Proceedings of the 14th INDIACom; INDIACom-2020; IEEE Conference ID: 49435
2020 7th International Conference on "Computing for Sustainable Global Development", 12th - 14th March, 2020
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Detection of Malicioius Content using AI

Yogesh Pingle
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
yogesh.pingle@vcet.edu.in

Sushmita Patil
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
sushmita.patil@gmail.com

Sneha N.Bhatkar
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
sneha24bhatkar@gmail.com

Shruti Patil
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
1999shrutipatil@gmail.com

***Abstract*—Most disruptive action which is performed on the Internet is phishing. Personal files or any business-related information will be at risk if a user gets attack by such actions. These attacks are getting increased day by day. Some attack is carried by inducing a URL which looks similar to a legitimate URL to steal the user's important files. Aim of the project is to detect malicious sites made by attackers to steal user's personal information in the aim of conducting illicit activities. Features from the submitted URL will be extracted. Then decision tree algorithm will use these features and will classify the site as malicious or genuine.**

Keywords— Information Gain (IG), Support vector machine (SVM), Uniform resource locator (URL) and Iterative Dichotomiser 3 (ID3)

I. INTRODUCTION

In recent years, Internet had an enormous growth and there has been also enormous growth of web service. Even web attacks have increased in large numbers and even improved in quality. One of the popular attacks which is growing since many years is Phishing. One of the malicious attacks, phishing is carried out to steal user's personal and important information such as bank details, passwords and other important files which may cause harm to user if used for illicit activities. Phishing is done with different communication forms such as instant messaging, email, SMS, etc. But mostly users get tricked by phishing attack is caused through uniform resource locator (URL) [3].

Business are the most prone to get attacked because if a user is tricked to access a malicious URL, it is easy way to access the user's information so that the attack can gain access to the business network. The RSA 2012 annual fraud report states that since 2011, 59% phishing attacks has increased in 2012. Loss of $1.5 billion has been estimated as the losses in 2012 globally. It has been estimated by 2013, the losses will be difficult to handle for the businesses.

Individual Internet users can also be prone to such attacks. .COM namespace, a top-level domain contained most unique domain names which is used for phishing website also having most numbers of attacks. This information is not good news, as many Internet users may have various accounts on different websites related to social media, banking, emails etc. Users doesn't know that this information stored are not safe on internet. The malicious links are placed on those genuine web pages which users unknowingly access it. Like genuine web pages, malicious web pages are created which take users to the attacker' server instead of genuine web server. Malicious or phishing URL have unique characteristics which are different from the genuine URLs.

Attacker attacks through the URL, which the user is not able to identify it.

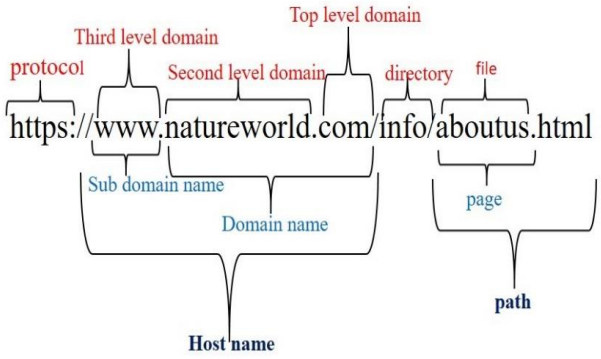

Fig. 1. Structure of URL

The above figure shows an example where URL structure for a website is defined. Protocol is used to access the page. Domain name is used to identify the server which owns that particular webpage. Then there is Sub domain, also called as registered domain name. There is suffix which is called top domain name. The Domain name part has to be registered under domain name Register. The part containing host name has sub domain name and a domain name. Attacker can easily

 ISSN 0973-7529; ISBN 978-93-80544-36-6

control the sub domain part. The URL have path which also can be controlled by the attacker. Attacker can register any domain name which has not been used before. The attack can change it anytime to create a new URL and will intelligently try to fool the users using convincing names [14].

