PHISHING WEBSITE DETECTION USING MACHINE LEARNING ALGORITHMS

[1]D.Mahesh,[2]M.Jhansi,[3]Asha Kiran Annaladasu

[1,2,3]Assistant Professor

Department of CSE

Visvesvaraya College Of Engineering & Technology,Ibrahimpatnam,Telangana

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 analysing 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.As a result, finding strategies that can automatically detect phishing websites and quickly manage zero-day phishing attempts is an open challenge in this field. The web page in the URL which hosts that contains a wealth of data that can be used to determine the web server's maliciousness.

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion [2]. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack.

Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high [3].

To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

1.2. DATASET

URLs of benign websites were collected from www.alexa.com and The URLs of phishing websites were collected from www.phishtank.com. The data set consists of total 36,711 URLs which include 17058 benign URLs and 19653 phishing URLs. Benign URLs are labelled as “0” and phishing URLs are labelled as “1”.

1.3. FEATURE EXTRACTION

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

1) Presence of IP address in URL: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) Presence of @ symbol in URL: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol [4].

3) Number of dots in Hostname: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of “amazon” word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

4) Prefix or Suffix separated by (-) to domain: If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

5) URL redirection: If “//” present in URL path then feature is set to 1 else to 0. The existence of “//” within the URL path means that the user will be redirected to another website [4].

6) HTTPS token in URL: If HTTPS token present inURL then the feature is set to 1 else to 0. 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-mpp-home.soft- hair.com [4].

7) Information submission to Email: Phisher might use “mail()” or “mailto:” functions to redirect the user’s information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

8) URL Shortening Services “TinyURL”: TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0

9) Length of Host name: Average length of the benign URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to

10)presence of sensitive words in URL: Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

11) Number of slash in URL: The number of slashes in benign URLs is found to be a 5; if number of slashes in URL is greater than 5 then the feature is set to 1 else to 0.

12) Presence of Unicode in URL: Phishers can make a use of Unicode characters in URL to trick users to click on it. For example the domain “xn-- 80ak6aa92e.com” is equivalent to "apple.com". Visible URL to user is "apple.com" but after clicking on this URL, user will visit to “xn--80ak6aa92e.com” which is a phishing site.

13) Age of SSL Certificate: The existence of HTTPS is very important in giving the impression of website legitimacy [4]. But minimum age of the SSL certificate of benign website is between 1 year to 2 year.

14) URL of Anchor: We have extracted this feature by crawling the source code oh the URL. URL of the anchor is defined by <a> tag. If the <a> tag has a maximum number of hyperlinks which are from the other domain then the feature is set to 1 else to 0.

15) IFRAME: We have extracted this feature by crawling the source code of the URL. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders [4]. Since border of inserted webpage is invisible, user seems that the inserted web page

is also the part of the main web page and can enter sensitive information.

16) Website Rank: We extracted the rank of websites and compare it with the first One hundred thousand websites of Alexa database. If rank of the website is greater than 10,0000 then feature is set to 1 else to 0.

1.4. EXISTING SYSTEM

An existing system surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber attacks are spread via mechanisms that exploit weaknesses found in end- users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently implemented phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

Disadvantages

1) The system less effective since it is not implemented for large number of datasets.

2) The system doesn't implement Data Preprocessing and not compared with number of classifiers.

1.5. PROPOSED SYSTEM

1) The Proposed system designs the following concepts which Presence of IP address in URL: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL

to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) Presence of @ symbol in URL: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol [4].

3) Number of dots in Hostname: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

4) Prefix or Suffix separated by (-) to domain: If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

5) URL redirection: If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website [4].

6) HTTPS token in URL: If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-wwwpaypal-it-mpp-home.soft-hair.com [4].

7) Information submission to Email: Phisher might use "mail()" or "mailto:" functions to redirect the user's information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

8) URL Shortening Services "TinyURL": TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0

9) Length of Host name: Average length of the benign URLs is found to be a 25, If URL's length is greater than 25 then the feature is set to 1 else to 0.

10) Presence of sensitive words in URL: Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm',

'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

Advantages

1) Proposes a Decision Tree Algorithm which implements for Presence of sensitive words in URL.

The proposed system incorporates which Phishes can make a use of Unicode characters in URL to trick users to click on it.

