DETECTION OF URL BASED PHISHING WEBSITES USING MACHINE LEARNING WITH PYTHON

1) Preliminary Investigation:

- **ABSTRACT:**

Phishing attacks are the easiest way to obtain sensitive information from innocent users. The purpose of phishers is to obtain important information such as usernames, passwords and bank account details. Network security personnel are looking for reliable and stable detection technology to detect phishing websites.
This article discusses machine learning techniques for detecting phishing URLs by extracting and analyzing various functions of legal and phishing URLs. Decision tree, random forest and support vector machine algorithms are used to detect phishing websites. The purpose of this article is to detect phishing URLs by comparing the accuracy, false positive rate and false positive rate of each algorithm, and narrow it down to the best machine learning algorithm.

- **Introduction:**

Nowadays, phishing has become a major area of concern for security researchers, because it is not difficult to create fake websites that look very close to legitimate websites. Experts can identify fake websites, but not all users can identify fake websites, and these users become victims of phishing attacks. The main purpose of the attacker is to steal bank account credentials. In American companies, because customers become victims of phishing, they lose $2 billion every year. In the third Microsoft Computing Security Index report released in February 2014, the global annual impact of phishing was estimated to be as high as $5 billion. Phishing attacks have become successful due to lack of user awareness. Since phishing attacks exploit the weaknesses found in users, it is difficult to mitigate them, but it is very important to enhance phishing detection techniques.

- **What is phishing website? :**

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is additionally referred to as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to seem legitimate via obfuscation and lots of other simple techniques including: fast-flux, during which proxies are automatically generated to host the web-page; algorithmic generation of latest URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack.
Heuristic based detection which incorporates characteristics that are found to exist in phishing attacks actually and may detect zero-hour phishing attack, but the characteristics aren't bound to always exist in such attacks and false positive rate in detection is extremely high. To beat 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 needs past data to form a choice or prediction on future data. Using this system, algorithm will analyze various blacklisted and bonafide URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

Keywords:

Phishing Detection, Feature Extraction, Phishing Website, Phishing Attacks

- **Methodology:**

Detecting and identifying Phishing Websites is really a complex and dynamic problem. Machine learning has been widely used in many areas to create automated solutions. The phishing attacks can be carried out in many ways such as email, website, malware, sms and voice. In this work, we concentrate on detecting website phishing (URL), which is achieved by making use of the Hybrid Algorithm Approach. Hybrid Algorithm Approach is a mixture of different classifiers working together which gives good prediction rate and improves the accuracy of the system.

Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we cannot differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification. Let us discuss each of them in details.

Naive Bayes Classifier:-

This classifier will assume that the existence of specific features in a class is not related to the existence of any other feature. If there is dependency among the features of each other or on the presence of other features, all of these will be considered as an independent contribution to the probability of the output. This classification algorithm is very much useful to large datasets and is very easy to use.

Random Forest: This classification algorithm are similar to ensemble learning method of classification. The regression and other tasks, work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees. This classifier accuracy for decision trees practice of over fitting the training data set.

Support vector machine (SVM): This is also one of the classification algorithm which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the n represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.

- # EXISTING SYSTEM

The existing anti-phishing approaches use the blacklist methods or features based machine learning techniques. Blacklist methods fail to detect new phishing attacks and produce high false positive rate. Moreover, existing machine learning based methods extract features from the third party, search engine, etc. Therefore, they are

complicated, slow in nature, and not fit for the real-time environment. To solve this problem, this paper presents a machine learning based novel anti-phishing approach that extracts the features from client side only. We have examined the various attributes of the phishing and legitimate websites in depth and identified 5 new outstanding features to distinguish phishing websites from legitimate ones

