Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739

Vol 3(1), May 2023, pp 63 - 73

DETECTION OF PHISHING URL LINKS USING MACHINE LEARNING

Dr. N. Priya[1], C.Tharuneya[2]

[1]Associate Professor, PG Department of Computer Science

Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women

[2] Student, PG Department of Computer Science

Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women

Email: drnpriya2015@gmail.com[1], p21cs027@sdnbvc.edu.in[2]

ABSTRACT:

The phishing attacks attempt to gain confidential data leads to scams and information leaks by clicking URL links that may look legitimate. The victims of the attacks can be innocent users whoare not aware of phishing. It is important to act against these attacks, and create awareness among people and there is a need for trustworthy detection techniques. There is a challenge in identifying phishing links and the majority of methods cannot accurately determine whether a new link is a phishing or not. In this project, machine learning algorithms are proposed XG Boost, Random Forest, Decision trees are used for the detection methods based on URL structure and features. This study is also compared the three machine learning models and also shows that XG Boost technique is better than the other models for predicting the URL link.

KEYWORDS: *Detection techniques, structure of URL, machine learning algorithms, URL behavior*.

INTRODUCTION:

In today's modern world, everybody has an electronic gadget with the internet in their hands, So there is an increase in online purchases, bank account openings, transactions, renewal of their identity cards, etc., To perform those actions they need to give their details like name, phone number, e-mail, card details. The source(sellers)contact through messages and the advertisement of the company may also send through messages, the messages will be in the form of URLs on clicking that link it will be directed to a website, sometimes the link may look like it is from the source, but it is not, clicking on that may lead to data leaks. By clicking on that hyperlink that is sent through emails or messages, the phishing attempt is originally launched it may ask to update or verify their details, If it is clicked the user's browser will lead users to a fake one that replicates the real website, this happens because the phisher creates a false URL similar to the original one, The main aim is to prevent the users from phishing attacks, securing the

sensitive information, to protect the users from attacks, several techniques were introduced. URL analysis: This involves analyzing the URL of any given website to check if it is legally correct or a phishing attempt. Phishing URLs often contain misspellings, extra characters, or other suspicious elements that canbe detected through URL analysis.

The method used to detecting these URLs is based on a certain structure that has been observed in legitimate URLs, the attributes, are analyzed, like, dots, symbols, slash, domain name, etc, and these behaviors are trainedin the machine learning algorithms, The process where it is about detecting phishing URLs using ML generally involves several steps. First, a dataset of already known phishing URLs and correct URLs is collectedand preprocessed to extract features such as domain age, URL length, and the if suspicious keywords are present or not. Next, these features are used to train a machine learning model, such as a Decision tree, Random forest, XGBoost are to classify URLs as either phishing or legitimate. it process like a collection of data, preprocessing, extracting the feature, the suitable algorithms are chosen, which helps topredict the legitimate and phishing URLs, Blacklists are the most popular method, although they have significant challenges when used to block new URLs, In order to detect phishing URL Links, techniques such as supervised learning, unsupervised learning, and semi-supervised learning algorithms are used. The system uses many features for feature extraction, however the accuracy of detection criticallydepends on feature knowledge prior analysis of phishing URLs. The detection using machine learning is a powerful tool for preventing phishing attacks and protecting users from identity theft and other forms of cybercrime so the certain features which suit for detection are selected and predict accurate results and improve the performance for detecting the phishing URL links and it aims to produce an accurate detection of phishing URL links, Phishing URL detection using ML is an effective approachto combating phishing attacks, as it can automate the detection process and quickly identify new threats.

[1] introduced the technique in detection of malicious URL using ML algorithms random forest and support vector machine, The new methods include extraction of features from dataset, Lexical features, Host-based Features. The HTML content of a website gives its content-based features, some of them are PctExHyperlinks, ExtFavicon, insecureForms, RelativeFormAction, ExtFormAction, AbnormalForm Action, JavaScriptUrlLength, which produces accuracy of 99.97%the features can be used to predict the phishing sites .

[2] Marcelo Ferreira**,** introduced the method on Batch learning algorithm, Batch learning is a type of ML algorithm that generally means training a model on a fixed dataset or batch of data. In batch learning,

the model is trained using all of the available data at once, rather than updating the model incrementally with new data used machine learning algorithms like Naïve bayes, where the features are independent of each other, where it calculates the conditional probability , Support Vector Machine for binary classification of data. Online learning algorithms are learnt, from training and predict it, the algorithms that generally learn by updating a vector with the malicious labels using only first order features and the training data are First Order ones. For Second order, these algorithms, with features like statistical features instead of using first order features, tries to increase the learning efficiency with high accuracy.

