International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

Phishing Website Detection Using Machine Learning

Megha Agarwal[1], Shruti Jani[2], Hansika Koli[3], Prof. Deepali Maste[4]

[1,2,3]*B.E. Student, Information Technology Engineering, Atharva College of Engineering, Mumbai, India*
[4]*H.O.D, Information Technology Engineering, Atharva College of Engineering, Mumbai, India*

Abstract - *Phishing internet sites are one of the internet protections issues that focus on human vulnerabilities rather than software program vulnerabilities. It can be defined because of the process of attracting online users to gain their touchy facts which include usernames and passwords. In this paper, we provide a sensible machine for detecting phishing websites. The gadget acts as a further functionality to an internet browser as an extension that routinely notifies the consumer whilst it detects a phishing internet site. The system is based on a device gaining knowledge of approach, particularly supervised mastering. We have decided on the XGBoost method because of its true overall performance in classification. Our focus is to pursue a better overall performance classifier by analyzing the features of phishing websites and choose the better aggregate of them to train the classifier. As a result, we finished our paper with accuracy of 99% accuracy with 48 features.*

Key Words: **XGBoost (Extreme Gradient Boosting), Classifier, Features, Phishing, Train, Accuracy.**

1. INTRODUCTION

Internet and cloud technology improvements in recent years have significantly increased electronic trade, or consumer-to-consumer online transactions. The resources of an enterprise are harmed by this growth, which permits unauthorised access to sensitive information about users. One well-known attack that manipulates users into accessing dangerous content and giving up their information is phishing. Most phishing websites use the same website interface and universal resource location (URL) as the legitimate websites.

1.1 Purpose

Internet consumers lose billions of dollars each year as a consequence of website phishing. Phishers prey on people's online security by stealing usernames, passwords, and financial account information. Due to the use of URL obfuscation to shorten the URL, link redirections, modifying links to make them appear trustworthy, and a long list of other techniques, detecting phishing websites is difficult. This made it necessary to convert from conventional programming methods to an approach based on machine learning.

When new phishing strategies are launched, phishing detection solutions do suffer from low detection accuracy and high false alarm rates. Additionally, since registering new domains has gotten simpler, the most popular methodology, the blacklist-based method, is ineffective at responding to phishing assaults that are on the rise. No comprehensive blacklist can guarantee a flawlessly up-to-date database.

1.2 Objective

- To create a reducing method for identifying dangerous URLs and warning users.
- To use ML approaches in the suggested approach to analyse real-time URLs and generate useful results.
- To implemented the idea of RNN, a well-known ML technique that can handle enormous volumes of data.
- The use of machine learning is crucial in preventing phishing attacks. This study investigates characteristics and techniques for machine learning-based detection.

2. LITERATURE REVIEW

MAHAJAN MAYURI VILAS, KAKADE PRACHI GHANSHAMSAWANT, PURVA JAYPRALASH and PAWAR SHILA [1] in their paper "Detection of Phishing Website Using Machine Learning Approach" , the goal of the study is to carry out ELM employing 30 different primary components that are characterized using ML. To prevent being discovered, most phishing URLs use HTTPS. Website phishing can be identified in three different ways. The first method evaluates several URL components; the second method assesses a website's authority, determines if it has been introduced or not, and determines who is in charge of it; the third method verifies a website's veracity.

In [2] MALAK ALJABRI and SAMIHA MIRZA proposed a paper "Phishing Attacks Detection using Machine Learning and Deep Learning Models" In this study, the highest correlated features from two distinct datasets were chosen. These features combined content-based, URL and domain-based features. Then, a comparison of the performance of a number of ML models was carried out. The results also sought to pinpoint the top characteristics that aid the algorithm in spotting phishing websites. The Random Forest (RF) method produced the best classification results for both datasets.

ADARSH MANDADI and SAIKIRAN BOPPANA in their study [3], the user-received URLs will be entered to the machine learning model, which will then process the input and report the results, indicating whether the URLs are phishing or not. SVM, Neural Networks, Random Forest, Decision Tree, XG boost, and other machine learning algorithms can all be used to categorize these URLs. The suggested method uses the Random Forest and Decision Tree classifiers. With an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers, respectively, the suggested technique successfully distinguished between Phishing and Legitimate URLs.