2. MODULES

2.1. Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Browse Website URLs and Train & Test Data Sets,View Trained and Tested Accuracy in Bar Chart,View Trained and Tested Accuracy Results,View Prediction of Website URL Type,View Website URL Type Ratio,Download Trained Data Sets,View Website URL Type Ratio Results,View All Remote Users.

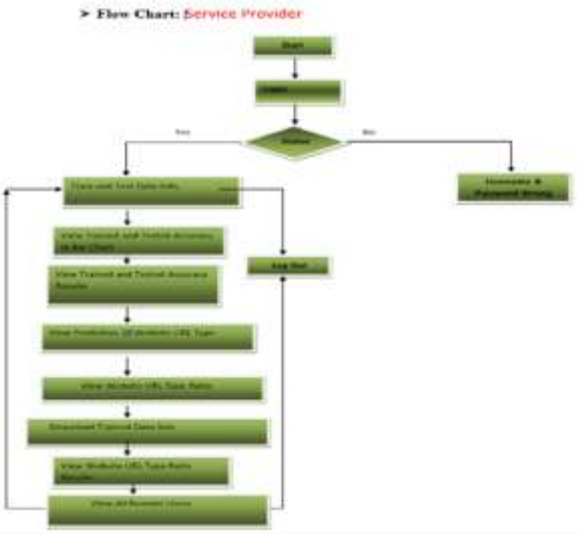


Fig.2(a) Flow chart of a service provider

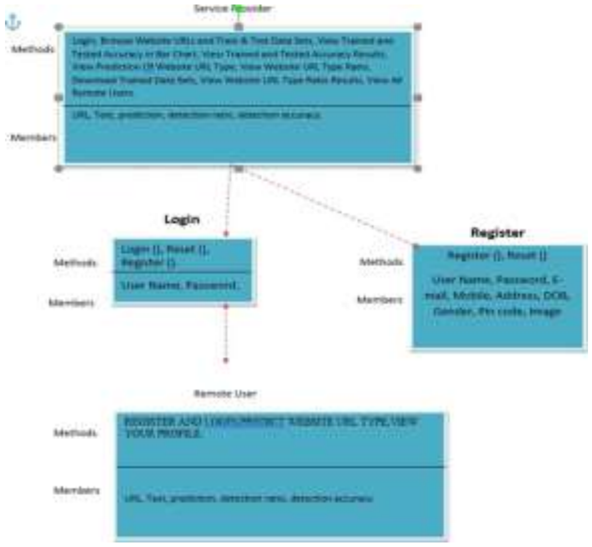

Service Provider
Methods
Login, Browse Website URLs and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Website URL Type, View Website URL Type Ratio, Download Trained Data Sets, View Website URL Type Ratio Results, View All Remote Users
Members
URL, Text, prediction, detection ratio, detection accuracy
Login
Methods
Login (), Reset (), Register ()
Members
User Name, Password.
Register
Methods
Register (), Reset ()
Members
User Name, Password, E-mail, Mobile, Address, DOB, Gender, Pin code, Image
Remote User
Methods
REGISTER AND LOGIN,PREDICT WEBSITE URL TYPE,VIEW YOUR PROFILE.
Members
URL, Text, prediction, detection ratio, detection accuracy

Fig.2(b) Class case model

2.1. View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorize the users.

2.1. Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like PREDICT WEBSITE URL TYPE, VIEW YOUR PROFILE.

➤ **Flow Chart: Remote User**

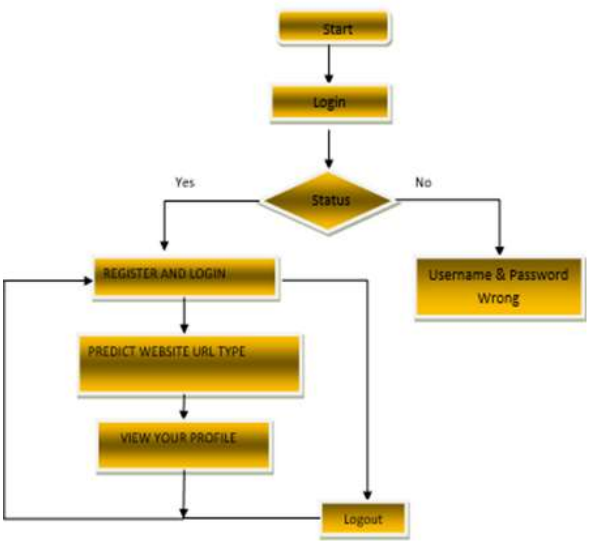


Fig.2(c) Flow chart of Remote User

3. MACHINE LEARNING ALGORITHMS

Three machine learning classification model Decision Tree, Random-forest and Support vector machine has been selected to detect phishing websites.

