Detection and Prevention of Phishing URL

[1]Prof. Satish Jaywant Manje, [2]Miss.Shubhada Dinanath Wavale, [3]Mr.Kshitij Pradeep Khopkar, [4]Mr.Mayur Suresh Satvilkar

[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]satishmanje93@gmail.com,[2]shubhadawavale@gmail.com,[3]kshitijkhopkar76@gmail.com, [4]mayursatvilkar10@gmail.com

Abstract- Now a day the misuse of the internet is ongoing with every other website, application and many more methods are being developed. The crucial data lost and use of that data for personal gain is ongoing day by day. The problem with phishing is large and there exist many solutions to minimize all vulnerabilities effectively. The phishing website is used the payment option or data transfer and the user falls in the trap as the website is designed in such a way to represent itself as legitimate. To detect such a website or URL using this model will give the results so that the prevention from such URL can be achieved. First by processing analyses of the features of URL, and secondly is by checking the website legitimacy by understanding where the webpage is being hosted and by whom is it managed, and third is by analyzing based on visual appearance to check the genuineness of website or URL. By the use of Machine Learning techniques and algorithms for evaluation of the features and prevention from such URL is achieved.

Keywords- Phishing, URL, RF (Random Forest), features

I. INTRODUCTION

Phishing techniques used so far imitate the features and characteristics of emails and it makes them look like the original ones. It appears like that of a legitimate URL source. The user considers this email has come from a real company or an organization. Thus, it lures the person to visit the website by link given in that phishing email. Also, the attacker makes the user fill up the personal information by giving warning messages so that they fill up the required particular information or data which can be used by attackers to misuse the data. They make such a situation that the user has to visit their fake website. Phishing is a cyber-crime, the main reason behind the attacker doing this kind of crime is that it is not hard to perform; it is effective and does not cost anything. [2]

It has become so easy to find anyone's email id now a day and you can send or receive an email from anyone as it is freely available across the world. [6] These attackers put very little cost and effort to get valuable information easily. These cybercriminals are interested in the data which consists of crucial information of the user such as OTP, password, CVV, credit/debit card number, medical data, sensitive data related to business, confidential data, personal information, etc. Sometimes these criminals also gather the information that can give them direct access to exploiting the weakness found in the system at the user's end. For example, a system or an account may be technically secure and safe enough for password theft. A phishing attack technique is consists of a situation in which an attacker sends the URL pretending to be someone or something he's not to get the sensitive personal information from the victim. The victims regarding their curiosity or urgency, enter the personal details, which they are likely to acquiesce.[3,8]

II. AIMS AND OBJECTIVE

a) Aim

The aim of preparing this paper is to make it easier for organizations and companies and the major implications of these web attacks affect the financial transactions over the internet. Phishing is one of the most used popular methods that are performed to gain the advantages of security flaws in the system. Detecting, blocking, and preventing a phishing attack is an extremely important aspect to preserve the personal security and confidentiality of an individual over the world of the internet.

b) Objective

An objective of this topic is to develop a technique or method that can be easily used by everyone to detect whether the website is legitimate or non-legitimate accurately in real-time. It provides knowledge and insight to inexperienced web users in identifying phishing URLs.

III. LITERATURE SURVEY

This section covers a literature survey on various relevant works. It contains a lot of research to prevent and detect phishing websites and phishing emails using different approaches. In this survey, a user or an organization needs to be aware of various types of phishing attacks, methods, and having prior knowledge is necessary for identifying these web pages in real-time, also all the approaches are mainly based on comparing the content of the Legitimate and Non-Legitimate websites or webpages.

Paper 1: Detection and blocking of the phishing websites manually in time.

Manually detecting phishing webpages is one of the most common approaches used by security providers. An architectural design model that measures and evaluates the cognitive behavior of an ACT_R is proposed by Williams and Li (2017). This is performed by analysis of the legitimacy and authenticity of web pages based on the security indicator of HTTP padlock. Greenstadt and Afroz have derived a technique/method called 'Phish Zoo' which uses website profiling along with profile matching in the phishing detection process.[2]

Paper 2: Blocking of phishing E-mails by various spam filter soft wares.

