International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.8 September - October – 2024
DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

Detection of URL Based Phishing Attacks Using Gradient Booster Classifier Algorithm

Ms. Silna.KV
Assistant Professor, Department of Computer Science and Engineering, Vimal Jyothi Engineering College, Chemperi, Kannur, Kerala, India

Abstract: Phishing is a social engineering cyber attack where criminals deceive users to obtain their credentials through a login form that submits the data to a malicious server.The Internet is getting stronger day by day and it makes our lives easier with many applications that are executed on cyberworld. Cyber attacks have increased gradually and identity thefts have emerged.It is a type of fraud committed by intruders by using fake web pages to access people's private information such as userid, password, credit card number and bank account numbers, etc.These scammers can also send e-mail from many important institutions and organizations by using phishing attacks which imitate these web pages and acts as if they are original.Traditional security mechanisms can not prevent these attacks because they directly target the weakest part of connection : end-users. Machine learning technology has been used to detect and prevent this type of intrusions.The antiphishing method has been developed by detecting the attacks made with the technologies used and combined the websites by phishing attacks into a dataset, then obtained some results using gradient booster classifier algorithms with this dataset.The experimental results showed that the proposed systems give very good accuracy levels for the detection of these attacks.

Index Terms: Phishing attack,machine learning technology,gradient booster classifier algorithm.

1. Introduction

The use of web services has increased significantly in recent years as a result of the ongoing digital transformation. Businesses that offer online services, such as e-banking, e-commerce, or SaaS (Software as a Service), encourage change.The COVID-19 pandemic has led to restrictions on the work-from-home model, which means that millions more employees, students, and teachers are now developing their work remotely. This has resulted in a significant increase in workload for services like email, student platforms, VPNs, and company portals[4]. Phishing attacks, in which perpetrators attempt to impersonate reputable websites in order to obtain login credentials or payment information, could affect even more targets.According to recent research, phishing, along with spam emails and websites, is one of the most important social engineering-based attacks that occurred during the COVID-19 epidemic A phisher typically creates a phony website in a phishing operation[3]. This webpage has a different but misleading URL and looks similar to the original webpage. They are able to access users' sensitive information in this way. A cautious user might identify that the URL is phishing-related and harmful.Phishers conceal their scams by using social engineering tactics and human vulnerabilities. As part of the false phishing process, emails sent by phishers appear to be from official accounts belonging to institutions and organizations. Clicking on these emails takes the user to a fraudulent website. The user's credentials are used by this website. This data is stored on some other server.

Web browsers that have implemented browser security features to alert users to potential phishing websites include Google Chrome, Mozilla Firefox, and Internet Explorer.The list-based technique, compares the requested URL to a phishing database, is one of the most widely used methods for phishing detection.The request is denied and the user is warned before accessing the website if a

DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

requested URL matches any record.The use of machine learning for automatic identification of phishing URLs has gained traction in re-search due to the aforementioned short- comings of blacklist-based approaches.Approaches that rely on the visual elements and content of the page necessitate going to the website in order to retrieve the source code and render it, which takes time.
Phishing prevention techniques and user training systems are inferior to phishing detection approaches that identify phishing on the server side. These systems can be accessed by specialized host-site software or through the client's web browser.The categorization of methods for detecting phishing attempts. Supervised and unsupervised learning techniques are the foundation of heuristic and machine learning approaches. To anticipate something, it needs characteristics or labels to learn about the surroundings. ML method and proactive phishing URL detection are comparable. URLs are analyzed and help a system determine whether a given URL is malicious or authentic. The classic ways to recognize phishing sites are the blacklist and whitelist techniques.
One of the most popular solutions for phishing detection is the list-based approach, which analyzes the requested URL against a phishing database.If a requested URL matches any record, the request is blocked, and a warning is displayed to the user before visiting the website.Due to the mentioned drawbacks with the blacklist-based methods, automatic detection of phishing URLs based on machine learning, have attracted attention in re- search.Methods based on the page content and visual features require visiting the website to collect the source code and render it, which is a time-consuming task. Proposed system presents a phishing URL data set using legitimate login websites to obtain the URLs from such pages. Then evaluate machine techniques for recommending the method with higher accuracy. Next, show how models trained with legitimate homepages struggle to classify legitimate login URLs, demonstrating our hypothesis about phishing detection and legitimate login URLs. Additionally, show that how the accuracy rate is displayed by the model. Numerous tools, techniques, and algorithms are employed in the detection of phishing emails. Both academic and commercial organizations employ them. There are numerous characteristics that set a phishing URL and a parallel page apart from malicious URLs. Let's use an example to show how the phishing attacker can choose a long, complex domain name in order to conceal the original name. This is quite clear to see. Occasionally, they refer to the IP address rather than the domain name. However, they can also use a shorter domain name that has nothing to do with the original, authentic website. There are numerous additional features that can be utilized for phishing detection in addition to the URL-based capability.