3.1 Decision Tree Algorithm

One of the most widely used algorithm in machine learning technology. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. In decision tree algorithm, gini index and information gain methods are used to calculate these nodes.

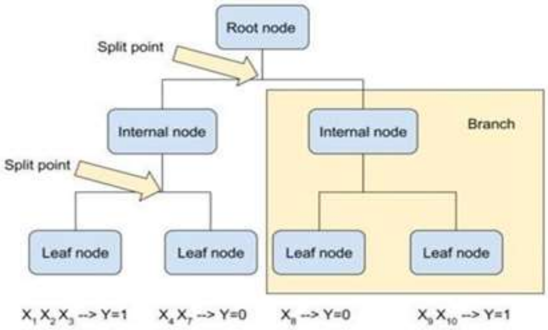


Fig. 3(a) Basic model of a Decision Tree Algorithm

It starts with a root node and ends with a decision made by leaves.Root nodes – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node.Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes. Sub-tree – just like a small portion of a graph is called sub-graph similarly a subsection of this decision tree is called sub-tree.Pruning is nothing but cutting down some nodes to stop overfitting

4.2 Random Forest Algorithm

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy.

Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random-forest algorithm also uses gini index and information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees.

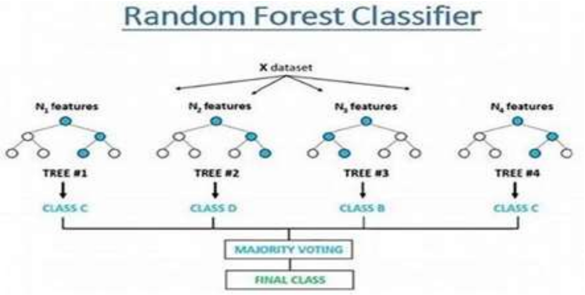


Fig. 3(b) Random Forest Classifier

Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

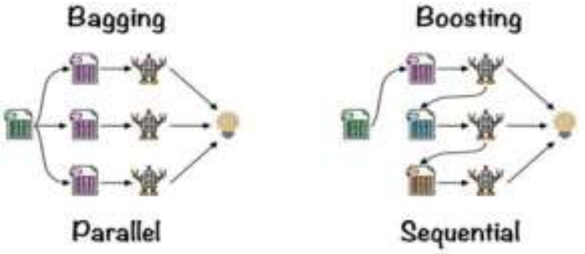


Fig 3(c). Example of Bagging & Boosting

3.3. Support Vector Machine Algorithm

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n- dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane.

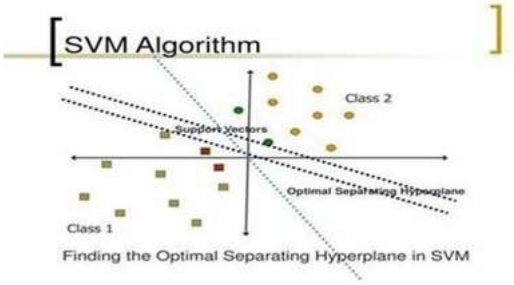


Fig 3(d) Graphical representation of SVM hyperplane

Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. In order to classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and non-linear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

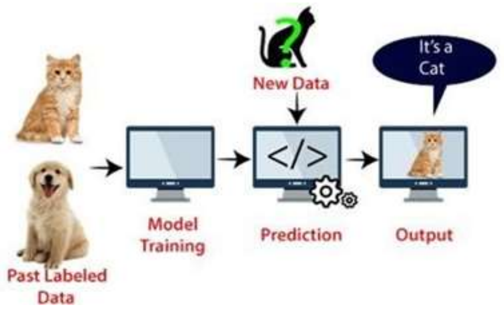


Fig. 3(e) Working of a SVM algorithm

SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm.

