

Detecting Phishing Website using Machine Learning

Lakshmishree¹ and Arpitha K²

¹Student, Dept. of Computer Science and Engg., BGSIT
Email: lakshmishree281@gmail.com

²Asst Prof., Dept. of Information Science and Engg., BGSIT
Email: arpithak@bgsit.ac.in

Abstract—Phishing means, it's a type of technique used by the criminals to get access the user credentials or sensitive data. In these days we all very much familiar about this phishing website, the reason is covid pandemic. Before also we have phishing websites but during this pandemic it's increasing. Because everything is done in virtual. Criminals fetch the user credentials for illegal activities. Mainly the phisher use email and SMS to send the phishing URL. They use logos and slogans of a legitimate website so that website or email looks like same as legitimate website if the user click on that link it redirected to another website means phishing website so user thing this is real website or legitimate website and start giving their personal information. This causes big problem in the society to solve this problem we bring a solution. For detecting phishing URL by machine learning technique.

I. INTRODUCTION

Machine learning is an important component of the growing field of data science through the use of statistical method, algorithms are trained to make classification or predictions. Here we used the fine-tuned features and rules for separating sites from legitimate website, here we using the Gradient boosting algorithm in this project to detect the phishing website and block the website. In earlier we are using traditional classification. This technique gives less accuracy.

Because every day phishers creating the number of new phishing website so we are using this model to reduce phishing website in earlier we are using the traditional method. Traditional method means list base technique in this method it contains two types list they are white list and black list. The white list contains only the legitimate website and the black list contains only the phishing websites then the black list website is obtained through either feedback of the user or third-party report, here the main drawback is, if the entered website present in that list in that time only we can tell it's a legitimate and phishing website otherwise we can't classify. In these days every day creating the number of new phishing website we can't classified in that time. And also, In the existing system they use another method for classified that is visual similarity technique. Visual similarity technique means comparing the figures, logos and slogans between legitimate website but here the phisher replicates the legitimate website logos and slogans, and also in the early days they use heuristic value for classification, these techniques are not satisfied for all the websites.

II. BACKGROUND WORK

The task is actualized utilizing four ML managed classification models. The four classification models are K-Nearest Neighbor, Kernel Support vector machine, Decision tree and Random Forest classifier. It was

discovered that the Random Forest classifier is most accurate for the chosen dataset and gives an accuracy score of 96.82% [1]. The proposed approach suggests utilizing practice swarm optimization (PSO) to weight various website features effectively to achieve higher accuracy when detecting phishing website. In particular the website based on how important they contribute towards recognizing the phishing from website [2]. Text based similarity approaches are relatively fast, but they are unable to detecting the phishing attack if the text is replaced with some image. Image processing-based approaches have high accuracy rate while they are complex in nature and are time-consuming [3]. To defend against phishing attacks and prevent financial losses. In this paper they used list-based methods and rule-based strategies to machine learning based approaches. In this proposed system they achieved 95% accuracy [4]. This tool should be improved constantly through continuous retraining. Actually, the availability of fresh and up-to - date training dataset which may acquire using our own tool [5].

III. SYSTEM ARCHITECTURE

- **Data acquisition:** Data set obtained or collected kaggle website and storing it in the database.
- **Data preprocessing:** Data contains a noise, that information should be treated before visualization or analysis.
- **Data Visualization:** Data visualization is an integral part of data analytics and machine learning. When there is a huge data set, manual analytics becomes almost impossible. Data Visualization plays a vital role in analysis in such situation. It involves uses of various plots- bar graphs, pie charts, box plots, line graph and many more.
- **Classification Techniques:** Based on the result obtained after classification, predict the website is phishing or legitimate website and accuracy is measured.

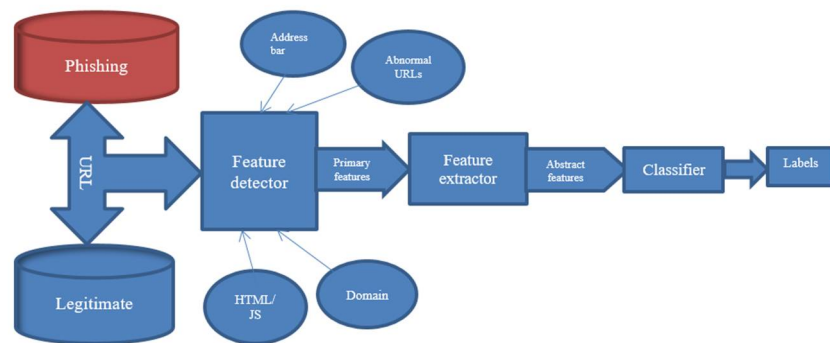


Figure 1: Block diagram

IV. IMPLEMENTATION

A. Importing Library

- “NumPy” is a package for scientific computing with python. This library is imported as np and will be used in the project.
- “Pandas” is for data manipulation and analysis. Pandas is an open source.
- “Matplotlib.pyplot” is a collection of command style function that make matplotlib work like MATLAB. It is imported as plot.
- “Seaborn” is a python data visualization library based on matplotlib for attractive and informative statistical graphics.

```

In [1]: #importing required libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn import metrics
import warnings
warnings.filterwarnings('ignore')
  
```

B. Importing data

Python code to import data from respective directory/file and assigning it to Data Frame df. The data stored in CSV format is being imported.

C. Data Visualization and data exploration needs to be carried out

- There are 11054 instances and 31 features in dataset.
- Out of which 30 are independent features whereas 1 is dependent feature.
- Each feature is in int datatype, so there is no need to use Label Encoder.
- There is no outlier present in dataset.
- There is no missing value in dataset.