For example, there is a URL name, http://paytm.com-webappuserid100456limited.active-userid.com/webapps/89980. It looks as a genuine paytm link.

TABLE I. STRUCTURE OF THE URL FOR ABOVE MENTIONED URL

Protocol	**http://**
Domain name	**Active-userid.com**
path	**//webapps/89980**
Sub domain 1	**com-webappuserid100456**
Sub domain 2	**paytm**

The above diagram shows the structure defined for the example mentioned above. The real domain name is active-userid.com. But the attacker tries to convince the user that it real website link of paytm.com by changing its sub domain. So users will trust whenever they see such link and in this ways share their credentials. There are different algorithms which help in solving this problem.

II. MOTIVATION

As we all know internet is growing rapidly because of which websites are becoming the intruder's main target. An intruder is someone who carryout malicious activities. He tries to insert this malicious content in a web page in order to do something bad or perform unwanted activities. If there is increase in the number of web pages, then there is also increase in a malicious web pages and eventually the phishing is rapidly becoming more genuine. Our aim is to provide a structure to detect a harmful web page. The main aim and goal of our system is that it will help the users to perform their activities easily without any fear. In this proposed system, 9 features are being extracted which will give more accuracy in finding a URL as malicious. There are two advantages of this system, one is that there are less chances of fraud or any type of harm, other is that user do not require any secure apps for their safety. As we have used the finest algorithm, latency time of the system has decreased and the results which we get our more efficient and web pages have become more user friendly.

There are many approaches which gives us the information on how this system works on web. Existing approaches like heuristic based approach, this phishing sites detection technique estimates and extracts phishing site related features from the URL such as Domain, obtained from the features detects the phishing sites. With this approach, when we want to scan a web page a signature database of known attacks is being built. When the signature in the database matches the heuristic patterns of the websites then the websites will be considered as phishing websites. New phishing sites and temporary phishing sites can be detected by this approach because it extracts features from the requested web page. Another approach is Machine learning, in this by making use of machine learning techniques such as Support Vector Machines (SVM), Decision tree algorithms, Random forest classification method etc. it utilizes many features of the URL and the websites [4]. ID3 decision tree algorithm is one of the algorithm to predict accurate results. The features which are being obtained are then combined to detect the phishing websites. But this approach has some limitations. First, techniques using machine learning may not work in such case that phishers will attack genuine domains and host malicious attacks on these servers. Second, due to text-based analysis mechanism, these detection techniques cannot detect the phishing websites which are purely made up of images.

III. LITERATURE SURVEY

The phishers aim is to trick the online users by making a fake URL which is similar to the authenticate URL of a company. To detect whether the URL is fake or authenticate, domain-related features of the URL are used. PrimaryDomain, SubDomains and PathDomain of the URL are checked to identify whether the website is fake or real. The ranking of site such as PageRank, AlexaRank, and AlexaReputation would help to detect phishing sites. It is used in the heuristic set to earn more accurate level detection [1].

To check whether the given URL is malicious or not, help of Machine learning algorithm is taken to develop an efficient classifier. Classification method used is decision tree. A tree form for classified samples is created. In a tree, every internal node represents an attribute, and the edges from the node divides the data which is based on the value of the attribute. The decision tree includes the decision area and leaf node. Condition of the samples is checked by the decision area and then it separates them both into leaf node or the new decision area. The advantage of the decision tree that it is very quick and implementing it is also easy. Decision tree model should have minimum complexity so that it's easy to classify the data. And also it has very good performance rate on large data sets [5].

There is one more algorithm called Random forest. Also denoted as RF, it is a popular packaged learning method for regression or classification. In Random forest classification method, it is done combination of many tree predictors, and the values of random vector are needed by every tree and those values are independently sampled. Every tree have a similar distribution. Mode or Mean is the result derived from the prediction of every tree. The disadvantage of Random Forest algorithm is that it is not easy to interpret, but has the capability to handle large datasets, and can handle multiple input features and also give prediction based only on input features. It has the highest accuracy prediction and also performs well in real world problems [4].