E-mail attacks are one of the most major sources leading users to phishing websites. For preventing spam email clicks, spam filtering is a great option. Spam filtering ensures a wide majority of malicious spam/fake email detection and is not delivered to anyone's inboxes. Roy et al. have developed a technique that uses good spam filters for detecting spam emails.[7]

Paper 3: Server-side detection, IEEE.

Hu et al. have put forward a method that analyses server logs and other information to identify phishing websites. When a user visits a phishing website, the browser contacts itself to the real one for providing resources. Then real legitimate website server registers this request in the log, thereafter it is used to identify illegitimate/malicious websites. Wu et al. have arisen with a method that uses the various techniques of fuzzy logic which combined with the techniques of ML and eliminates the use of the Boolean algorithm in the method. They use the domain name, sub-domain name, and also the lifetime of the website in the authentication process.[5]

IV. EXISTING SYSTEM

Prevention and detection of phishing websites is the analysis of the detecting phishing sites, mails, messages, emails, etc. The paper detects the attacks and takes preventions from phishing sites. This existing system of Anti-phishing approaches uses feature-based ML techniques or Blacklist Methods. These methods are Fails to recognize new phishing attacks and also get a high positive rate. The machine learning-based existing techniques extract features and data from the search engine, third party, etc. Therefore, these methods are intricate, slow, and imperfect for the real-time environment. [2, 6]

The solution for this problem is, this paper has represented ML-based Novel Anti Phishing apps that extract features of the client-side. The various attributes of malicious and non-malicious websites in-depth are examined and identified five new amazing features to differentiate the phishing websites from non-phishing ones. [4] Phishing websites are common entry points in the online human and social engineering attacks, including various ongoing web scams.

V. COMPARATIVE STUDY

SR NO.	PAPER TITLE	AUTHOR NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Detection and blocking the phishing websites manually in time.	Li and Williams (2017), Afroz and Greenstadt	An artificial neural network-based Back-propagation algorithm	It helps users to find more accurate data.	Security awareness training is not continuous.
2.	Blocking of the phishing E-mails by various spam filter soft wares	Roy et al., Pandey and Ravi	Optical Flow Analysis	It helps users to find phishing emails.	These spam filters are used to block genuine messages.
3.	Server-side Phishing Detection	Wu et al., Hu et al.	Hungarian method	It detects the phishing server easily.	They underperform in slow internet connections

Table No.1: Comparative analysis

VI. PROBLEM STATEMENT

The current existing systems are fully functional but when security flaws, safety, reliability of an individual is concerned; it is very disappointing to provide efficiency and after evaluating the performance and accuracy of the systems. The prime concerns of every user are safety, security issues, confidentiality, time concerns, and personal data loss to anyone.

VII. PROPOSED SYSTEM

The idea of the proposed system solution has to make an effective system that will give maximum accuracy. Here decision tree classifier will be used for data fitting and a random forest algorithm will be used for classification.

The planned system has the following features:

a) Monitor all "HTTP" traffic of the end-user system by creating a browser extension. The benefits of extension over any software or application are that the system will be based on the uniquely in the same time at real-time also quite agile in sending the output.

b) Comparing each URL domain with the white-list of trusted domains and also the black-list of illegitimate domains.

c) If the domain of the URL is found under the white list, mark the URL as innocent (Exact Matching), else go further and use the other approaches.

d) The whole website analysis would now be done by considering various details. The set of features selected are:

Website protocol (secure or insecure), length of the URL, number of hyphens (-) in URL, number of @ symbol in URL, number of dots in the URL, using the direct IP address or not, daily page view, registration, and expiration date of website, daily unique visitor, favicon icon similarity and Google indexing.

