INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

PHISHING DETECTION SYSTEM THROUGH HYBRID MODEL IN MACHINE LEARNING BASED ON URL

[1]B. Vasantha, [2]K. Poojitha, [3]L. Kruthika, 4M. Rohini, [5]Y. Rachana

[1]Assistant Professor, Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad, INDIA.

[2345]UG,Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad , INDIA.

Abstract:

Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Although phishing was first used in 1996, it has become the most severe and dangerous cybercrime on the internet. Phishing utilizes email distortion as its underlying mechanism for tricky correspondences, followed by mock sites, to obtain the required data from people in question. Different studies have presented their work on the precaution, identification, and knowledge of phishing attacks; however, there is currently no complete and proper solution for frustrating them. Therefore, machine learning plays a vital role in defending against cybercrimes involving phishing attacks. The proposed study is based on the phishing URL-based dataset extracted from the famous dataset repository, which consists of phishing and legitimate URL attributes collected from 11000+ website datasets in vector form. After preprocessing, many machine learning algorithms have been applied and designed to prevent phishing URLs and provide protection to the user. This study uses machine learning models such as decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and proposed hybrid LSD model, which is a combination of logistic regression, support vector machine, and decision tree (LR+SVC+DT) with soft and hard voting, to defend against phishing attacks with high accuracy and efficiency. The canopy feature selection technique with cross fold valoidation and Grid Search Hyperparameter Optimization techniques are used with proposed LSD model. Furthermore, to evaluate the proposed approach, different evaluation parameters were adopted, such as the precision, accuracy,

ISSN: 2057-5688

recall, F1-score, and specificity, to illustrate the effects and efficiency of the models.

1 INTRODUCTION

Phishing imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organisation. This makes the user to forcefully visit the phishing website through the links given in the phishing email. These phishing websites are made to mock the appearance of an original organisation website. The phishers force user to fill up the personal information by giving alarming messages or validate account messages etc so that they fill up the required information which can be used by them to misuse it. They make the situation such that the user is not left with any other option but to visit their spoofed website. [8] Phishing is a cyber crime, the reason behind the phishers doing this crime is that it is very easy to do this, it does not cost anything and it effective. The phishing can easily access the email id of any person it is very easy to find the email id now a day and you can sending an email to anyone is freely available across the world. These attakers put very less cost and effort to get valuable data quickly and easily. The phishing frauds leads to malware infections, loss of data, identity theft etc. The data in which these cyber criminals are interested is the crucial information of a user like the password, OTP, credit/ debit card numbers CVV, sensitive data related to business, medical data, confidential data etc.

Sometimes these criminals also gather information which can give them direct access to the social media account their emails. [3] A lot of software / approaches and algorithms are used for phishing detection. These are used at academic and commercial organisation levels. A phishing URL and the parallel page have many features which are different from the malignant URL. Let us take an example to hide the original domain name the phishing attacker can select very long and confusing name of the domain. This is very easily visible. Sometimes they use the IP address instead of using the domain name. On the other hand they can also use a shorter domain name which will not be

ISSN: 2057-5688

relevant to the original legitimate website. Apart from the URL based feature of phishing detection there are many different features which can also be used for the detection of Phishing websites namely the Domain-Based Features, Page-Based Features and Content-Based Features. [16]

2 METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict phishing. We shall also explain our proposed methodology to detect phishing website. In section A we shall explain various classifiers and methods which can be used to check the phising and legtiminate website. In section B we shall explain our proposed system.

Detecting and identifying Phishing Websites is really a complex and dynamic problem. Machine learning has been widely used in many areas to create automated solutions.The phishing attacks can be carried out in many ways such as email, website, malware,sms and voice.In this work, we concentrate on detecting website phishing (URL), which is achieved by making use of the Hybrid Algorithm Approach. Hybrid Algorithm Approach is a mixture of different classifiers working together which gives good prediction rate and improves the accuracy of the system. Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification.

The dataset of phishing and legitimate URL's is given to the system which is then pre-processed so that the data is in the useable format for analysis. The features have around 30 characteristics of phishing websites which is used to differentiate it from legitimate ones. Each category has its own characteristics of phishing attributes and values are defined. The specified characteristics are extracted for each URL and valid ranges of inputs are identified. These values are then assigned to each phishing website risk. For each input the values range from 0 to 10 , while for output range is from 0 to 100. The phishing attributes values are represented with binary no 0

ISSN: 2057-5688

and 1 which indicates the attribute is present or not. After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms are already explained in previous section. After this we use a hybrid classification in which we combine two of the classifier namely Naive Bayes and Random forest to predict the accuracy of the detection of the phishing URL, hence we get our desired result. This is also called a hybrid approach to test the data, in this method we propose to use the combination of two classifiers, as mentioned above. We shall then test the data and evaluate the prediction accuracy which shall be more than the existing system. We shall now see the different classifiers and discuss the hybrid combination used for our proposed system.

