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.2021.35400

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

Feature based Phishing Website Detection using Random Forest Classifier

Sonali Kadam[1], Shristy Kumari[2], Subhu Trivedi[3], Vanshika Shah[4]
[1, 2, 3, 4]*Computer Engineering Department, Bharati Vidyapeeth's College of Engineering for Women,Pune, Maharashtra, Savitribai Phule Pune University, Pune*

***Abstract:** In today's world, one of the most vulnerable security threat which poses a problem to the internet users is phishing. Phishing is an attack made to steal the sensitive information of the users such as password, PIN, card details etc., In a phishing attack, the attacker creates a fake website to make the users click it and steal the sensitive information of users. . In this paper, we propose a feature-based phishing detection technique that uses uniform resource locator (URL) features. This paper focuses on the extracting the features which are then classified based on their effect within a website. The feature groups include address- bar related features, abnormal- based features, HTML – JavaScript based features and domain based features. We plan to use machine learning and implement some classification algorithms and compare the performance of these algorithms on our dataset.*

***Keywords:** Phishing Websites, URL-based features, Machine Learning.*

I. INTRODUCTION

Computer and Internet are now playing a major role in every aspect of human lives which mainly includes data, transactions, information storage and its retrieval. As we all are living in the 20th century, everything is being performed online like online shopping, bill payments etc. through various online platforms like websites, android applications etc. which has led us to move towards the digital era. Digitalization is basically the use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business. Thus, increase in the digitalization has led to the increase of fraudulent activities in various sectors. There are various types of frauds which occur on online platform and one such attack is stealing user's data through phishing websites. Phishing is a malicious attack in online theft to steal the user's private information. That is a kind of scam in which unauthorized user tries to gain user private data and thus user falls into such traps. Phishing is the technique of extracting user credentials and sensitive data from users by masquerading as a genuine website. In phishing, the user is provided with a mirror website which is identical to the legitimate one but with malicious code to extract and send user credentials to phishers. Phishing attacks can lead to huge financial losses for customers of banking and financial services. The motive of our paper is to propose a structure that is safe for identifying phishing websites in less time with high accuracy. The goal of our project is to implement a machine learning solution to the problem of detecting phishing and malicious web links. The end result of our project will be a software product which uses machine learning algorithm to detect malicious URLs.

II. RELATED WORK

Phishing is a security attack which is the most common anddangerous attack to gain account details, confidential information, card details or the password of a user to conduct illegal transaction.

Vaibhav Patil, Pritesh Thakkar and Chirag Shah et al. [1] proposes a combined solution that uses three algorithms – whitelist and blacklist, heuristics and visual similarity. These approaches provide three-level security blocks and hence, this system is more effective and accurate.

Anu Vazhayil, Vinaya Kumar R and Soman KP et al. [2] focuses on a combination of CNN with the Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to derive the accuracy in classifying the phishing URLs. LSTM extracts sequential information and CNN helps to extract special information among the characters. CNN used to learn the special co-relationship among the characters.

Aburrous et al. [5] proposed a smart structure for phishing webpage finding. They anticipated a model that depends on fuzzy logic united with data mining approach to study the techniques by telling the illegal websites aspects by classifying the phishing types.

Arade et al. [3] implemented an innovative kind of intelligent aspects depending on string matching to evaluate the addresses in the database of the implemented system and webpage address. The problem in this study is with the chanceof taking place false positive incidence i.e., legal webpages can be assumed as legal webpages.

©IJRASET: All Rights are Reserved

A model is proposed for detecting phishing webpages which is implemented by Shahriar & Zulkernine [8] with the dependability of suspected pages. A finite state machine is developed to evaluate webpage performance by tracing webpage GUI the submission with the resultant reaction in thestudy done by them.

The research conducted by Alqahtani [11] produced a novelmethod for detecting phishing websites. This method was referred to as Phishing Websites Classification using Association Classification (PWCAC). PWCAC was novel as it makes use of the association rule induction technique to categorize whether a website is genuine or a phishing website.

