Applied GIS

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

PHISHING WEBSITE DETECTION USING LIGHT GBM AND SVM ALGORITHM

P.ANIL JAWALKAR[1], NEESHA ROY[2], PATIL TANISHA[3], VARSHINI.PURELLA[4], T.GEETHIKA[5]

[1]Assistant Professor, Department of Information Technology, Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Hyderabad, TS, India.

[2,3,4,5] UG Students, Department of Information Technology, Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Hyderabad, TS, India.

ABSTRACT--Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as using light gbm and svm algorithm.

Keywords: URL, SVM, Light GBM, Cyber security, phishing website.

I INTRODUCTION:

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world. Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualizes and associations have

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection challenge while using artifcial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition. Opposite to our recognition system since it is a less greedy in terms of CPU and memory unlike other proposed systems as it needs only six features completely extracted from the URL as input. In this paper, after a summary of this feld key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last but not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

OBJECTIVE OF THE PROJECT

Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

II RELATED WORK

[1] Andrei Butnaru et al., used a supervised Machine Learning algorithm to block phishing attacks, based on novel mixture phishing attacks and compare with Google Safe browsers.

[2] Valid Shahrivari et al., proposed a one of the most successful techniques for identifying these

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

malicious works is Machine Learning. It is because of most Phishing attacks have same features which can be noticed by Machine learning techniques. In this many machine learning-based classifiers are used for forecasting the phishing websites. The main advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a forceful tool.

[3] Ammara Zamir et al., proposed a framework for identifying phishing websites using heaping model. Information gain, gain ratio, Relief-F, and recursive feature elimination (RFE) are some of the feature selection algorithms that can be used to analyse Phishing characteristics. The greatest and weakest traits are combined to create two features. Bagging is used in principal component analysis using several Machine learning algorithms, including random forest [RF] and neural network [NN]. Two heaping representations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.

[4] Nguyet Quang Do, Ali Selamat et al., conducted a study on phishing detection and proposed a four different deep learning technique, includes deep neural network (DNN), convolution neural networks (CNN), Long Short-term memory (LSTM), and gated recurrent unit (GRU). To analyse behaviour of these deep learning architectures, extensive experiments were carried out to examine the impact of parameter tuning on the performance accuracy of the deep learning models. In which each model shows different accuracies from different models.

[5] Ashit Kumar Dutta proposed a URL detection procedure based on Machine Learning methods. An RNN is used for identifying the phishing URL. It is evaluated with 7900 malicious and 5800 genuine sites, respectively. The outcome of this method shows a good concert compare to recent tactics.

EXISTING SYSTEM:

Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers.

PROPOSED SYSTEM:

Phishing attacks have evolved in terms of sophistication and have increased in sheer number in recent years. This has led to corresponding developments in the methods used to evade the detection of phishing attacks, which pose daunting challenges to the privacy and security of the users of smart systems. This study uses LightGBM and features of the domain name to propose a machine-learning-based method to identify phishing websites and maintain the security of smart systems. Domain name features, often known as symmetry, are the property wherein multiple domain-name-generation algorithms remain constant. The proposed model of detection is first used to extract features of the domain name of the given website, including character-level features and information on the domain name. The features are filtered to improve the model's accuracy and are subsequently used for classification. The results of experimental comparisons showed that the proposed model of detection, which integrates two types of features for training, significantly outperforms the model that uses a single type of feature. The proposed method also has a higher detection accuracy than other methods and is suitable for the real-time detection of many phishing websites.