In [4] HEMALI SAMPAT, MANISHA SAHARKAR, AJAY PANDEY AND HEZAL LOPES have proposed a system for Detection of Phishing Websites using Machine learning. Their proposed method uses both Classification and Association algorithms to optimise the system, making it faster and more effective than the current approach. The proposed system's inaccuracy rate is reduced by 30% by combining these two algorithms with the WHOIS protocol, making it an effective technique to identify phishing websites.

SAFA ALREFAAI, GHINA ÖZDEMIR and AFNAN MOHAMED [5] used Machine Learning is being used to detect phishing websites. They used Kaggle data with 86 features and 11,430 total URLs, half of which are phishing and half of which are legitimate. They trained their data using Decision Tree (DT), Random Forest (RF), XGBoost, Multilayer Perceptrons, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting, with X G Boost.

In [6] , SUNDARA PANDIYAN S, PRABHA SELVARAJ, VIJAY KUMAR BURUGARI, JULIAN BENADIT P and KANMANI P employed a wide range of techniques, including Decision Tree, Random Forest, Multi-Layer Perceptrons, XG Boost Classifier, SVM, Light BGM Classifier, and Cat Boost Classifier. Our team discovered that Light GBM had the best precision, with an average accuracy of about 85.5%. One class SVM, on the other hand, has the lowest precision, at about 79.6%.

3. PROPOSED SOLUTION

Each type of phishing differs slightly in how the procedure is carried out to deceive the unwary customer. When a hacker sends a potential user an email with a link that takes them to phishing websites, this is known as an email phishing attack.

We use different machine learning models trained over features like if URL contains @, if it has double slash redirecting, page rank of the URL, number of external links embedded on the webpage, etc. Neural network perceptron on data provided by Machine Learning and were able to achieve a better accuracy. This approach could get up to 92% true positive rate and 0.4% false positive rate.

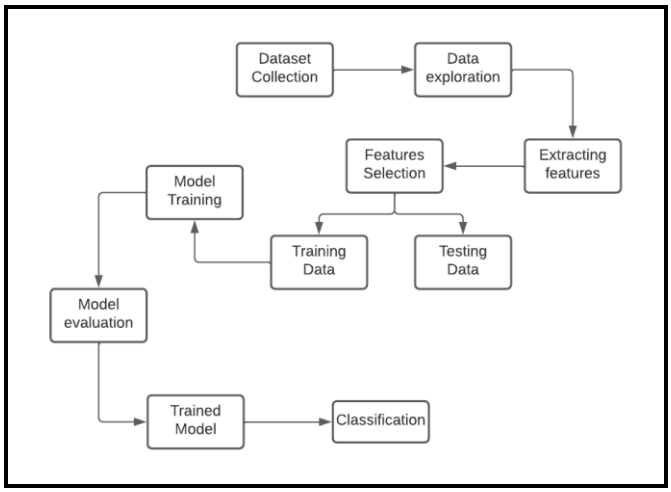

Fig -2.1: Flowchart

4. METHODOLOGY

Data collection, cleaning, and consolidation into a single file or data table are all steps in the process of data preparation, which is done largely for analytical purposes as shown in Fig2.1 . The following are the main activities we utilise for data preparation: data reduction, data transformation, data integration, and data discretization.

The crucial libraries, including XGBoost, Numpy, Matplotlib, Pandas, and Numpy, are loaded first. The dataset from Kaggle is then imported after the libraries have been imported. "Phishing Legitimate full" is the name of the dataset that we have selected. We divided the dataset into training and testing sets after importing it using train test split from sklearn. 20% of the dataset is used for testing, while 80% is used for the training set.

We have set up a model that uses five distinct algorithms, including Logistic Regression, KNeighborsClassifier, Random Forest, Decision Tree, and XGBoost, to compare the accuracy of various techniques. We work on model fitting, which makes predictions, to achieve the desired result, and then we work on model evaluation. For this evaluation, test data is utilised. We compare the accuracy of each method using several algorithms, such as confusion matrix, to obtain the best result.