A hybrid phishing detection method was developed by Ali & Ahmed [9]. This method is a hybridization of an evolutionary algorithm and a deep neural network. The implemented evolutionary algorithm in their research work was a genetic- algorithm (GA) technique which was used to find the highly informative features from the original feature sets.

A Deep Belief Network (DBN) was also implemented to detect phishing websites by Verma et al. [12]. The DBN model extracts deep hierarchical representation from the given dataset by using Restricted Boltzmann machines (RBM) to develop its model. Finally, the model was fine-tuned by supervised gradient descent (i.e., a logistic regression classifier) in order to classify the input based on the last hidden layer output. The performance of the developed model was evaluated using the accuracy metric and it was able to achieve a 94.426% accuracy.

Srushti Patil and Sudhir Dhage et al. [10] use different methods like Anti Phishing solutions. Anti-Phishing solutions include various approaches. Heuristic Approach is used for classifying the URLs. Features are extract and they are classified by using ML methods. Various approaches are collaborated to check whether the website is illegal or legitimate.

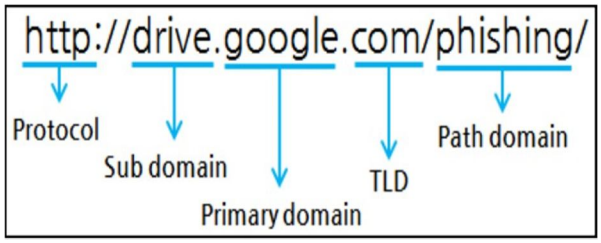


Fig. 1 URL Structure

In this study, the subdomain, primary domain, and TLD are collectively referred to as the domain. Fig. 1 depicts the individual components of a URL.

The protocol refers to a communication protocol for exchanging information between information devices; e.g., HTTP, FTP, HTTPS, etc. Protocols are of various types and are used in accordance with the desired communication method.

The subdomain is an ancillary domain given to the domain and has various types depending on the services provided by the domain page. The domain is the name given to the real Internet Protocol (IP) address through the Domain Name System (DNS). The primary domain is the most important partof a domain. The TLD is the domain in the highest position in the domain name hierarchy architecture; e.g., .com, .net, .kr, .jp, etc. [10]. We define features of each component of the URL; these features are used for phishing site detection.

III.PROPOSED APPROACH

The proposed algorithm depends on the ML process and real-time phishing detection. By using various features phishing URLs are extracted. For a machine learning classification, the extracted features are used to detect phishing websites in real-time. After analysis, the survey was done which is due to comparing various classification algorithms.

A. URL Structure

A URL is a protocol that is used to indicate the location of data on a network. The URL is composed of the protocol, subdomain, primary domain, top-level domain (TLD), and path domain.

B. URL Features

The dataset plays an important role to extract the phishing features for each URL under four categories: Addressed basedfeatures, Abnormal features, HTML, JavaScript features and Domain features. These features have 30 phishing websites characteristics which are helping to differentiate from a legitimate website.

©IJRASET: All Rights are Reserved

1) *Address-bar based Features:* The features which are related to the address of an URL are referred as address bar- related features. These includes the length of the host URL, number of dots and slashes, special characters, HTTP and SSL check, @symbol, and IP Address.
2) *Abnormal based Features:* The URL features which relates to anomalies or discrepancies between the W3C objects and Web Identity are known as abnormal based features. These features are mostly related to the source code of the web page. These features will play an important role in identifying the phishing websites. These features include Request URL (RURL), URL of an anchor (AURL), Server Form Handler (SFH).
3) *HTML and JavaScript based Features:* The features which are related to HTML tags and JavaScript functions are treated as HTML and JavaScript based features. These features include Redirect Page, Disabling Right Click and Using Pop-up window and onMouseOver.
4) *Domain based features:* The URL features related to domain name-based information are known as Domain based features. These features include Alexa Page Rank, Age of the Domain, DNS Record and Website Traffic.