In this paper, Sonam Saxena et al [9], describes various ways in which phishing is done and various contributions done for classification of malicious URLs. This paper gives a survey on successful techniques and different ways for classification of URLs. In this system, use of PhishTank database is done to give input to the system. PhishTank database is an anti-phishing website or a repository online users track suspected phishes or malicious URLs and submit it so that track of suspicious URLs is mentioned and even other user can verify it. The PhishTank database includes different features like date of reporting the malicious website, URL, name of the company or client and etc. The PhishTank database checks any feature or attribute is missing. System will check for attributes. Then the list of URL is separated from database. Then evaluation of URL is done by using some features from URL. Then with the help of threshold values, 2D vector is transformed into binary form. Then the rule mining algorithm are applied to generate the associating rules, two important algorithm are used, one is FP tree and another is Apriori. Association rules are kept differently so that it can be used for classification of URLs so recognize malicious and genuine URLs. The main agenda of this survey was to check the issues of web security which gets spoiled due to this phishing attack and various techniques are explored to the classification of URL.

S. Jagadeesan et al [10], describes in this paper that by with use of URL information, we can detect whether the website is malicious or not without opening it. In this proposed method, the author has worked with two datasets which were extracted from UCI Machine Learning Repository. First dataset consist of features related to URL and many URLs both phishing and non-phishing. Even the second dataset consist many features and the URLs are of three types, phishing, non-phishing and suspicious. When this dataset are extracted, then data slicing is performed. Data slicing is done by diving the data into two parts which is training dataset and testing dataset. Training dataset is where it is used to train model. Testing dataset is done after training the model is completed and then accuracy of the model is tested. Author has used two algorithms for classification, Random forest and SVM. In Random forest, randomly number of classification trees are created. The trees are created with use of various samples from the dataset according to its features. When the trees are formed, classification is done by checking the results for every tree and then allocating the results to the class which decides the most number of trees. In SVM, hyper plane are created so that it creates boundaries for various classes. During testing, author comes to conclusion that Random forest is better than SVM as the accuracy was more precise of the Random forest.

Lekshmi AR et al [11], in this paper the author goes through different techniques used for classification of URL for detection of phishing websites. Also reviews the phases like feature representation and extraction phase. The author explain different methods for detection of URL. First method is, blacklist approach which is common method in which a database is used consisting list of URLs where they are suspected to be malicious. Whenever a user visits a new URL, the blacklist will perform a search in database to find whether the URL is mentioned, if it is then it will be detected. But it has drawback that the attacker can modify the URL in less period of time. It confuses the host with other domain or large host names. And due to this it has low false positive rates. Other method is heuristic or rule based approach; where it obtains the features of the malicious site and then detects it. Each time a URL is detected, features are extracted from that website and stored as signatures list and the applied for next time. Website with malicious content are occurring, different patterns are obtained. For Behavior based detection, it has a data collector, interpreter and a matcher. When all types of information is collected by the data module into intermediate result and then the matcher matches it with behavior patterns. Data mining algorithm used for heuristic approach is Naïve Bayes theorem. This approach requires more of time, money and work force to extract the distinguished patterns. Next approach is Machine

Learning approach, where the system is trained itself to learn without interference of human. A training model is developed with the help of different algorithms which is trained by set of training data. When a new input appears, it gives some prediction with the help of training model. Predictions are checked by the accuracy. Until the satisfied results, the trained model is developed.

