International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2173124

Phishing Website Detection Based on URL

Salvi Siddhi Ravindra[1], Shah Juhi Sanjay[1], Shaikh Nausheenbanu Ahmed Gulzar[1], Khodke Pallavi[2]
[1]Computer Engineering Department, Shah & Anchor Kutchhi Engineering College Mumbai, Maharashtra, India
[2]Professor, Computer Engineering Department, Shah & Anchor Kutchhi Engineering College Mumbai, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 589-594

Publication Issue :
May-June-2021

Article History
Accepted : 20 June 2021
Published : 30 June 2021

ABSTRACT

In today's era, due to the surge in the usage of the internet and other online platforms, security has been major attention. Many cyberattacks take place each day out of which website phishing is the most common issue. It is an act of imitating a legitimate website and thereby tricking the users and stealing their sensitive information. So, concerning this problem, this paper will introduce a possible solution to avoid such attacks by checking whether the provided URLs are phishing URLs or legitimate URLs. It is a Machine Learning based system especially Supervised learning where we have provided 2000 phishing and 2000 legitimate URL dataset. We have taken into consideration the Random Forest Algorithm due to its performance and accuracy. It considers 9 features and hence detects whether the URL is safe to access or a phishing URL.

Keywords : URLs, Phishing, Legitimate, Machine Learning.

I. INTRODUCTION

In today's fast-paced world technology has become an essential part of everyone's life. Technology has been greatly escalating and thereby making our experiences comfortable. Nowadays our presence and business have been dependent on the internet and various online platforms. People perform various activities in their day-to-day life that includes accessing online shopping websites, banking websites, educational websites, and social media. Nonetheless, all these websites ask for our data and some of them consist of sensitive information that may be bank details or card details. And as a result of all this, hackers have found an easy way of attacking other personal information and tracking their behaviour. There are many types of attacks including Man-in-the-middle Attacks, Dos Attacks, SQL injection, Phishing Attacks, and many more. Out of all these websites, phishing has been considered a great threat to a user's vital information. The social engineering trick is used to manipulate the users and thereby duping them with the legitimate-looking URL which is a fake URL. It is difficult for a naive user to spot whether the URL is legitimate or fake. Research has shown that there has been a great boost in phishing attacks. Some researchers have a heuristic-based approach while some use a Machine learning-based approach. Machine learning has two different approaches i.e., Supervised Learning and Unsupervised Learning. In this paper, we have focused on supervised learning where we have

provided a 2000 dataset of phishing URLs and a 2000 dataset of legitimate URLs from phishtank.com. Also, we have made use of the Random Forest Algorithm due to its high accuracy, robustness, and good performance. And based on characteristic classification the system will differentiate the provided URL and will conclude whether the given URL is legitimate or a phishing URL. By which the user will be able to figure out that he might endanger his information if he visits that particular URL. And hence this system helps in guarding and thereby providing a possible solution towards the issue.

II. LITERATURE SURVEY

1) In this survey paper they have mentioned how phishing attacks appear, how the phishers use email or message, as evidence to target the individual or business by sending the link to victim people and deceive them with a large no of phishing emails or messages every day, so many of the corporations or individual are not able to recognize them all. so, here they have mentioned various types of phishing attacks like Learning Model Algorithm, Naive Bayes Algorithm, Decision tree, SVM (Support Vector Machine), Artificial Neural Network and many more. They have also mention phishing detection approaches like the Heuristic-based Approach, Fuzzy-based Approach, Machine Learning Approach, Image-based Approach, and so on [1].
2) This approach makes use of the Naïve Bayes, SMO, J48 algorithm are used for feature selection. There are several separate processes. The first process is to extract the properties of URLs and generate a matrix then secondary process uses the attribute-based feature selection technique to specify the prominent properties after using the attribute-based technique, the new dataset is used as input data to the Machine Learning Algorithm to analyze the website is legitimate or not. Based on the classification method on J48(Decision Tree), Naive Bayes, and SMO (Sequential Minimal Optimization). Here SMO & J48 shows their best accuracy output result was as Naive Bayes performed poorly and it is the least recommended method among all the methods [2].
3) This paper illustrates various types of Phishing Techniques and Anti-Phishing technique because phishing attack is one of a version of harmful content which has found recently a wide circulation in an information field of the modern switched communication systems. So, to identify the website is legitimate or not so there is some feature through which we can identify that the website is legitimate or not. If we enter the URL first it will be checked in the blacklist or whitelist if it is a blacklist that means it is a phishing URL else it is a legitimate URL (whitelist).[3]
4) The gap mentioned in "Phishing Websites Detection using Machine Learning" is that they have used a small dataset of 1300 URLs where we overcome this by using 4000 datasets from phishtank.com.[4]
5) "Phishing Website Detection based on Machine Learning: A Survey" is a survey paper that discusses different types of attacks and anti-phishing approaches. Also, some defense techniques for phishing are mentioned.[5]

