Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

HYBRID APPROACH FOR PHISHING WEBSITE DETECTION USING M L.

M.PRAVEENA, HOD, Department of COMPUTER SCIENCE, SRI DURGA MALLESWARA SIDDHARTHA MAHILA KALASALA, Vijayawada, A.P., India.

V.SIVA KRISHNAVENI, LECTURER, Department of Computer Science, SRI DURGA MALLESWARA SIDDHARTHA MAHILA KALASALA, Vijayawada, A.P., 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

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. Alot of useful data, information and datacan 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 informationsimilar 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 lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishingattack 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 artificial 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 requireas 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.

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 detectphishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

LITERATURE SURVEY

Rashmi Karnik et al., proposed a model ofclassification method, kernel-based approach. In this we categories phishing. This method produces estimated accuracy of 95% in detecting the phishing and malware sites.

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

Vahid Shahrivari et al., proposed a one of the most successful techniques for identifying these 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 forforecasting 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.

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 heapingrepresentations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.

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-termmemory (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 learningmodels. In which each model showsdifferent accuracies from different models.

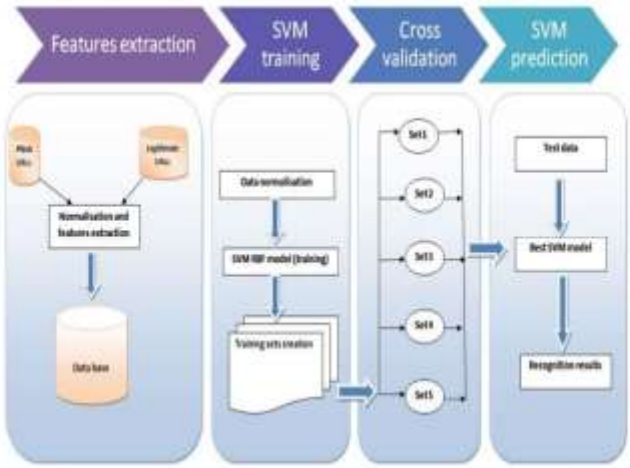
EXISTING SYSTEM:

Phishing is an internet scam in which an attacker sends out fake messagesthat 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 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. The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively

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 ofphishing 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 multipledomain-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 theproposed 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 higherdetection accuracy than other methods andis suitable for the real-time detection of many phishing websites.



3 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. 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

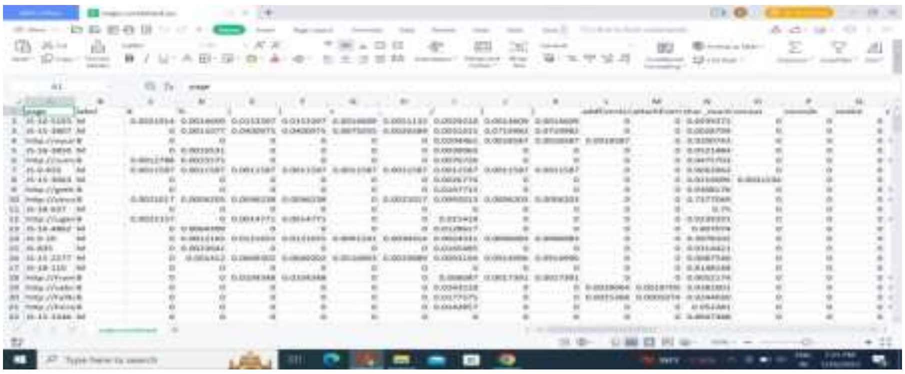
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