Table I
Attributes and values for url features

Feature category	Attributes	Values
Features Of Address Based	having_IP_Address	{ 1,0 }
	URL_Length	{ 1,0,-1 }
	Shortining_Service	{ 0,1 }
	having_At_Symbol	{ 0,1 }
	double_slash_redirecting	{ 1,0 }
	Prefix_Suffix	{ -1,0,1 }
	having_Sub_Domain	{ -1,0,1}
	SSLfinal_State	{ -1,1,0 }
	Domain_registeration_length	{ 0,1,-1}
	Favicon	{ 0,1 }
	Non-standard port	{ 0,1 }
	HTTPS_token	{ 1,0 }
Features Which Are Abnormal	Request_URL	{ 1,-1 }
	URL_of_Anchor	{ -1,0,1 }
	Links_in_tags	{ 1,-10 }
	SFH	{ -1,1 }
	Submitting_to_email	{ 1,0 }
	Abnormal_URL	{ 1,0 }
JavaScript, HTML Features	Redirect	{ 0,1 }
	on_mouseover	{ 0,1 }
	RightClick	{ 0,1 }
	popUpWidnow	{ 0,1 }
	Iframe	{ 0,1 }
Domain Based Features	age_of_domain	{ -1,0,1 }
	DNSRecord	{ 1,0 }
	web_traffic	{ -1,0,1 }
	Page_Rank	{ -1,0,1 }
	Google_Index	{ 0,1 }
	Links_pointing_to_page	{ 1,0,-1 }
	Statistical_report	{ 1,0 }

Each type has its phishing characteristics i.e. attribute and values are defined. In this dataset, input attributes can take 3 different values which are 1, 0, and -1. Output attributes can take 2 different values which are 1, and -1.

C. Algorithms

To determine a classifier with the best performance and accuracy for using URL-based features, we employed several machine learning algorithms: support vector machine (SVM), naive Bayes, decision tree and random forest.

1) *Support Vector Machine (SVM):* Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. After that, we perform classification by finding the hyper-plane that differentiates the two classes verywell. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

©IJRASET: All Rights are Reserved

2) *Naive Bayes:* Naive Bayes is a classifier that can achieve relatively good performance on classification tasks. It is based on the elementary Bayes theorem. It is one of most successful learning algorithms for text categorization [16]. On account of the conditional model‘s feature, naive Bayes is effectively trained in supervised learning. It provides the advantage of learning essential parameters using small training samples.
3) *Decision Tree:* Decision tree is a classification method that was introduced in 1992 by Quinlan [4]. It creates a tree form for classifying samples. Each internal node of the tree corresponds to a feature, and the edges from the node separate the data based on the value of the feature [7]. Decision tree includes a decision area and leaf node. The decision area checks the condition of the samples and separates them into each leaf node or the next decision area. The decision tree is very fast and easy to implement; however, it has the risk of overfitting.
4) *Random Forest:* Random forests are the classifiers that combine many tree possibilities, where each trees are depends on the values of a random vector sampled independently. then,all trees in the forest will have same allotment. To construct a tree, we assume that n is the number of training observations and p is the number of variables (features) in a training set. Todetermine the decision node at a tree we choose k « p as the number of variables to be selected. We select a bootstrapsample from the n observations in the training set and use the rest of the observations to estimate the error of the tree in testing phase. hence, we randomly choose ‘k’ variables as a decision at certain node in the tree and calculate the best split based on the k variables in the training set. Trees are always grown and never pruned compared to other tree algorithms. Random forests can handle large number of variables in a dataset. Also, during the forest building process they generate an internal unbiased estimate of the generalization error. Additionally, they can estimate missing data closely. A major disadvantage of random forests algorithm is it does not gives precise continuous forecast.

