IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

# Detection of Phishing Websites using Machine Learning

B.Ramesh Babu, A.Sai Jahnavi, D.Tejaswi, K.Ravaliaka, G.Pradeep Reddy

Assistant Professor, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India
UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India
UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India
UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India
UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

**ABSTRACT:** Phishing is a kind of cyber-attack,which has an huge impact on people where the user is directed to create fake websites without security and try to grab their confidential and one's personal data which includes passwords of different social media accounts, bank account details, atm pin-card details, credit card details etc. Hence protecting confidential data from those insecure or malwares or websites phishing is difficult. So Machine learning is a tool for the study of data analysis and scientific study of data algorithms, which has come into existence in recent times for opposing such phishing websites.

This paper gives the explanation on how to make use of ML and its data algorithm techniques in identifying such phishing website attacks and reports their positives and negatives. In General ,there are many ML algorithms that have been explored to select and declare the right choice that act as anti-phishing website tools. We have designed a Phishing Classification system which extracts the important features that are meant to define and defeat such common phishing attack detection approaches. We make use of numeric representation of data along with the comparative and systematic approaches of classical machine learning algorithms and deep learning algorithms techniques like Random Forest Classifier, K-Nearest Neighbours, Standard vector Machine, XGbooster, Logistic Regression based features selection which contains and explains the metadata of URLs and make use of such data to determine whether the website is phishing or legitimate.

**KEYWORDS:** Cyber-attack, anti-phishing tools, classification system, Machine Learning

## I.INTRODUCTION

In recent days cyber-attacks are increasing day by day which was not done or happened before.Phishing attack is one among those number of cyber attacks. In phishing, phishers hunts the end-users by making and allowing them to click on the hyper-links which was not secure and to grab the confidential data and make them lose their personal information, banking details and credit card details, and secure passwords. In this attack the attackers impersonate themselves as a trusted entities such as employees of the particular organization or technical-support team from the organization or service providers so that the end-users can blindly trust them. It is mainly done through the end-users emails asking them to update the system, or saying that account has been suspended temporarily to click on the link to activate etc. The main goal of cyber attack phishing is to allow end-users to share their confidential data.

## II.LITERATURE SURVEY

This paper explores detailed literature available in prominent journals, conferences, and chapters. This paper explores relevant articles from Springer, IEEE, Elsevier, Wiley, Taylor & Francis, and other well-known publishers. This literature review is formulated after an exhaustive search on the existing literature published in the last 10 years.

A phishing attack is one of the most hazardious threat for an organisation. Initially, these were done on telephone networks also known as Phone Phreaking which is the reason the term "fishing" was replaced with the term "Phishing", *ph* replaced *f* in fishing. From the reports of the anti-phishing working group (APWG) [1], it can be confirmed that phishing was discovered in 1996 when America-on-line (AOL) accounts were attacked by social engineering. Phishing turns into a danger to numerous people, especially individuals who are unaware of the dangers while being in the internet world. In light of a report created by the Federal Bureau of Investigation (FBI) [4], from October-2013 to February-2016, a phishing attack caused severe damage of 2.3 billion dollars. In general, users tend to overlook the URL of a website. At times, phishing tricks connected through phishing websites can be effectively

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

prevented by seeing whether a URL is of phishing or an authentic website. For the situation where a website is suspected as a targeted phish, a client can escape from the criminal's trap.

The conventional approaches for phishing attack detection give low accuracy and can recognize only about 20% of phishing attacks. Machine learning approaches give good outcomes for phishing detection but are time-consuming even on the small-sized datasets and not scale-able. Phishing recognition by heuristics techniques gives high false-positive rates. Client mindfulness is a significant issue, for resistance against phishing attacks. Fake URLs are utilized by phisher, to catch confidential private data of the targeted victim like bank account data, personal data, username, secret password, etc.

Previous work on phishing attack detection has focused on one or more techniques to improve accuracy however, accuracy can be further improved by feature reduction and by using an ensemble model. Existing work done for phishing attack detection can be placed in two categories:

- Deep learning for phishing attack detection
- Machine learning for phishing attack detection

## III.PROPOSED SYSTEM

### Data Pre-Processing

A dataset typically contains some noises,null values,outliers,lack of values and some unstructed layout which are incapable for Machine learning and they need to be removed .

### Dataset Description

The set of phishing URLs are collected from open source service called PhishTank. This service provides a set of phishing URLs in several formats like csv, json etc. which gets updated hourly. To download the data we used the following link: https://www.phishtank.com/developer_info.php. From this dataset, 5000 random phishing URLs are collected to train the ML models.