III.PURPOSE

1. To preserve the confidentiality.
2. To protect the user from phishing websites.
3. To develop a user-friendly environment.
4. To prevent or mitigate harm or destruction of computer networks, applications, devices, and data

IV. METHODS AND MATERIAL

In a study of phishing, it is attacked through Email, Messages, or any communication media through which the particular link has to be clicked. In this proposed system we deal with User Interface (UI) to detect websites based on URLs. While addressing Machine Learning Algorithm is being approached followed by feature classification moreover based on that websites would be distinguished as phishing or authorized websites. Firstly, user will go with copying and pasting the URL in the provided UI and then if the link is safe or legitimate the user will be addressed directly to that particular website. If the link fails to be safe the UI will pop up the message.

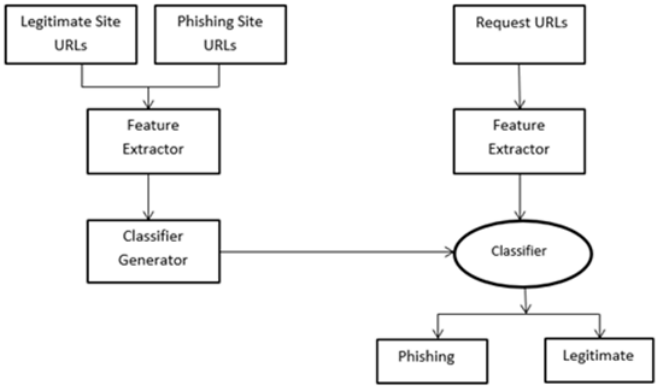

Fig. 1. Block Diagram

This is the block diagram where Feature Extraction is performed on the URL and by classifier URL is being differentiated.

Here comes the Modules, first module is Feature Extraction where we undergo the 9 features.

1. Length of URL
2. URL has HTTP
3. URL has Suspicious Char
4. Prefix or Suffix
5. Number of dots
6. Number of slashes
7. URL has Phishing terms
8. Length of Subdomain
9. URL contains IP Address

The next Module is Algorithm, here Random Forest Algorithm is applied which

1. unexcelled inaccuracy
2. runs efficiently
3. accepts thousands of input variables
4. Maintains accuracy
5. saved for future use
6. helps in balancing error

Further in Algorithm TP, FP, FN, TN values are calculated:

1. **True Positive (TP)** : Values that are *positive* and predicted *positive*.
2. **False Positive (FP):** Values that are *negative* but predicted to *positive*.
3. **False Negative (FN):** Values that are *positive* but predicted to *negative.*
4. **True Negative (TN):** Values that are *negative* and predicted to *negative.*

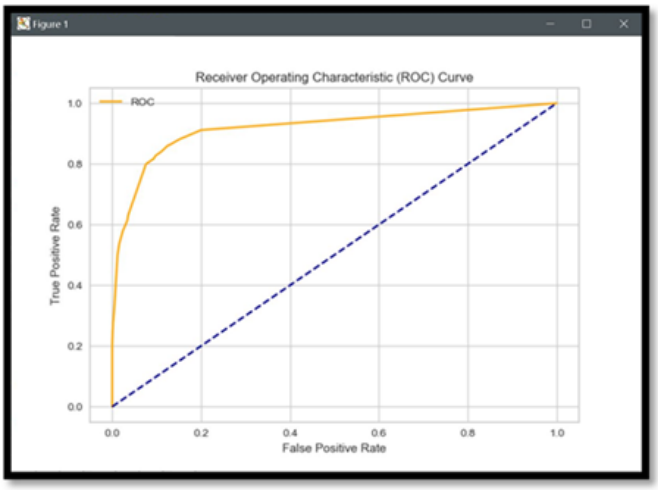

Fig.2. True Positive & False Positive Rate w.r.t. ROC

Calculations:

Accuracy: (TP+TN)/(TP+FP+TN+FN) (85%)

Precision: TP/(TP+FP)

Precision: TP/(TP+FN)

F-measure: Precision Recall/(Precision + Recall)

Input:

-D: the training data set,

-A: the feature space {A1, A2,...,AM},

-Y: the feature space {y1, y2,...,yq},

-K: the number of trees,
-m: the size of subspaces.
Output:
- A random forest μ value.

