A Review on Phishing Website Detection

Mhasin Adil, Salma Mahmoud and Ashraf Alzubier

December 21, 2019

A Review on Phishing Website Detection

Mhasin Adil
Master student
international university of Africa
Khartoum Sudan
+249922279296, SDN. 249
Mhasinadil2225@gmail.com

Salma mahmoud
Master student
international university of Africa
Khartoum Sudan
+249916263969, SDN. 249
SalmaMahmoud143@yahoo.com

Ashraf Alzubier
The head of IT department
international university of Africa
Khartoum Sudan
+249126888870, SDN. 249
ashrafalzubier@gmail.com

ABSTRACT

Internet has become a useful part of our regular day to day life. The Web browsing is the main internet service. Many people use browser to perform various activities like online shopping, online bill payment, online mobile recharge, banking transaction .Due to wide use of this service's customer face various security threats like cybercrime. Phishing is a form of web threat, Web Phishing lures the user to interact with the fake website. The main objective of this attack is to steal the sensitive information from the user. This review paper contend several Type of phishing website, some of phishing Detection methods and Technology used in this Detection methods, also describe Data set Types and used to develop anti-phishing model and Evaluation it.

Keywords

Detection methods; Phishing Target; Phishing Detection; Feature Extraction; Phishing Website; Phishing Attacks.

1. INTRODUCTION

Phishing is a criminal mechanism employing both social engineering and technical tricks to steal consumers' personal identity data and financial account such as credentials, username, password and account numbers [1]. Typically phishing attack exploits the social engineering to lure the victim through sending a spoofed link by redirecting the victim to a fake web page. The fake webpage is created similar to the legitimate webpage[2]. Thus, rather than directing the victim request to the real web server, it will be directed to the attacker server. Phishing costs Internet users billions of dollars per year and the current solutions of antivirus, firewall and designated software do not fully prevent the web spoofing attack. The implementation of Secure Socket Layer (SSL) and digital certificate also does not protect the web user against such attack

2. Type of phishing:

The author's in [3-5] explain various types of phishing attack as following:

1. **Website phishing attack:** Website phishing typically begins by creating website that imitates legitimate website due to the internet users believes on appearance of the website for its identification.
2. **Email phishing attack:** Attacker sends email to users about the need to verify account information, system failure requiring users to re-enter their information, fictitious account charges, undesirable account changes, new free services requiring quick action and many other scams are sent to large number of people.
3. **Web Trojans:** This attacks pop ups invisibly when users are attempt to login in for trusted website. Attacker collects the user's credentials locally and transmits them to the phishers.
4. **Content-injection phishing:** attack, attacker replaces part of content of a legitimate site with fake content designed to misdirect to user into giving up their personal information to attacker.

These types of phishing attack done in deferent form[6, 7]:

- **Creating fake URL:** Attackers usually try to make (URL) of phishing sites look similar to legitimate sites to misguide internet users.
- **Misspelled URLs:** attackers make more spelling mistake if user are not careful, they will think that they are on "apple" site.
- **Creating anchor text:** it is Similar to the URL feature, but here the links within the webpage may point to a domain different from the domain which is typed in the URL address bar.
- **Fake SSL lock:** Now a day it is cheap and easy enough for the attacker to obtain SSL certificates for their malicious sites, therefore users lose one of the methods for identifying trusted sites from phishing targets.
- **URL manipulating using java script:** The attacker will insert a string to be used in the webpage and treated by the user's browser as code and when the browser loads the page, the malicious script executes without the user even knowing that such an attack has taken place.

3. The phishing Detection methods:

From a technical point of view, the following categories explain various anti phishing solutions[8, 9].

1. **Blacklist method:** This is most commonly used approach in which list of phishing URL is stored in database and then if URL is found in database. It's easy and faster to implement. The update of list is necessary to counter new attack.
2. **Heuristic based method:** it is extension of blacklist and able to detect new attack as use features extracted from phishing site to detect phishing attack. But limitation is cannot detect all new attack and easies to bypass once attacker know algorithm or features used.
3. **Visual similarity:** This approach deceive user by extracting image of legitimate site. But limitation of

this is image comparison takes more time as well as more space to store image.