VIII. ALGORITHM

Generalize basic idea of the working of the proposed system algorithm is given as follow:

URL detection Algorithm

Step 1. Start

Step 2. Take the dataset and concatenate them.

legitimate_urls=pd.read_csv("legitimate-urls.csv")

phishing_urls=pd.read_csv("phishing-urls.csv")

Step 3. Train the dataset:-

data_ train, data_ test, labels_ train, labels_ test =train_ test_ split(urls_ without_ labels, labels, test_ size=0.20, random_state=100)

train_0_dist = 711/1410

print(train_0_dist)

train_1_dist = 699/1410

print(train_1_dist)

test_0_dist = 306/605

print(test_0_dist)

test_1_dist = 299/605

print(test_1_dist)

Step 4. Fitting the data in model

DTmodel=DecisionTreeClassifier(random_state=0)

DTmodel.fit(data_train,labels_train)

Step 5. Use the algorithm to classify the URL's as legit or phishing

RFmodel = RandomForestClassifier()

RFmodel.fit(data_train,labels_train)

rf_pred_label = RFmodel.predict(data_test)

Step 6. Check the accuracy score of model

accuracy_score(labels_test,rf_pred_label)

imp_rf_model=RandomForestClassifier(n_estimators=100 ,max_depth=30,max_leaf_nodes=10000)

imp_pred_label=imp_rf_model.predict(data_test)

Step 7. Open the browser and execute the model.

Step 8. Enter the URL to check.

Step 9. Display the result as a warning.

Step 10. Stop.

IX. MATHEMATICAL MODEL

1. Random Forest

- A random forest algorithm is one of the types of supervised learning algorithms.
- Random forest merges the decisions of multiple decision trees to find a solution, which gives an output that is an average of all these decision trees.
- This method is applied to the labeled dataset (Phishing/Antiphishing) to "learn" how to classify the unlabeled dataset.
- A Gini Index is used when the random forest is performed on classification data.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

- Here, c defines the number of classes and p_i defines the relative frequency of the class which is observed in the dataset.
- The above formula uses the probability and class to find out the Gini index of each of the branches on a tree node, to determine which of the branches are more certain to occur.
- RF is implemented by following:

importing the libraries from sklearn of the alogorithm.
from sklearn.ensemble.import RandomForestClassifier

RFmodel = RandomForestClassifier()

RFmodel.fit(data_train,labels_train)

rf_pred_label = RFmodel.predict(data_test)

2. Decision Tree Classifier:

- A decision tree model is another type of supervised learning algorithm in which the data is divide according to given conditions.
- The tree is described by two entities that are nodes and leaves.
- In this classifier, a class label is assigned or given to each leaf node.
- The non-terminal nodes, which mostly include the root node and other attributes use a condition to dissociate the records that have different values.

from sklearn.tree import DecisionTreeClassifier
DTmodel=DecisionTreeClassifier(random_state=0)
DTmodel.fit(data_train,labels_train)
pred_label=DTmodel.predict(data_test)

3. Model Evaluation

- Evaluation metrics are used to evaluate the performance of the system which are accuracy (ACC), precision (Prec), recall (Rec), and f-score.
- Accuracy measures the correctly predicted URLs ratio. Precision measures the fraction of URLs that are correctly predicted as phishing. Recall metric measures the fraction of phishing websites identified by the system. F-score is another performance metric that measures the mean value of recall and precision metrics.
- To describe the performance of a classification model, a confusion matrix is used on a testing dataset for which the true real values are given.