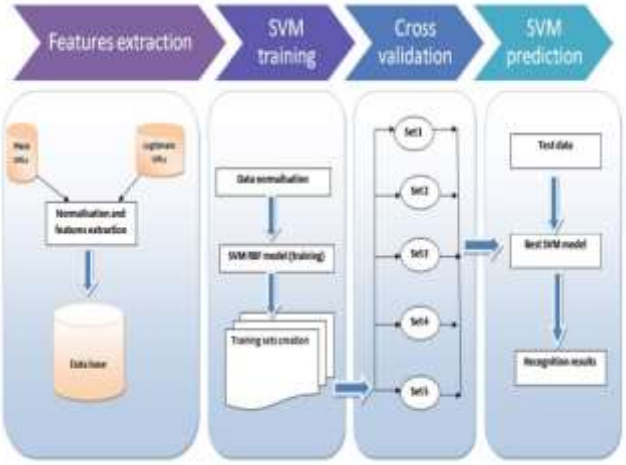


Fig.1. Phishing website process.

III METHODOLOGY

In this segment we going to learn about the classifiers used in machine learning to envisage phishing. Here we intend to explain our proposed methodology to detect phishing website. In this we divided into 2 parts one for classifiers and another to explain our proposed system.

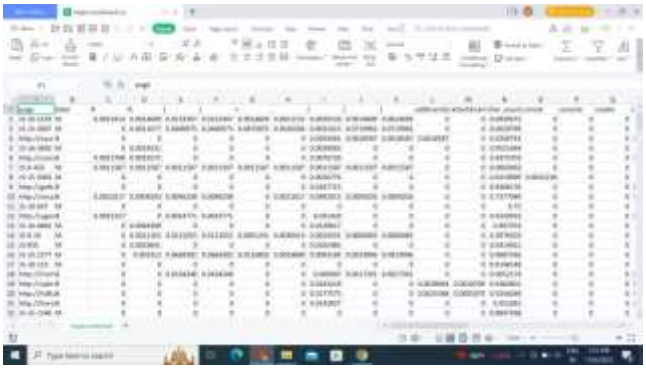
ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

Machine learning classifiers and methods to perceive the phishing website Distinguishing and recognizing phishing websites is really an intricate and energetic problem. Machine learning has been extensively used in numerous areas to produce automated results. Phishing attacks can take numerous forms, including dispatch, website, malware, and voice. This paper focuses on detecting website phishing (URL) using the Hybrid Algorithm Approach. It is a mix of different classifiers that work together to improve the system's accuracy and estimate rate. Depending on the application and the nature of the dataset used we can use any classification algorithms. As there are various applications, we cannot discriminate which of the algorithms are superior or not.

Support Vector Machine (SVM): This is also one of the supervised and simple to use classification algorithms. It can be used in both classification and regression applications; however, classification applications are preferred. SVMs differ from other classification algorithms in that they employ the distance between the nearest data points of all classes to determine the decision boundary. The maximum margin classifier or maximum margin hyper plane is the decision boundary created by SVMs. The classification is based on the differences between the classes, which are data set points in various planes.

Data set:

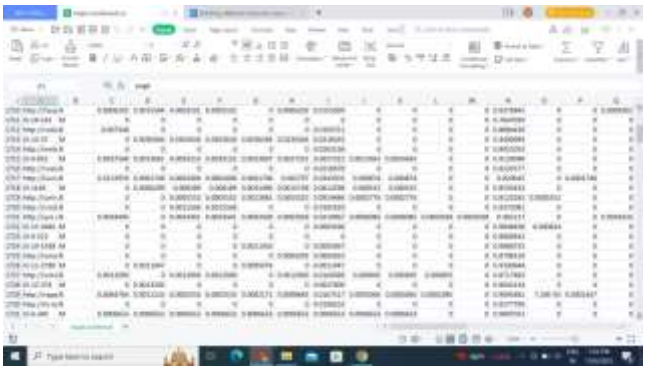
Phishing continues to prove one of the most successful and effective ways for cybercriminals to defraud us and steal our personal and financial information. Our growing reliance on the internet to conduct much of our day-to-day business has provided fraudsters with the perfect environment to launch targeted phishing attacks. The phishing attacks taking place today are sophisticated and increasingly more difficult to spot. A study conducted by Intel found that 97% of security experts fail at identifying phishing emails from genuine emails.



The provided dataset includes 11430 URLs with 87 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

the content of their correspondent pages, and 7 are extracted by querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.