4. **URL based method:** it is Lexical feature Lexical features are the textual properties of the URL itself, not the content of the page it points to.

a. The features of the URL

A. The authors in [10, 11] explain some features related to URL based method.

i. **Using the IP Address:** If an IP address is used as an alternative of the domain name in the URL Rule: IF The Domain Part has an IP

Address → Phishing

Otherwise→ Legitimate

ii. **Long URL to Hide the Suspicious Part:** Phishers can use long URL to hide the doubtful part in the address bar. Rule: IF URL length is ≤ 75 → legitimate

Otherwise → Phishing.

iii. **Adding Prefix or Suffix Separated by (-) to the Domain:** The dash symbol is rarely used in legitimate URLs

Rule: IF Domain Name Part Includes (-) Symbol → Phishing

Otherwise → Legitimate

iv. **Submitting Information to Email:** To that end, a server-side script language might be used such as "mail()" function in PHP

Rule: IF Using ""mail()\" or \"mailto:\" Function to Submit User Information" → Phishing

Otherwise → Legitimate

v. **Using Pop-up Window:** personal information was asked to be filled in through these pop-up windows.

Rule: IF Popup Window Contains Text Fields→ Phishing

Otherwise → Legitimate

The authors in [11] used some additional Features for URL based method

i. **Number of Dots & Slashes:** If {No. of Dots >= 5 → feature = Phishy

Otherwise → feature = Legitimate

if {No. of Slashes >= 5 → feature = Phishy

Otherwise → feature = Legitimate

ii. **Having @ symbol:** If {URL having @ symbol → feature = True

Otherwise → feature = False

iii. **Special Character:** IF a URL contains any of this characters such as dash (-), underscore (_), comma (,), and semicolon (;) will be phishy.

iv. **HTTP & SSL check:** For the security impression of trustworthy and authorized websites use SSL certification secured encryption transaction (https ://). Generally legitimate websites transfer their confidential information using https: // protocol over internet. Since it is found that even phishy websites use the https://, we further need to check for the trusted issuer and the SSL certificate age. If {Use http is trusted age >= 2 years → feature =low

Using http and issuer is not trusted → feature = Moderate

Otherwise → feature = High

v. **Request URL:** Generally the page content such as video, audio, images etc. are loaded from within the Domain as in address bar. We have to check for the presence of domain in the URL in <Src =>.

If {Request URL % < 20% → feature = Legitimate

Request URL %> = 20% &< 50% → feature = Suspicious

Otherwise → feature = phishy

vi. **Google Page Rank:** If {Googlepage rank> 5 → feature = Legitimate Googlepage rank> = 3 &<5 → feature = Suspicious

Googlepage rank< 3 → feature = phishy

vii. **Age of Domain:** If {Age of Domain >= 2 years → feature = Legitimate

Age of Domain> = 1 &< 2 → feature = Suspicious

Otherwise → feature = Phishy

B. The authors in [12] add **Content based approach**:

This approach detects fake website by inspecting the similarity between legitimate and fake website. The similarity between two website is calculated based on the similarity of web page content

C. The authors in [13]add new feature for URL based method

i. **Host based analysis:** Host-based features explain "where" phishing sites are hosted, "who" they are managed by, and "how" they are administered. Phishing Web sites may be hosted in less reputable hosting centers, on machines that are not usual Web hosts, or through not so reputable registrars.

ii. **WHOIS properties:** gives details about the date of registration, update and expiry, who is the registrar and the registrant.

iii. **Geographic properties:** give details about the continent/country/city to which the IP address belongs.

iv. **Blacklist membership:** are precompiled lists or databases that contain IP addresses, domain names or URLs of malicious sites the web users should avoid. Blacklists: is a low overhead operation and inexpensive. The Limitations is can't detect new phishing website and you must update the list.

- DNS-Based Blacklists: Users submit a query representing the IP address or the domain name in question to the blacklist provider's special DNS server.
- Browser Toolbars**:** Browser toolbars provide a client side defense for users. Before a user visits a site, the toolbar intercepts the URL from the address bar and cross references a URL blacklist.

- Network Appliances: Dedicated network hardware is another popular option for deploying blacklists. These appliances serve as proxies between user machines within an enterprise network and the rest of the Internet.

v. **Page/Popularity Based Property:** Various popularity features are as follows:

- PageRank: It is one of the methods Google uses to determine a page's relevance or importance. The maximum PR of all pages on the web changes every month when Google does its re-indexing.
- Traffic Rank details: ranks various websites according to the Internet traffic based.