IV. EVALUATION AND RESULTS

To conduct classifier training and evaluation through an experimental dataset, we collected the URLs of phishing and legitimate sites. The dataset consists of 11,055 entries with 6157 phishing instances and 4898 legitimate instances. Each instance consists of 30 features comprising of variousattributes typically associated with phishing or suspicious webpages. The comparison based on the performance of all the classifiers is described in table below on the phishing dataset. We have evaluated these algorithms on 2,764 test samples using various performance metrics.

Table II
Experimental results of algorithms

ALGORITHM	ACCURACY
Support Vector Machine	92.58%
Naïve Bayes Classifier	87.15%
Decision Tree Classifier	91.28%
Random Forest Classifier	96.49%

V. CONCLUSIONS

In this paper, we proposed a URL feature-based phishing detection technique. The method combines URL-based features used in previous studies with new features by analysing phishing site URLs. Additionally, we generated classifiers through several machine learning algorithms and determined that the best classifier was random forest. It showed a high accuracy of 95.89% and a low false-positive rate. The objective of this technique is to detect fake or illegal websites and notify the user in advance to prevent users from getting their private information to be misused. In future work we intend to build a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. For future enhancements, we aim to build the phishing detection system as a scalable web service which will incorporate onlinelearning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

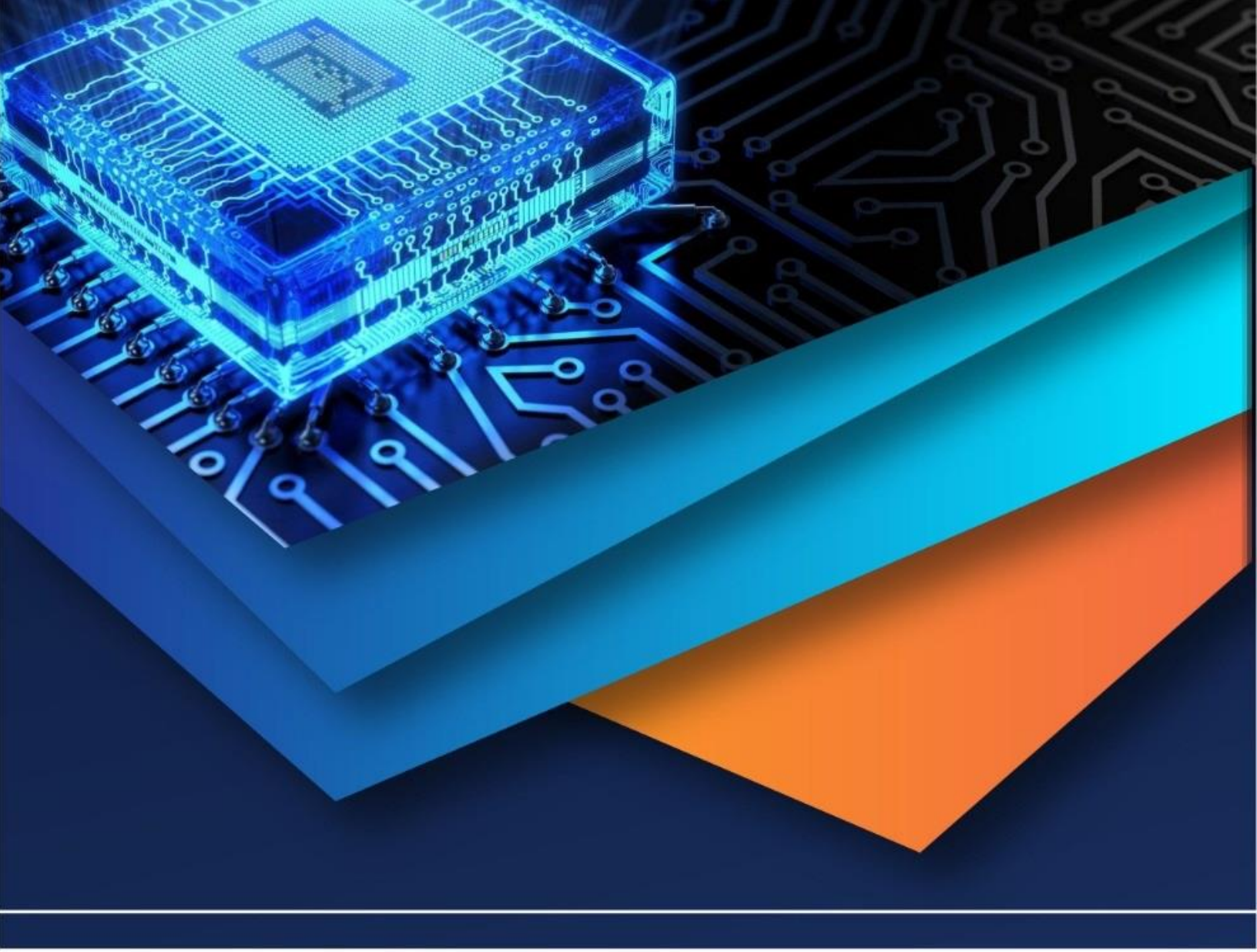
REFERENCES

[1] Vaibhav Patil, Pritesh Thakkar, Cjirag Shah, Tushar Bhat, Prof. S. P.Godse, “Detection and Prevention of Phishing Websites using Machine Learning Approach”, 3rd ed., vol. 2. Oxford: Clarendon, 1892,pp.68–73.