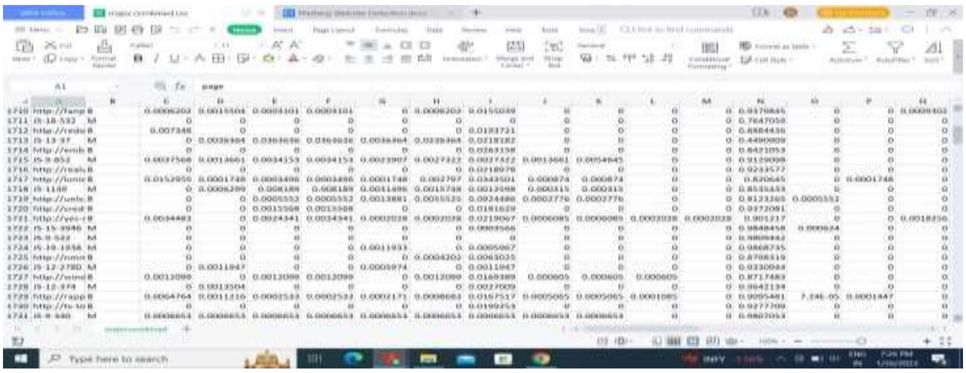
various applications, we cannot discriminate which of the algorithms are superior ornot.

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, classificationapplications 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 setpoints 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 perfectenvironment to launch targeted phishingattacks. The phishing attacks taking place today are sophisticated and increasingly more difficult to spot. A study conducted by Intel found that 97% of security expertsfail at identifying phishing emails from genuine emails.





Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

4 IMPLEMENTATION

In this section, we will going discuss about the actual steps which were executed while doing the experiment. We shall discuss the stepwise procedure used to analyse the data and to predict the phishing. We have used unstructured data which consists only url. There are 11064 urls obtained from the internet. Which consists of both phishing and genuine url where most of urls obtained are phishing.

1. First, we have unstructured data of urls from Phish tank website.

2. In Preprocessing, feature generation is done where eight features are generated from unstructured data. These features are length of url, http tokens, suspiciouscharacter, prefix/suffix, number of slashes, phishing term, length of subdomain, url IPaddress.

3. Next, a structured dataset is constructed, with binary values (0,1) for each feature, which is then sent to the various classifiers.

4. Next we train the two different classifiers and compare their performance on the basis of accuracy two classifiersnamely SVM, lightgbm.

5. Then classifier detects the given url based on the training data that is if the site is phishing it shows error and if legitimate it opens that page in browser.

6. We compare the accuracy of different classifiers and found lightgbm is the best classifier which gives the maximum accuracy.

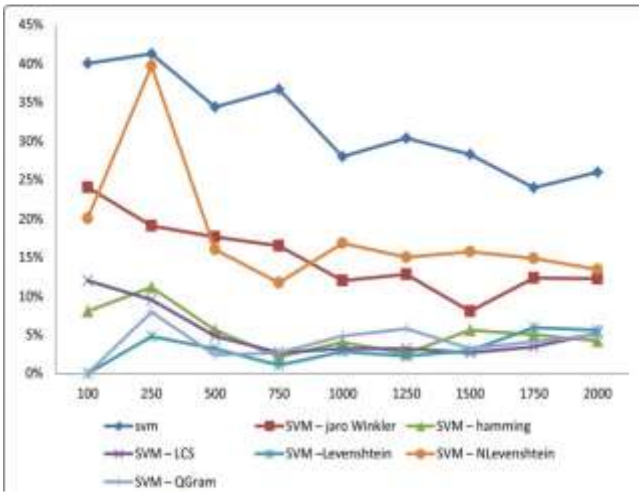


Fig.2. Graphical representation

Affirmed by the results of our tests, we have demonstrated the potential impactof the use of the similarity distance on the detection of phishing websites. Indeed, in three tests performed on four, theintroduction of the distance of similarity has significantly improved the recognition rate of our detection system. In the same context, the only case where the similarity did not have a positive impact on the phishing websites recognition rate is the test with Probabilistic Neural networks that records the worst recognition rate among all of our tests. This impact is mostobvious in tests performed using the SVM method since the use of the Hamming distance as one of the input characteristics of our system has improved the recognition rate of 21.8%.

5 CONCLUSION

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy usingrandom 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.

6 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. 7,2019R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

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

[5] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based PhishingDetection.13 June 2021

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

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

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