D. The authors in [14] explain two additional features :

i. Feature Set Requirements:

a. Generalizability: Accumulating ground truth phishing and legitimate data is challenging. Phishing websites have very short lifetimes [10] and can display different content depending on a browser's user-agent or user's geographic location.

b. Adaptability: Several automated classification techniques rely on a static set of features learned from a training set such as the bag-of-words model or "term frequency-inverse document frequency" (TF-IDF) computation.

c. Usability: It is desirable that features are computable on an end user system without relying on online access to centralized servers or proprietary data (e.g. Google PageRank).

d. Computational Efficiency: Features must be quickly computable to allow integration with real time detection systems that do not impact users' web surfing experience.

ii. Computing Features:

a. URL: First we define nine statistical features related to the lexical composition of URLs.

b. Term usage consistency: The second set of features (f2) captures the consistency of term usage between different types (controlled vs. uncontrolled; constrained vs. unconstrained) of data sources in the page.

c. Usage of starting and landing mld: Legitimate websites are likely to register a domain name reflecting the brand or the service they represent. However, phishers often use domain names having no relation with their target.

d. RDN usage: We define 13 features (f4) related to RDN usage consistency.

e. Webpage content: Finally, five features (f5) count the number of terms in the text and the title (2), and the number of input fields, images and IFrames (3) in the page.

E. The authors in [15] also add new Feature

i. **Original Feature:** There are some features in the phishing URL, such as special characters. We definite these features in URL as an original feature as follows:

a. There are special characters in URL, such as @, Unicode, and so on. Those special characters are not allowed in a normal URL.

b. There are too many dots or less than four dots in normal URL.

c. The age of the domain is too short. For example, the age of the normal domain is more than 3 months.

ii. **Interaction Feature:** There are some features in graph $G=\ (V,)$,such as access frequency

a. In-degree of URL node from REF is very small. In general, the normal websites do not link to phishing sites. The phishing sites are directly accessed.

b. Out-degree of URL node is very small. In order to get personal private information, the phishing sites are usually terminal websites and do not link to the other sites.

c. The frequency of URL from AD is one. In general, one user accesses the phishing site only one time and the user cannot access the phishing site more than one time.

d. When AD accesses URL, user browser type UA is not the main browser. Well-known browser vendors often have a built-in filtering phishing site plug-in. A user who uses unknown browsers is more likely to access the phishing sites.

e. There is no cookie in user. The phishing site does not leave its cookie in user.

4. Technology used in phishing Detection methods:

The authors in [13, 16-20] explains various technologies for website phishing detection:

1. Based on genetic algorithm

i. Phase 1 Feature Extraction: In this phase 10 features are extracted in their paper.

ii. Phase 2 Pre-processing: During this phase the value of each feature is classified into Phishing, legitimate or suspicious class.

iii. Phase 3 Weight adjustment: The aim of the weight adjustment is to find the best weights that can classify the website accurately and genetic algorithm is used for weight adjustment.

iv. Phase 4 Results: In this phase best weights derived in third phase are used to calculate the fitness of the URLs in data set and classified into legitimate or phishy by comparing the fitness with the threshold value.

2. Based on associative classification data mining

i. When the end user clicks on a link within an email or browses the internet

ii. User will be directed to a website that could be original or fake. Therefore this website is basically the test data.

iii. A script which is written in PHP is embedded within the browser and starts processing to extract the features of the test data and saves them in a data structure.

iv. Then the intelligent model will be active within the browser to predict the type of the website based on rules learnt from previous websites and the rules of the classifier are utilized to predict the type of the test data based on features similarity.

v. When the browsed website is identified as original, no action will be taken. But, when the website turned to be fake, then user will be warned by the intelligent method that he is under risk.

3. Intelligent Phishing website detection and categorization model

i. **Feature extractor:** used to extract the terms from the webpages and then converts the terms to a group of 32-bit global IDs as the feature of the data collection after that, for training samples these integer vectors are transformed into term frequency features and collected in the database.

ii. **Classifier training module:** ten heterogeneous classifiers are built based on the characteristic of different features.

iii. **Ensemble classification module:** Ensemble classification method is used to combine all the prediction results from heterogeneous classifiers.

iv. **Cluster training module:** Hierarchical clustering algorithm is applied on the term frequency vectors with the TF-IDF weighting scheme.