The legitimate URLs are obatined from the open datasets of the University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign url dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

- **Address Bar based Features**: In this category 9 features are extracted.

Domain of URL, Redirection'//' in URL,IP Address in URL, 'http/https' in Domain name, '@' Symbol in URL, Using URL Shotening Service, Length of URL,Prefix or Suffix"-" in Domain,Depth of URL

- **Domain based Features**:In this category 4 features are extracted.

DNS Record, Age of Domain, Website Traffic, End Period Of Domain

- **HTML & Javascript based Features**: In this category 4 features are extracted.

Iframe Redirection, Disabling RightClick, Status Bar Customization,Web Forwarding

So, all together 17 features are extracted from the 10,000 URL dataset.

### Handling Missing data

After importing our dataset ,the missing data can be handled either by replacing the data with mean or median or byremoving those particular data.

### Splitting the set of data

The data is splitted into 75% of data set into a training data after checking whether the data has missing values and remaining data is given to testing.

### Correlation

We used correlation matrix and constructed a correlation heatmap and found the correlation is weak with target variable.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

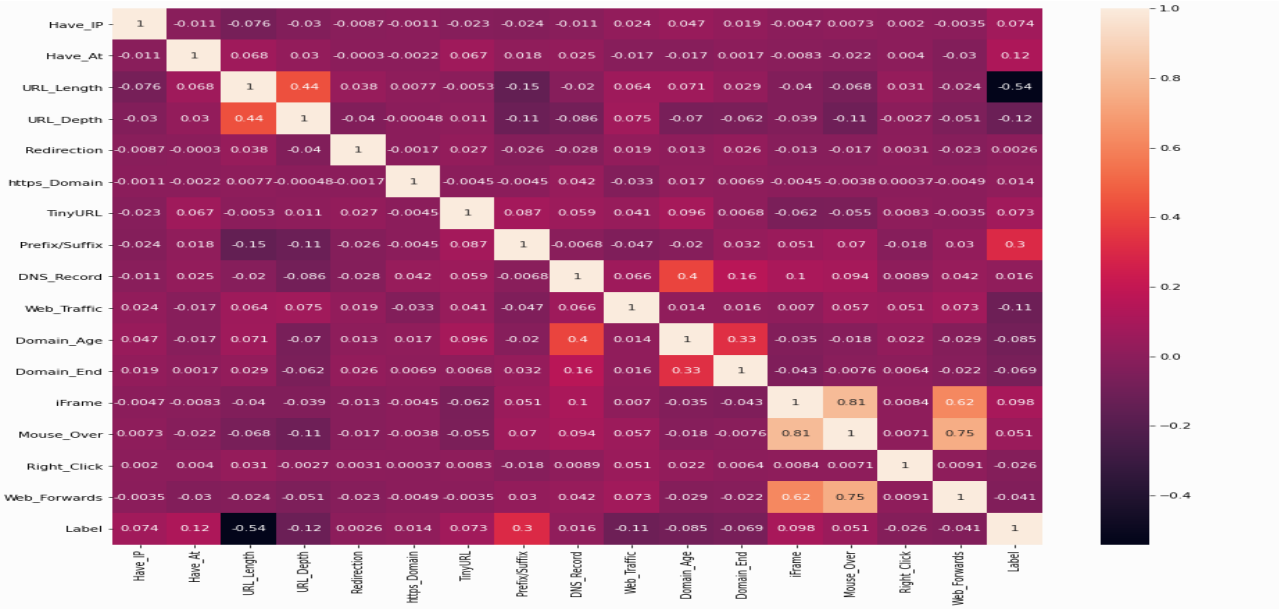


Fig:1 correlation heatmap

**Algorithms Used**

**Logistic Regression**

Logistic regression is a type of supervised algorithm that is used to predict the probability of a target variable.

**Random Forest Classifier**

A random Forest classifier is technique of machine learning which is used to solve classification and regression problems it finds solutions to many complex problems that combines many classifier by using ensemble learning.

