iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35671

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Phishing Website Detection using Machine Learning

Prof Rajendra Kankrale[1], Ganesh Lengre[2], Mahendra Jadhav[3], Neeraj Durge[4]
[1, 2, 3, 4]*Dept. of Information Technology) Sanjivani College of Engineering, Kopargaon, India*

Abstract: ***Phishing is not a legal activity, in this people are misguided into the wrong websites using various types of fraudulent methods. The main objective of this phishing websites is to seize all personal information or financial details for own personal benefits or misuse. As the technology facilitates, the phishing approaches used needs to get progressed and therefore there is urgent need for better security and better mechanisms to prevent as well as detect this phishing approaches. The primary focus of this paper to put forward a model as a solution to detect phishing websites by using the URL detection using Random Forest Algorithm. There are basically three major phases such.***
Index Terms: Component, formatting, style, styling, insert

I. INTRODUCTION

The Internet is widely used among people and it has become an inseparable part of our life. Therefore, huge amounts of data are exchanged. Those users could be more or less experienced using the web. But, nevertheless, nobody is safe from the huge threat that is available there outside. Those threats are phishing websites that are hard to differentiate from the original ones. These websites are used to collect personal and confidential user data that usually should be protected. Later, information is misused and people are experiencing consequences. Some of the consequences could be identity loss or financial debts. More than 60,000 phishing websites were reported in March 2020. 96 percentage of all targeted attacks are intended for intelligence-gathering. 1 in every 3 companies that suffered a ransomware attack paid the ransom and the average ransom demand is nearly dollar 84,000. 22 percentage of all data breaches in 2020 involved phishing attacks. Almost one third of all data breaches in 2017 were due to phishing attacks. Approximately 55 percentage of phishing websites in 2019 used SSL certificates. Research also shows that 33 percentage of people closed their business after a phishing attack.

The problem with phishing attacks is not only that they are increasing, but also, they are improving and becoming more sophisticated. Due to that, it is necessary to develop systems that will help in detection of these phishing websites to prevent negative outcomes. Therefore, in this work we want to develop an intelligent system that will be used to detect phishing websites. We are going to use machine learning algorithms for classification such as logistic regression, Support Vector Machine(SVM) and Random Forest (RF).

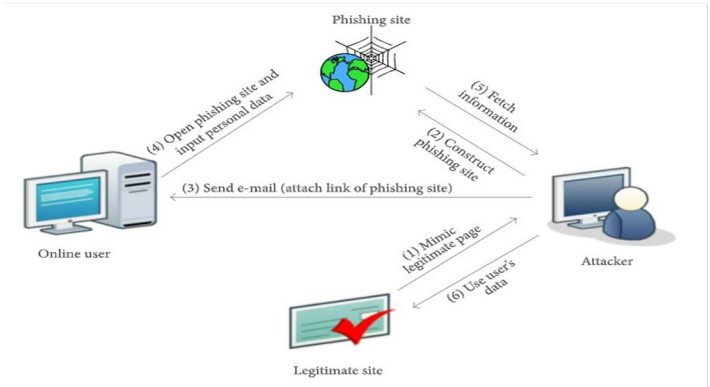


Fig. 1. Phishing Mechanism.

The rest of the work is organized as follows: in section two, we are giving overview of several works related to the phishing websites detection. Section three and four provide details about the database and methodologies, respectively. Results are presented in section five. We conclude our work with section six.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

II. LITERATURE SURVEY

Web phishing is a serious security threat that is presentin the Internet. Sensitive financial and personal informationare taken from the users thanks to the phishing websites.

These websites look like legitimate websites and they areused to gather private data. Therefore, phishing attacks use weaknesses of the user and it is hard to reduce those, but itis important to work on detection techniques improvement. In this section we present several works related to detection of phishing websites. Ankit Kumar Jain and B. B. Gupta et al. [1] proposed a novel classification approach that use heuristic based feature extraction approach. In this, they have classifiedextracted features into three categories such as URL Obfus- cation features, Third-Party-based features, Hyperlink-based features. Moreover, proposed technique gives 99.55 percentage accuracy. Drawback of this is that as this model uses third party features, classification of website dependent on speedof third-party services. Also this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has a limitation of more execution time for the websites with more number of links. A. Mishra and B.