4. Based on heuristics anti-phishing detection

i. Domain check module: This module will compare the domain name which user is trying to navigate by, with the domain names that are stored by user's browser. If they have certain similarity then system will give warning to user.

ii. URL check module: This module includes three steps, step_1 Check whether the URL that will be navigated to contains suspicious username; step_2. Check whether the host name or domain name in the URL didn't been hidden. Step_3. Check the page which will being navigated to request from a standard port.

iii. Email check module: This module checks whether the directed links link to email address, whether the current email domain name is empty also whether an email domain name is from a known website.

iv. Password check module: This module checks whether current page contains fields such as „password‟ or "pass‟ or "pwd‟, if the page contains these fields and the fields hasn't been encrypted the system will give a warning to user.

v. Link check module: This module checks whether current pages contains suspicious links and suspicious link refers to a link which triggered a warning when it was passing domain check and URL check.

vi. Image check module: This module will compare images in current page with images from pages which are accessed before and compute their hash values. If an image in current page is having same value from one image accessed before, the system will give a warning.

5. Based on Neuro-Fuzzy algorithm

- In this study, five inputs which are five tables where features are extracted and stored for reference and these includes: Legitimate Cluster training module: Hierarchical clustering algorithm is applied on the term frequency vectors with the TF-IDF weighting scheme.

6. Based on heuristics anti-phishing detection

Domain check module: This module will compare the domain name which user is trying to navigate by, with the domain names that are stored by user's browser. If they site rules: legitimate site rule is a summary of law which covers phishing laws, User behavior profile: It is list of people's behavior when interacting with legitimate and phishing website and Phish tank: it is a free community website operated by open domain names where suspicious websites are verified and voted as phished by the community experts or user specific side. Among the five inputs, 288 features are extracted which are used as training and testing input data into the Neuro-Fuzzy system for generating Fuzzy IF…THEN rules, and for discriminate between phishing, suspicious and legitimate websites.

7. Based on SVM classifier

Support vector machine is supervised machine learning algorithm which can be used for classification and regression. In this study first a given webpage is parsed into a DOM (Document Object Model) tree to allow easier processing for further step. DOM which is a World Wide Web Consortium (W3C) standard is a platform and language neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of document. Therefore after the DOM tree is constructed they will check whether a page contains any text inputs, since a fake page always requires users to input credentials.

8. Machine learning algorithms:

In this study the Comma Separated Values (CSV) file format was used. The input file to the WEKA was obtained by a MATLAB program by appending 'YES' in place of decision vector '1' (phish) and 'NO' in place of decision vector '0' (benign) of the dataset generated by MATLAB from input URL list after Feature Extraction they used four type of Machine learning classification algorithms to classify the URL [13]

The four machine learning algorithms considered for processing the feature set are:

i. **Naive Bayes:** it is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes's rule) with strong independence (naive) assumptions. It takes only one pass over the training set and is computationally very fast

ii. **J48 decision tree:** it is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data.

iii. **K-NN:** It is based on closest training examples in the feature space. An object is classified by a majority vote of its neighbors.

iv. **SVM:** The SVM performs classification by finding the hyper plane that maximizes the margin between two classes. The vectors that define the hyper plane are the support vectors.

- Place of decision vector ‘0’ (benign) of the dataset generated by MATLAB from input URL list.

5. Data set Types

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. The term data set originated with IBM, where its meaning was similar to that of file. In an IBM mainframe operating system, a data sets a named collection of data that contains individual data units organized (formatted) in a specific, IBM-prescribed way and accessed by a specific access method based on the data set organization. Types of data set organization include sequential, relative sequential, indexed sequential, and partitioned. Access methods include the Virtual Sequential Access Method (VSAM) and the Indexed Sequential Access Method (ISAM).

The authors in [11] used Dataset of URL that is fed to the model at the initial stage, Which contains 200 Legitimate as well as phishy websites URLs, {collected from the Phish Tank and yahoo directory}, Which will be used to Train the Machine Learning Algorithm and Test the performance of it.