D. Splitting the data into train and test sets. The training data contains 80% of data and testing data contains 20% of data

E. Applying algorithms on the split data

- Gradient Boosting Classifier
- Random Forest
- Decision Tree

F. Predict the entering website is legitimate or phishing.

V. RESULTS

```
In [36]: # displaying total result  
sorted_result
```

Out[36]:

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	Random Forest	0.967	0.971	0.993	0.990
2	Decision Tree	0.959	0.963	0.991	0.993

Figure 2: Accuracy of the entire algorithms

In the below Figure 3: shows the training and testing accuracy of the gradient boosting algorithm. It gives 97.4% accuracy in the testing.

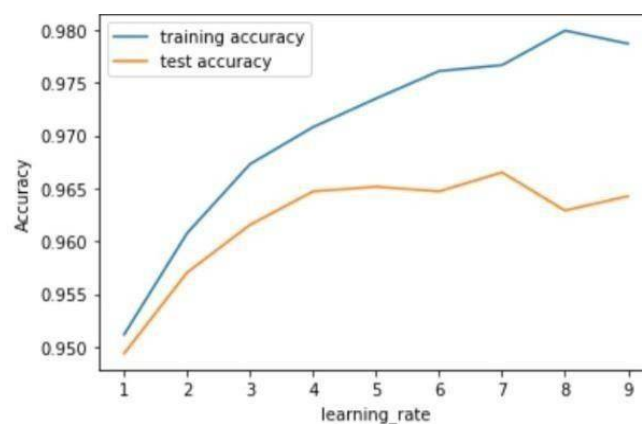


Figure 3: Gradient boosting algorithm graph

In this below Figure 4: shows the random forest algorithm training and testing data accuracy it gives 96.5%.

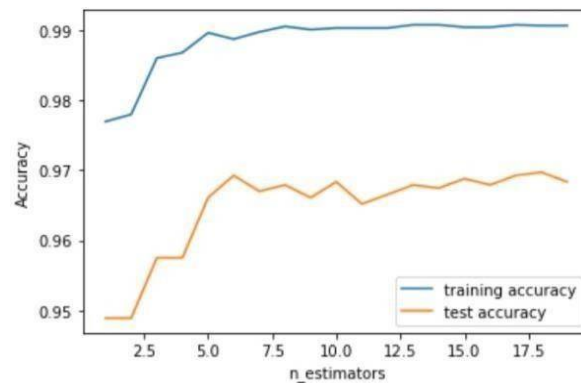


Figure 4: Random forest algorithm graph

Decision tree algorithm gives the 95.9% accuracy on training data set it shown in the figure 5.

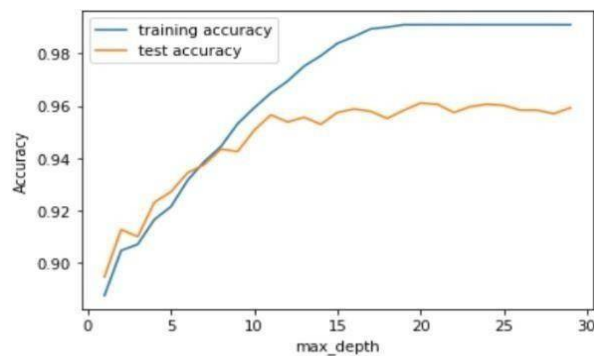


Figure 5: Decision tree algorithm graph

Here we implement the front end using HTML and CSS. The result displayed in that web page there we can insert URL and it classify its safe website or unsafe website. Then if we want to continue that option also inserted in the webpage.

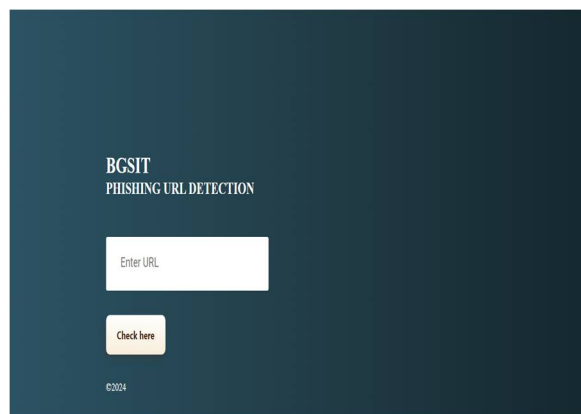


Figure 6: web page here we give the url link for classification

VI. CONCLUSION

The proposed system aim is to defining the some of the features of URL rules, means how the legitimate URL looks like and phishing URL looks like. Here we are using some of the machine learning algorithms and artificial neural network. These techniques better performance and accuracy as compared to other techniques.

FUTURE ENHANCEMENT

Upon exclusive study and research, one of the challenges faced by our research was the unavailability of reliable training datasets. Even we collect all reliable data it is difficult to shape a dataset that covers all possible features. We proposed some new features; experimentally assign new rules to some well-known features. If the phisher creates phishing website apart from this feature, we can't predict it is a phishing website or legitimate website but in feature if we get a more features for predicting websites. Feature classifications helps to determine the website better and get more accuracy.

REFERENCE

- [1] L. McCluskey, F. Thabtah, and R. M. Mohammad, "Intelligent rulebased phishing websites classification," *IET Inf. Secur.*, vol. 8, no. 3, pp. 153–160, 2014.
- [2] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, 2014.
- [3] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," *Internet Technol. ...*, pp. 492–497, 2012.
- [4] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," *Appl. Soft Comput. J.*, vol. 48, pp. 729–734, 2016.
- [5] N. Abdelhamid, "Multi-label rules for phishing classification," *Appl. Comput. Informatics*, vol. 11, no. 1, pp. 29–46, 2015.