Waleed Ali**,** proposed the method based on Wrapper based feature section and filter based, Features are selected based on statistical measures to evaluate features in the filter-based evaluation techniques. Information gain (IG) is one among the very common techniques, wrapper-based strategies make use of an inductive classifier to get the main usage of the features subset, then also to eliminate redundant value, and further a search algorithm is used to find all possible features and then carry out the evaluation of every subset by evaluating a model on that subset. It usually provides with the best features set and achieve good performance for that classifier and also it provides with the performance of machine learning classifiers, using WEKA software the performances of ML classifiers under the wrapper-based features selection is used, the information gain and the principal component analysis, correct classification rate, Algorithms like Back-Propagation Neural Network (BPNN), K-Nearest Neighbour (kNN), Naïve Bayes Classifier (NB), the wrapper based feature selection produces 97.1% of accuracy. The learning in the is carried out in two phases in BPNNs: the forward pass phase and the backward pass phase, training input pattern is used and then the received output is compared to the desired output pattern in order to compute an error.

[3] Hung Le, Quang Pham**,** proposed URLNet, a CNN based deep neural network for Malicious URL Detection, the advanced word-embedding techniques which are particularly useful to deal with rare words, a problem usually observed in malicious URL Detection tasks building Character level CNNs for Malicious URL Detection .

[4] Arun Kulkarni ,Leonard L. Brown, proposed four classifiers using MATLAB scripts, decision tree, Naïve Bayesian classifier, support vector machine (SVM), and neural network, SVM uses a nonlinear mapping that transforms original training data into a higher dimension and finds hyper planes that partition data samples respectively. Neural networks tend to learn with a training dataset and make decisions. The three layers are the input layer, hidden layer, and the output layer. Which provides accuracy of 91.5%The number of units in the input layer equals to the number offeatures, and the number of units in the output layer equals to the number of classes

Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

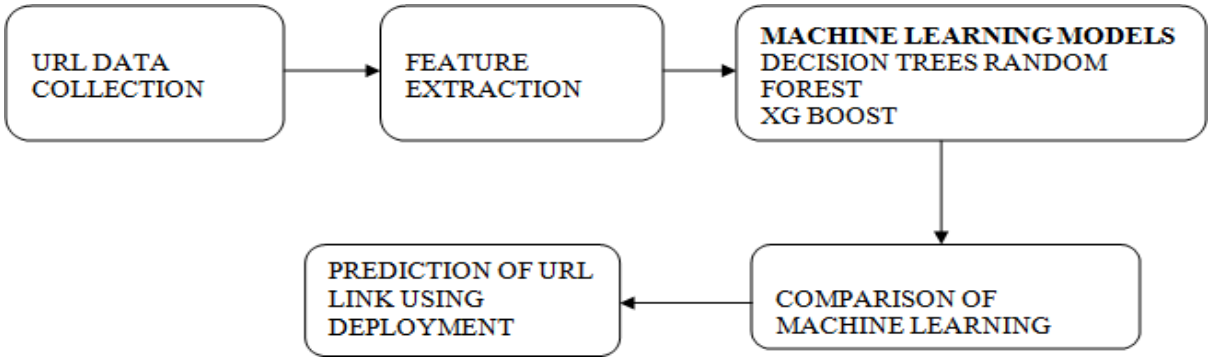
DATASET :

The dataset has been collected from "phish tank website" for phishing URLs which has 5000 URLs and legitimate URL collected from kaggle where 5000 URL are collected.

PROPOSED METHOD:

Our study detects the phishing URL based on extracting the features from URL,where most important features where observed from previous studies. Then shape of the dataset is analyzed, and checks if the null values are present in the dataset and then the extracted features are trained to the machine learning models.

ARCHITECTURE DIAGRAM:



FEATURE EXTRACTION:

- Feature extraction is an important step in phishing detection, where relevant information is extracted from the raw data to enable the identification of phishing attacks
- Domain-based features: These features are based on the domain name of the URL, such as the domainage, registration information,

THE FEATURES ARE:

LONG URL: Long URL to Hide the Suspicious Part, If the length of the URL is greater than or equal 54 characters then the URL classified as phishing.

@ SYMBOL: Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

REDIRECTION USING'//': The existence of "//" within the URL path means that the user will be redirected to another website"http://www.legitimate.com//http://www.phishing.com". By examining the location where the "//" appears it can found that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the "//" should appear in seventh position.

NUMBER OF DOTS: Sub-Domain and Multi Sub-Domains The legitimate URL link has two dotsin the. If the number of dots is equal to three then the URL is classified as "Suspicious" since it has one sub-domain. However, if the dots are greater than three it is classified as "Phishy" since it will have multiple sub-domains.

TINY URL: URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL.

HTTPS TOKEN: The Existence of "HTTPS" Token in the Domain Part of the URL.The phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/

IFRAME: IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the "frameBorder" attribute which causes the browser to render a visual delineation

PREFIX OR SUFFIX: The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage

IP ADDRESS: If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information.

