International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 3, pp 144-148, March-2022

https://doi.org/10.47001/IRJIET/2022.603019

Based on URL Feature Extraction Identify Malicious Website Using Machine Learning Techniques

[1]Khushbu Digesh Vara, [2]Vaibhav Sudhir Dimble, [3]Mansi Mohan Yadav, [4]Aarti Ashok Thorat

[1,2,3,4]Navsahyadri Education Society's Group of Institutions, Pune, India

***Abstract* - A phishing attack is the simplest way to obtain sensitive information from users. The aim of the phishers is to acquire critical information like username, password, bank account details and other personal information. With the development of Internet technology, network security is under different threats. Especially attackers can spread malicious uniform resource locators (URLs) to carry out attacks such as phishing and spam. The research on malicious URL detection is significant for defending against this attack. Some existing detection methods are easy to cover by attackers. We design a malicious URL detection model based on Machine Learning Techniques to solve these problems. Cyber security persons are now looking for reliable and stable detection techniques for phishing websites detection. This propose system deals with machine learning technology for the detection of phishing URLs by extracting and analysing various feature of legitimate and phishing URLs. Decision Trees, random forest and support vector machine algorithms are used to detect phishing websites or unsecure websites. The aim of the paper is to detect phishing URLs as well as cut down to the best machine learning algorithm by comparing the accuracy rate, false positive and false negative rate of each algorithm. This paper analyses the structural feature of the URL of the Phishing websites extracts 12 kinds of features and uses four machine learning algorithms for training and use the best-performing algorithm as our model to identify unknown URLs.**

Keywords: Phishing attack, Machine learning, Cyber Security, Website Classification.

I. INTRODUCTION

Hackers frequently use spam and phishing to trick customers into clicking malicious URL, the Trojans might be implanted into the victims' computers, or the victims' sensitive information might be leaked. The technology of malicious URL detection can assist customers pick out malicious URL and save you customers from being attacked by malicious URL. Traditionally, studies on malicious URL detection adopt blacklist-based techniques to detect malicious URL. This technique has a few particular advantages. It has excessive speed, has low false-high quality rate, and is simple to realize. However, nowadays, the domain generation algorithm (DGA) can generate thousands of various malicious domain names each day, which cannot be detected correctly by the conventional blacklist-based techniques.

Researchers were the use of a machine learning approach to discover malicious URL. However, those techniques frequently want to extract the capabilities manually, and attackers can design these functions to avoid being identified. Faced with today's complex community environment, designing a extra effective malicious URL detection model will become a research The human-understandable URLs identify focus. Since phishing attack exploits the weaknesses observed in users, it's far very hard to mitigate them however it may be very critical to enhance phishing detection techniques.

The popular approach to discover phishing web sites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is likewise recognized as "blacklist" approach. To cover blacklists attackers makes use of creative strategies to fool customers through editing the URL to seem valid through obfuscation and lots of other easy strategies together with: fast-flux, wherein proxies are mechanically generated to host the web-page; algorithmic technology of latest URLs; etc. Major disadvantage of this approach is that, it cannot discover zero hour phishing attack. Heuristic primarily based totally detection which consists of characteristics which can be observed to exist in phishing attacks in fact and may discover zero-hour phishing attack, however the characteristics are not assured to usually exist in such attacks and false effective rate in detection could be very high.

To overcome the drawbacks of blacklist and heuristics primarily based totally approach, many safety researchers now focused on machine learning strategies. Machine learning generation includes a many algorithms which require past information to decide or prediction on future records. Using this technique, set of rules will examine numerous blacklisted and valid URLs and their functions to correctly discover the phishing web sites together with zero- hour phishing web sites.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 3, pp 144-148, March-2022

https://doi.org/10.47001/IRJIET/2022.603019

II. LITERATURE SURVEY

The human-understandable URLs identify billions of websites hosted over the present-day internet. Adversaries who try to get unauthorized access to confidential data may use malicious URLs and present them as legitimate URLs to naive users. Such URLs that act as a gateway for unsolicited activities are called malicious URLs. These malicious URLs can cause unethical activities such as theft of private and confidential data and ransom ware installation on user devices, resulting in massive losses every year globally. Even security agencies are cautious about malicious URLs as they can compromise sensitive and confidential data of government and private organizations. With the advancement of social networking platforms, many allow their users to publish unauthorized URLs. Many of these URLs are related to business promotion and self-advertisement, but some of these unprecedented resource locators can pose a vulnerable threat to naive users. The naive users who use the malicious URLs will face serious security threats initiated by the adversary.