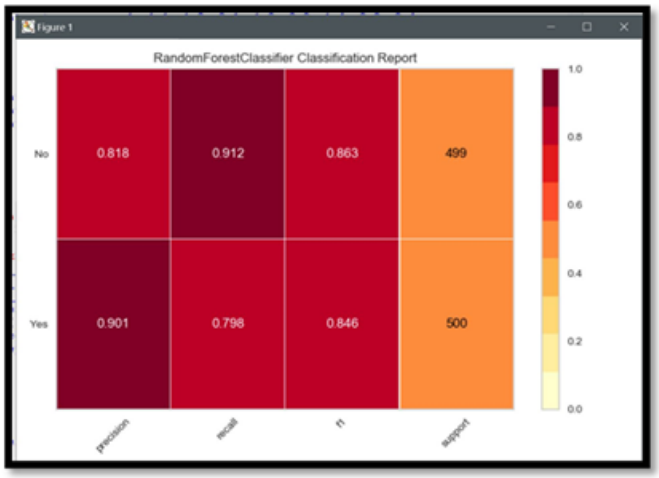

Fig.3. Random Forest Classifier

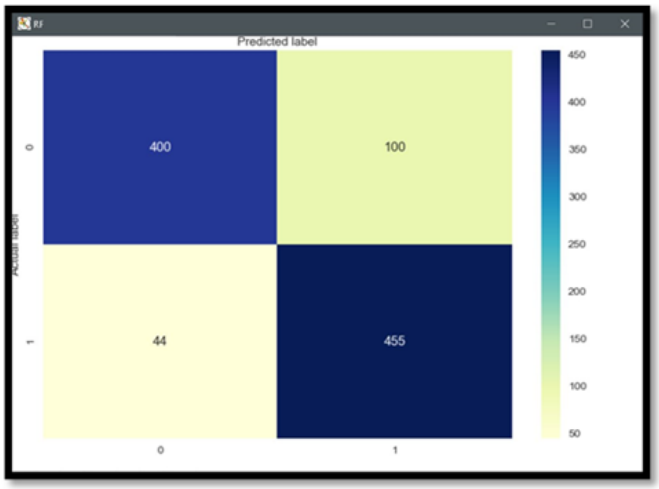

Fig. 4. Random Forest Graph
(Confusion Matrix)

The third module is User Interface where the user will paste the URL.
Input: URL
Output: Phishing site or not

V. EXPECTED OUTCOMES

In the proposed idea, the application will be effectively trained using the Machine Learning technology under supervised learning to predict whether the given URL is phishing or legitimate. It follows a feature classification approach with the GUI. Based on the given features and the dataset of Phishing as well as legitimate URLs the application will successfully give the output as a phishing website for the malicious mail and will block the particular website. And for the legitimate URLs, it will divert the user to that particular website. This approach follows the Random Forest algorithm where the accuracy is 86% for the proposed system.

VI. BENEFIT TO SOCIETY

Nowadays due to the increase in cyberattacks, and types of attacks such as phishing of websites, SQL injection, etc. The information of people is at stake and in danger. So, there is a need to protect the user's vital information, such as his bank details, card details. And due to these attacks, a user might get phished and he may lose his credentials. So, to save from these attacks there's a need to build a system that would save the user and his data. Hence this proposed idea will restrict the hacker from hacking. The user will be made aware of Phishing and legitimate mails by which he can save himself. Also, this system can be used in industries, businesses, schools, colleges. The user might receive a malicious URL from anywhere, his Emails, Websites, SMS, etc. So, this is where the user can check for the URL where he can determine the legitimate and phishing URLs.

VII. FUTURE SCOPE

Furthermore, the proposed idea can be improvised by detecting the clickable pictures, malicious QR code, etc. The limitation is that all features are discrete. The other limitation is that the URL is to be copied and we have to search in the application then it will predict whether it is legitimate or not rather than redirecting the URL link to the application.If the URL

is not there in the training and testing data set then it is difficult to predict that the URL is legitimate or not.

VIII. REFERENCES

[1]. Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

[2]. Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2010. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q2_2 014.pdf

[3]. Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_report_q2_2010.pdf

[4]. Huang, Huajun, Junshan Tan, and Lingxi Liu. "Countermeasure techniques for deceptive phishing attack." New Trends in Information and Service Science, 2009. NISS'09. International Conference on. IEEE, 2009.

[5]. Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[6]. Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[7]. Wikipedia. (2015. March) Uniform Resource Loactor. Avaliable: http://en.wikipedia.org/wiki/Uniform_resource_locator

[8]. Kausar, Firdous, et al. "Hybrid Client Side Phishing Websites Detection Approach." International Journal of Advanced Computer Science and Applications (IJACSA) 5.7 (2014).

[9]. Sunil, A. Naga Venkata, and Anjali Sardana. "A pagerank based detection technique for phishing web sites." Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.

[10]. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." Information Security, IET 8.3 (2014): 153-160.

[11]. Singh, C., & `Meenu., "Phishing Website Detection Based on Machine Learning: A Survey", IEEE 6th International Conference on Advanced Computing & Communication Systems, Gorakhpur, India, 2020, 978- 1-7281-5197-7.

[12]. Aydin, M., Butun, I., Bicakci, K., & Baykal, N., "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs",IEEE 10th Annual Computing and Communication Workshop and Conference, Ankara, Turkey,Goteborg, Sweden, Guzelyurt, Cyprus, 30-May- 2020, 978-1-7281-3783-4.

[13]. A, A. A., & K, P. “Towards the Detection of Phishing Attacks”, IEEE 4th International Conference on Trends in Electronics and Informatics, Coimbatore,India, July 27-2020, 978-1-7281-5518-0

[14]. Arun Kulkarni & Leonard L. Brown, ” Phishing Websites Detection using Machine Learning ”, International Journal of Advanced Computer Science and Applications , Tyler, TX, 2019.

[15]. El Aassal, A., Baki, S., Das, A., & Verma, R. M.,"An In-Depth Benchmarking and Evaluationof Phishing Detection Research for Security Needs", IEEE Access,Houston, U.S., 5-Feb-2020, 2969780.

[16]. Korkmaz, M., Sahingoz, O. K., & Diri, B., "Detection of Phishing Websites by Using Machine Learning- Based URL Analysis",IEEE IIT - Kharagpur, Istanbul, Turkey, 1-Jul-2020, 49239.

[17]. Mohammed Hazim Alkawaz, Stephanie Joanne Steven and Asif Iqbal Hajamydeen,"Detecting Phishing Website Using Machine Learning",IEEE International Colloquium on Signal Processing & its Applications, Selangor, Malaysia, 28-Feb-2020, 978-1-7281-5310-0.

[18]. Mohammed Zakariah, "Classification of large datasets using Random Forest Algorithm in various applications: Survey", International Journal of Engineering and Innovative Technology, Riyadh, Kingdom of Saudi Arabia, September 2014, 2277-3754.

[19]. Mehmet Korkmaz, Ozgur Koray Sahingoz & Banu Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis", from IEEE Xplore, Istanbul/Turkey, July 2020, 49239.

[20]. Bhagyashree E. Sananse & Tanuja K. Sarode, "Phishing URL Detection: A Machine Learning and Web Mining- based Approach", International Journal of Computer Applications, Mumbai, August 2015, 0975 – 8887.

[21]. Rishikesh Mahajan & Irfan Siddavatam, "Phishing Website Detection using Machine Learning Algorithms", International Journal of Computer Applications, Mumbai, October 2018,328541785.

[22]. Jin-Lee Lee, Dong-Hyun Kim & Chang-Hoon, Lee, "Heuristic-based Approach for Phishing Site Detection Using URL Features", Conf. on Advances in Computing, Electronics and Electrical Technology, USA, 2015, 978-1-63248-056-9-84.

Cite this article as :

Salvi Siddhi Ravindra, Shah Juhi Sanjay, Shaikh Nausheenbanu Ahmed Gulzar, Khodke Pallavi, "Phishing Website Detection Based on URL", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 589-594, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT2173124
Journal URL : https://ijsrcseit.com/CSEIT2173124