2. PROBLEM DEFINITION

The major problem in cyber attacks is one of the most challenging tasks,to notify that and to avoid cyber attacks. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized. Several conventional techniques for detecting phishing website have been suggested to cope with this problem. Detecting phishing websites is a challenging task, as most of these techniques are not able to make an accurate decision.The intelligent programmers can generate URLs automatically and powerful tools can be utilized. The main limitations of random forest classifier algorithm is that a large number of trees can make the algorithm too slow and ineffective for real time predictions. To propose an intelligent model for detecting phishing web pages based on machine learning techniques using gradient booster classifier algorithm.

3.PROPOSED SYSTEM

Phishers try to click the URL of the site their victims to enter in their attacks. Identity hunters use some differences to change the appearance of the URL structure in various ways from the legitimate site. What are going to do here is that by doing extensive research on the malicious URL and use some properties to classify the web page and analyze the URLs are detected. Some features

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.8 September - October – 2024
DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

are defined below for the malicious URL.

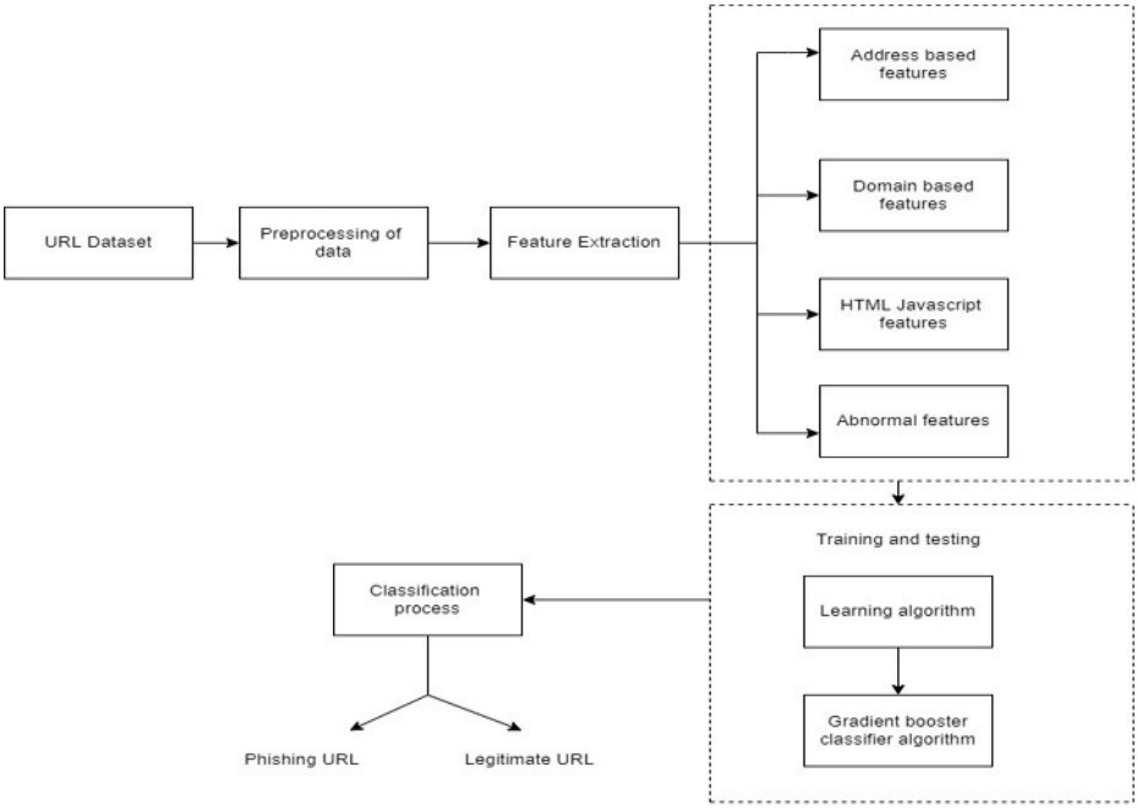

Figure 4.1 *Block diagram of the proposed system*

3.1.1 Machine Learning Techniques

This system, which is supported by menu script, compares several machine learning models used in the categorization of phishing websites.The gradient booster classifier algorithm is a machine learning technique that is used.This algorithm can be used to detect phishing URLs using categorization, with the output being classified as either authentic or phishing URLs.

3.1.2 Phishing URL Detection

Phishers try to deceive their victims by social engineering or creating mock sup websites to steal information such as account ID, username and password from individuals and organizations.

3.1.3 Gradient Booster Classifier Algorithm

Gradient boosting is a method standing out for its prediction speed and accuracy, par- ticularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results and already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model.

Gradient boosting machines are a family of powerful machine-learning techniques that have shown considerable success in a wide range of practical applications. They are highly customization to the particular needs of the application, like being learned with respect to different loss functions. This article gives a tutorial introduction into the method- ology of gradient boosting methods with astrong focus on machine learning aspects of modeling. A theoretical information is complemented with descriptive examples and illustrations which cover all the stages of the gradient boosting model design. Considerations on handling the model complexity are discussed. Three practical examples of gradient boosting applications are presented and comprehensively analyzed.Friedman's gradient boost algorithm shown in Algorithm 3.1.