	Predicted positive class	Predicted negative class
Actual positive class	TP	FN
Actual negative class	FP	TN

Table No.2: Confusion Matrix

- Here, TP (True Positive) rate measures that a model correctly predicts an URL as fake, TN (True Negative) rate measure that an URL is wrongly classified as legitimate, FP (False Positive) rate that measures an URL is incorrectly classified as phishing and FN (False negative) rate measures that an URL is incorrectly classified as legitimate while it is phishing.
- The mathematical equations of the performance metrics are as follows:

1. $ACC = \frac{TP+TN}{(TP+FN+TN+FP)}$
2. $Prec = \frac{TP}{(TP+FP)}$
3. $Rec = \frac{TP}{(TP+FN)}$
4. $F-score = 2 \times \left(\frac{Prec \times Rec}{Prec+Rec}\right)$

X. SYSTEM ARCHITECTURE

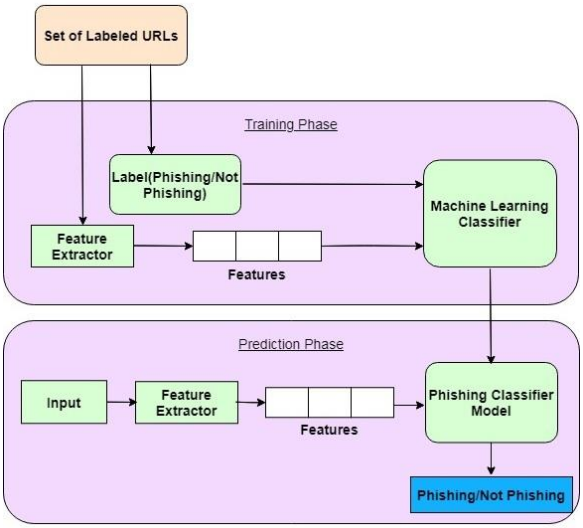

Fig No.1: System Architecture

Description:

There are two types of phases in system namely training phase and prediction phase.

1) Training phase:

In training phase, system learns that how to perform detection by providing it the already detected dataset of the URLs. So that main output of the training phase will be provide to the prediction phase of the system. In this phase, a set of labeled URLs with labels of phishing or non-phishing is given to the labeled data and feature extractor. Then the system performs feature extraction operation on this set and generates various features. These features are afterward given to the Machine learning classifier which trains the algorithm to perform operation on provided input and gives the output of the training phase.

2) Prediction phase:

Prediction phase is the main phase of the system which will be actually working on new data given by user. In the prediction phase, user gives a fresh new input URL to the system. This input is passed to the feature extraction model which separates the features. Then the features are provided to the phishing classifier model which also has the knowledge from the machine learning classifier of the training phase. Then this phishing classifier model generates the output and show the result whether the URL is phishing or non-phishing.

XI. ADVANTAGES

1) Everyone can use this system to achieve safe and secure surfing over the internet.

2) Payments can be done securely over the internet on secure and genuine e-banking sites.

3) This helps to build good customer relationships which would benefit the customers as well as the E-commerce websites.

4) Better performance is achieved by using ML-based algorithms in the system instead of other traditional algorithms.

5) Users can also purchase things, sells online, and safely roam on the internet without any hesitation.