B. Gupta et al. [2] proposed a novel anti phishing approach that extracts features from client-side only. Proposed approachis fast and reliable as it is not dependent on third party butit extracts features only from URL and source code. In this paper, they have achieved 99.09 percentage of overall detection accuracy for phishing website. This paper have concluded that this approach has limitation as it can detect webpage written inHTML .Non-HTML webpage cannot detect by this approach. Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang et.al [3]The algorithms we applied in our paper are logistic regression,(SVM)Support Vector Machine and Random Forest (RF) for obtaining phishing website datasets. Eric Medvet, Engin Kirda and Christopher Kruegel . Authors [4] presents anapproach to detect phishing email attacks using machine learn-ing. This is used to perform the semantic analysis of the textto detect malicious intent. A support vector machine (SVM) isa supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. In light of the job of each word in the sentence, this strategy recognizes whether the sentence is an inquiry or an order. Matthew Dunlop, Stephen Groat, and David Shelly [5] Supervised machine learning is used to generate the blacklist of malicious pairs. Authors defined algorithm for detecting phishing emails and Netcraft Anti-Phishing Toolbar is used to verify the validity of a URL. This algorithm is implemented with Python scripts and dataset Nazario phishing email set is used. Results of Netcraft and SEAHound are compared and obtained precision 98 and 95 percentage respectively

III. URLS AND ATTACKER'S TECHNIQUES

Attackers use different types of techniques for not to be detected either by security mechanisms or system admins. In this section, some of these techniques will be detailed. To understand the approach of attackers, firstly, the components of URLs should be known. The basic structure of a URL is depicted in Fig. 3. In the standard form, a URL starts withits protocol name used to access the web page. After that, the subdomain and the Second Level Domain (SLD) name, which commonly refers to the organization name in the server hosting, is located and finally the Top-Level Domain (TLD) name, which shows the domains in the DNS root zone of the Internet takes place. The previous parts compose the domain name (host name) of the web page; however, the inner addressis represented by the path of the page in the server and withthe name of the page in the HTML form. Although SLDname generally shows the type of activity or company name, an attacker can easily find or buy it for phishing. The nameof SLD can only be set once, at the beginning. However, an unlimited number of URLs can be generated by an attacker with extending the SLD by path and file names, because the inner address design directly depends on attackers. The unique (and critical) part of a URL is the composition ofSLD and TLD, which is named as domain name. Therefore, cybersecurity companies make a great effort to identify the fraudulent domains by name, which are used for phishingattacks. If a domain name is identified as phishing, the IP address can be easily blocked to prevent from accessing the web pages located in it.

To increase the performance of the attack and steal more sensitive information, an attacker mainly uses some important methods to increase the vulnerability of victims such as the useof random characters, combined word usage, cybersquatting, typosquatting, etc. Therefore, detection mechanisms shouldtake into consideration of these attack methods.

Fig. 2. URL components.

IV. BACKGROUND THEORY

Phishing attack can be implemented in various form like Email phishing, Website phishing, spear phishing, Whaling, Tab napping, Evil twin phishing etc.To avoid this phishing attack various anti-phishing solutions should be use. There are various anti phishing solutions such as Blacklist, heuristic,visual similarity, machine learning etc.

A. Blacklist Method

This is most commonly used approach in which list ofphishing URL is stored in database and then if URL is foundin database, it is known as phishing URL and gives warning otherwise it is called legitimate. This approach is easy and faster to implement as it see URL is in db or not. But limitations is small change in URL is sufficient to bypass the list based technique and Frequent update of list is necessaryto counter new attack.

B. Heuristic Based Method

This 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 easiesto bypass once attacker know algorithm or features used. In addition, this has poor detection because site may or may not have common features.

C. Machine Learning

This approach works efficiently in large dataset. This also removes drawback of existing approach and able to detect zeroday attack .Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99size of train-ing data, feature set, and type of classifier. Limitation of thisis it fails to detect when attacker use compromised domainfor hosting their site.Many of research have been performedin this area of phishing detection. Most research has workedon improving accuracy of phishing website detection using different classifiers. Various Classifiers used are KNN, SVM, Decision tree, ANN, Na¨ıve Bayes, PART, ELM and Randomforest. Among all of this tree based classifiers and RF isbest as increase dataset as per my literature survey. Therefore, proposed approach will be on phishing website detection usingtreebased classifiers.

V. ALGORITHMS

A. Support Vector Machine

Support Vector Machine (SVM) is a relatively simple Su- pervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data.y. SVM is a classification technique based on the statis-tical learning, which successfully utilized in many applications of nonlinear classification and large datasets and issues. SVM classifiers employ the hyperplane to isolate categories. Every hyper-plane is determined by its direction (w), the precise position in space or a threshold is (b), (xi) denotes the input array of constituent N and indicates the category. A set of the training case is shown in fig 3 (x1,y1),(x2,y2),...,(xk,yk);xi ∈R K represents the number of training dataset and d denotes the number of the dimensions of the Logistic regressionlogistic regressioninput dataset. The function of decision is specifiedas follows: f(x,w,b) = sign((w,xi)+b),w ∈ R,b. ∈R. One of the advantages of employing the SVM for training the systemis its ability to work with multi-dimensional data. SVM is a classifier that takes a given labeled training data as input and outputs an optimal hyperplane which classifies new examples.SVM makes a hyperplane between data sets by maximizing the margin as shown in Fig. 3