The main idea behind this algorithm is to build models sequentially and these subsequent modelstry

DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

to reduce the errors of the previous model.The first step in gradient boosting is to build a base model to predict the observations in the training dataset.

i. Training

Datasets are uploaded by the user are used to train the system. These datasets are in the form of excel-sheets or .csv files. These are the inputs to the system. The system then extracts various features from the URLs present in these files. The features to be extracted are given in table 3.1. These extracted features are then used to create a decision tree, using gradient booster classifier algorithm as explained.An gradient booster decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch denotes an outcome of the test, and each leaf node represents a class. This decision tree is then used to form the rule sets. Rule sets are a set of rules that are used to classify new samples into already existing classes. After the rule sets are generated, they are stored in the database.

ii. Testing

After the training phase completes, the testing phase begins. The testing phase is the part in whichthe users can enter in random URLs to check whether they are legitimate sites or phishing sites. This checking is done using the already generated rule sets which are stored in the database. The gradient booster decision tree is used here to check for the legitimacy of a site. The entered URL undergoes multiple decision tests, based on the features shown in the table above, in the back end and then the output given is whether the site is a legitimate site or a phishing site.

4. IMPLEMENTATION

4.1 Data Collection

Is a collection of properties of urls. Dataset consist of 31 features of properties of urls shown in Table 4.1.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 do- main (TLD), and path domain. In this study, the subdomain, primary domain, and TLD are collectively referred to as the domain.

4.2 Data Pre-processing

Data Processing is the task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative.

4.3 Feature Extraction

It is provided with a dataset consisting of legitimate and phishing websites. Feature extractor extracts a set of features from it and feeds it to the classifier generator.It refers to the processof transforming raw data into numerical features that can be processed while preserving the information in the original data set.