XII. DESIGN DETAILS

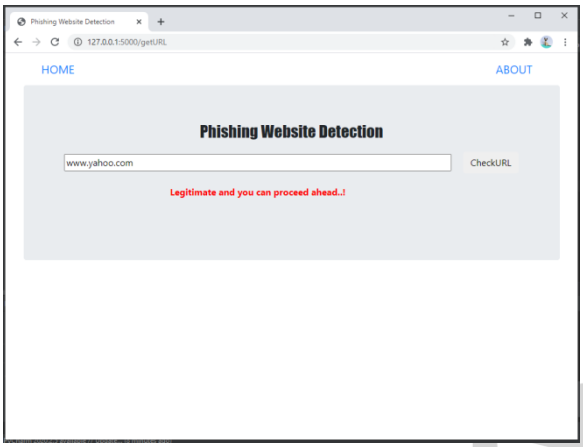

Fig No.2: Legitimate Result

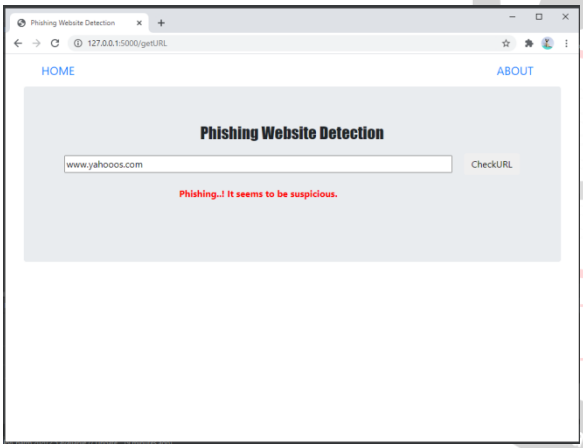

Fig No.3: Phishing Result

The web application is created to check the legitimacy of an URL. Users first have to open the application on web browser and provide the URL in input section. Then click the button of CheckURL to detect the result given by system and wait for system response. In the figures given above two results are shown. Figure no. 1 shows that the input URL is legitimate and user can proceed and figure no. 2 shows that the input URL is phishing URL and it seems to be suspicious.

XIII. CONCLUSION

Thus, we have tried to implement the paper "Vaibhav Patil, Tushar Bhat, Pritesh Thakkar, Chirag Shah", "Detection and Prevention of Phishing Websites using Machine Learning Approach" in IEEE 2018 and according to the Application various techniques for machine-learning-based have been studied and two of them are applied in this paper and the accuracy is measured. In this paper, a Machine learning algorithm is applied progressively and successively to predict the phishing of an URL of a website. Thus, the proposed system enables internet users to have safe browsing and safe transactions. It helps users to save their important private details that should not be leaked. User only needs to provide an active internet connection to check the legitimacy of an URL. Providing our proposed system to users in the form of an extension makes the process of delivering our system much easier. A challenge in this domain is that criminals are constantly making new strategies to counter our defense measures. To succeed and achieve efficient results, there is a need for algorithms that continually adapt to new threats, examples, new phishing techniques and features of phishing URLs. And thus, online learning algorithms are used. This paper provides a system that gives a result with maximum accuracy. Using random forest with a decision tree classifier enhances the evaluation metrics of the system and provides an efficient protection system for the internet user. In the future system is designed to work efficiently and successfully to detect false-negative and false-positive results, more accuracy in the detection and new problems in the features of an URL.

REFERENCE

[1] Moises Guy Harrison; Steven Feuerstein (2008). *MySQL Stored Procedure Programming*. O'Reilly Media

[2] Pandey M, Ravi V (2013) Text and data mining to detect phishing websites.In: Das S, Suganthan PN, Panigrahi BK, Dash SS Swarm, evolutionary, and memetic computing. SEMCO 2013. https://doi.org/10.1007/978-3- 319-03756-1_50

[3] 2017 Internet Crime Report. (n.d.). Retrieved https://pdf.ic3.gov/2017_IC3 Report.pdf.

[4] Greenstadt R, Afroz S, (2011) PhishZoo: detecting and blocking phishing websites by ML. IEEE (2011) https://doi.org/10.1109/ICSC.2011.52.

[5] Ankit Kumar Jain, B. B. Gupta, "Phishing Detection: Analysis of Visual Similarity-Based Approaches", IEEE 2017, https://doi.org/10.1155/2017/5421046

[6] Yuxiang G, Futai Z, Bei P, Linsen L, Li P, (2016) Website phishing detection by graph mining. 2016 IEEE (ICCC). https://doi.org/10.1109/compcomm.

[7] Mandal K, Kunar S, Roy S, Sau S, Patra A, (2013) An efficient spam filtering method for email phishing. http://www.ajer.org/papers/v2(10)/F02106373.pdf

[8] Buber E, Sahingoz OK, Diri B, Demir O, (2019) Machine learning (ML) based phishing detection of URLs. Expert System Applied 117:345–357. https://doi.org/10.1016/j.eswa.2018.09.029