The authors in [13] collected URLs of benign websites from www.alexa.com www.dmoz.org and personal web browser history. The phishing URLs were collected from www.phishtak.com. The data set consists of 17000 phishing URLs and 20000 benign URLs. They obtained PageRank of 240 benign websites and 240 phishing websites by checking PageRank individually at PR Checker. Also they collected WHOIS information of 240 benign websites and 240 phishing websites.

The authors in [14] obtained URLs from two sources in order to gather ground truth data of phishing and legitimate webpages. Neither dataset contains personal data. Both datasets are available on request for research use. The phishing URL sets (Phish) were obtained through the community website Phish Tank. They conducted three different collection “campaigns”. The first resulted in phish Train which was used for training the phishing detection classifier. The second, collected at a later point in time, resulted in phish Test which was used as the test set. The last, phish Brand, was used for evaluating our target identification scheme.

6. Evaluation

Evaluation is the systematic and assessment of information related to the outcomes, operation or processes of policy structure, organization or relationship. It is necessary to ensure the accountability, effectiveness and sustainability for a project, making it a necessary part of project management.

After you determine your general approach and establish an evaluation framework, you will need to choose your evaluation tools.

Your evaluation tool should reflect the overall objective of your evaluation and the indicators you are trying to measure. Detailing process lends itself to qualitative tools while large-scale aggregate efficiency evaluations need quantities tool. Many evaluations will require a mixed-method approach utilizing quantitative and qualitative tools to satisfy an array of audiences.

The authors in [13] [21, 22] evaluated the performance based on Detection Accuracy, True Positive Rate and False Positive Rate and Features by using quality metrics, Confusion matrix. The evaluation do for Classification algorithm agents each and one of Classification algorithm.

Table 1. Explain the Evaluation proses

<table>
<tr><th>Evaluation for
Evaluation Tools</th><th colspan="2">Classification algorithm agents each</th><th colspan="2">One of Classification algorithm</th></tr>
<tr><td rowspan="2">quality metrics</td><td colspan="2">Test Detection Accuracy Vs.</td><td colspan="2">Test Detection Accuracy Vs.</td></tr>
<tr><td>TP Rate, FP Rate</td><td>Features</td><td>TP Rate, FP Rate</td><td>Features</td></tr>
<tr><td rowspan="2">Confusion matrix</td><td colspan="2">True Positive Rate and False Positive Rate Vs.</td><td colspan="2">True Positive Rate and False Positive Rate Vs.</td></tr>
<tr><td>Accuracy</td><td>Features</td><td>Accuracy</td><td>Features</td></tr>
</table>

7. DISCUSSION

Phishing attack had four types as showing in figure 1, Website phishing attack, Email phishing attack, Web Trojans and Content-injection phishing. This type of phishing attack can detecting by used one of phishing Detection methods we mentioned it in this paper Blacklist method, Heuristic based method, Visual similarity, Content based approach and URL based method this method can classified in 3 criteria’s. Also this paper discuss Technology used in phishing Detection methods Based on Neuro-Fuzzy algorithm, Based on heuristics anti-phishing detection, Based on associative classification data mining, Based on genetic algorithm, Based on Machine learning and Intelligent Phishing website detection and categorization model. After that we speak about Data set Types Dataset had two type URLs of Legitimate website and URLs of Phishing website. And finally we speak about evaluation proses by using quality metrics, Confusion matrix. The performance evaluated based on True Positive Rate and False Positive Rate, Detection Accuracy.