4.4 Classification

In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. In the terminology of machine learning, classification is considered an instance of super- vised learning, i.e. learning where a training set of correctly identified observations is available. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. Classification is a process of categorizing a given set of data into classes, it can be performed on both structured or unstruc- tured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.8 September - October – 2024
DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

Features			
Number	*URL Properties*	*Number*	*URL Properties*
1	*UsingIP*	*17*	*LongURL*
2	*LongURL*	*18*	*InfoEmail*
3	*ShortURL*	*19*	*AbnormalURL*
4	*Symbol@*	*20*	*WebsiteForwarding*
5	*Redirecting*	*21*	*StatusBarCust*
6	*PrefixSuffix-*	*22*	*DisableRightClick*
7	*SubDomains*	*23*	*UsingPopup*
8	*HTTPS*	*24*	*IframeRedirection*
9	*DomainRegLen*	*25*	*AgeofDomain*
10	*Favicon*	*26*	*DNSRecording*
11	*NonStdPort*	*27*	*WebsiteTraffic*
12	*DomainURL*	*28*	*PageRank*
13	*RequestURL*	*29*	*GoogleIndex*
14	*AnchorURL*	*30*	*LinksPointingPage*
15	*LinksInScriptTags*	*31*	*StatsReport*
16	*ServerFormHandler*		

Table 4.1 *Data set URL Properties.*

4.5 Training

Train the features of url properties to set the value 1 is true and -1 is false.

4.6 Testing

The testing is a data set utilized to provide an impartial final design fit evaluation on the training set of data. In this stage, use the groups that were trained in the previous step that was trainedin , and the features were extracted by learning the network when the data set passes from url features, proposed system used 31 features of the data for testing.

4.7 Prediction

After the previous operations, phishing websites are detected and classified and give prevention method.

5.RESULTS

In the classification with the Gradient booster classifier algorithm,obtained a high accuracy of the performance was shown in the training model.training model for each class that was classified in training data.The training data's overall performance accuracy is 98.65%. Additionally, it calculated the test data's overall performance accuracy, which came out to be 98.65%, and the validation loss, which was only 0.018%.

Recent machine learning proposals have reported good performance trained with Pandas data sets.Train the network and save the trained network so that the training process is not repeated and so the time taken for training is provided, after which the network is also tested by testing data and showing the accuracy. In the last stage, through which of the url are detected and classified. The result of the project is 98.65% accurate output, which is much greater compared to existing other systems.So the performance of the implemented model is greater than other systems.Even though Gradient Booster Classifier Algorithm takes more time for training,the testing takes a comparatively far lesser computation times. The classification was performed and testing the dataset.

5.1 Performance Evaluation

Train the model and save the trained model so that the training process is not repeated andso the

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.8 September - October – 2024
DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

time taken for training is provided, after which the network is also tested by testing data and showing the accuracy. The prepared system was used for the training and testing of datasets containing many URLs. This led to the calculation of the TP, TN, FP and FN values. Thereon, the Specificity, Sensitivity, Accuracy and Error rate were calculated with the above values.

i. Recall : The proportion of positive samples is calculated according to the total number of positive samples in the correct classification used.
ii. Precision: It is a measure of accuracy in all estimated classes.
iii. F- Measure: It is the harmonic mean of Recall and Precision values. The purpose here is to measure the performance value shown by the classifiers. It is mostly used to compare classifiers. It is preferable to be high.
iv. Accuracy:Here is how often the classifier gets the correct predictions.

5.2 Evaluating Gradient Boosting Classifier using confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data.Confusion matrix and its usage for evaluating the classifier model from the evaluation of the Decision Tree algorithm.In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one; in unsupervised learning it is usually called a matching matrix.
Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa – both variants are found in the literature. The name stems from the fact that it makes it easy to see whether the system is confusing two classes