The verification of URLs is essential to ensure that users should be prevented from visiting malicious websites. Many mechanisms have been proposed to detect malicious URLs. One of the essential features that a tool should possess is to allow the benign URLs requested by the client and prevent the malicious URLs before reaching the user. This is achieved by notifying the user that it was a malicious website and a caution should be exercised. To achieve this, a system should take semantic and lexical properties of every URL rather than relying on syntactic properties of the URLs. Traditional methodologies such as Blacklisting, Heuristic Classification can detect these URLs and block them before reaching the user.

Back-listing is one of the primary and trivial mechanisms in detecting malicious URLs. Generally, Black-List is a database that contains a list of all URLs which are previously known to be malicious. A database lookup is performed every time the System comes across a new URL. The unique URL will be matched and tested with every previously known malicious URL in the black list. The update must be made in the black list whenever the System comes across a new malicious URL. The technique is repetitive, time-consuming, and computationally intensive with ever-increasing new URLs.

The other existing approach, Heuristic classification, improves the Black-Listing. Here, the signatures are matched and tested to find the correlation between the new URL and the signature of the existing malicious URL. Even though both Black-Listing and Heuristic Classification can effectively classify the malign and benign URLs, they cannot cope with the evolving attack techniques. Recent statistics imply a 20-25% growth in the attacks yearly, and the threats coming from the newly created URLs are on the rise. One severe limitation of these techniques is that they are inefficient in classifying the newly generated URLs.

One of the other collaborative work has been initiated by top-tier Internet companies such as Google, Facebook, and many start-up companies to build a single platform that works all together for one cause of preventing the naive users from the malicious URLs. Many of these web-based companies use exhaustive databases that can store as many as millions of URLs and regularly refine these URL sets. Unblock ad blocker is an excellent example here to mention. Even though it is a manual procedure to update periodically, the performance was good, and the database contains up-to-date URLs. But is this the feasible solution to all the problems? The answer is NO. Despite greater accuracy, the need for human intervention to update and maintain the URL list is one of the major limiting factors in this method.

We propose a novel approach using sophisticated machine learning techniques that Internet users could use as a common platform to counter these limitations. In this paper, we offer a method to detect malicious URLs. Various feature sets for URL detection have also been presented to be used with Support Vector Machines (SVM). The feature set comprises 18 features, such as token count, average path token, most extensive path, most significant token, etc. We also propose a generic framework that can be used at the network edge. That would safeguard the naive users of the network against cyber attacks.

III. PROPOSED SYSTEM

In this approach, a machine learning-based website URL phishing detection system has been proposed which is more accurate and fast as compared to other methods. The detailed System is as explained below. The block diagram of the proposed system is as shown in the given Fig.

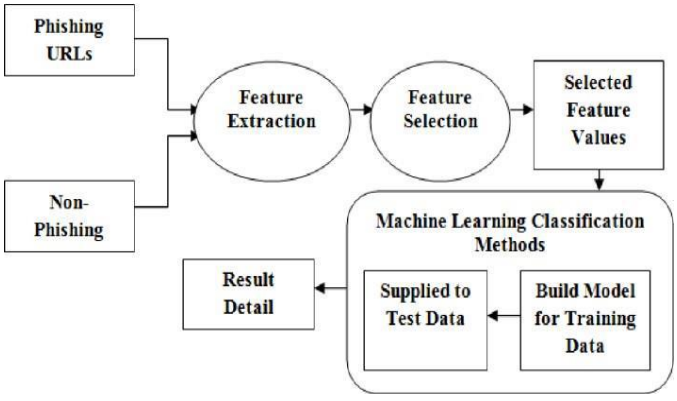

Figure 1: Block Diagram of proposed System

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 3, pp 144-148, March-2022

https://doi.org/10.47001/IRJIET/2022.603019

In this approach, a machine learning-based website URL phishing detection system has been proposed which is more accurate and fast as compared to other methods. URLs of benign websites were collected from www.alexa.com and The URLs of phishing websites were collected from www.phishtank.com. The feature extracted from the URL is Lexical features, WHOIS-based features, PageRank, Alexa Rank, Phish tank-based features. The classification of features is performed using Random Forest and a content-based algorithm. The performance of the System is evaluated using accuracy. System design is the process of defining the architecture, modules, and data for a system to satisfy specified requirements. It includes a diagrammatical representation of the modules, data flow, and use cases of the System. In this chapter, we try to model the Emotion-based Movie Recommendation system in a way that will help design and shape the further process.