In this paper [12], the author has given wide survey on various works done to detect malicious URL. An approach is mentioned where feature selection algorithms are developed to minimize the features of dataset to obtain larger order execution. Dataset rom UCI machine learning repository is used. Different data mining classification algorithms are performed to compare the results and the results showed that some show more accurate algorithms. Classification algorithms like Bayesian network, lazy k star, logistic model tree (LMT) and ID3 gives more accurate results and reduces malicious dataset.

In this paper [13] et al, the author has given a literature review on detecting the emails which are malicious. To detect this malicious emails, use of hybrid features are made. Hybrid features are collection of two features that are content based and header information.

IV. Existing System

To get away from such malicious attacks, there has been research done in this field where analysis is done to check whether the URL is malicious. URL analysis or link analysis means the information of URL which is used in an email is used to detect the email for malicious attack. This analysis is done to check if the link displayed in the email matches the original website URL which the user visits when clicked on that particular, and the patterns of the URLs in the email is examined to check and compare the features of the malicious or phishing URLs. The disadvantage of method on analyzing on URL is the vulnerability to the malicious URLs or emails holding the URLs in different forms. The phishers use auto-generated system to develop new URL every time. Also, this URL analysis is a heuristic based approach but disadvantage by using heuristics that they produce high FPR. (It wrongly labels emails as malicious or phishing).

A popular website Phish Tank, contains a blacklist, uses a wisdom-of crowd's approach so that it can collect malicious websites. In order to check whether the mentioned website is legitimate or phishing, users are told to report which sites they feel is phishing to the PhishTank website and then it is decided by people's vote whether the submitted names of the website are phishing scams or not. Since October 2006, PhishTank has received more than 7 million votes [9]. In blacklisting, client and server side, both are necessary. Implementation of client's component can be completed with the help of email or a browser plug-in which will communicate with the server component. Server component is a public website which contains list of phishing websites.

V. PROPOSED SYSTEM

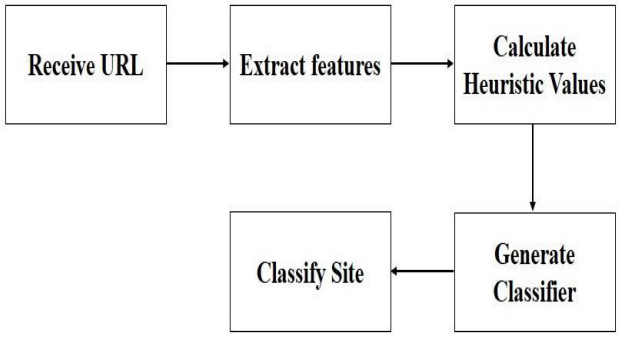

Fig. 2. Diagram for Proposed System

Detection of Malicious content on World Wide Web has become very complex due to the growth in advanced technologies and also growth in web attacks. The present state of cybercrime has made it easy for Hackers to do attacks with shorter lifecycles, diminishing blacklist effectiveness.

In modern times as the techniques for Detection of Malicious content have advanced, various methods present some advantages as well as issues. Data mining techniques are most popular and best techniques for this detection. Therefore, there are many systems implemented in this field. As a result, we need a system with,

1. Appropriate methodology
2. Less processing time
3. Good value of evaluation metrics

The proposed system focuses on yielding accurate results regarding the decision about Phishing site or legitimate by dividing the system in training phase and detection phase. In training phase, the given URL is compared with the dataset and then with the help of 9 features, total score will be calculated.

Total score is calculated with the help of ID3 Decision tree algorithm. URLs are further transferred to the feature extractor which excerpts values from the features through the pre-defined URL based features. Features will be extracted with the use of association rule of mining. This rule is used to ascertain the URL type when a user accesses it. This rule is used to make sure the type of URL when accessed by user. This excerpted feature are kept as an input and then transmitted to the classifier generator, input feature is built. Whether the given site is malicious or not will be carried out by classifier in detection phase. Feature extraction excerpts the values from the features and give this value as an input to the classifier. Classifier will recognize the phishing site and will alert the user. It uses less processing of time and fast retrieval of data. Heuristic based approach along with decision tree algorithm is used in this system to enhance the accuracy of mentioned system [4][6].