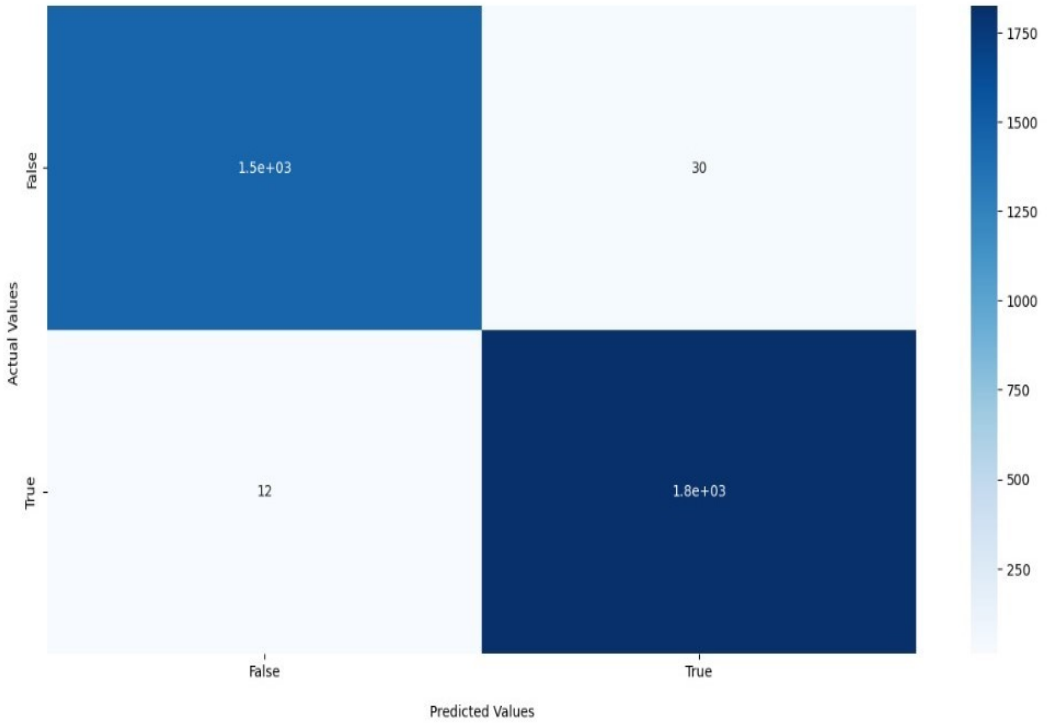

Figure 5.1 Confusion matrix

6. CONCLUSION

As a conclusion of this work, found that the gradient booster classifier algorithm works better than with others at relatively high accuracy rates. The models can be enhanced by using 31 features in

DOI:10.46647/ijetms.2024.v08i05.009 ISSN: 2581-4621

the system as apart from the URL based features also can be used here. Finally, identified whether the page is phishing or not.For detection of phishing sites URL of the web pages are mainly used. By using this data some features are produced and these features are used for detection of whether the web page is phishing or not. To predict this using gradient booster classifier algorithms.The method combines URL-based features by analyzing phishing site URLs. The proposed technique in turn provides security for personal information and reduces damage caused by phishing attacks because it detects new and temporary phishing sites gives accuracy about 98.65% percentage.

REFERENCES

[1] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre(2020), "Impact of current phishing strategies in machine learning models for phishing detection," IEEE Access, vol. 8, pp. 2021, pp. 87–96 .

[2] A. K. Jain and B. B. Gupta, (2020), "PHISH-SAFE: URL features-based phishing detection system using machine learning,,", IEEE Access, 2018, pp. 467–474

[3] Swati Singh, Sheifali Gupta(2018), "Machine learning based phishing detection from URLs, ",, Vol. 10, 06-Special Issue, 2018

[4] aizhou Li , Jianhui Lin , Jinrong Liu and Yandong Zhao(2020), "Phishing attacks and defenses," ",Int J Theor Phys, 2020;56(10):3029–49 .

[5] jangsheng Gui a, Jingyi Feia, ZixianWua, XiapingFub , Alou Diakitea (2018) "Detection of phishing websites from URLSs by using Classification Techniques on WEKA",arXiv preprint arXiz08.016679.

[6] APWG, "Anti phishing work group," http://www.antiphishing.org/.

[7] OpenDNS, "Phishtank," http://www.phishtank.com.

[8] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker, "Spamscatter: Characterizing internet scam hosting infrastructure," in *USENIX Security*, 2007.

[9] D. K. McGrath and M. Gupta, "Behind phishing: An examination of phisher modi operandi," in *In Proc. of the USENIX Workshop on Large Scale Explo its and EmergentThreats(LEET), San Francisco, CA*, 2008.

[10] R. S. Rao, T. Vaishnavi, and A. R. Pais, ''CatchPhish: Detection of phishing Websites by inspecting URLs,'' *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 2, pp. 813–825, Feb. 2020.