SYSTEM DESIGN

SYSTEM ARCHITECTURE

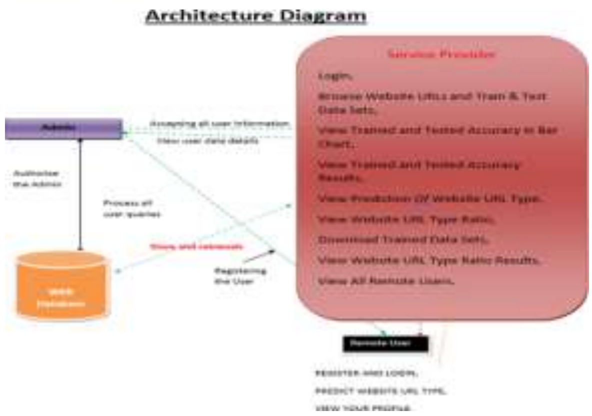

Architecture Diagram
Service Provider
Login,
Browse Website URLs and Train & Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View Prediction Of Website URL Type,
View Website URL Type Ratio,
Download Trained Data Sets,
View Website URL Type Ratio Results,
View All Remote Users.
Admin
Accepting all user information
View user data details
Authorize the Admin
Process all user queries
Store and retrievals
Registering the User
Web Database
Remote User
REGISTER AND LOGIN,
PREDICT WEBSITE URL TYPE,
VIEW YOUR PROFILE.

Fig. 6(b) Architecture Diagram

EXECUTION & OUTPUTS

Fig. 9(a) Login page of Remote User

Fig. 9(b) Login Page of Service Provider

Fig.9(c) Webpage to predict the URL links

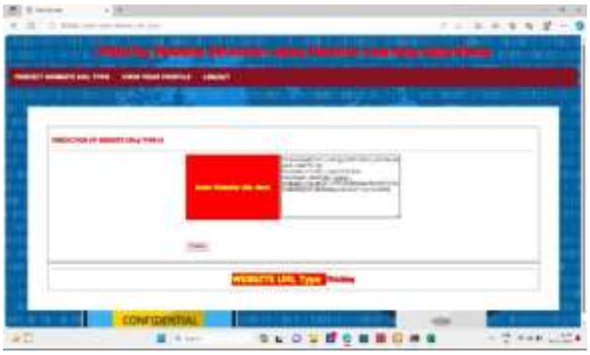


Fig. 9(d) Result/Type of the URL link

Fig.9(e) History of predicted URL links

Fig.9(f) Credentials of the User

Fig.9(g)Credentials of all Remote Users

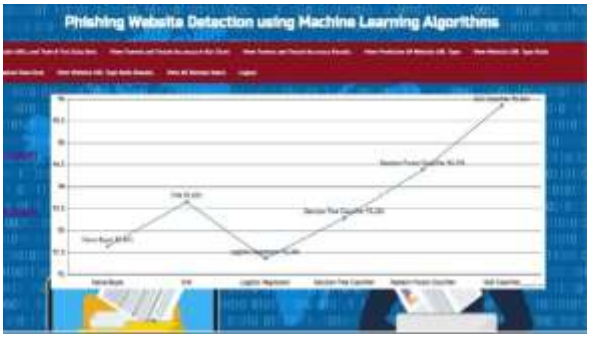


Fig. 9(h) Line graph of Accuracy of ML algorithms

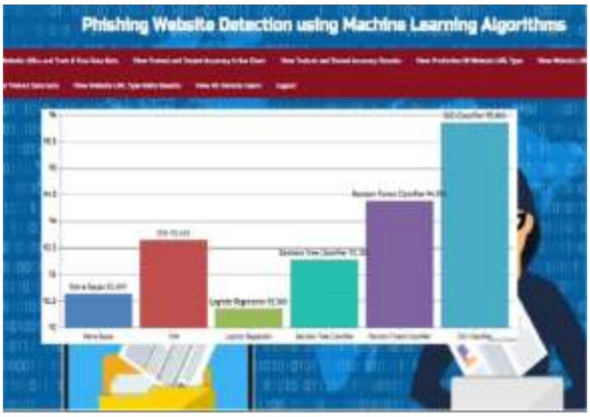
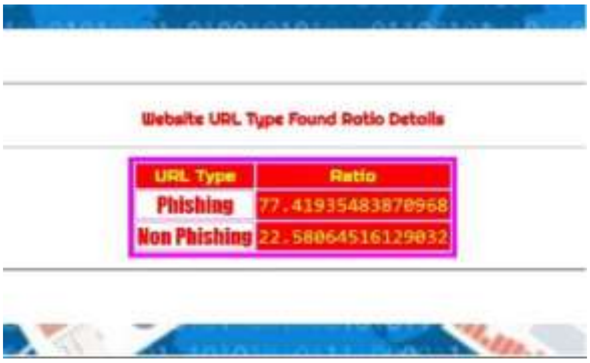


