10, 2016 International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NCREIS - 2021 Conference Proceedings Volume 9, Issue 13 Special Issue - 2021 160
22. Rishikesh Mahajan, and Irfan Siddavatam, "Phishing website detection using machine learning algorithms," International Journal of Computer Applications(0975-8887), vol. 181, no. 23, 2018
23. Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, and Bindhumadhava BS, "Phishing website classification and detection using machine learning," International Conference on Computer Communication and Informatics(ICCCI), 2020
24. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, "Detection of phishing websites by using machine learning-based URL analysis," 11nth International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2020
25. Mohammad Nazmul Alam, Dhiman Sarma et al., "Phishing attacks detection using machine learning approach," 3 rd International Conference on Smart Systems and Inventive Technology(ICSSIT), 2020
26. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, "Intelligent phishing website detection using Random Forest classifier," International Conference on Electrical and Computing Technologies and Applications(ICECTA), 2017
27. Structure of a URL – image, https://towardsdatascience.com/phishingdomain-detection-with-ml5be9c99293e5 [18] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, "Phishing websites features"