A. Working

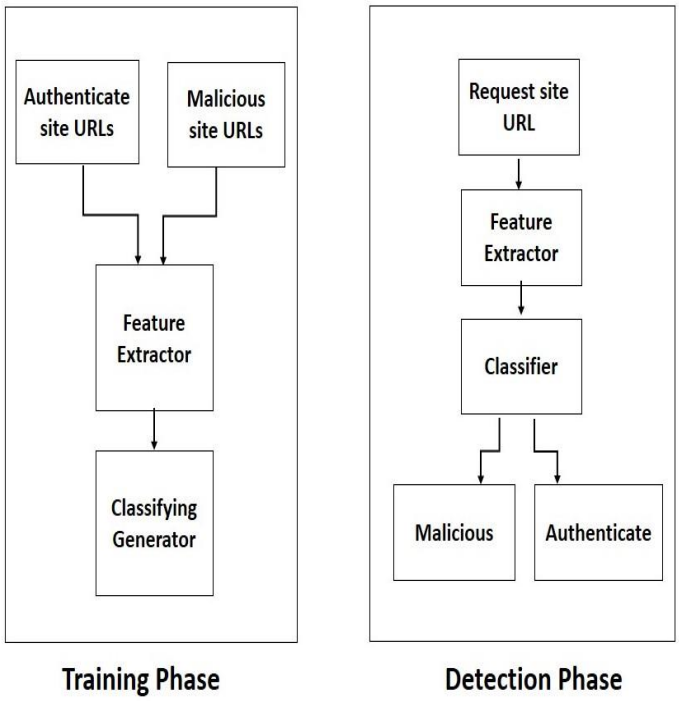

Fig. 3. Diagram shows working of the system

- Step 1: In this at first, users searches for query.
- Step 2: Then search engine tries to find items in a database like keywords or characters specified by the user, used specifically for discovering specific sites on the Internet.
- Step 3: The list of these URLs searched by users are stored in the database.
- Step 4: For each URL, 9 factors are calculated and stored as total-score in the database.
- Step 5: Total-score is calculated using ID3 algorithm like,
  a) It creates a tree form for classifying URLs.
  b) In that tree, internal node coincide to a feature and the edges from the node divides the data related to the feature value.
  c) It contains a decision area and leaf node.
  d) Decision area checks the conditions of the URLs and classify into leaf node or the next decision area.

e) URLs are checked on basis of the 9 factors. This factors are being given some score through which the total-score is calculated by ID3 and then stored in the database.

- Step 6: In this the training phase collects the legitimate and phishing URLs.
- Step 7: This URLs are further transferred to the feature extractor which excerpts the values from the features through the pre-defined URL based features.
- Step 8: This excerpted feature are stored as an input and then transmitted to the classifier generator, input features and the machine algorithm will be produced.
- Step 9: Classifier will discover whether the mention site is malicious site or not in Detection phase.
- Step 10: Feature extraction excerpts the values from the features and give this value as an input to the classifier.
- Step 11: Classifier recognizes the phishing site and alerts the user.

B. Extraction of Features

The main aim of the attacker is to craft a URL in a way to fool users to click on it and at same time avoid detection by detection systems. Data mining is a method which is used to get patterns in large data sets. The main aim of data mining method is to take out information from data set and reconstruct it into clear and understandable structure. One of the method used in data mining is association rule mining, and its aim is to find associations among items in a set, by mining required knowledge from the database. The process of searching the website is generated using association rules in which the different heuristics are used to extract various information [8].

There are many malicious URLs found on the internet, where most have them have features such as long URLs, more number of dots, and presence of words such as verify, online, secure, etc. to fool users into thinking that it is a legitimate website. Commonly, the malicious detection systems, work on URL features check length on domain name, path length, number of dots or special characters, presence of brand name, etc. For example, the attacker creates a malicious URL, http:/slin.ch/STD/Standardbank.co.za/index.php, to appear as a webpage from standard bank website.[13]

This website can be detected using features such as length of hostname, length of URL and no. of special characters in URL. If the malicious website created is www.slin.ch, then the chances of the online user clicking it is lesser. Therefore, to make a successful attack, the URL created by the attacker must long enough to include a deceptive name (like verify, account, etc.) or brand names to fool users into clicking it.

The Features considered to detect the malicious site are:

1. IP Address: If the given URL contains IP Address as part of it, then it is considered as Phishing otherwise legitimate. If an IP Address is attached with URL, then it a hint that some attacker is trying to access the system and steal sensitive information through the attack. Therefore, A URL contains IP Address, then the system will specify it as legitimate or malicious.
2. URL length: URL is a format of text string used by users to discover a network resource on the Internet. If the length of the URL is accepted till 75, then the length more than 75 will be detected as suspicious or malicious. Example, if the length of the URL is <54, them it is Legitimate, or length of the URL is >=54 and <=74 then it is Suspicious.
3. Suspicious Character: If the domain name of the URL includes '@' symbol then it considered as Malicious otherwise legitimate.
4. Prefix and Suffix: Whether the URL is malicious or not, will be confirmed when there is existence of more than one hyphen in the host name of the URL. In many cases, genuine URLs are found to have one hyphen. If the domain of the URL contains '-', then it will be considered as Malicious.
5. Number of dots: If the amount of dots in the domain is specified by certain number, the URL is classified as phishing. For example, If the amount of the dots in the domain part of the URL<3, then it is Legitimate or if number of the dots in the domain part = 3, then it is Suspicious means detected malicious Phishing.
6. Length of subdomain: Legitimate websites will have only one top level domain and few subdomains, but a fraud website will have many subdomains to fool the users. For example, www.networksolutions.com.012892378267.239821432.mobi/login.secure is a link. So attackers try to make it less obvious by using really long URLs hoping that the users won't have time to check and just click on it and won't check the dots. Check for the first dot from Right to Left. Comma or slashes. You should spot the first dot just before .mob and the second dot shows that the actual domain is **239821432.mobi**. So all the other characters to the left of **.239821432.mobi** are subdomains served by **239821432.mobi.** The point is that phishing websites use similar or genuine looking sub-domains to try to trick the users [7].
7. Number of slash: Attackers tries to fool web users by imitating the suspicious URL look genuine. One of the process used in scamming is the process of additional of slashes in URL. The current study, therefore, checks the amount of slashes added in the URLs as feature to find malicious websites and examining the amount of slashes in the original URL and the malicious URLs. If there are amount of slashes in a given URL exceeds number specified to detect, then it is detected as Malicious. For example, if the total number of slash is more than 2, then it is considered as malicious or legitimate.
8. Http Protocol: If the http is not present in the URL, then it is considered as Malicious.

9. Phishing term: Attackers use a lot of terms which are included in a authenticate website such as secure, verified, or banking terms, etc. in the intention to deceive users. If such words are mentioned in the URL domain, then the system will detect it as malicious.

C. Algorithm

ID3 algorithm, also known as Interative Dichotomiser 3 is a popular algorithm in machine learning and data mining and it is often used because it is easy to use and is very effective. These Algorithm is developed by J.Rose Quinlan in 1986 on Concept Learning System (CLS) algorithm. This algorithm constructs a decision tree from some calculated data or significant symbolic data in order to arrange them and estimate the categorization of new data. The information should have different characteristics with varied values. Meanwhile, this data also has to belong various predefined, separate classes (which is Yes or No). Information gain (IG) is used by the decision tree for choosing the attribute for decision making [14].

The tree is created by choosing the most important and prominent feature which shows the structure of the model for the detection part. Mechanism which has the highest success rate is based on training dataset. Training set has numerous amount of URLs taken from different data sources.

VI. CONCLUSION

We developed a malicious URL detection technique that used URL based features. As we know traditional approach fall short in detection of malicious URL.

This method is a combination of URL based features that are used in previous studies with some additional features for the purpose of analyzing the web page link. Here in this system, we proposed the feature set that can easily classify the URLs, whether it is malicious URL or not. The motive of this technique is to reduce damage caused by phishing attacks and provide better security for personal information. We tested several machine learning algorithm and determined that best classifier was ID3. Our method achieved an accuracy for detecting malicious activity. This technique helps the user to not be a victim of malicious URL. The motive of dividing the project into small modules so that the implementation can be easier.

In future work, we will use new machine learning algorithm that will provide better result by utilizing the given feature set. Addition to that another question is how to handle huge amount of URLs in small time span so we will apply algorithm to reduce features. We will test phishing activity detection technique that utilize not only URL Characteristics but also HTML features of web page for better process of performing the task .

REFERENCES

[1] Nilima Ramdas Narad, Sandeep U. Kadam, "Web Phishing Detection System: Bayesian and Clustering Approach", International Journal of Computer Applications, Volume 145, No. 10, July 2016.

[2] Dhanashree Pawar, Dhanalakshmi Sherbhai, Akshata Shelar, Neha singh, :Detection of Phishing site Using Efficient Approach", IOSR Journal of Enginnering (IOSRJEN), ISSN (e): 2250-3021, ISSN (p): 22788719, Volume 7, PP 55-57.

[3] Gyan Kamal, Monotosh Manna, "Detection of Phishing Websites Using Naïve Bayes Algorithm", International Journal of Recent Research and Review, Vol XI, ISSUE 4, December 2018.

[4] Frank Vanhoenshoven Gonzalo Naples, Rafael Falcon, Koen Vanhoop and Marion Koppen, "Detecting Malicious URLs using Machine Learning Techniques", 978-1-5090-4240-1, 2016 IEEE.

[5] Akshay Sushena Manjeri, Kaushik R, Ajay MNV, Priyanka C. Nair, "A Machine Learning Approach for Detecing Malicious Websitesusing URL Features", ICECA 2019, IEEe Conference Record 45616: IEEE Xplore ISBN: 978-1-7281-0167-5.

[6] Denbanchakkarawarthi G., Partan AS. Sachin Lal, Surya A, "Classification of URL into Malicious or Benign using Machine Learning Apporach", IJARCCE, Vol 8, Issue 2, February 2019.

[7] https://easykey.uk/computer-safety/how-scammers-use-sub-domains

[8] https://hcis-journal.springeropen.com/articles/10.1186/s13673-0160064-3

[9] Sonam Saxena, Amit Shrivastava, Vijay Birchha, "A Proposal on Phishing URL Classification for Web Security", International Journal of Computer Applications (0975-8887), Volume 178 - No.39, August 2019.

[10] S. Jagadeesan, Anchit Chaturvedi, Shashank Kumar, "URL Phishing Analysis using Random Forest", International Journal of Pure and Applied Mathematics, Volume 118, No. 20, 2018, 4159-4163, ISSN: 1314-3395 (on-line version)

[11] Lekshmi AR, Seena Thomas, "Detecting Malicious URL Using Machine Learning Techniques: A Comparative Literature Review", International Research Journal of Engineering and Technology (IRJET), Volume: 06, Issue:06, June 2019.

[12] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machime Learning", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue- 2S11, September 2019.

[13] Jagruti Patel, Sheetal Mehta, "A Literature Review on Phishing Email Detection Using Data Mining", International Journal of Engineering Sciences and Research Technology, (IJESRT), ISSN: 2277-9655, Scientific Journal Impact Factor: 3.449, (ISRA), Impact Factor: 2.114, March 15.

[14] www.towardatascience.com