[2] Anu Vazhayil, Vinaya Kumar R and Soman KP, “Comparative Srudy Of The Detcetion Of Malicious URLs Using Shallow and Deep Netoworks “, 9th ICCCNT2018 July 10-12,2018,IISC,Bangluru,India.
[3] M. S. Arade, P. Bhaskar, and R. Kamat, “Antiphishing model with URL and image based web page matching”, Int. J. Comput. Sci. Technol. IJCST, vol. 2,no. 2, pp. 282-286, 2011.

©IJRASET: All Rights are Reserved

[4] Amani Alswailem, Bashayr Alabdullah and Norah Alrumayh , "Detecting Phishing Websites Using Machine Learning", 978-1- 7281- 0108-8/19/$31.00 2019 IEEE.
[5] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert Syst. Appl., vol.37, no. 12,pp.7913-7921, 2010.
[6] Varsharani Ramdas Hawanna, V. Y. Kulakarni and R.A. Rane, "A Novel Algorithm to Detect Phishing URL's", 978-1-5090-2080- 5/16/2016 IEEE.
[7] Gayatri. S, "Phishing Website Classifier Using Polynomial Neural Networks in Genetic Algorithm", 2017 4th International Conference On Signal Processing, Comminunications and Networking (ICSCN- 2017), March 16-18, 2017, Chennai, India.
[8] H. Shahriar and M. Zulkernine, "Trustworthiness testing of phishing websites:A behaviour model-based approach", Spec. Sect. SS Trust. Softw. Behav. SS Econ. Comput. Serv., vol. 28, no. 8, pp. 1258-1271, Oct. 2012.
[9] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithmbased feature selection and weighting," IET Inf. Secur., vol. 13, no. 6, pp. 659–699, 2019.
[10] Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an AntiPhishing Framework", 2019 5th International Conference On Advanced Computing & Communication System(ICACCS), pp. 1-6..
[11] M. Alqahtani, "Phishing Websites Classification using Association Classification (PWCAC)," 2019 Int. Conf. Comput. Inf. Sci., pp. 1–6, 2019.
[12] M. K. Verma, S. Yadav, B. K. Goyal, B. R. Prasad, and S. Agarawal, "Phishing Website Detection Using Neural Network and Deep Belief Network," Recent Find. Intell. Comput. Tech. Adv. Intell. Syst. Comput., vol. Springer, no. Singapore, pp. 293– 300, 2019.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)