B. Logistic regression

It is a form of regression analysis. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and commu- nicate between variables. Logistic Regression The logistic regression technique includes dependent variable which canbe signified in the binary (0 or 1, true or false, yes or no) values, means that the result could only be in either one form of two. For example, it can be applied when we need to find the probability of positive or fail event. Here, the same method is used with the additional sigmoid function, and the value of Y ranges from 0 to 1.Consider a model with two predictors, x1 and x2; these may be constant variables or indicator functions for binary variables (taking value 0 or 1). Fig 4 represents the comparison method. 1/(1+e-x).Linear and Logistic regression are the furthermost basic form of regression which are usually used. The crucial difference between these two is that Logisticregression is used when the dependent variable is binary in nature. In difference, Linear regression is used when the dependent variable is continuous and nature of the regression line is linear. Regression is a method is used to predict the value of a response (dependent) variables, from one or more predictor variables, where the variable is numeric.

There are several forms of regression such as linear, multiple, logistic, polynomial, non-parametric, etc.

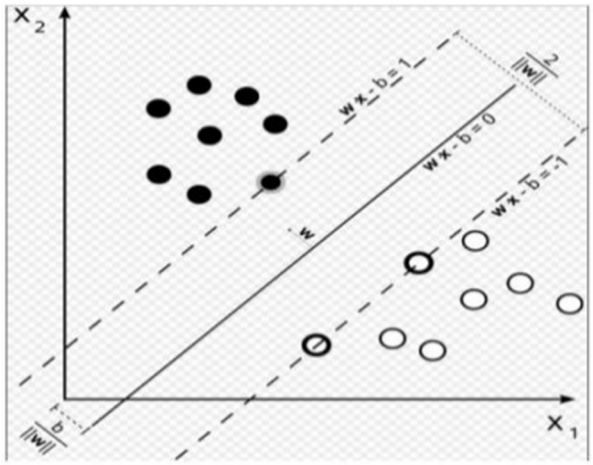


Fig. 3. SVM for phishing websites classification.

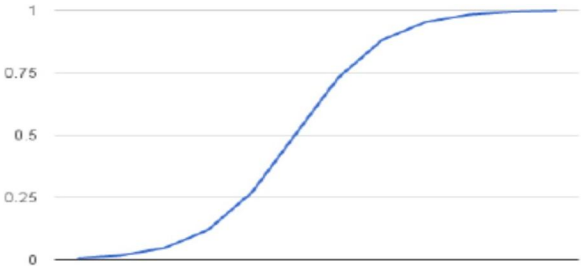


Fig. 4. Logistic Function.

C. Random Forest

One of the best way to classify the features in datasetsextracted from the website which we are mining is usingRandom Forest classifier. Random Forest is an ensemble ma- chine learning algorithm. It reduces variance and improves performances. It trains each of the decision tree by slightly taking different set of observations. In this we are splitting the nodes of each tree considering a limited number of the features. For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes. nij= wjcj-wleft(j)Cleft-Wright(j)C nij= the importance of node j wj = weighted number of samples reaching node j Cj= the impurity value of node j left(j) = child node from left split on node j right(j) = child node from right split on node j

VI. SYSTEM ARCHITECTURE

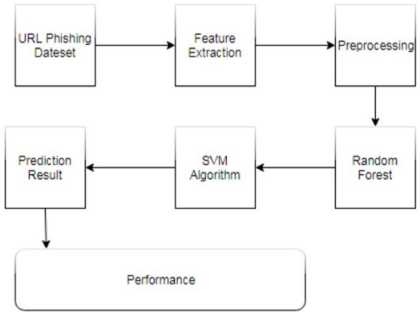


Fig. 5. System Architecture.

As shown in fig 5 System is divided into seven parts i.e: 1 Url phishing dataset 2 Feature extraction 3 Preprocessing 4 Random forest 5 SVM algorithm 6 Prediction Result 7Prefromance.