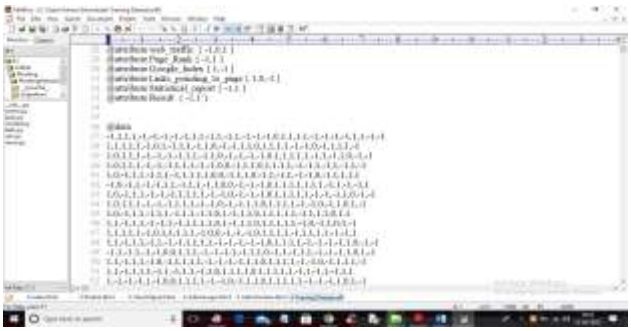


IV IMPLEMENTATION

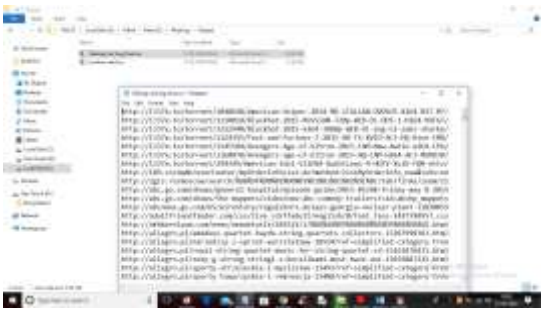
Detection of Phishing Website using SVM & Light GBM

In this project we are implementing SVM and Light GBM machine learning algorithms o detect phishing website URLS. We are training all this algorithms with normal and phishing URLS and build a trained model and this train model will be applied on new TEST URL to detect whether its normal or phishing URL.

In this project you asked to use UCI machine learning phishing dataset but this dataset contains only 0's and 1's values like below screen



From above dataset ML algorithms can get trained but we can't understand anything so I am using REAL WORLD URL dataset which contains normal and phishing URLS like below screen



In above screen you can see our dataset contains 2 folders called benign (phishing URLS) and valid (normal URL) and this are real world URLS and we will train all algorithms with above dataset and then when we input any test URL then ML model will predict as normal or phishing. To run this project double click on 'run.bat' file to start python DJANO server like below screen

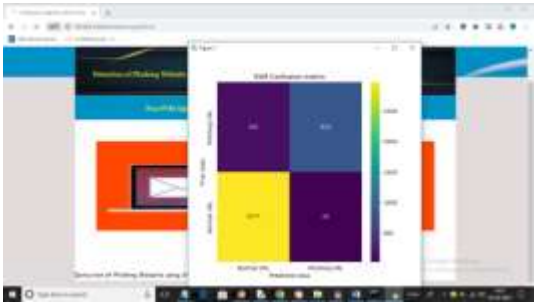
ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

In above screen DJANGO webserver started and now open browser and enter URL http://127.0.0.1:8000/index.html and press enter key to get below output

In above screen click on 'Admin Login Here' link to get below login screen

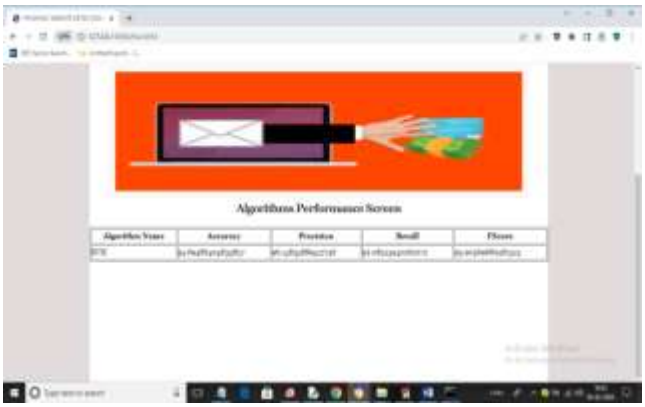
In above screen enter username and password as 'admin' and 'admin' and then press button to get below output

In above screen click on 'Run SVM Algorithm' link to train SVM algorithm and get below output

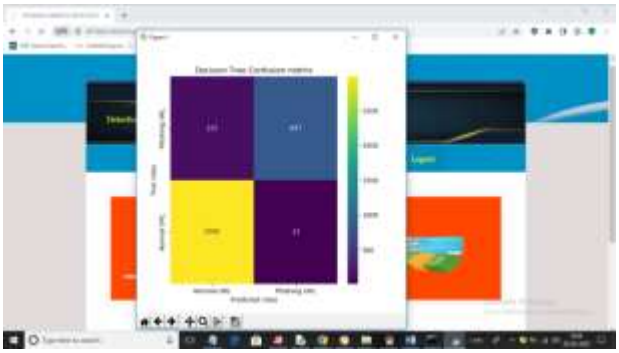


In above screen we can see SVM confusion matrix where x-axis represents predicted class and y-axis represents TRUE class and we can see SVM predict 2977 records correctly as NORMAL and only 145 are incorrect prediction and it predict 824 records as PHISHING URL and only 26 are incorrect prediction and now close above graph to get below output

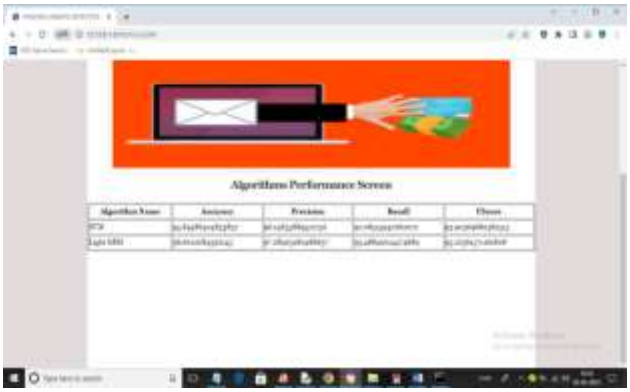
ISSN: 1832-5505
Vol-11 Issue-04 oct 2023



In above screen with SVM we got 95% accuracy and now click on 'Run Light GBM Algorithm' link to get below output



In above screen we can see Decision Tree confusion matrix graph and now close above graph to get below output



In above screen with Light GBM also we got 96% accuracy and now click on 'Test Your URL' link to get below screen

In above screen enter any URL and then press button and then Light GBM will predict whether that URL IS normal or phishing

In above screen I entered URL as https://mail.google.com and then press button to get below output

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

In above screen in blue colour text we can see given URL predicted as GENUINE (normal) and now test other URL. Similarly now I will enter Google.com in below screen

In above screen I gave URL as Google.com and below is the output

In above screen Google.com also predicted as Genuine. Now in below screen from internet I am taking one phishing URL and then input to my application to get prediction

In above screen blue colour URL is the phishing URL and I will input that to my application in below screen and below is the phishing URL from internet

'https://in.xero.com/3LQDhRwfvoQfeDtlDMqkk1JWSqC4CMJt4VVJRsGN'

In above screen I entered same URL and press button to get below output

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

In above screen in blue colour text we can see application detected PHISHING in given URL and similarly you can enter any URL and detect it as NORMAL or phishing

V CONCLUSION

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

Feature Analysis

The features of the domain name used here can be obtained only by using known strings of domain names without obtaining information related to user privacy, such as traffic in the network. Features of the domain name can be divided into two categories according to the acquisition method: features of the characters used in the domain name and features of information on the domain name. The features of information on the domain name can be obtained through the corresponding website or other query websites to this end, whereas the features of the characters used in the domain name can be obtained through a local feature-extraction algorithm without visiting the website.

VI REFERANCES

[1] Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)

[2] Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.

[3] Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)

[4] R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

[5] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)

[6] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards

ISSN: 1832-5505
Vol-11 Issue-04 oct 2023

Lightweight URL-Based Phishing Detection.13 June 2021

[7] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021

[8] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.

[9] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.

[10] Valid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

[11] A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.

[12] 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). Springer.

[13] 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.