In the training phase, we should use the labeled data in which there are samples such as phish area and legitimate area. If we do this then classification will not be a problem for detecting the phishing domain. To do a working detection model it is very crucial to use data set in the training phase. We should use samples whose classes are known to us, which means the samples whom we label as phishing should be detected only as phishing. Similarly the samples which are labeled as legitimate will be detected as legitimate URL. The dataset to be used for machine learning must actually consist these features.There so many machine learning algorithms and each algorithm has its own working mechanism which we have already seen in the previous chapter. The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts its accuracy. Each of the algorithms which explain in the earlier section has some disadvantages hence it is not recommended to use one machine learning algorithm to detect the phishing website [10]

3 IMPLEMNTATION

System design is used for understanding the construction of system. We have explained the flow of our system and the software used in the system in this section. To get structured data we do feature generation of the data at the preprocessing stage. We have used

ISSN: 2057-5688

techniques like XG Boost, Naive Bayes,SVM, Meta classifiers and stacking classifier to detect the phishing and legitimate websites.

Data set: The data of urls is obtained from Phishtank website,where Phishtank is an anti-phishing site.It contains 2905 urls which is in unstructured form. Our main objective is to detect whether the url is phishing or legitimate based on the features extracted.In Preprocessing we have done feature extraction where The URLs are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features.The features have assigned binary values 0 and 1 which indicates that feature is present or not as shown in figure below. The extracted feature values are stored as input and passed to the classifiers. A structured dataset is given to the classifiers. We use four methods classification namely: XG Boost, SVM, Naive Bayes and stacking classifier for detection of url as phishing or legitimate. Now the classifier will find whether a requested site is a phishing site. When there is a page request , the URL of the requested site is radiated to the feature extractor. It extracts the feature values through the predefined URL-based features. These feature values are act as a input for the classifier. After this we will come to know if the site is phishing or not

Table 1: URL features

Sr. No	Feature name	Description
1	IP address	Whether domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has _@‘, _//‘
4	Prefix and suffix	Whether URL has _-‘
5	Length of subdomain	Length of subdomain
6	Number of ‘/’	Number of ‘/’ in URL
7	HTTPS protocol	Whether URL use https.
8	Phishing words in URL	Whether url has phishing terms
9	Number of ‘.’	Number of dots ‘.’ in url

In this section we shall discuss about the actual steps which were implemented while doing the m experiment. We shall explain the stepwise procedure used to analyse the data and to predict the phising . The system consists of the following main steps, We have used unstructured data which consists only urls.There are 2905 urls obtained from Phishtank website which consists of both phishing and legitimate url where most of urls obtained are phishing.

1. We have collected unstructured data of urls from Phishtank website. 2. In

ISSN: 2057-5688

preprocessing ,feature generation is done where nine features are generated from unstructured data. These features are length of url,url has http,url has suspicious character,prefix/suffix,number. of dots,number of slash,url has phishing term,length of subdomain,url contains ip address.

2. After this a structured dataset is created in which each feature contains binary value(0,1) which is then passed to the different classifiers.

3. Next we train the four different classifiers and compare their performance on the basis of accuracy four classifiers used are XG Boost,SVM,Naive Bayes and Stacking,where stacking uses XG Boost and SVM as its base classifier and Random Forest as its meta classifier. 5. Then classifier detects the given url based on the training data that is if the site is phishing it shows a pop-up and if legitimate it opens that page in browser.

6. We compare the accuracy of different classifiers and found XG Boost and Stacking are the best classifiers which gives the maximum accuracy.

7. Below are the screen shots for the implementation process.

We have got the desired results of testing the site is phishing or not by using four different classifiers. Refer the graph below for the exact results. Refer the graphs in Fig.15 and Fig.16 for the results. In the graph, shown in Fig. 15 shows the AUC, precision, recall and the F1 score obtained by using different classifiers. The graph shown in Fig 16. explains about the accuracy obtained by using different classifiers in the histogram graphical representation

4 CONCLUSION

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside, but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we

ISSN: 2057-5688

proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning algorithms.We have found that our system provides us with 85.5 % of accuracy for XG Boost Classifier, 86.3% accuracy for SVM Classifier, 80.2 % accuracy for Naïve Bayes Classifier and finally 85.6 percentage of accuracy when using Stacking Classifier.Hence we found that the best among all the above classifiers is SVM and Stacking Classifier which shows maximum accuracy. The proposed technique is much more secured as it detects new and previous phishing sites.

5 REFERENCES

[1] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th
International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). pringer.

[2] Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. Computers & Security, 83, 246-267.

[3] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. computers & security, 84, 256-275.

[4] Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016, June). Know your phish: Novel techniques for detecting phishing sites and their targets. In 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) (pp. 323-333). IEEE.

[5] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.

[6] Rathod, J., & Nandy, D. Anti-Phishing Technique to Detect URL Obfuscation.

[7] Hodžić, A., Kevrić, J., & Karadag, A. (2016). Comparison of machine learning techniques in phishing website classification. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256).

ISSN: 2057-5688

[8] Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.

[9] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.

[10] Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798-805.

[11] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[12] Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications, 42(3), 1166-1177.

[13] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Phishing page detection via learning classifiers from page layout feature. EURASIP Journal on Wireless Communications and Networking, 2019(1),43.

[14] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012, December). An assessment of features related to phishing websites using an automated technique. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 492-497). IEEE.

[15] https://www.researchgate.net/publication/226420039-Detection-ofPhishing-Attacks-A-Machine-Learning-Approach

[16] https://www.proofpoint.com/us/threat-reference/phishing

[17] https://towardsdatascience.com/phishing-domain-detection-with-ml5be9c99293e5

[18] https://en.wikipedia.org/wiki/Phishing

[19] https://www.techrepublic.com/article/how-to-tackle-phishing-withmachine-learning/

[20] https://www.irjet.net/archives/V5/i3/IRJET-V5I3580.pdf

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688