Fig. 9(i) Bar graph of Accuracy of ML algorithms



Website URL Type Found Ratio Details

URL Type	Ratio
Phishing	77.41935483870968
Non Phishing	22.58064516129032

Fig. 9(j) Ratio of URL Type

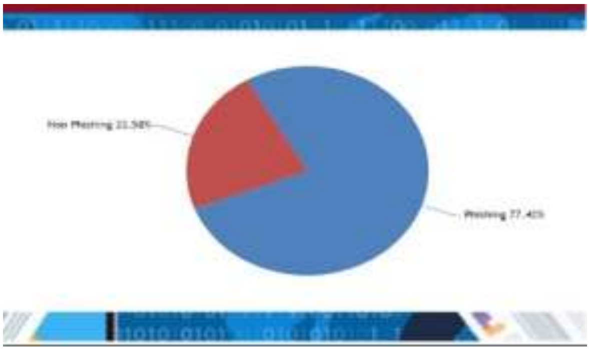


Fig. 9(k) Pie Chart of URL Type

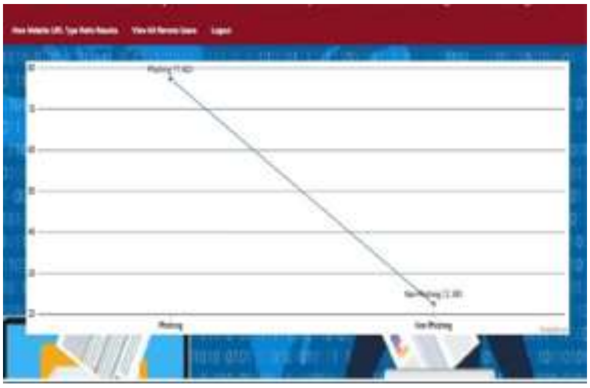


Fig. 9(l) Line graph of URL Types

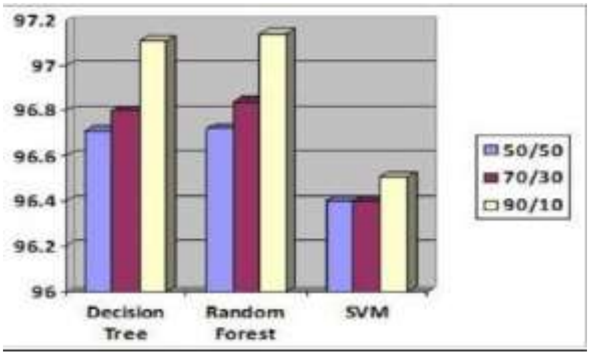


Fig. 9(m) Bar graph of Execution rate of ML algorithms

5. 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.

Phishing website attacks are a massive challenge for researchers, and they continue to show a rising trend in recent years. Blacklist/whitelist techniques are the traditional way to alleviate such threats. However, these methods fail to detect non-blacklisted phishing websites (i.e., 0-day attacks). As an improvement, machine learning techniques are being used to increase detection efficiency and reduce the misclassification ratio. However, some of them extract features from third-party services, search engines, website traffic, etc.

FUTURE SCOPE

In future work, we plane to include some new features to detect the phishing websites that contain malware. As we said in “Limitations” section, our approach could not detect the attached malware with phishing webpage. Nowadays, blockchain technology is more popular and seems to be a perfect target for phishing attacks like phishing scams on the blockchain.

Blockchain is an open and distributed ledger that can effectively register transactions between receiving and sending parties, demonstrably and constantly, making it common among investors. Thus, detecting phishing scams in the blockchain environment is a defiance for more research and evolution. Moreover, detecting phishing attacks in mobile devices is another important topic in this area due to the popularity of smart phones, which has made them a common target of phishing offenses.

A new dataset is constructed to measure the performance of the phishing detection approach, and various classification algorithms are employed. Furthermore, the performance of each category of the proposed feature set is also evaluated. According to the empirical and comparison results from the implemented classification algorithms, the XGBoost classifier with integration of all kinds of features provides the best performance. It acquired 1.39% false-positive rate and 96.76% of overall detection accuracy on our dataset. An accuracy of 98.48% with a 2.09% false-positive rate on a benchmark dataset.

REFERENCES

- Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.
- https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref
- Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
- Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
- http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/
- http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/
- https://www.kdnuggets.com/2016/07/support-vectormachines-simple-explanation.html
- www.alexa.com [9] www.phishtank.com