**Support Vector Machine**

SVM is a supervised machine learning algorithm which is used for classification and regression model to analyse the data

**K-Nearest Neighbours**

KNN algorithm is used to stores the entire dataset and uses it to represent the data.It is mostly used for classification problems

**Naive Baye's**

Naïve Bayes is a effective classification alogorithm which is used to make quick prediction on high dimensional data or used to predict the value based on probability of an object

**XGBooster**

XGboost is an implementation of gradient booster decision tress which is used to increase speed and performance.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

**Data Visualisation**

**Plotting bargraphs**

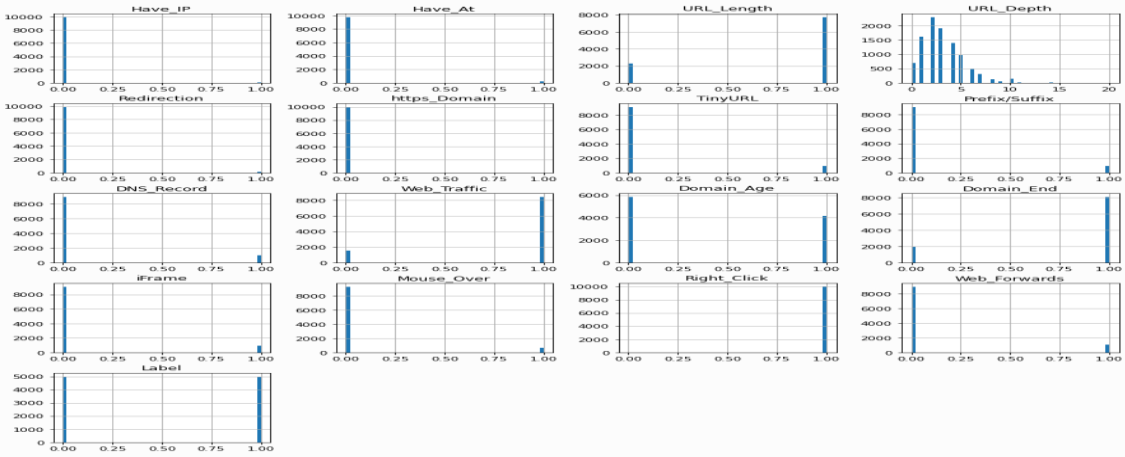


Fig:2 Bargraphs of the attributes

The fig 2 illustrates different bargraphs of the data where we can observe the attributes distributions.

**Result Analysis**

After applying six classification algorithms on the data Random Forest Classifier and Xgbooster gave the accurate results with the accuracy of 86.6.Since Xgbooster has a good performance and speed compared to all algorithms we choose that.

| | ML Model | Test Accuracy |
|---|---|---|
| 0 | Logistic Regression | 0.800 |
| 1 | Random Forest | 0.869 |
| 2 | SVM | 0.804 |
| 3 | KNN | 0.847 |
| 4 | Naivebayes | 0.794 |
| 5 | XGBooster | 0.867 |

Fig:3 Result analysis of algorithms

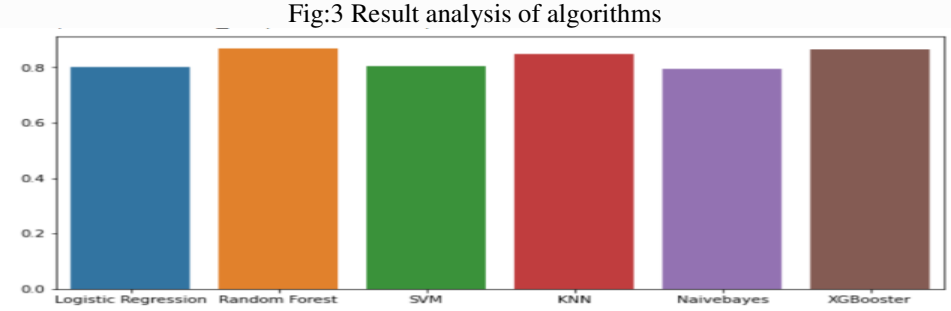


Fig:4 Bargraphs of algorithms

## V. RESULTS

According to the dataset if the prediction value is equal to 1 then the website is a phishing website and if the predicted value is equal to o the the website is legitimate