5. SCOPE

Internet customers lose billions of dollars every year due to website phishing. Phishers prey on people's online security by stealing their usernames, passwords, and financial account information. The COVID-19 epidemic has increased technology use across all industries, leading to a transition from offline to online spaces for tasks like scheduling official meetings, going to classes, buying, and making payments. This means that phishers will have more chances to carry

out assaults that harm the victim's finances, psychological well-being, and professional prospects.

This process can be made much more difficult by the introduction of browser extensions or sophisticated GUIs that analyse URLs to determine whether they are legitimate phishing sites. We are currently getting closer to launching the browser extension for this project. can even test out the GUI option. The further characteristics can be updated as soon as possible. We are eager to create a complete programme that, rather than requiring verification, immediately disables the website.

6. RESULT

To acquire useful results, we've compared a number of algorithms. There are numerous algorithms that can be used to identify phishing websites; however, after reviewing numerous research articles, we settled on five algorithms to test the model.

6.1 Model Comparison

Table -1:

ML MODELS			
		Train Accuracy	Test Accuracy
1	LogisticRegression	0.946	0.941
2	Decision Tree	0.966	0.956
3	Random Forest	0.967	0.964
4	KNeighborsClassifier	1.000	0.881
5	XGBoost	1.000	0.990

6.2 Model Output

```
XGBoost: Accuracy on the Model:  0.9905
XGBoost: Accuracy on training Data: 1.000
XGBoost : Accuracy on test Data: 0.991
              precision    recall  f1-score   support

           0       0.90      0.85      0.87       979
           1       0.86      0.91      0.89      1021

    accuracy                           0.88      2000
   macro avg       0.88      0.88      0.88      2000
weighted avg       0.88      0.88      0.88      2000

[[830 149]
 [ 89 932]]
```

Fig -5.2.1: Output

XGBoost is a distributed gradient boosting library that was created to be incredibly powerful, versatile, and portable. The machine learning techniques are implemented using the Gradient Boosting framework. XGBoost (also known as GBDT or GBM), a parallel tree boosting technique, is available to swiftly and accurately address a number of data science problems.

The term "XGBoost" refers to a proficiency configuration. Generally speaking, it is not feasible to rely only on one machine learning model. Through outfit education, a tactical strategy to handling the prophetic power of integrating different students is offered. A single model that displays the total outcome from numerous models is the end result.We trained our data using Logistic Regression, KNeighborsClassifier, Random Forest, Decision Tree, and XGBoost with X G Boost achieving the highest accuracy of 99.05.

7. CONCLUSIONS

To the best of our knowledge, this study is the first analysis to include the findings of all other studies into the detection of phishing websites using machine learning techniques. The suggested research on phishing uses a categorical paradigm, where phishing websites are thought to automatically classify websites into a given range of sophisticated values depending on a variety of factors and the grandeur variable.

The website functionality is used by ML-based phishing approaches to collect information that could be used to classify websites for the purpose of identifying phishing sites. Developing focused anti-phishing approaches and methods as well as minimizing their inconvenience are two ways to prevent phishing.

We achieved 99.05% detection accuracy using XG Boost algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data.

REFERENCES

[1] M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.

[2] M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models," 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034.

[3] A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine

Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

[4] Hemali Sampat, Manisha Saharkar, Ajay Pandey and Hezal Lopes, "Detection of Phishing Website Using Machine Learning," 2018 International Research Journal of Engineering and Technology (IRJET),2018, e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[5] S. Alrefaai, G. Özdemir and A. Mohamed, "Detecting Phishing Websites Using Machine Learning," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022, pp. 1-6, doi: 10.1109/HORA55278.2022.9799917.

[6] Sundara Pandiyan S, Prabha Selvaraj, Vijay Kumar Burugari, Julian Benadit P, Kanmani P, Phishing attack detection using Machine Learning, Measurement: Sensor Volume 24,2022,100476,ISSN 2665-9174

BIOGRAPHIES

Megha Agarwal
I.T Engineer (2019-2023) from Atharva College of Engineering, Malad, Mumbai

Shruti Jani
I.T Engineer (2019-2023) from Atharva College of Engineering, Malad, Mumbai

Hansika Koli
I.T Engineer (2019-2023) from Atharva College of Engineering, Malad, Mumbai