- **PROPOSED SYSTEM:**

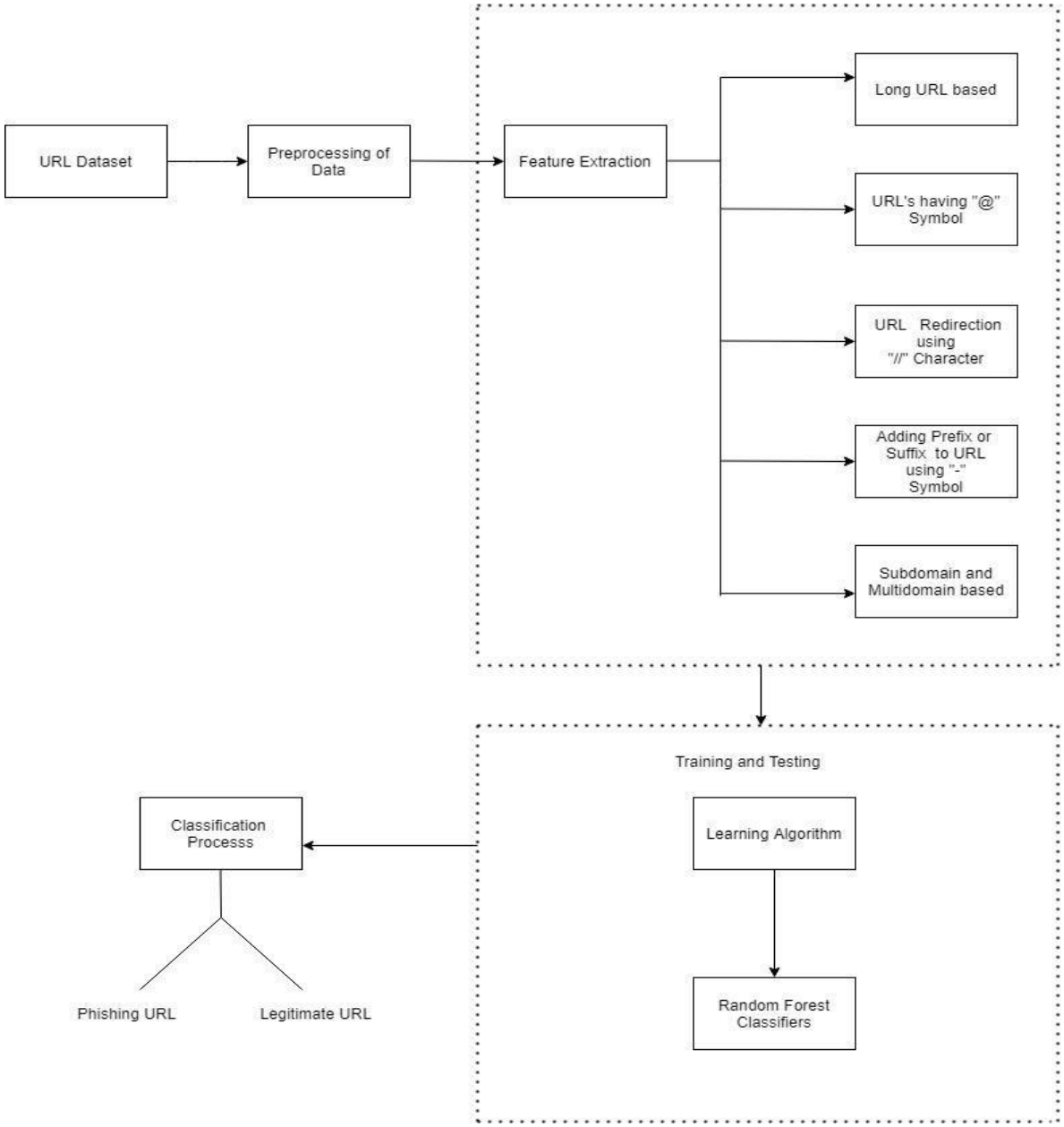
The dataset of phishing and legitimate URL's is given to the system which is then pre-processed so that the data is in the useable format for analysis. The features have around 300 characteristics of phishing websites which is used to differentiate it from legitimate ones.

Each category has its own characteristics of phishing attributes and values are defined. The specified characteristics are extracted for each URL and valid ranges of inputs are identified. These values are then assigned to each phishing website risk. For each input the values range from 0 to 10, while for output range is from 0 to 100. The phishing attributes values are represented with binary no 0 and 1 which indicates the attribute is present or not.

After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms are already explained in previous section. After this we use a classification named Random forest to predict the accuracy of the detection of the phishing URL, hence we get our desired result. This is also called a random approach to test the data, in this method we propose to use the classifier, as mentioned above.

We shall then test the data and evaluate the prediction accuracy which shall be more than the existing system. We shall now see the different classifiers and discuss the hybrid combination used for our proposed system.

- **Proposed System block diagram**



REQUIREMENT AND SPECIFICATIONS

2.1 Functional Requirements:

Graphical User interface with the User.

2.2 Software Requirements:

For developing the application the following are the Software Requirements:

Operating system : Windows 7, Windows 8 , Windows 10

Coding Language : Python (3.7.4)

Technologies : Anaconda Navigator.

IDE : Jupyter Notebook(6.0.3), Python IDLE (3.7.4)

2.3 Hardware Requirements:

For developing the application the following are the Hardware Requirements:

Processor : Pentium IV or higher

RAM : 4 GB

Space on Hard Disk : minimum 20GB

- **Problem statement:**

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside, but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning algorithms. We have found that our system provides us with 85.6 % of accuracy for Random Forest Classifier. The proposed technique is much more secured as it detects new and previous phishing sites.

- **Future scope**

In future if we get structured dataset of phishing we can perform phishing detection much faster than any other technique. In future we can use a combination of any other two or more classifier to get maximum accuracy. We also plan to explore various phishing techniques that uses Lexical features, Network based features, Content based features, Webpage based features and HTML and JavaScript features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.