IV. FEATURE EXTRACTION

We have implemented a C# program to extract features from URL. Below are the features that we have extracted for the detection of phishing URLs.

1) Presence of IP address in URL: Most the benign sites do not use IP addresses as an URL. If IP address is present in URL, then the feature is set to 1 else set to 0.

2) Presence of @ symbol in URL: If @ symbol is present in URL then the feature is set to 1 else set t.0.

3) Number of dots in URL: Phishing URLs have many dots in URL. Maximum number of dots in benign URLs is 3. When the Number of dots in URLs is more than 3 then the feature is set to 1 else to 0. Let's see example, http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of “amazon” word is to trick users to click on it.

4) Domain separated by prefix or suffix by symbol (-): If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash (-) symbol is rarely used in legitimate URLs. Phishers add a dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example, the Actual site is http://www.onlineamazon.com but phishers can create another fake website like http://www.onlineamazon.com to confuse innocent users.

5) URL redirection: If “//” is present in the URL path then the feature is set to 1 else to 0. The existence of “//” within the URL path means that the user will be redirected to another website.

6) Present of HTTPS token in URL: Some of Phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://httpswwwpaypal-it-mpp-home.soft- hair.com. If HTTPS token present in URL, then the feature is set to 1 else to 0. To download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.

7) Information submission to Email: Phishers might use “mail ()” or “maillot:” functions to redirect the user’s information to his personal email. If such functions are present in the URL then the feature is set to 1 else to 0.

8) URL Shortening Services “Tiny URL”: Tiny URL service allows the phisher to hide long phishing URLs by making them short. The goal is to redirect the user to phishing websites. If the URL is composed using shortening services (like bit.ly) then the feature is set to 1 else 0.

9) Length of Hostname: The average length of the legitimate URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to 0

10) Presence of sensitive words in URL: Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'by ISAPI', 'website', 'signing', 'mail', 'install', 'toolbar', 'backup', 'PayPal', 'password', 'username', etc.

11) Number of slashes in URL: The number of slashes in benign URLs is found to be a 5; if a number of slashes in URL is greater than 5 then the feature is set to 1 else to 0.

12) Presence of Unicode in URL: Phishers can make use of Unicode characters in URL to trick users to click on them. For example the domain “xn80ak6aa92e.com” is equivalent to "apple.com". Visible URL to the user is "apple.com" but after clicking on this URL, the user will visit to “xn--80ak6aa92e.com” which is a phishing site.

13) The Age of SSL Certificate: The existence of HTTPS is most important in giving the impression of website legitimacy. But minimum age of the SSL certificate of benign websites is between 1 year to 2 years.

14) URL of Anchor: We have extracted this feature by crawling the source code of the URL. URL of the anchor is defined by the tag. If the tag has a maximum number of hyperlinks that are from the other domain then the feature is set to 1 else to 0.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 3, pp 144-148, March-2022

https://doi.org/10.47001/IRJIET/2022.603019

15) IFRAME (The Inline Frame Element): We have extracted this feature by crawling the source code of the URL. This tag is used to add another web page to the existing main webpage. Phishers can make use of the "frame" tag and make it invisible i.e. without frame borders. Since the border of the inserted webpage is invisible, the user seems that the inserted webpage is also part of the main web page and can enter sensitive information.

16) Website Rank: We extracted the rank of websites and compare it with the first One hundred thousand websites of the Alexa database. If the rank of the website is greater than 10, 0000 then the feature is set to 1 else to 0.

V. MACHINE LEARNING ALGORITHM

5.1 Decision Tree Algorithm

One of the most widely used algorithms in machine learning technology. The decision tree algorithm is easy to understand and also easy to implement. The decision tree begins its work by choosing the best splitter from the available attributes for classification which is considered as a root of the tree. The algorithm continues to build a tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to the class label. In the decision tree algorithm, gain index and information gain methods are used to calculate these nodes.

5.2 Random Forest Algorithm

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on the concept of decision tree algorithm. Random forest algorithm creates the forest with a number of decision trees. A high number of the tree gives high detection accuracy.

The creation of trees is based on the bootstrap method. In the bootstrap method features and samples of the dataset are randomly selected with replacement to construct the single tree. Among randomly selected features, a random forest algorithm will choose the best splitter for the classification, and as the decision tree algorithm; the Random forest algorithm also uses Gini index and information gain methods to find the best splitter. This process will get continue until a random forest creates n number of trees.

Each tree in the forest predicts the target value and then the algorithm will calculate the votes for each predicted target. Finally, random forest algorithm considers high voted predicted target as a final prediction.

5.3 Support Vector Machine Algorithm

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n-dimensional space, and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as a hyperplane.

Support vector machine seeks for the closest points called support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then constructs separating line which bisects and perpendicular to the connecting line. In order to classify data perfectly, the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In a real scenario, it is not possible to separate complex and nonlinear data, to solve this problem support vector machine uses kernel trick which transforms lower-dimensional space to higher-dimensional space.

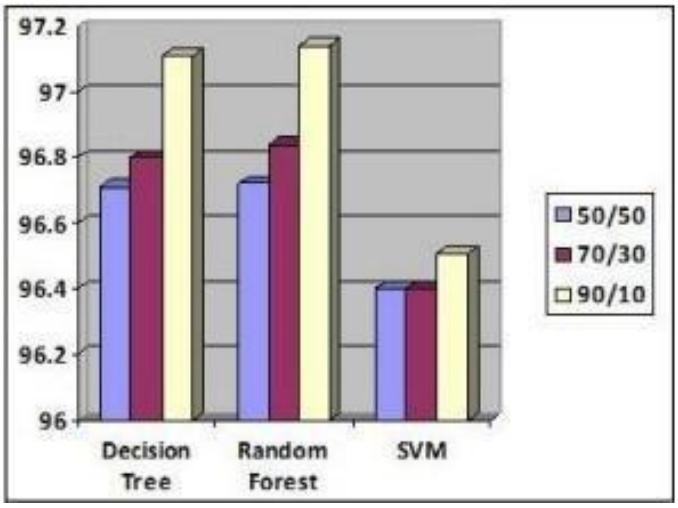

Figure 2: Detection accuracy comparison

VI. CONCLUSION

This paper aims to enhance detection methods to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using a random forest algorithm with the lowest false positive rate. Also, the result shows that classifiers give better performance when we used more data as training data.

In the future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

ISSN (online): 2581-3048
Volume 6, Issue 3, pp 144-148, March-2022
https://doi.org/10.47001/IRJIET/2022.603019

VII. FUTURE WORK

In this approach, we have come up with a new method for identifying phishing websites. Our goal is not just to identify phishing websites, but also to provide with the possible targeted domain.

In the future, we can use a combination of any other two or more classifiers to get maximum accuracy. We also plan to explore various phishing techniques that use Lexical features, Network-based features, Content-based features, Web page-based features, and HTML and JavaScript features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.

REFERENCES

[1] Justin. Ma, Lawrence. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: learning to detect malicious websites from suspicious URLs," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2009, pp. 1245–1254.

[2] Mohammed Al-Janabi, Ed de Quincey, Peter Andras, "Using Supervised Machine Learning Algorithms to Detect suspicious URLs in online social networks" Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017.

[3] Pde las Cuevas, Z. Chelly, A. Mora, J. Merelo, and A. Esparcia Alcazar, "An improved decision system for URL accesses based on a rough feature selection technique," in Recent Advances in Computational Intelligence in Defense and Security. Springer, 2016, pp. 139–167.

[4] A.Mora, P. De las Cuevas, and J. Merelo, "Going a step beyond the black and white lists for URL access in the enterprise by means of categorical classifiers," ECTA, pp. 125–134, 2014.

[5] M.-Y. Kan and H. O. N. Thi, "Fast webpage classification using URL features," in Proceedings of the 14th ACM international conference on information and knowledge management. ACM, 2005, pp. 325–326.

[6] E. Bayan, M. Henninger, L. Marian, and I. Weber, "Purely URL-based topic classification," in Proceedings of the 18th international conference on World wide web. ACM, 2009, pp. 1109–1110.

[7] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious URLs," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009, pp. 1245–1254.

[8] "Learning to detect malicious URLs," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 30, 2011.

[9] P. Zhao and S. C. Hoi, "Cost-sensitive online active learning with application to malicious URL detection," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 919–927.

[10] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina, a content-based approach to detecting phishing web sites" Proceedings of the 16th international conference on World Wide Web - WWW 07, pp. 639-648, 2007.

Citation of this Article:

Khushbu Digesh Vara, Vaibhav Sudhir Dimble, Mansi Mohan Yadav, Aarti Ashok Thorat, "Based on URL Feature Extraction Identify Malicious Website Using Machine Learning Techniques" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 6, Issue 3, pp 144-148, March 2022. Article DOI https://doi.org/10.47001/IRJIET/2022.603019

****