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

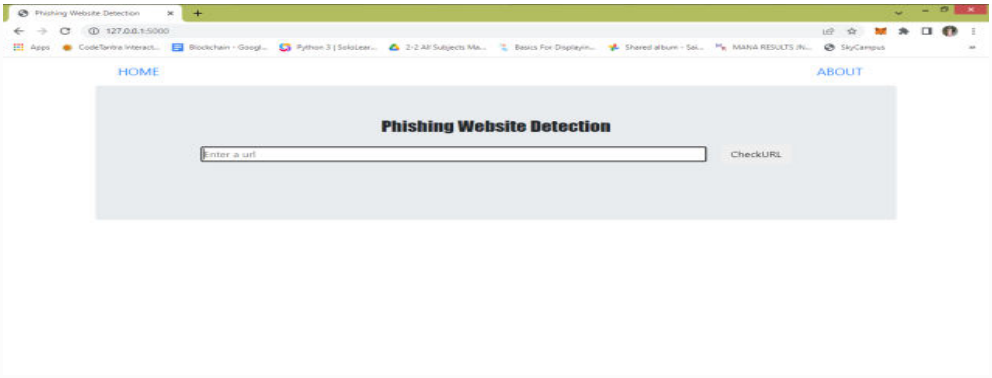


Fig:5 Output Screen for Home Page

In the home page if we enter the url in the given label it predicts the output to either phishing or legitimate.The home page also dispalys the about page which shows the details of phishing.

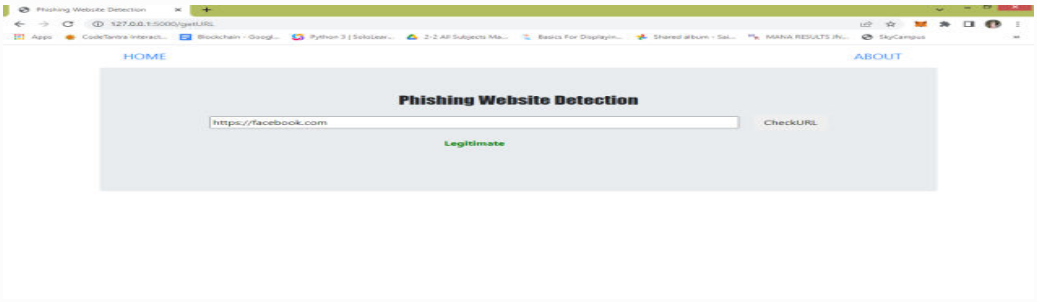


Fig:6 Prediction of legitimate website

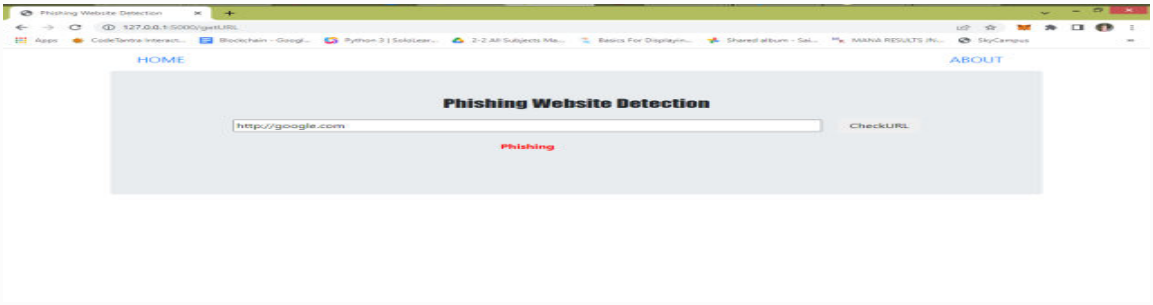


Fig:7 Prediction of Phishing Website

## VI.CONCLUSION

I have utilised six supervised machine learning algorithms of classification where random forest classifier and XGbooster gave the highest accuracy among all.The accuracy for both algorithms is almost equal and it is 86.7%.Based on the given input of the domain it examines all the features and returns the predicted values as 1 or 0 which indicates phishing or legitimate.

## REFERENCES

[1].Afroz S, Greenstadt R (2011) PhishZoo: detecting phishing websites by looking at them. In: 2011 IEEE fifth international conference on semantic computing, Palo Alto, CA, pp 368–375. https://doi.org/10.1109/ICSC.2011.52. Accessed 29 Aug 2020

[2].APWG trends report q1 2019. (n.d.). Retrieved from https://docs.apwg.org/ reports/apwg_trends_report_q1_2019.pdf. Accessed 29 Aug 2020.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005075|

[3].Armano G, Marchal S, Asokan N 2016) Real-time client-side phishing prevention add-on. In: 2016 IEEE 36th international conference on distributed computing systems (ICDCS), pp 777–778. https://doi.org/10.1109/icdcs.2016.44

[4].Futai Z, Yuxiang G, Bei P, Li P, Linsen L (2016) Web phishing detection based on graph mining. In: 2016 2nd IEEE international conference on computer and communications (ICCC). https://doi.org/10.1109/compcomm.2016.7924867

[5].Join the fight against phishing. (n.d.) retrieved from https://www.phishtank.com/ . Accessed 29 Aug 2020

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details