DOMAIN REGISTRATION LENGTH: Based on the fact that a phishing website lives for a short period of time, trust worthy domains is regularly paid for several years in advance. By analyzing the dataset, the longest fraudulent domains have been used for one year only.

Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

WEB TRAFFIC: This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short periodof time, they may not be recognized by the Alexa database. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as "Phishing".

METHODOLOGY:

There are various algorithms and a large number of data types for the detection of phishing in the academic literature and commercial products range. Any phishing URL and the corresponding page tends to have several features which can then be differentiated from a malicious URL. Features extracted from URL Based Features.

1. Domain-Based Features
2. Page-Based Features
3. Content-Based Features

The machine algorithms are of three types, Unsupervised, supervised, reinforcement learning. Supervised machine learning algorithms are Random Forest, Decision Trees and XGBoost, where labeled data given to the machine for prediction.

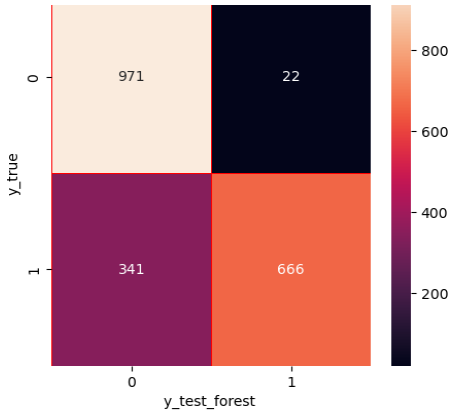
PROPOSED ALGORITHM:

The identification of phishing URL Links, helps to predicts the links earlier which is safe or not to click and enables safe browsing, The method used for detection of phishing links from all kind of sources like links from social media, messages from phone, website links,e-mails,bank transaction. The proposed Algorithm are XGBoost, Random Forest, Decision trees, which helps to predict the URLs.

Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

RANDOM FOREST:

It works by building a group of decision trees, each of which is thereby trained using a random subset of the input data and output features. By combining all of the forest's predictions, the final outcome is reached. The collection of dataset based on URLs that are classified as either phishing or legitimate in order to apply Random Forest for phishing URL detection to identify characteristics in each URL that can be used to differentiate between phishing and trustworthy URLs. The trees are created where each trees has built with the feature like the length of the URL, the presence of specific keywords, and the use of sub domainsare some features where tree is built with the rules and these rules are obtained from features and train a Random Forest classifier on the labeled dataset after the features have been extracted. This method chooses the random features and builds a tree. Each tree in the forest makes a prediction for the desired value, and an algorithm then calculates the votes for each target prediction. At the end, the random forest algorithm uses the target with the highest number of votes as its final prediction. Then for each input URL,the output prediction will be overall outcomes of each tree. It can handle large dimension of data and it is useful to interpret the results.The confusion matrix of the random forest given below.

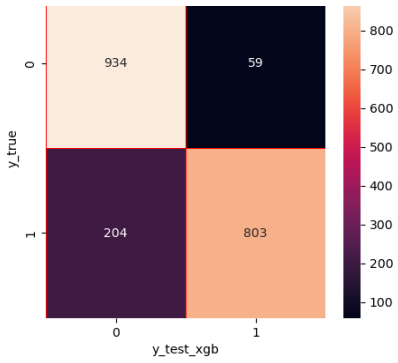


XGBOOST:

XGBoost expands to extreme Gradient Boosting. XGBoost is actually an implementation of boosted decision for improving the performances. During training, the algorithm assigns weights to each training example based on how difficult it is to classify correctly. The algorithm then learns to assign higher weights to examples that were misclassified by the previous trees, so that the next tree can focus on correcting those errors. It is a highly flexible algorithm that can handle a wide range of input features,

Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

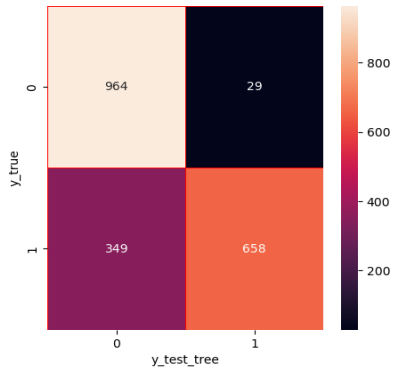
and it is able to model complex relationships between features. It also has built-in regularization techniques that can help prevent over fitting, which is important for any machine learning model. In XGBoost, the decision trees are built in a sequential manner, where each tree tries to correct the mistakes of the previous tree. In phishing URL detection, XGBoost can be used to classify URLs as either legitimate or phishing based on a set of predefined features. These features could include characteristics such as the length of the URL, the presence of certain keywords or symbols, and the similarity of the URL to a known legitimate website.The XGBoost algorithm works by iteratively adding decision trees to the ensemble, where each tree is built to correct the mistakes of the previous tree. During each iteration, the algorithm calculates the gradient, which measures the difference between the predicted and actual labels. The gradient represents the direction of the steepest ascent. The XGBoost algorithm then uses the gradient. The tree is built by recursively splitting the data into smaller subsets based on the feature values, with each split optimizing the loss function.Once the decision trees are built, XGBoost combines the predictions of all the trees to make a final prediction. This is done by assigning a weight to each tree based on its performance, where trees that make more accurate predictions are assigned higher weights. XGBoost has several advantages over other ML algorithms, such as its ability to handle large datasets with a large number of features; XGBoost is a powerful ML algorithm that can be used for phishing URL detection by building an ensemble of decision trees that combine to make a final prediction. The algorithm iteratively adds decision trees to the ensemble and assigns weights to each tree based on its performance. This results in a highly accurate model that can handle large datasets with complex relationships between the features and the labels, the confusion matrix of XGBoost given below



Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

DECISION TREES:

The decision trees are built by creating the trees of each features independently, The decision tree algorithm would use these features to create a set of rules that would be applied to new URLs to determine their likelihood of being phishing URLs. The algorithm would start at the root node of the tree and evaluate each feature until a final decision is made. For example, the decision tree may start by evaluating the length of the URL. If the URL is longer than a certain threshold, it may move down one branch of the tree, while if it is shorter, it may move down another branch. The algorithm would continue evaluating features and moving down branches until it reaches a leaf node, which would represent a final decision on whether the URL is legitimate or phishing. Once the decision tree is trained on a dataset of known phishing and legitimate URLs, it can be used to classify new URLs as either phishing or legitimate with a high degree of accuracy. However, it is important to continually update and retrain the decision tree as new phishing techniques and URL structures emerge. During training, the algorithm learns to split the data into smaller and smaller subsets based on the values of the input features. Once the tree is built, a new URL can be classified by traversing the tree from the root to a leaf node, where the leaf node represents the predicted class label. Once the tree is built, a new URL can be classified by traversing the tree from the root to a leaf node, where the leaf node represents the predicted class label. the confusion matrix of decision trees is given below.



Svādhyāya - International Journal of Transdisciplinary Research and Development (SIJTRD) ISSN Online : 2583-1739
Vol 3(1), May 2023, pp 63 - 73

IMPLEMENTATION AND RESULT:

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 80:20

MODEL	ACCURACY
Random forest	81.85%
Decision tree	81.11%
XG Boost	83.68%

CONCLUSION:

The aim is to predict whether a given URL is legitimate or a part of phishing site using machine learning algorithms in order to provide an early automatic detection of phishing links and the phishing attacks on innocent peoples can be prevented, enables safe browsing. Phishing attacks are becoming increasingly sophisticated: Phishing attacks are becoming more sophisticated, with attackers using social engineering techniques and other methods to make their messages appear legitimate .Education and awareness are critical for preventing phishing attacks: To prevent phishing attacks, individuals and organizations must be educated on the risks and trained on how to identify and respond to phishing attempts. This includes techniques such as checking the sender's email address, looking for spelling and grammar errors in messages, and avoiding clicking on suspicious links. Phishing can be defined as a cybercrime procedure that utilizes both social building and specialized deception to take individual sensitive data. Besides this, it is also considered as a type of fraud. Experimentations against recent dependable phishing data sets that utilizes different classification algorithm have been performed thereby receiving different learning methods. The base of these experiments is the accuracy measure. The aim of this research work is to predict whether a given URL is a genuine website or a phishing one. It turns out that in the given experiment that XGBoost based classifiers are the best classifier with great accuracy in classification as in 86.68% for the given dataset of phishing site. As part of future work this model can be used to other Phishing dataset with larger size than the one used now and then by testing the performance of those classification algorithms in terms of classification accuracy.

REFERENCES:

1. Cho Do Xuan,Hoa Dinh Nguyen, Malicious URL Detection based on Machine Learning (IJACSA) International Journal of Advanced Computer Science and Applications, Volume- 11, No. 1(2020)

2. Arun D. Kulkarni ,Leonard L. Brown, Phishing Websites Detection using Machine Learning, (IJACSA) International Journal of Advanced Computer Science and Applications, Volume-10, No. 7, 2019

3. Marcelo Ferreira, Malicious URL Detection using Machine Learning Algorithms, Proceedings of the Digital Privacy and Security Conference 2019 ,10.11228/dpsc.01.01.012

4. Waleed Ali,Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection,(IJACSA) International Journal of Advanced Computer Science and Applications, Vol.8, No. 9, 2017

5. Hung Le, Quang Pham, Doyen Sahoo, Steven C.H. Hoi, URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection Conference'17, July 2017, arXiv:1802.03162v2 [cs.CR] 2 Mar 2018.