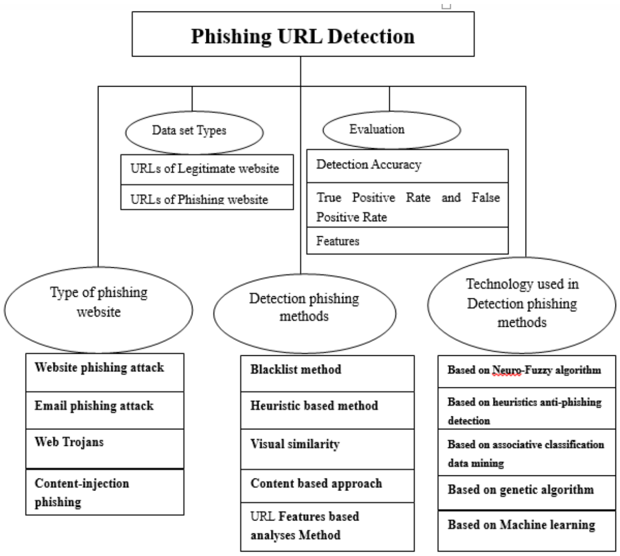

Figure 1. LITERATURE REVIEW

8. CONCLUSION

Based on the literature review discussed in this paper, the four type of phishing attack done in deferent form like creating fake URL, Misspelled URLs, Creating anchor text, Fake SSL lock and URL manipulating using java script and The Detection methods can classified in 3 criteria's Lexical based feature this Lexical based Features are textual properties of the URL itself and it easy to implement, Host based feature this criteria explain "where" phishing sites are hosted, "who" they are managed by, and "how" they are administered and Content based feature This criteria detects fake website by inspecting the similarity between legitimate and fake website. It's better to collect a large dataset to get good result when you develop anti-phishing model. Detects fake website by inspecting the similarity between legitimate and fake website. It's better to collect a large dataset to get good result when you develop anti-phishing model.

9. REFRENCE

1. Workman, M., *Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security.* Journal of the American Society for Information Science and Technology, 2008. **59**(4): p. 662-674.
2. Jain, A.K. 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, 2016. **2016**(1): p. 9.
3. Shukla, Z., K. Zala, and R. Kotak, *A Survey of Website Phishing Detection Techniques.* 2018.
4. Suganya, V., *A review on phishing attacks and various anti phishing techniques.* International Journal of Computer Applications, 2016. **139**(1): p. 20-23.
5. Gupta, S., A. Singhal, and A. Kapoor. *A literature survey on social engineering attacks: Phishing attack*. in *2016 international conference on computing, communication and automation (ICCCA)*. 2016. IEEE.
6. Alsharnouby, M., F. Alaca, and S. Chiasson, *Why phishing still works: User strategies for combating phishing attacks.* International Journal of Human-Computer Studies, 2015. **82**: p. 69-82.
7. Floderus, S. and L. Rosenholm, *An educational experiment in discovering spear phishing attacks*. 2019.
8. Pujara, P. and M. Chaudhari, *Phishing Website Detection using Machine Learning: A Review.* 2018.
9. Thaker, M., et al. *Detecting Phishing Websites using Data Mining*. in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2018. IEEE.
10. Sampat, H., et al., *Detection of Phishing Website Using Machine Learning*. 2018, IRJET.
11. Solanki, J. 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*. 2015.
12. Shukla, Z., K. Zala, and R. Kotak, *A Survey of Website Phishing Detection Techniques.* January 2018.
13. James, J., L. Sandhya, and C. Thomas. *Detection of phishing URLs using machine learning techniques*. in *2013 International Conference on Control Communication and Computing (ICCC)*. 2013. IEEE.
14. Marchal, S., et al. *Know your phish: Novel techniques for detecting phishing sites and their targets*. in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. 2016. IEEE.
15. Yi, P., et al., *Web phishing detection using a deep learning framework.* Wireless Communications and Mobile Computing, 2018. **2018**.
16. Suleman, M.T. and S.M. Awan, *Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms.* Automatic Control and Computer Sciences, 2019. **53**(4): p. 333-341.
17. Gawade, N., et al., *URL Based Analysis for Phishing Detection Using Data Mining.* Available at SSRN 3267436, 2018.
18. Lee, J.-L., D.-H. Kim, and L. Chang-Hoon. *Heuristic-based approach for phishing site detection using url features*. in *Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology-CEET*. 2015.
19. Fehringer, G. and P. Barraclough, *Intelligent security for phishing online using adaptive neuro fuzzy systems.* International Journal of Advanced Computer Science and Applications, 2017. **8**(6): p. 1-10.
20. Li, Y., L. Yang, and J. Ding, *A minimum enclosing ball-based support vector machine approach for detection of phishing websites.* Optik, 2016. **127**(1): p. 345-351.
21. Choi, H., B.B. Zhu, and H. Lee, *Detecting Malicious Web Links and Identifying Their Attack Types.* WebApps, 2011. **11**(11): p. 218.
22. Jain, A.K. and B.B. Gupta, *A machine learning based approach for phishing detection using hyperlinks information.* Journal of Ambient Intelligence and Humanized Computing, 2019. **10**(5): p. 2015-2028.