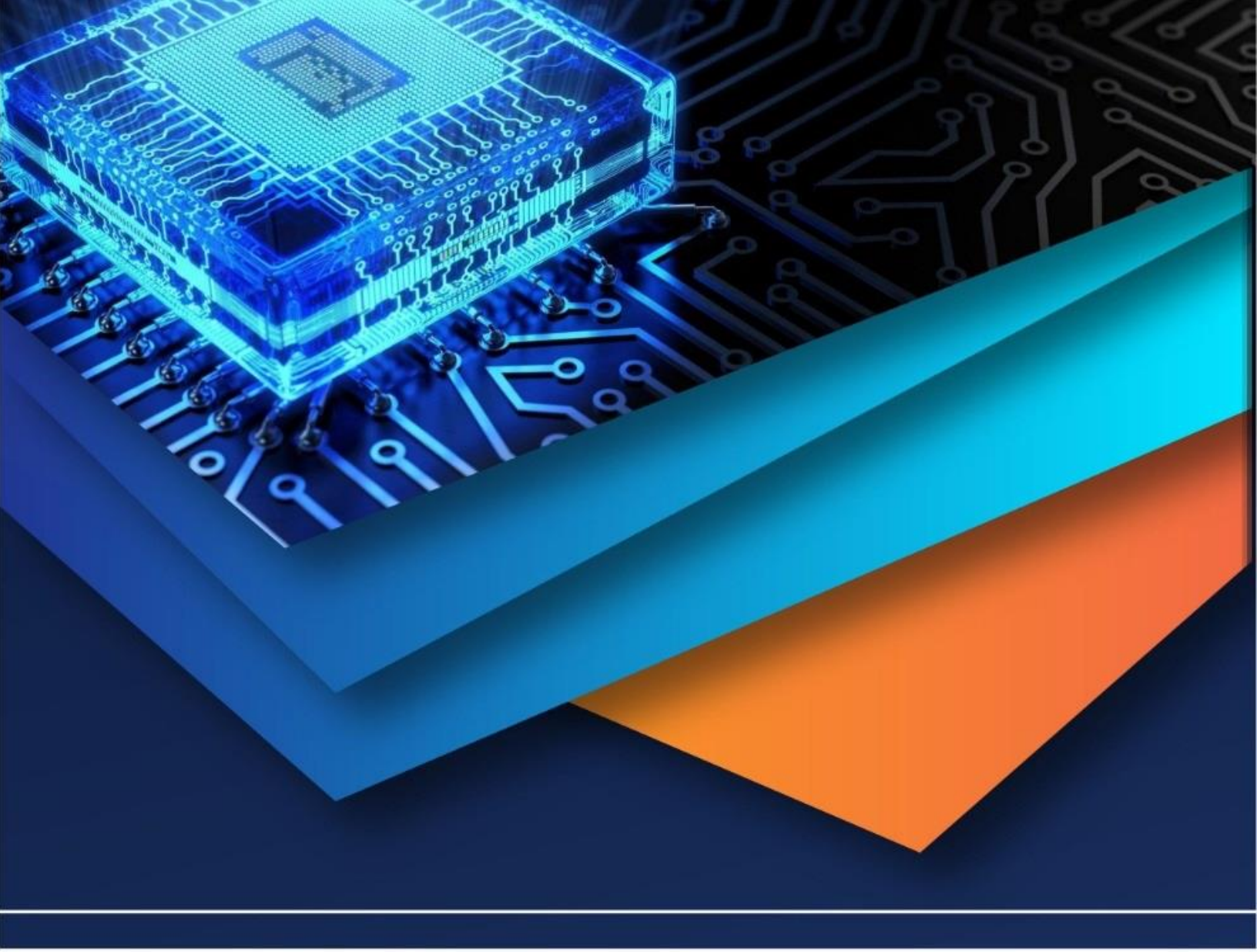
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

VII. CONCLUSION

Phishing is an appalling threat in the web security domain. In this attack, the user inputs his/her personal information toa fake website which looks like a legitimate one. We have presented a survey on phishing detection approaches based on visual similarity. This survey provides a better understanding of phishing website, various solution, and future scope inphishing detection.Phishing is a way to obtain user's private information via email or website. As usage of internet is very vast, almost all things are available online now it is either aboutshopping cloths, electronic gadgets, crockery or to payment ofmobile, TV electricity bill. Rather than standing out in line for hours, people are being aware of using online method.Due to this phisher has wide scope to implement phishing scam. As there is lot of research work done in this area, thereis not any single technique, which is enough to detect alltypes of phishing attack. As technology increases, phishing attackers using new methods day by day. This enables us to find effective classifier to detection of phishing.

REFERENCES

[1] Ankit Kumar Jain and B. B. Gupta, "Phishing DetectionAnalysis of Visual Similarity Based Approaches", Hindawi 2017
[2] A. Mishra and B. B. Gupta, "Hybrid Solution to Detect and Filter Zero-day Phishing Attacks", ERCICA 2014.
[3] Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, "Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features", INCS 2013.
[4] Eric Medvet, Engin Kirda and Christopher Kruegel,"VisualSimilarity-Based Phishing Detection", ACM 2015.
[5] Matthew Dunlop, Stephen Groat, and David Shelly,"GoldPhish Using Images for Content-Based Phishing anal- ysis", IEEE 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY